Genomic and Metabolomic Determinants of Neurological and Psychiatric Traits

Dina Vojinović

The work presented in this thesis was conducted at the Genetic Epidemiology Unit, Department of Epidemiology, Erasmus University Medical Center, Rotterdam, the Netherlands.

The Erasmus Rucphen Family (ERF) study as a part of EUROSPAN (European Special Populations Research Network) was supported by European Commission FP6 STRP Grant No. 018947 (LSHG-CT-2006-01947) and also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013)/ Grant Agreement HEALTH-F4-2007-201413 by the European Commission under the programme "Quality of Life and Management of the Living Resources" of 5th Framework Programme (no. QLG2-CT-2002-01254). High-throughput analysis of the ERF data was supported by joint grant from Netherlands Organization for Scientific Research and the Russian Foundation for Basic Research (NWO-RFBR 047.017.043), and Russian Federal Agency of Scientific Organizations projects VI.53.2.2 and 0324-2015-0003.

The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organization for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. The generation and management of GWAS genotype data for the Rotterdam Study is supported by the Netherlands Organisation of Scientific Research NWO Investments (nr. 175.010.2005.011, 911-03-012), the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO) Netherlands Consortium for Healthy Aging (NCHA), project nr. 050-060-810. The authors are grateful to the study participants, the staff from the Rotterdam Study and the participating general practitioners and pharmacists.

The research described in this thesis was supported by a grant of the Dutch Heart Foundation (CVON 2012B003). Other financial support leading to this thesis: the CoSTREAM project (www.costream.eu, grant agreement No 667375), Biobanking and Biomolecular Resources Research Infrastructure (BBMRI)-NL (184.021.007), the European Union's Horizon 2020 research and innovation programme Marie Skłodowska-Curie Research and Innovation Staff Exchange (RISE) under the grant agreement No 645740 as part of the Personalized pREvention of Chronic DIseases (PRECeDI) project.

Printing of this thesis was financially supported by: Department of Epidemiology, Erasmus University Medical Center, Rotterdam, Erasmus University, Rotterdam, Alzheimer Nederland, and Chipsoft. Financial support by the Dutch Heart Foundation for the publication of this thesis is gratefully acknowledged.

gesubsidieerd door Hartstichting

alzheimer nederland

ChipSoft

Layout: Optima Grafische Communicatie
Cover design: Erwin Timmerman, Optima Grafische Communicatie and Dina Vojinović
Printing: Optima Grafische Communicatie

ISBN: 978-94-6361-159-6

Genomic and Metabolomic Determinants of Neurological and Psychiatric Traits

Erfelijke en metabole determinanten van neurologische en psychiatrische aandoeningen

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

Prof.dr. R.C.M.E. Engels

en volgens besluit van het College voor Promoties.
De openbare verdediging zal plaatsvinden op

woensdag 12 december 2018 om 9.30 uur

door

Dina Vojinović
geboren te Čačak, Servië

Erasmus University Rotterdam

PROMOTIECOMMISSIE

Promotor:	Prof.dr.ir. C.M. van Duijn
Overige leden:	Prof.dr. S. Debette Prof.dr. M.A. Ikram Prof.dr. P.J. Koudstaal
Copromotor:	Dr. N. Amin

Paranimfen:	Ashley van der Spek Hata Čomić

CONTENT

PUBLICATIONS AND MANUSCRIPTS BASED ON THE STUDIES DESCRIBED IN THIS THESIS

Chapter 2.1

Dina Vojinovic, Hieab H. Adams, Xueqiu Jian, Qiong Yang, Albert Vernon Smith, Joshua C. Bis, Alexander Teumer, Markus Scholz, Nicola J. Armstrong, Edith Hofer, Yasaman Saba, Michelle Luciano, Manon Bernard, Stella Trompet, Jingyun Yang, Nathan A. Gillespie, Sven J. van der Lee, Alexander Neumann, Shahzad Ahmad, Ole A. Andreassen, David Ames, Najaf Amin, Konstantinos Arfanakis, Mark E. Bastin, Diane M. Becker, Alexa S. Beiser, Frauke Beyer, Henry Brodaty, R. Nick Bryan, Robin Bülow, Anders M. Dale, Philip L. De Jager, Ian J. Deary, Charles DeCarli, Debra A. Fleischman, Rebecca F. Gottesman, Jeroen van der Grond, Vilmundur Gudnason, Tamara B. Harris, Georg Homuth, David S. Knopman, John B. Kwok, Cora E. Lewis, Shuo Li, Markus Loeffler, Oscar L. Lopez, Pauline Maillard, Hanan El Marroun, Karen A. Mather, Thomas H. Mosley, Ryan Muetzel, Matthias Nauck, Paul A. Nyquist, Matthew S. Panizzon, Zdenka Pausova, Bruce M. Psaty, Ken Rice, Jerome I. Rotter, Natalie Royle, Claudia L. Satizabal, Reinhold Schmidt, Peter R. Schofield, Pamela J. Schreiner, Stephen Sidney, David J. Stott, Anbupalam Thalamuthu, Andre G. Uitterlinden, Maria C. Valdés Hernández, Meike W. Vernooij, Wei Wen, Tonya White, A. Veronica Witte, Katharina Wittfeld, Margaret J. Wright, Lisa R. Yanek, Henning Tiemeier, William S. Kremen, David A. Bennett, J. Wouter Jukema, Tomas Paus, Joanna M. Wardlaw, Helena Schmidt, Perminder S. Sachdev, Arno Villringer, Hans Jörgen Grabe, WT Longstreth, Cornelia M. van Duijn, Lenore J. Launer, Sudha Seshadri, M Arfan Ikram, Myriam Fornage. **Genome-wide association study of 23,500 individuals identifies 7 loci associated with brain ventricular volume.** Accepted for publication in Nature Communications

Chapter 2.2

Dina Vojinovic, Caroline Hayward, Jennifer A. Smith, Wei Zhao, Jan Bressler, Stella Trompet, Chloé Sarnowski, Murali Sargurupremraj, Jingyun Yang, Paul R.H.J. Timmers, Narelle K. Hansell, Ari Ahola-Olli, Eva Krapohl, Joshua C. Bis, Daniel E. Gustavson, Teemu Palviainen, Yasaman Saba, Anbu Thalamuthu, Sudheer Giddaluru, Leonie Weinhold, Najaf Amin, Nicola Armstrong, Lawrence F. Bielak, Anne C. Böhmer, Patricia A. Boyle, Henry Brodaty, Harry Campbell, David W. Clark, Baptiste Couvy-Duchesne, Philip L De Jager, Jeremy A. Elman, Thomas Espeseth, Jessica D. Faul, Annette Fitzpatrick, Scott D. Gordon, Thomas Hankemeier, Edith Hofer, M. Arfan Ikram, Peter K. Joshi, Rima Kaddurah-Daouk, Jaakko Kaprio, Sharon LR Kardia, Katherine A. Kentistou, Luca Kleineidam, Nicole Kochan, John Kwok, Markus Leber, Teresa Lee, Terho Lehtimäki, Anu Loukola, Anders Lundquist, Leo-Pekka Lyytikäinen, Karen Mather, Grant W. Montgomery, Simone Reppermund, Richard J. Rose, Suvi Rovio, Perminder Sachdev, Matthias Schmid, Helena Schmidt, Andre G. Uitterlinden, Eero Vuoksimaa, Michael Wagner, Holger Wagner, David R. Weir, Margaret

J. Wright, Miao Yu, Lars Nyberg, Alfredo Ramirez, Stephanie Le Hellard, Peter Schofield, David Ames, Reinhold Schmidt, Danielle Dick, David Porteous, William S. Kremen, Bruce M. Psaty, Olli Raitakari, Nicholas G. Martin, James F. Wilson, David A. Bennett, Stephanie Debette, J. Wouter Jukema, Thomas H Mosley, Jr, Sudha Seshadri, Cornelia M. van Dujin. **Genetic determinants of general cognitive function and their association to circulating metabolites: a cross-omics study.** (In preparation)

Chapter 2.3

Riccardo E. Marioni*, Allan F. McRae*, Jan Bressler*, Elena Colicino*, Eilis Hannon*, Shuo Li*, Diddier Prada*, Jennifer A Smith*, Letizia Trevisi*, Pei-Chien Tsai*, Dina Vojinovic*, Jeannette Simino, Daniel Levy, Chunyu Liu, Michael Mendelson, Claudia L. Satizabal, Qiong Yang, Min A. Jhun, Sharon L. R. Kardia, Wei Zhao, Stefania Bandinelli, Luigi Ferrucci, Dena G. Hernandez, Andrew B. Singleton, Sarah E. Harris, John M. Starr, Douglas P. Kiel, Robert R. McLean, Allan C. Just, Joel Schwartz, Avron Spiro III, Pantel Vokonas, Najaf Amin, M. Arfan Ikram, Andre G. Uitterlinden, Joyce B. J. van Meurs, Tim D. Spector, Claire Steves, Andrea A. Baccarelli, Jordana T. Bell, Cornelia M. van Duijn, Myriam Fornage, Yi-Hsiang Hsu, Jonathan Mill, Thomas H. Mosley, Sudha Seshadri, Ian J. Deary. **Meta-analysis of epigenome-wide association studies of cognitive abilities.** Mol Psychiatry. 2018 Jan 8. [Epub ahead of print]

*These authors contributed equally to this work

Chapter 2.4

Dina Vojinovic, Hieab H.H. Adams, Sven J. van der Lee, Carla A. Ibrahim-Verbaas, Rutger Brouwer, Mirjam C.G.N. van den Hout, Edwin Oole, Jeroen van Rooij, Andre Uitterlinden, Albert Hofman, Wilfred F.J. van IJcken, Annemieke Aartsma-Rus, GertJan B. van Ommen, M. Arfan Ikram, Cornelia M. van Duijn, Najaf Amin**. The dystrophin gene and cognitive function in the general population.** Eur J Hum Genet. 2015 Jun;23(6):837-43.

Chapter 2.5

Vedrana Milic Rasic, Dina Vojinovic, Jovan Pesovic, Gordana Mijalkovic, Vera Lukic, Jelena Mladenovic, Ana Kosac, Ivana Novakovic, Nela Maksimovic, Stanka Romac, Slobodanka Todorovic, Dusanka Savic Pavicevic. **Intellectual ability in Duchenne muscular dystrophy and dystrophin gene mutation location.** Balkan J Med Genet. 2015 Apr 10;17(2):25-35.

Chapter 3.1

Dina Vojinovic, Maryam Kavousi, Mohsen Ghanbari, Rutger W.W. Brouwer, Jeroen G.J. van Rooij, Mirjam C.G.N. van den Hout, Robert Kraaij, Wilfred F.J. van IJcken, Andre G. Uitterlinden, Cornelia M. van Duijn, Najaf Amin. **Whole-genome linkage scan combined**

with exome sequencing identifies novel candidate genes for carotid intima-media thickness. Accepted for publication in Frontiers in Genetics

Chapter 3.2

Dina Vojinovic*, Sven J. van der Lee*, Cornelia M.van Duijn, Meike W. Vernooij, Maryam Kavousi, Najaf Amin, Ayşe Demirkan, M. Arfan Ikram, Aad van der Lugt, Daniel Bos. **Metabolic profiling of intra- and extracranial carotid artery atherosclerosis.** Atherosclerosis. 2018 May;272:60-65.

* These authors contributed equally to this work

Chapter 3.3

Dina Vojinovic, Marita Kalaoja, Stella Trompet, Krista Fischer, Martin J. Shipley, Shuo Li, Aki S. Havulinna, Markus Perola, Veikko Salomaa, Qiong Yang, Naveed Sattar, Pekka Jousilahti, Najaf Amin, Ramachandran S. Vasan, M. Arfan Ikram, Mika Ala-Korpela, J. Wouter Jukema, Sudha Seshadri, Johannes Kettunen, Mika Kivimaki, Tonu Esko, Cornelia M. van Duijn. **Circulating metabolites and risk of stroke in seven population-based cohorts.** (In preparation)

Chapter 3.4

Dina Vojinovic*, Djawad Radjabzadeh*, Alexander Kurilshikov*, Najaf Amin, Cisca Wijmenga, Lude Franke, Andre G. Uitterlinden, Alexandra Zhernakova, Jingyaun Fu**, Robert Kraaij**, Cornelia M. van Dujin**. **Relationship between gut microbiota and circulating metabolites in population-based cohorts.** (In preparation)

*These authors contributed equally to this work

**These senior authors contributed equally to this work

Chapter 4.1

Dina Vojinovic, Nathalie Brison, Shahzad Ahmad, Ilse Noens, Irene Pappa, Lennart C Karssen, Henning Tiemeier, Cornelia M. van Duijn, Hilde Peeters, Najaf Amin. **Variants in *TTC25* affect autistic trait in patients with autism spectrum disorder and general population.** Eur J Hum Genet. 2017 Aug;25(8):982-987.

Chapter 4.2

Alejandro Arias-Vásquez*, Alexander J. Groffen*, Sabine Spijker*, Klaasjan G. Ouwens*, Marieke Klein*, Dina Vojinovic*, Tessel E. Galesloot, Janita Bralten, Jouke-Jan Hottenga, Peter J. van der Most, V. Mathijs Kattenberg, Rene Pool, Ilja M. Nolte, Brenda W.J.H. Penninx, Iryna O. Fedko, Conor V. Dolan, Michel G. Nivard, Anouk den Braber, Cornelia M. van Duijn, Pieter J. Hoekstra, Jan K. Buitelaar, Bart Kiemeney, Martine Hoogman, Christel M. Middeldorp, Harmen H.M. Draisma, Sit H. Vermeulen, Cristina Sánchez-Mora, J. Antoni

Ramos-Quiroga, Marta Ribasés, The EAGLE-ADHD Consortium, Catharina A. Hartman, J.J. Sandra Kooij, Najaf Amin, August B. Smit**, Barbara Franke**, Dorret I. Boomsma.** **STXBP5 Antisense RNA 1 gene and adult ADHD symptoms.** (Submitted)

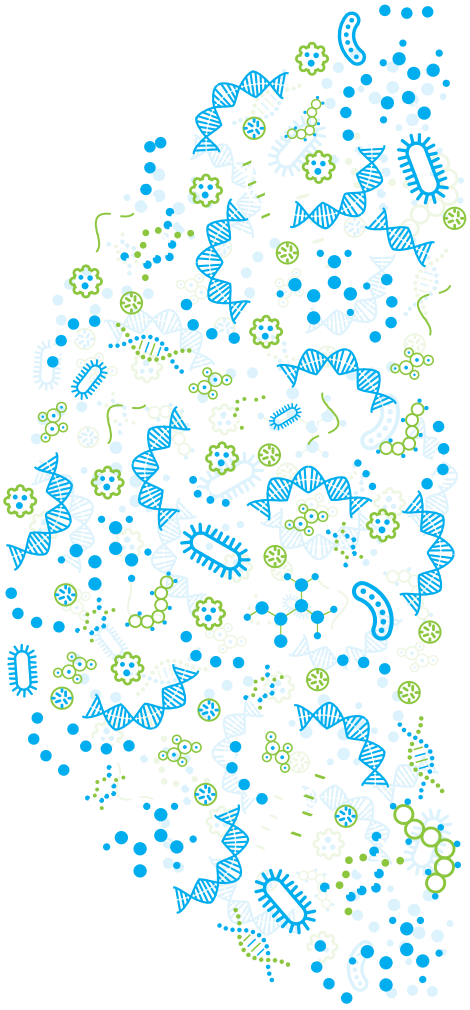
* These authors contributed equally to this work

** These authors share final responsibility

Chapter 5.2

Dina Vojinovic*, Anna Puggina*, Christian van der Werf, Carla G. van El, Olga C. Damman, Najaf Amin, Ayse Demirkan, Bruno H. Stricker, Muir Gray, Stefania Boccia, Martina C. Cornel, Cornelia M. van Duijn, Anant Jani**. A model for mass personalization in cardiology: standard outcomes-based systems that can deliver personalized care.** (Submitted)

* These authors contributed equally to this work



Chapter 1

General introduction

The human brain is the most complex organ in the human body. The adult human brain comprises billions of neuronal and glial cells interconnected via trillions of synapses.[1,2] It is responsible for motor functions, processing sensory information, language, cognitive processes, and function of other organs. Pathology of the brain may occur prenatal, in early childhood or adolescence up to senescence. Disorders of the brain comprise a heterogeneous group of neurological and psychiatric disorders and are an important cause of disability and death worldwide.[3-5] These disorders are the result of a combination of genetic, environmental, and lifestyle factors. The focus of research presented in this thesis are most common neurological disorders from an epidemiological perspective. They include late-onset neurodegeneration and cerebrovascular pathology and the most common neurodevelopmental disorders including attention deficit hyperactivity disorder (ADHD) and autism spectrum disorder (ASD). I have also studied Duchenne muscular dystrophy, a recessive inherited disorder.

Expanding our knowledge on the molecular processes and pathways of these disorders and early pathology may facilitate development of new prevention and treatment strategies. The early changes manifested prior to the onset of clinical symptoms of the disease are usually approached as heritable quantitative measures and referred to as endophenotypes.[6-8] Endophenotypes can be measured accurately on a continuous scale, overcoming the problem of defining the arbitrary boundary between the presence and absence of subclinical disease in controls.[9] For long, cognitive ability has been studied as endophenotype of neurodegenerative and psychiatric disorders,[10-15] whereas more recently brain volumetric and vascular measures depicted by state-of-the-art imaging techniques have been studied as endophenotype of neurodegeneration and neurovascular pathology.[16-18]

LATE ONSET NEUROLOGICAL DISORDERS AND RELATED ENDOPHENOTYPES

The most common presentation of cerebrovascular pathology is stroke, a neurological disorder of sudden onset. Risk factors come in many varieties, including genetic factors and various modifiable risk factors. Beyond a large number of rare monogenic disorders underlying stroke,[9] 32 risk loci encompassing common and less-frequent variants have been associated with stroke in a study of 520,000 subjects.[19] These provide additional insights into stroke pathophysiology.[19] Several biological pathways including enlarged heart, decreased cardiac muscle contractility, and oxaloacetate metabolism emerged as relevant for any stroke, whereas various cardiac pathways, muscle-cell fate commitment, and nitric oxide metabolism are implicated in cardioembolic stroke.[19] A significant proportion of stroke risk also resides in modifiable risk factors including hypertension,

diabetes mellitus, cardiovascular disease, and smoking.[20,21] As management of these risk factors demonstrated reduction of stroke burden, additional research efforts to identify high-risk patients have been sought to improve the chances of success. Several studies performed to date searched for novel metabolic disturbances and identified various small circulating compounds to be associated with stroke.[22-27] The most comprehensive study to date is conducted in China Kadoorie Biobank, involving patients with both ischemic stroke (IS) (n = 1,146) and intracerebral hemorrhage (ICH) (n = 1,138).[26] The study reported association between lipoproteins and lipids with IS, but not with ICH. Additionally, the study reported association of glycoprotein acetyls and several non-lipid related metabolites with both IS and ICH.[26] To date, the studies in Europeans are based on relatively small samples.[25,27] A study involving 268 patients with incident stroke revealed no metabolites associated with stroke,[25] whereas another study reported association between lysophosphatidylcholine and stroke recurrence.[27] This asks for larger metabolomics studies of stroke in persons of European origin as presented in this thesis.

Cerebrovascular disease is also an important cause of dementia and cognitive decline.[28] A large number of genes have been implicated in dementia, predominately Alzheimer's disease (AD) but also frontotemporal dementia and Lewy body dementia.[29-32] The growing interest in early prevention of AD and cognitive decline, brought research of cognition in in the spotlight. Also there has been major progress in finding genes for cognitive function as endophenotype for various neurological and psychiatric disorders.[10-13,15] The major cognitive domains that have been studied in relation to these disorders include memory, language, executive function, and visuospatial ability.[10-13,15] Although the search for genes implicated in specific domains of cognition yielded some genes (e.g. *CADM2*, *HS3ST4*, *SPOCK3*),[33,34] the gene discovery improved its success when using general cognitive function, which captures all cognitive subdomains and shows a high correlation with intelligence and education.[35,36] General cognitive function is determined by environmental and genetic factors. Heritability estimates are reported to be more than 50% in adolescence and adulthood twin sample and 20-30% of variance is attributed to common variants.[35,37-39] Recent efforts identified more than 140 genomic regions encompassing common variants.[39] Furthermore, recent effort also reported evidence for a shared genetic origin with body mass index, waist to hip ratio, high-density lipoprotein levels, and cardiovascular diseases.[39] Even though these are drivers of the human metabolism, we have not linked yet genetic determinants of general cognitive function to circulating metabolites. Furthermore, most studies conducted to date included participants of European ancestry and a question to answer is whether the findings are generalizable to other ethnic groups. In this thesis I aim to find genetic determinants of general cognitive function, evaluate their generalizability to other ethnic groups and explore metabolic pathophysiology underlying established genetic variants. Despite all

efforts to date, common variants explain only a small proportion of cognitive test scores. Furthermore, diverse environmental factors have been implicated to influence cognitive function and the complex balance between genes and environment to cognitive function is poorly understood.[36,40]As studying epigenetic modifications may provide insights into molecular mechanisms underlying cognitive function, in this thesis we made an attempt to identify DNA methylation signatures of cognitive function.

At present, imaging is emerging as an endophenotype used in large-scale research of neurodegenerative disorders and stroke.[9] Finding genetic loci that influence this endophenotype may lead to identification of genes underlying related disorders. Studying brain structures using magnetic resonance imaging (MRI),[41,42] carotid intima-media thickness measured by carotid ultrasound and carotid artery calcification measured through computerized tomography (CT)[9,43] will expand our knowledge and provide novel insights into the pathophysiology of related disorders. In this thesis, I aim to explore genetic determinants of lateral ventricular volume, a measure of neurodegeneration, and intima-media thickness of carotid artery. Further, I aim to study metabolic determinants of carotid artery calcification, a measure of atherosclerosis.

NEURODEVELOPMENTAL DISORDERS

The most common neurodevelopmental disorders are ASD and ADHD.[44]

ASD is characterized by deficits in social communication and social interaction and restricted and repetitive patterns of activities and behavior.[45] The importance of genetic etiology is highlighted by heritability estimates ranging from 37% to 90%.[46-49] Progress in understanding genetic architecture of ASD has been made by identifying rare and de novo structural and sequence variation.[50,51] From a genetic perspective, ASD is an interesting disorder, as novel mutations have been implicated in patients that are not found back in either parent.[50] These variants have been identified in family-based studies.[52] Although most of the genetic risk for ASD is attributed to common variants, only a few genetic regions were successfully linked to ASD in family-based and population-based studies including unrelated patients and controls.[49,53-56] Despite a substantial increase in sample size, the most recent effort including over 16,000 individuals with ASD failed to identify common genetic variants associated with ASD asking for other approaches.[57] In this thesis, besides assessing the effect of single variants on ASD, I aim to evaluate the joint effect of multiple single genetic variants in a gene in a gene-based association analysis in patients with ASD.

ADHD is characterized by age-inappropriate inattentiveness, increased impulsivity and hyperactivity.[58] Heritability estimates in childhood are reported to be 70-80%, whereas estimates in adults show moderate heritability of 30-40%.[59] Several candidate genes have been associated with ADHD.[60] Although 10-28% of genetic risk is attributed to common variants,[61,62] the first risk loci with a high frequency have been reported recently.[63] Several of these loci are located near or in genes implicated in neurodevelopmental processes including *FOXP2* and *DUSP6*.[63] ADHD has been regarded as the extreme end of continuous distribution of inattentiveness and/or hyperactivity,[64-66] just like hypertension is the extreme of the continuous distribution of blood pressure in the population. As ADHD diagnosis is the extreme end of a continuous ADHD symptom scores[67] and genetic factors for ADHD diagnosis and ADHD symptoms showed an overlap,[67] novel more powerful approaches involving continuous measures in population-based setting could provide an opportunity to discover additional common variants and detect genes underlying ADHD. I aim to use this approach in order to evaluate contribution of common genetic variants in ADHD symptoms.

Furthermore, I have also studied Duchenne muscular dystrophy (DMD), the most common form of muscular dystrophy during childhood caused by mutations in dystrophin gene (*DMD*).[68] This fatal disease leads to progressive muscular weakness and less well described non-progressive central nervous system manifestations. As the risk of cognitive impairment is increased among the patients with DMD and higher occurrence of various neurodevelopmental disorders such as ASD and ADHD is also reported,[69-75] I address the question in this thesis whether *DMD* gene has an effect in general populations.

MOLECULAR APPROACHES USED IN THIS THESIS

To improve our understanding of the pathogenesis and heterogeneity in diseases and to facilitate development of personalized and more precise prevention and treatment, various omics approaches may be used to study changes underlying diseases at the molecular level. Omics approaches refer to large-scale high throughput technologies.[76,56] These technologies cover different molecular layers from the level of DNA (genomics) to DNA methylation/histone modification (epigenomics), RNA (transcriptomics), proteins (proteomics), and metabolites (metabolomics) as depicted in **Figure 1**. Furthermore, omics technologies also address microorganisms colonizing human body (microbiomics).

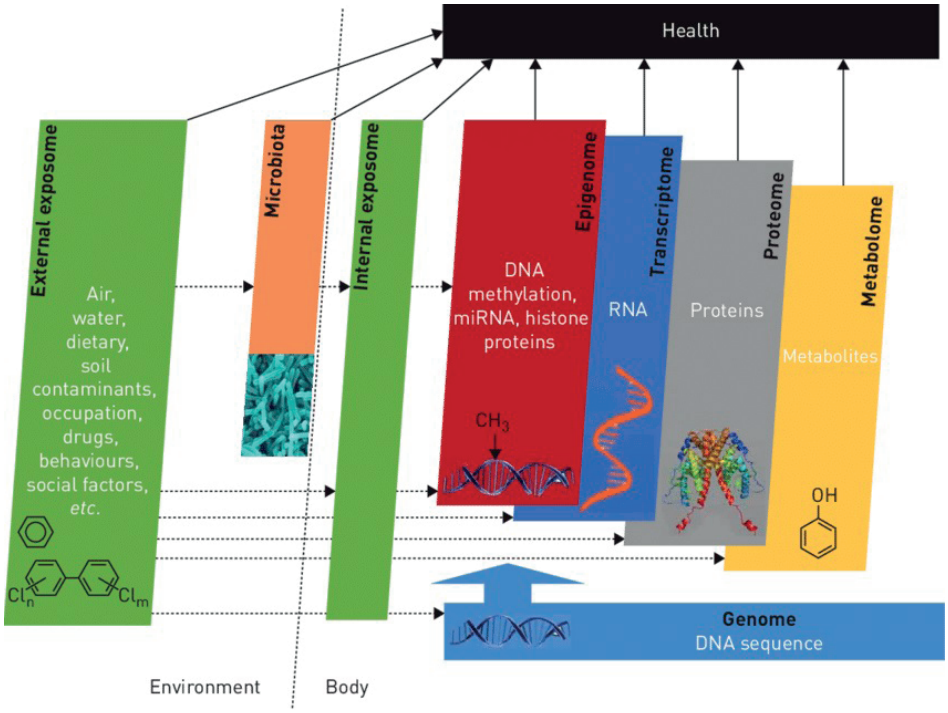


Figure 1. Diagram of omics layers and some of the interactions between them. Source: Siroux *et al.*[77]

In this thesis, I concentrated on several omics approaches including genomics, epigenomics, metabolomics, and microbiomics in relation to neurological and psychiatric disorders.

Genomics

The human genome captures all variations in our DNA, the blueprint of our proteins. Focusing on whole human genome, genomics provides important insights into genetic architecture of complex disorders, which involve effects of rare and common variants and variants conveying a small or large effect on pathology. Genetic determinants including single nucleotide polymorphisms (SNPs) or structural variation (SV) can be found in either protein-coding regions and may impact sequence of the protein or in non-coding regions more likely affecting gene expression and splicing processes.[78-80] Contribution of genetic variants commonly occurring in general population (minor allele frequency (MAF) > 5%) is often assessed by genome-wide association studies (GWAS).[81] The genetic variants often have a small effect on the trait. Although their individual effect is not informative, the joint effect is for a large part determining the risk of common diseases, as predicted by RA Fisher even before the structure of DNA was unraveled.[82] Thus, common variants provide important insights into the biology,

unravelling the pathological pathways, and jointly improve the proportion of variance explained by genetic factors, surpassing that of important epidemiological factors such as that of body mass index (BMI) on lipid levels.[83,84] Availability of relatively inexpensive SNP arrays and the possibility of imputing variants using large reference panels such as 1000 Genomes and Haplotype Reference Consortium (HRC), enabled the number of genetic variants for association testing to be increased and facilitated meta-analyses of studies using different arrays. [85-87] This resulted in mega GWAS of large sample size (currently up to a million).[88] A typical GWAS design involves hypothesis-free discovery study followed by replication of the associations in an independent sample.[89,90] Both

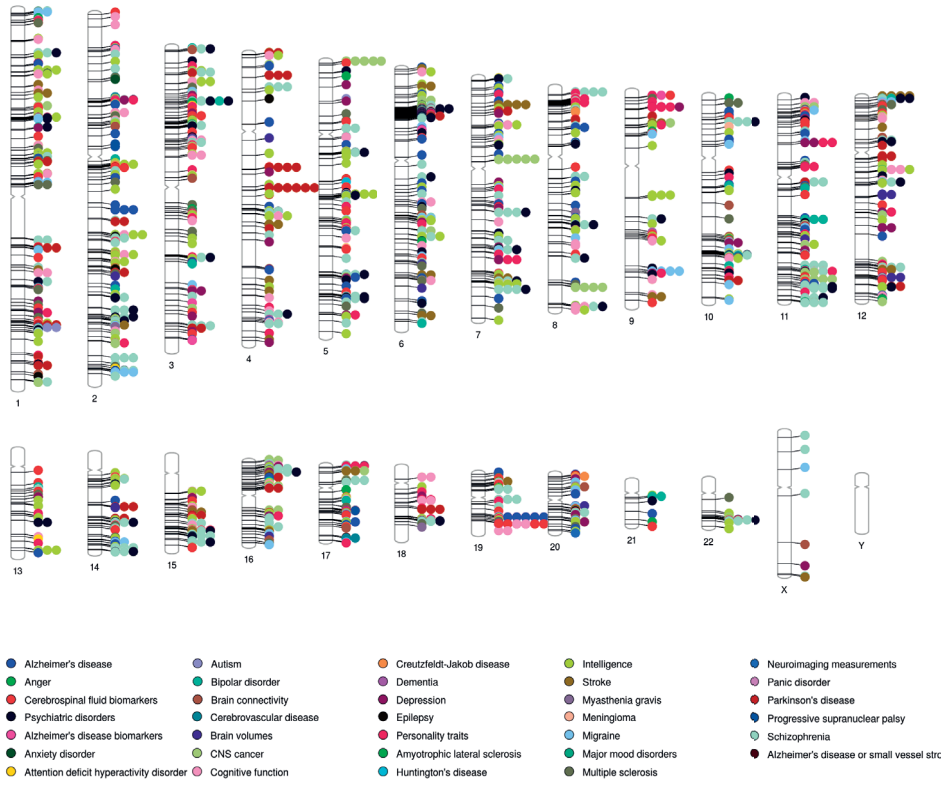


Figure 2. Associations of neurological and psychiatric disorders with SNPs accross the genome (GWAS catalog as of April 2018).[97] The genome is displayed divided into separate chromosomes. Color denotes disorders.

the discovery and replication are subjected to a stringent level of significance, adjusting for the large number of tests with the low *a priori* probability of association.[91] To date, more than 800 associations have been reported between the SNPs and neurological and psychiatric disorders (GWAS catalog as of April 2018) (**Figure 2**). Identified associations

not only confirmed previously identified genes (e.g. *APOE* locus was firstly identified in AD families followed by association analyses and later on replicated in GWAS)[92-94] but also identified novel genetic regions.[95] Additionally, GWAS provided opportunity to explore genetic architecture between the various complex disorders with methods such as LD score regression.[96]

The big data meta-analyses allowed to include more low-frequency and rare variants (MAF < 5%) in the GWAS.[98] However, there is a limit in that very rare variants are difficult to impute.[98] Thus, GWAS is unable to systematically explore the contribution of the rare variants which could also contribute to the genetic architecture and explain "missing heritability".[99,100] More importantly, these rare variants are key to personalized and precision prevention, e.g. as occurred in the prevention of breast cancer in *BRCA1/2* carriers through preventive mastectomy[101] and early mortality in carriers of *LDLR* mutations through treatment with statins starting in early adolescence.[102] Development of next-generation sequencing technologies including whole-genome sequencing (WGS) and whole-exome sequencing (WES) allowed detection of low-frequency or rare variants with large or moderate effects.[79,103] Applied to neurological and psychiatric disorders, some of the examples of success to date include discovery of rare coding variant in *TREM2* associated with AD,[104,105] rare variant in *VPS35* associated with Parkinson disease,[106,107] and several rare variants underlying the genetic etiology of ASD.[108] The development of dedicated rare variant arrays (e.g. the exome arrays), allowed the application of GWAS for rare variants in large datasets, i.e. as was successfully done for AD.[109] With increasing application to other disorders, more discoveries are underway, using both classical family-based methods as well as GWAS methodology.[110]

Epigenomics

Epigenomics focuses on genome-wide characterization of chemical modifications of DNA or DNA-associated proteins such as DNA methylation or histone modification.[111] Those modifications of DNA and histones play important role in the regulation of gene expression without changing the DNA sequence and are influenced by both genetic and environmental factors.[112] The most studied and best characterized epigenetic modification is DNA methylation -- addition of methyl group to the CpG sites of the DNA molecule. DNA methylation is essential for regulating X chromosome inactivation, genomic imprinting, and tissue-specific gene expression.[113,114] The pattern of DNA methylation established either during development[114] or late in life can have consequences within the brain. Abnormal methylation in *FMR1* gene causes mental retardation (Fragile X Syndrome),[115] whereas improper methylation of a single imprinted allele causes mental impairment (Prader-Willi Syndrome).[116,117] Late in life, environmental risk factors may have major impact, e.g. smoking and obesity-related pathologies are known to be

major determinants of expression.[118] With the development of epigenome-wide studies (EWAS), an opportunity to study DNA methylation pattern underlying complex neurological and psychiatric disorders has become available. Although methylation may be tissue-specific, there are many instances reported where there is a high correlation between the methylation in the brain and in blood.[119-121] Alteration of DNA methylation pattern has been observed in both psychiatric disorders such as schizophrenia and bipolar disorder and neurodegenerative disorders such as dementias.[122,123] Even though our understanding of the role of epigenetics in etiology of neurological and psychiatric disorders is still limited and may involve not only methylation but also acetylation in the brain,[124,125] epigenomics holds great potential for identifying useful biomarkers that could contribute to unraveling underlying mechanisms of these disorders. In addition to the etiological significance of methylation, one may speculate that methylation may possibly lead not only to timely diagnosis but also defining preclinical stages of disorders.

Metabolomics

The rapid development of new technologies enabled quantification of substrates and products of metabolism referred to as metabolites.[126] These low molecular weight compounds are influenced by genetic factors, lifestyle factors, pharmacological treatments, mechanisms of disease, and microbiota.[126,127] Last but not least, metabolites may reflect the disease process and may be a cause rather than a consequence of disease.

Identifying the metabolites and metabolic pathways has a potential to provide new insights into pathophysiology and for discovery of new diagnostic markers for disease risk that could facilitate the development of novel and precise diagnostic tools, and treatment and preventive strategies.[128,129] Metabolic profiling of biological fluids, including blood, urine, and cerebrospinal fluid, and tissues holds great potential for investigation of neurological and psychiatric disorders. Thousands of metabolites may be detected by targeted approaches, whereas this number increases if untargeted approaches are applied.[130] Although metabolite processes may be tissue specific, there is growing interest in vascular origin of neurodegeneration and cerebrovascular pathology. To date, metabolic profiling has been reported for various psychiatric disorders such as schizophrenia, bipolar disorder, and neurological conditions including AD and stroke.[24,131-137] However, not all studies performed to date were well powered, emphasizing need to explore metabolomics profiles in large epidemiological follow-up studies.

Microbiomics

Microbiomics focuses on microorganisms colonizing different parts of human body, such as skin (skin microbiota), the mouth (oral microbiota), the gut (gut microbiota) and

so on. The gut harbors thousands of microbial species which are considered to be a central signaling hub that integrates environmental inputs summarized as exposome (e.g. diet, life style, medication) with genetic and immune signals to affect the host's metabolism.[138] Gut microbiota is responsible for several functions including food digestion, vitamin and short chain acid (SCFA) production, amino acid synthesis, activation

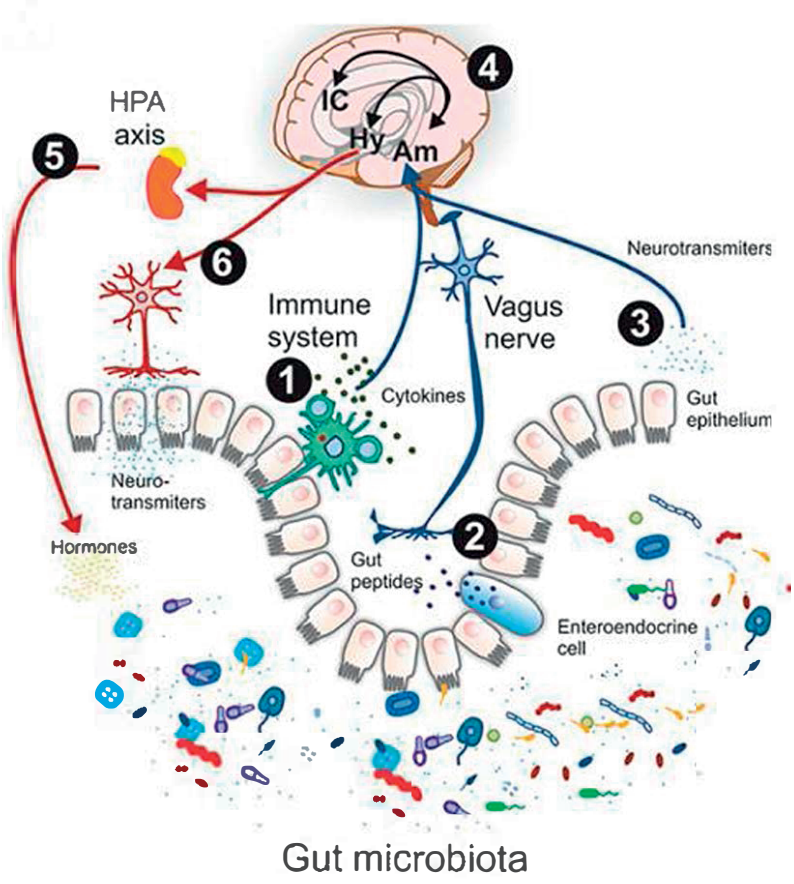


Figure 3. Bidirectional interaction between the gut microbiota and the central nervous system involving direct and indirect endocrine, immune and neural pathways. For instance: (1) cytokines released by lymphocytes which may sense the gut lumen can have endocrine or paracrine actions, (2) gut peptides released by enteroendocrine cells may activate sensory neuronal terminals, such as on the vagus nerve, (3) microbiota metabolites (neurotransmitters or its precursors) may reach the gut epithelium having endocrine or paracrine effects. (4) Centrally, after brainstem relays (e.g. nucleus *tractus solitarii*) a neural network involving the amygdala (Am) and the insular cortex (IC) integrates visceral inputs. Consistently hypothalamic (Hy) activation initiates: (5) corticosteroids release (results of the hypothalamic-pituitary-adrenal (HPA) axis activation) which modulates gut microbiota composition, (6) neuronal efferent activation ("anti-inflammatory cholinergic reflex" and/or sympathetic activation) liberating neurotransmitters that may affect the gut microbiota composition. Source: Montiel-Castro et al.[150]

of certain drugs, signaling molecules and anti-microbial compounds production, bile acid biotransformation and development of our immune system.[138-141] With advances in technology of microbial phenotyping methods, gut microbiota has been implicated in various neurological and psychiatric disorders and has been linked to cognitive ability, neurodevelopmental disorders (e.g. ASD), and neurodegenerative disorders (e.g. Parkinson disease, Alzheimer's disease).[142-148] The gut-brain axis has been long recognized. As depicted in **Figure 3** it involves metabolic and immune signals from the gut to the brain and vice versa from the brain to the gut and direct nerve innervation (nervus vagus).[148,149] Understanding mechanisms of complex nature of host-microbiome metabolism may help develop new strategies for preventing and treating diseases. In this thesis, we aim to explore link between gut microbiota and the metabolome.

AIM OF THIS THESIS

The aim of this thesis is to identify genomic and metabolomic determinants underlying neurological and psychiatric diseases and their related endophenotypes.

In **Chapter 2** omics studies of neurodegeneration are described. **Chapter 2.1** explores genetic determinants of brain structures determined by brain MRI. More specifically, I examine contribution of common genetic variants underlying lateral ventricular volume. Subsequently, other endophenotypes of neurological and psychiatric disorders are explored. Firstly, **Chapter 2.2** addresses common genetic determinants of general cognitive function and furthermore explores metabolic pathophysiology underlying established genetic variants implicated in cognitive ability. Then **Chapter 2.3** provides insights into complex DNA methylation signatures in relation to cognitive function. Finally, **Chapter 2.4** and **Chapter 2.5** apply candidate gene approach to study effect of rare variants mapped to a dystrophin gene on cognitive ability in general population and to determine whether the location of mutations in dystrophin gene and its impact on specific dystrophin isoforms has an effect on cognitive ability.

Chapter 3 addresses determinants of neurovascular pathology. In **Chapter 3.1**, contribution of rare genetic variants underlying carotid intima-media thickness is studied. Carotid intima-media thickness is a marker of subclinical atherosclerosis that predicts future cardiovascular events. **Chapter 3.2** addresses associations of metabolites measured by state-of-the-art metabolomics and carotid artery calcification, whereas **Chapter 3.3** focusses on metabolomic determinants of stroke in large prospective population-based studies including participants of European ancestry. **Chapter 3.4** provides insights into the relationship between gut microbiota and circulating metabolites.

Chapter 4 focusses on genetic determinants of neurodevelopmental disorders. **Chapter 4.1** explores genetic determinants in ASD, whereas the contribution of common genetic variants in ADHD symptoms is evaluated in **Chapter 4.2.**

Finally, **Chapter 5** summarizes the main findings of this thesis and provides suggestions for future research. **Chapter 5.1** describes major findings and in **Chapter 5.2** information derived from the genomic research of cardiovascular disorders is used to develop translational models aiming at effective prevention programs, earlier diagnosis and prognosis, and individualized treatments.

REFERENCES

1. Azevedo, F.A.C. *et al.* Equal Numbers of Neuronal and Nonneuronal Cells Make the Human Brain an Isometrically Scaled-Up Primate Brain. *Journal of Comparative Neurology* **513**, 532-541 (2009).
2. Qureshi, I.A. & Mehler, M.F. An evolving view of epigenetic complexity in the brain. *Philosophical Transactions of the Royal Society B-Biological Sciences* **369**(2014).
3. Whiteford, H.A., Ferrari, A.J., Degenhardt, L., Feigin, V. & Vos, T. The Global Burden of Mental, Neurological and Substance Use Disorders: An Analysis from the Global Burden of Disease Study 2010. *Plos One* **10**(2015).
4. Wittchen, H.U. *et al.* The size and burden of mental disorders and other disorders of the brain in Europe 2010. *Eur Neuropsychopharmacol* **21**, 655-79 (2011).
5. Group, G.B.D.N.D.C. Global, regional, and national burden of neurological disorders during 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet Neurol* **16**, 877-897 (2017).
6. Burggren, A. & Brown, J. Imaging markers of structural and functional brain changes that precede cognitive symptoms in risk for Alzheimer's disease. *Brain Imaging and Behavior***.8**, pp (2014).
7. Vermeer, S.E. *et al.* Silent brain infarcts and the risk of dementia and cognitive decline. *N Engl J Med* **348**, 1215-22 (2003).
8. Bearden, C.E. & Freimer, N.B. Endophenotypes for psychiatric disorders: ready for primetime? *Trends Genet* **22**, 306-13 (2006).
9. Markus, H.S. Stroke genetics. *Hum Mol Genet* **20**, R124-31 (2011).
10. Bennett, D.A., De Jager, P.L., Leurgans, S.E. & Schneider, J.A. Neuropathologic intermediate phenotypes enhance association to Alzheimer susceptibility alleles. *Neurology* **72**, 1495-1503 (2009).
11. Kohler, S. *et al.* Progression to dementia in memory clinic patients without dementia: A latent profile analysis. *Neurology***.81**, pp (2013).
12. Gur, R.E. *et al.* Neurocognitive endophenotypes in a multiplex multigenerational family study of schizophrenia. *Am J Psychiatry* **164**, 813-9 (2007).
13. Hu, M.R. *et al.* Semantic fluency and executive functions as candidate endophenotypes for the early diagnosis of schizophrenia in Han Chinese. *Neuroscience Letters* **502**, 173-177 (2011).
14. Glahn, D.C. *et al.* Neurocognitive Endophenotypes for Bipolar Disorder Identified in Multiplex Multigenerational Families. *Archives of General Psychiatry* **67**, 168-177 (2010).
15. Shang, C.Y., Gau, S.S., Id, Shang, C.Y.O.h.o.o. & Gau, S.S.O.h.o.o. Visual memory as a potential cognitive endophenotype of attention deficit hyperactivity disorder. *Psychological Medicine***.41**, pp (2011).
16. Meschia, J.F., Worrall, B.B. & Rich, S.S. Genetic susceptibility to ischemic stroke. *Nature Reviews Neurology* **7**, 369-378 (2011).
17. Weinstein, G. *et al.* Brain Imaging and Cognitive Predictors of Stroke and Alzheimer Disease in the Framingham Heart Study. *Stroke* **44**, 2787-2794 (2013).
18. Burggren, A. & Brown, J. Imaging markers of structural and functional brain changes that precede cognitive symptoms in risk for Alzheimer's disease. *Brain imaging behav.* **8**, 251-61 (2014).
19. 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).
20. Andersen, K.K., Olsen, T.S., Dehlendorff, C. & Kammersgaard, L.P. Hemorrhagic and ischemic strokes compared: stroke severity, mortality, and risk factors. *Stroke* **40**, 2068-72 (2009).
21. Boehme, A.K., Esenwa, C. & Elkind, M.S. Stroke Risk Factors, Genetics, and Prevention. *Circ Res* **120**, 472-495 (2017).

22. Wang, D., Kong, J., Wu, J., Wang, X. & Lai, M. GC-MS-based metabolomics identifies an amino acid signature of acute ischemic stroke. *Neurosci Lett* **642**, 7-13 (2017).
23. Lee, Y., Khan, A., Hong, S., Jee, S.H. & Park, Y.H. A metabolomic study on high-risk stroke patients determines low levels of serum lysine metabolites: a retrospective cohort study. *Mol Biosyst* **13**, 1109-1120 (2017).
24. Jung, J.Y. *et al.* 1H-NMR-based metabolomics study of cerebral infarction. *Stroke* **42**, 1282-8 (2011).
25. Floegel, A. *et al.* Serum metabolites and risk of myocardial infarction and ischemic stroke: a targeted metabolomic approach in two German prospective cohorts. *Eur J Epidemiol* **33**, 55-66 (2018).
26. Holmes, M.V. *et al.* Lipids, Lipoproteins, and Metabolites and Risk of Myocardial Infarction and Stroke. *J Am Coll Cardiol* **71**, 620-632 (2018).
27. Jove, M. *et al.* Metabolomics predicts stroke recurrence after transient ischemic attack. *Neurology* **84**, 36-45 (2015).
28. Knopman, D.S. Dementia and cerebrovascular disease. *Mayo Clin Proc* **81**, 223-30 (2006).
29. Van Cauwenberghe, C., Van Broeckhoven, C. & Sleegers, K. The genetic landscape of Alzheimer disease: clinical implications and perspectives. *Genet Med* **18**, 421-30 (2016).
30. Olszewska, D.A., Lonergan, R., Fallon, E.M. & Lynch, T. Genetics of Frontotemporal Dementia. *Curr Neurol Neurosci Rep* **16**, 107 (2016).
31. Vergouw, L.J.M. *et al.* An update on the genetics of dementia with Lewy bodies. *Parkinsonism Relat Disord* **43**, 1-8 (2017).
32. Meeus, B., Theuns, J. & Van Broeckhoven, C. The genetics of dementia with Lewy bodies: what are we missing? *Arch Neurol* **69**, 1113-8 (2012).
33. Ibrahim-Verbaas, C.A. *et al.* GWAS for executive function and processing speed suggests involvement of the CADM2 gene. *Mol Psychiatry* **21**, 189-197 (2016).
34. Debette, S. *et al.* Genome-wide studies of verbal declarative memory in nondemented older people: the Cohorts for Heart and Aging Research in Genomic Epidemiology consortium. *Biol Psychiatry* **77**, 749-63 (2015).
35. Davies, G. *et al.* Genetic contributions to variation in general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE consortium (N=53 949). *Molecular Psychiatry* **20**, 183-192 (2015).
36. Deary, I.J., Johnson, W. & Houlihan, L.M. Genetic foundations of human intelligence. *Hum Genet* **126**, 215-32 (2009).
37. Haworth, C.M. *et al.* The heritability of general cognitive ability increases linearly from childhood to young adulthood. *Mol Psychiatry* **15**, 1112-20 (2010).
38. Davies, G. *et al.* Genome-wide association study of cognitive functions and educational attainment in UK Biobank (N=112 151). *Mol Psychiatry* **21**, 758-67 (2016).
39. Davies, G. *et al.* Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nat Commun* **9**, 2098 (2018).
40. Anstey, K.J. Optimizing cognitive development over the life course and preventing cognitive decline: Introducing the Cognitive Health Environment Life Course Model (CHELM). *International Journal of Behavioral Development* **38**, 1-10 (2014).
41. Stein, J.L. *et al.* Identification of common variants associated with human hippocampal and intracranial volumes. *Nat Genet* **44**, 552-61 (2012).
42. Adams, H.H.H. *et al.* Novel genetic loci underlying human intracranial volume identified through genome-wide association. *Nature Neuroscience***.19**, pp (2016).

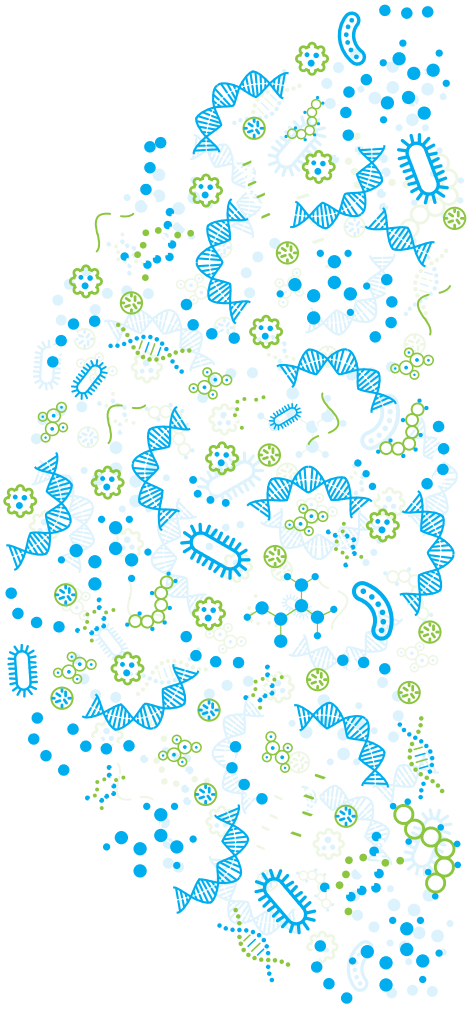
43. Wendell, C.R. *et al.* Carotid atherosclerosis and prospective risk of dementia. *Stroke* **43**, 3319-24 (2012).
44. Thapar, A., Cooper, M. & Rutter, M. Neurodevelopmental disorders. *Lancet Psychiatry* **4**, 339-246 (2017).
45. Huguet, G., Ey, E. & Bourgeron, T. The genetic landscapes of autism spectrum disorders. *Annu Rev Genomics Hum Genet* **14**, 191-213 (2013).
46. Hallmayer, J. *et al.* Genetic Heritability and Shared Environmental Factors Among Twin Pairs With Autism. *Archives of General Psychiatry* **68**, 1095-1102 (2011).
47. Bailey, A. *et al.* Autism as a Strongly Genetic Disorder - Evidence from a British Twin Study. *Psychological Medicine* **25**, 63-77 (1995).
48. Sandin, S. *et al.* The Familial Risk of Autism. *Jama-Journal of the American Medical Association* **311**, 1770-1777 (2014).
49. Gaugler, T. *et al.* Most genetic risk for autism resides with common variation. *Nature Genetics* **46**, 881-885 (2014).
50. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-U136 (2014).
51. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368-372 (2010).
52. McClellan, J. & King, M.C. Genetic heterogeneity in human disease. *Cell* **141**, 210-7 (2010).
53. Wang, K. *et al.* Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* **459**, 528-533 (2009).
54. Ma, D.Q. *et al.* A Genome-wide Association Study of Autism Reveals a Common Novel Risk Locus at 5p14.1. *Annals of Human Genetics* **73**, 263-273 (2009).
55. Weiss, L.A., Arking, D.E. & Consortium, J.H.A. A genome-wide linkage and association scan reveals novel loci for autism. *Nature* **461**, 802-U62 (2009).
56. Anney, R. *et al.* Individual common variants exert weak effects on the risk for autism spectrum disorders. *Hum Mol Genet* **21**, 4781-92 (2012).
57. Anney, R.J.L. *et al.* Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. *Molecular Autism* **8**(2017).
58. American Psychiatric Association Diagnostic and Statistical Manual of Mental Disorders4th edn, revised. *American Psychiatric Press: Washington, DC* (2000).
59. Brikell, I., Kuja-Halkola, R. & Larsson, H. Heritability of attention-deficit hyperactivity disorder in adults. *Am J Med Genet B Neuropsychiatr Genet* **168**, 406-413 (2015).
60. Hawi, Z. *et al.* The molecular genetic architecture of attention deficit hyperactivity disorder. *Molecular Psychiatry* **20**, 289-297 (2015).
61. Cross-Disorder Group of the Psychiatric Genomics, C. *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet* **45**, 984-94 (2013).
62. Anttila, V. Analysis of shared heritability in common disorders of the brain. *Science* **360**.6395 (2018).
63. Demontis, D., Waters, R.K. & Martin, J. Discovery of the first genome-wide significant risk loci for ADHD. *bioRxiv* (2017).
64. Lubke, G.H., Hudziak, J.J., Derks, E.M., van Bijsterveldt, T.C. & Boomsma, D.I. Maternal ratings of attention problems in ADHD: evidence for the existence of a continuum. *J Am Acad Child Adolesc Psychiatry* **48**, 1085-93 (2009).

65. Levy, F., Hay, D.A., McStephen, M., Wood, C. & Waldman, I. Attention-deficit hyperactivity disorder: a category or a continuum? Genetic analysis of a large-scale twin study. *J Am Acad Child Adolesc Psychiatry* **36**, 737-44 (1997).
66. Larsson, H., Anckarsater, H., Rastam, M., Chang, Z. & Lichtenstein, P. Childhood attention-deficit hyperactivity disorder as an extreme of a continuous trait: a quantitative genetic study of 8,500 twin pairs. *Journal of Child Psychology and Psychiatry* **53**, 73-80 (2012).
67. Middeldorp, C.M. *et al.* A Genome-Wide Association Meta-Analysis of Attention-Deficit/Hyperactivity Disorder Symptoms in Population-Based Pediatric Cohorts. *J Am Acad Child Adolesc Psychiatry* **55**, 896-905 e6 (2016).
68. Hoffman, E.P., Brown, R.H. & Kunkel, L.M. Dystrophin - the Protein Product of the Duchenne Muscular-Dystrophy Locus. *Cell* **51**, 919-928 (1987).
69. Cyrulnik, S.E., Fee, R.J., De Vivo, D.C., Goldstein, E. & Hinton, V.J. Delayed developmental language milestones in children with Duchenne's muscular dystrophy. *Journal of Pediatrics* **150**, 474-478 (2007).
70. Wicksell, R.K., Kihlgren, M., Melin, L. & Eeg-Olofsson, O. Specific cognitive deficits are common in children with Duchenne muscular dystrophy. *Developmental Medicine and Child Neurology* **46**, 154-159 (2004).
71. Hinton, V.J., De Vivo, D.C., Nereo, N.E., Goldstein, E. & Stern, Y. Poor verbal working memory across intellectual level in boys with Duchenne dystrophy. *Neurology* **54**, 2127-2132 (2000).
72. Mento, G., Tarantino, V. & Bisiacchi, P.S. The Neuropsychological Profile of Infantile Duchenne Muscular Dystrophy. *Clinical Neuropsychologist* **25**, 1359-1377 (2011).
73. D'Angelo, M.G. *et al.* Neurocognitive Profiles in Duchenne Muscular Dystrophy and Gene Mutation Site. *Pediatric Neurology* **45**, 292-299 (2011).
74. Hendriksen, J.G.M. & Vles, J.S.H. Neuropsychiatric disorders in males with Duchenne muscular dystrophy: Frequency rate of attention-deficit hyperactivity disorder (ADHD), autism spectrum disorder, and obsessive-compulsive disorder. *Journal of Child Neurology* **23**, 477-481 (2008).
75. Wu, J.Y., Kuban, K.C., Allred, E., Shapiro, F. & Darras, B.T. Association of Duchenne muscular dystrophy with autism spectrum disorder. *J Child Neurol* **20**, 790-5 (2005).
76. Valdes, A.M., Glass, D. & Spector, T.D. Omics technologies and the study of human ageing. *Nat Rev Genet* **14**, 601-7 (2013).
77. Siroux, V., Agier, L. & Slama, R. The exposome concept: a challenge and a potential driver for environmental health research. *Eur Respir Rev* **25**, 124-9 (2016).
78. Gonzaga-Jauregui, C., Lupski, J.R. & Gibbs, R.A. Human genome sequencing in health and disease. *Annu Rev Med* **63**, 35-61 (2012).
79. Metzker, M.L. Sequencing technologies - the next generation. *Nat Rev Genet* **11**, 31-46 (2010).
80. Zhang, F. & Lupski, J.R. Non-coding genetic variants in human disease. *Hum Mol Genet* **24**, R102-10 (2015).
81. Visscher, P.M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**, 5-22 (2017).
82. Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* **40**, 695-701 (2008).
83. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565-9 (2010).
84. Aulchenko, Y.S. *et al.* Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat Genet* **41**, 47-55 (2009).

85. Bush, W.S. & Moore, J.H. Chapter 11: Genome-wide association studies. *PLoS Comput Biol* **8**, e1002822 (2012).
86. Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
87. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-83 (2016).
88. Evangelou, E. *et al.* Genetic analysis of over one million people identifies 535 novel loci for blood pressure. *bioRxiv* (2017).
89. Pe'er, I., Yelensk, R., Altshuler, D. & Daly, M.J. Estimation of the multiple testing burden for genome-wide association studies of nearly all common variants. *Genetic Epidemiology* **32**, 381-385 (2008).
90. Clayton, D.G. *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genetics* **37**, 1243-1246 (2005).
91. Broer, L. *et al.* Distinguishing true from false positives in genomic studies: p values. *Eur J Epidemiol* **28**, 131-8 (2013).
92. Corder, E.H. *et al.* Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**, 921-3 (1993).
93. Pericak-Vance, M.A. *et al.* Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. *Am J Hum Genet* **48**, 1034-50 (1991).
94. Saunders, A.M. *et al.* Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. *Neurology* **43**, 1467-72 (1993).
95. Lambert, J.C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics* **45**, 1452-U206 (2013).
96. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).
97. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* **45**, D896-D901 (2017).
98. Iglesias, A.I. *et al.* Haplotype reference consortium panel: Practical implications of imputations with large reference panels. *Human Mutation* **38**, 1025-1032 (2017).
99. Price, A.L., Spencer, C.C. & Donnelly, P. Progress and promise in understanding the genetic basis of common diseases. *Proc Biol Sci* **282**, 20151684 (2015).
100. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456**, 18-21 (2008).
101. Olopade, O.I., Grushko, T.A., Nanda, R. & Huo, D. Advances in breast cancer: pathways to personalized medicine. *Clin Cancer Res* **14**, 7988-99 (2008).
102. Neil, A. *et al.* Reductions in all-cause, cancer, and coronary mortality in statin-treated patients with heterozygous familial hypercholesterolaemia: a prospective registry study. *Eur Heart J* **29**, 2625-33 (2008).
103. Guerreiro, R., Bras, J., Hardy, J. & Singleton, A. Next generation sequencing techniques in neurological diseases: redefining clinical and molecular associations. *Human Molecular Genetics* **23**, R47-R53 (2014).
104. Guerreiro, R. *et al.* TREM2 variants in Alzheimer's disease. *N Engl J Med* **368**, 117-27 (2013).
105. Jonsson, T. *et al.* Variant of TREM2 associated with the risk of Alzheimer's disease. *N Engl J Med* **368**, 107-16 (2013).
106. Zimprich, A. *et al.* A mutation in VPS35, encoding a subunit of the retromer complex, causes late-onset Parkinson disease. *Am J Hum Genet* **89**, 168-75 (2011).
107. Vilarino-Guell, C. *et al.* VPS35 mutations in Parkinson disease. *Am J Hum Genet* **89**, 162-7 (2011).

108. Krumm, N., O'Roak, B.J., Shendure, J. & Eichler, E.E. A de novo convergence of autism genetics and molecular neuroscience. *Trends in Neurosciences* **37**, 95-105 (2014).
109. Sims, R. *et al.* Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. *Nature Genetics* **49**, 1373-+ (2017).
110. Sanders, S.J. *et al.* Whole Genome Sequencing in Psychiatric Disorders: the WGSPD Consortium. *bioRxiv* (2017).
111. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biology* **18** (2017).
112. Smith, Z.D. & Meissner, A. DNA methylation: roles in mammalian development. *Nature Reviews Genetics* **14**, 204-220 (2013).
113. Moore, L.D., Le, T. & Fan, G. DNA methylation and its basic function. *Neuropsychopharmacology* **38**, 23-38 (2013).
114. Smith, Z.D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat Rev Genet* **14**, 204-20 (2013).
115. Verkerk, A.J.M.H. *et al.* Identification of a Gene (Fmr-1) Containing a Cgg Repeat Coincident with a Breakpoint Cluster Region Exhibiting Length Variation in Fragile-X Syndrome. *Cell* **65**, 905-914 (1991).
116. Buiting, K. Prader-Willi Syndrome and Angelman Syndrome. *American Journal of Medical Genetics Part C-Seminars in Medical Genetics* **154c**, 365-376 (2010).
117. Moore, L.D., Le, T. & Fan, G.P. DNA Methylation and Its Basic Function. *Neuropsychopharmacology* **38**, 23-38 (2013).
118. Zeilinger, S. *et al.* Tobacco Smoking Leads to Extensive Genome-Wide Changes in DNA Methylation. *Plos One* **8** (2013).
119. Ewald, E.R. *et al.* Alterations in DNA methylation of Fkbp5 as a determinant of blood-brain correlation of glucocorticoid exposure. *Psychoneuroendocrinology* **44**, 112-22 (2014).
120. Horvath, S. *et al.* Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol* **13**, R97 (2012).
121. Masliah, E., Dumaop, W., Galasko, D. & Desplats, P. Distinctive patterns of DNA methylation associated with Parkinson disease: identification of concordant epigenetic changes in brain and peripheral blood leukocytes. *Epigenetics* **8**, 1030-8 (2013).
122. Mill, J. *et al.* Epigenomic profiling reveals DNA-Methylation changes associated with major psychosis. *American Journal of Human Genetics* **82**, 696-711 (2008).
123. Fransquet, P.D. *et al.* Blood DNA methylation as a potential biomarker of dementia: A systematic review. *Alzheimers & Dementia* **14**, 81-103 (2018).
124. Konsoula, Z. & Barile, F.A. Epigenetic histone acetylation and deacetylation mechanisms in experimental models of neurodegenerative disorders. *J Pharmacol Toxicol Methods* **66**, 215-20 (2012).
125. Lu, X., Wang, L., Yu, C.J., Yu, D.H. & Yu, G. Histone acetylation modifiers in the pathogenesis of Alzheimer's disease. *Frontiers in Cellular Neuroscience* **9**(2015).
126. Johnson, C.H., Ivanisevic, J. & Siuzdak, G. Metabolomics: beyond biomarkers and towards mechanisms. *Nat Rev Mol Cell Biol* **17**, 451-9 (2016).
127. Menni, C., Zierer, J., Valdes, A.M. & Spector, T.D. Mixing omics: combining genetics and metabolomics to study rheumatic diseases. *Nat Rev Rheumatol* **13**, 174-181 (2017).
128. Kaddurah-Daouk, R., Kristal, B.S. & Weinshilboum, R.M. Metabolomics: a global biochemical approach to drug response and disease. *Annu Rev Pharmacol Toxicol* **48**, 653-83 (2008).
129. Guest, P.C., Guest, F.L. & Martins-de Souza, D. Making Sense of Blood-Based Proteomics and Metabolomics in Psychiatric Research. *Int J Neuropsychopharmacol* **19**(2016).
130. Markley, J.L. *et al.* The future of NMR-based metabolomics. *Curr Opin Biotechnol* **43**, 34-40 (2017).

131. McIntyre, R.S. *et al.* Advancing biomarker research: utilizing 'Big Data' approaches for the characterization and prevention of bipolar disorder. *Bipolar Disorders* **16**, 531-547 (2014).
132. Paredes, R.M. *et al.* Metabolomic profiling of schizophrenia patients at risk for metabolic syndrome. *International Journal of Neuropsychopharmacology* **17**, 1139-1148 (2014).
133. Pickard, B.S. Schizophrenia biomarkers: Translating the descriptive into the diagnostic. *Journal of Psychopharmacology* **29**, 138-143 (2015).
134. Guest, P.C., Guest, F.L. & Martins-de Souza, D. Making Sense of Blood-Based Proteomics and Metabolomics in Psychiatric Research. *International Journal of Neuropsychopharmacology* **19** (2016).
135. Tynkkynen, J. *et al.* Association of branched-chain amino acids and other circulating metabolites with risk of incident dementia and Alzheimer's disease: A prospective study in eight cohorts. *Alzheimers Dement* (2018).
136. Sethi, S. & Brietzke, E. Omics-Based Biomarkers: Application of Metabolomics in Neuropsychiatric Disorders. *Int J Neuropsychopharmacol* **19**, pyv096 (2015).
137. van der Lee, S.J. *et al.* Circulating metabolites and general cognitive ability and dementia: Evidence from 11 cohort studies. *Alzheimers Dement* (2018).
138. Thaiss, C.A., Zmora, N., Levy, M. & Elinav, E. The microbiome and innate immunity. *Nature* **535**, 65-74 (2016).
139. Vernocchi, P., Del Chierico, F. & Putignani, L. Gut Microbiota Profiling: Metabolomics Based Approach to Unravel Compounds Affecting Human Health. *Front Microbiol* **7**, 1144 (2016).
140. Conlon, M.A. & Bird, A.R. The impact of diet and lifestyle on gut microbiota and human health. *Nutrients* **7**, 17-44 (2014).
141. Holmes, E., Li, J.V., Athanasiou, T., Ashrafian, H. & Nicholson, J.K. Understanding the role of gut microbiome-host metabolic signal disruption in health and disease. *Trends Microbiol* **19**, 349-59 (2011).
142. Hsiao, E.Y. *et al.* Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. *Cell* **155**, 1451-63 (2013).
143. Sampson, T.R. *et al.* Gut Microbiota Regulate Motor Deficits and Neuroinflammation in a Model of Parkinson's Disease. *Cell* **167**, 1469-1480 e12 (2016).
144. Rogers, G.B. *et al.* From gut dysbiosis to altered brain function and mental illness: mechanisms and pathways. *Mol Psychiatry* **21**, 738-48 (2016).
145. Vogt, N.M. *et al.* Gut microbiome alterations in Alzheimer's disease. *Scientific Reports* **7**(2017).
146. Fernandez-Real, J.M. *et al.* Gut Microbiota Interacts With Brain Microstructure and Function. *Journal of Clinical Endocrinology & Metabolism* **100**, 4505-4513 (2015).
147. Cattaneo, A. *et al.* Association of brain amyloidosis with pro-inflammatory gut bacterial taxa and peripheral inflammation markers in cognitively impaired elderly. *Neurobiology of Aging* **49**, 60-68 (2017).
148. Fung, T.C., Olson, C.A. & Hsiao, E.Y. Interactions between the microbiota, immune and nervous systems in health and disease. *Nature Neuroscience* **20**, 145-155 (2017).
149. Powell, N., Walker, M.M. & Talley, N.J. The mucosal immune system: master regulator of bidirectional gut-brain communications. *Nat Rev Gastroenterol Hepatol* **14**, 143-159 (2017).
150. Montiel-Castro, A.J., Gonzalez-Cervantes, R.M., Bravo-Ruiseco, G. & Pacheco-Lopez, G. The microbiota-gut-brain axis: neurobehavioral correlates, health and sociality. *Front Integr Neurosci* **7**, 70 (2013).



Chapter 2

Omics of neurodegeneration

Chapter 2.1

Genome-wide association study of 23,500 individuals identifies 7 loci associated with brain ventricular volume

Dina Vojinovic, Hieab H. Adams, Xueqiu Jian, Qiong Yang, Albert Vernon Smith, Joshua C. Bis, Alexander Teumer, Markus Scholz, Nicola J. Armstrong, Edith Hofer, Yasaman Saba, Michelle Luciano, Manon Bernard, Stella Trompet, Jingyun Yang, Nathan A. Gillespie, Sven J. van der Lee, Alexander Neumann, Shahzad Ahmad, Ole A. Andreassen, David Ames, Najaf Amin, Konstantinos Arfanakis, Mark E. Bastin, Diane M. Becker, Alexa S. Beiser, Frauke Beyer, Henry Brodaty, R. Nick Bryan, Robin Bülow, Anders M. Dale, Philip L. De Jager, Ian J. Deary, Charles DeCarli, Debra A. Fleischman, Rebecca F. Gottesman, Jeroen van der Grond, Vilmundur Gudnason, Tamara B. Harris, Georg Homuth, David S. Knopman, John B. Kwok, Cora E. Lewis, Shuo Li, Markus Loeffler, Oscar L. Lopez, Pauline Maillard, Hanan El Marroun, Karen A. Mather, Thomas H. Mosley, Ryan Muetzel, Matthias Nauck, Paul A. Nyquist, Matthew S. Panizzon, Zdenka Pausova, Bruce M. Psaty, Ken Rice, Jerome I. Rotter, Natalie Royle, Claudia L. Satizabal, Reinhold Schmidt, Peter R. Schofield, Pamela J. Schreiner, Stephen Sidney, David J. Stott, Anbupalam Thalamuthu, Andre G. Uitterlinden, Maria C. Valdés Hernández, Meike W. Vernooij, Wei Wen, Tonya White, A. Veronica Witte, Katharina Wittfeld, Margaret J. Wright, Lisa R. Yanek, Henning Tiemeier, William S. Kremen, David A. Bennett, J. Wouter Jukema, Tomas Paus, Joanna M. Wardlaw, Helena Schmidt, Perminder S. Sachdev, Arno Villringer, Hans Jörgen Grabe, WT Longstreth, Cornelia M. van Duijn, Lenore J. Launer, Sudha Seshadri, M. Arfan Ikram, Myriam Fornage

This chapter is accepted for publication in Nature Communications.

The supplemental information for this paper is available at https://drive.google.com/drive/folders/1-2X-Nx3tdaeX4E0_bAWANKo7X_umEx2X?usp=sharing

ABSTRACT

The volume of the lateral ventricles (LV) increases with age and their abnormal enlargement is a key feature of several neurological and psychiatric diseases. Although lateral ventricular volume is heritable, a comprehensive investigation of its genetic determinants is lacking. In this meta-analysis of genome-wide association studies of 23,533 healthy middle-aged to elderly individuals from 26 population-based cohorts, we identify 7 genetic loci associated with LV volume. These loci map to chromosomes 3q28, 7p22.3, 10p12.31, 11q23.1, 12q23.3, 16q24.2, and 22q13.1 and implicate pathways related to tau pathology, S1P signaling, and cytoskeleton organization. We also report a significant genetic overlap between the thalamus and LV volumes ($\rho_{genetic}$ = -0.59, *p*-value = 3.14×10^{-6}), suggesting that these brain structures may share a common biology. These genetic associations of LV volume provide insights into brain morphology.

INTRODUCTION

The volume of lateral ventricles increases in normal aging.[1-4] The enlargement of lateral ventricles has also been suggested in various complex neurological disorders such as Alzheimer's disease, vascular dementia and Parkinson's disease[5-8] as well as psychiatric disorders such as schizophrenia and bipolar disorder.[9-11] Furthermore, ventricular enlargement has been associated with poor cognitive functioning and cerebral small vessel disease pathology.[12-14] Even though it might be intuitive to interpret ventricular expansion primarily as an indicator of brain shrinkage after the onset of the disorder, recent studies have provided evidence against this notion.[15,16] The size of lateral ventricles' is influenced by genetic factors with heritability estimated to be 54%, on average,[16] but changing with age, from 32-35% in childhood to about 75% in late middle and older age.[16] Even though the size of surrounding gray matter structures is also heritable,[17-19] ventricular volume is reported to be genetically independent of other brain regions surrounding the ventricles.[20] Similarly, ventricular enlargement in schizophrenia does not appear to be linked to volume reduction in the surrounding structures.[15]

Elucidating the genetic contribution to inter-individual variation in lateral ventricular volume can thus provide important insights and better understanding of the complex genetic architecture of brain structures and related neurological and psychiatric disorders. Candidate gene studies have identified single nucleotide polymorphisms (SNPs) mapping to Catechol-O-Methyltransferase (*COMT*) and Neuregulin 1 (*NRG1*) genes as associated with larger lateral ventricular volume in patients with the first episode of non-affective psychosis.[21,22] However, a comprehensive investigation of the genetic determinants of lateral ventricular volume is lacking.

Here, we perform a genome-wide association (GWA) meta-analysis of 23,533 middle-aged to elderly individuals from population-based cohorts participating in the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium in order to identify common genetic variants that influence lateral ventricle volume. We apply a commonly used two-stage GWA design followed by a joint analysis approach that combines information across the stages and provides greater power.[23] We identify 7 genetic loci associated with lateral ventricular volume and report genome-wide overlap with thalamus volume.

RESULTS

Genome-wide association results

The overview of study design is illustrated in **Supplementary Fig. 1**. The GWA results from 12 studies were combined in stage 1 and subsequently evaluated in an independent sample from 14 studies in stage 2. Finally, the results of stage 1 and stage 2 analyses were combined in stage 3. Detailed information on study participants, image acquisition and genotyping is provided in **Supplementary Note 1** and **Supplementary Data 1-3**.

The results of the stage 1 meta-analysis (N = 11,396) are illustrated in **Supplementary Fig. 2**. The quantile-quantile plot suggests that potential population stratification and/or cryptic relatedness are well controlled after genomic correction (λ = 1.04) (**Supplementary Fig. 2**, **Supplementary Table 1**). The stage 1 meta-analysis identified 146 significant variant associations, mapping to three chromosomal regions at 3q28, 7p22.3, and 16q24.2 (**Table 1**). All 146 stage 1 significant associations replicated in the stage 2 meta-analysis (N = 12,137) with the same direction of effect at Bonferroni adjusted significance (p-value = 5×10^{-3}, **Supplementary Data 4**), except one SNP (p-value = 7.6×10^{-3}). Subsequently, the results from all individual studies were combined in the stage 3 GWA meta-analysis (N = 23,533). The quantile-quantile plot showed again adequate control of population stratification or relatedness (**Supplementary Fig. 3**). The combined stage 3 GWA meta-analysis identified 314 additional significant associations mapping to four additional chromosomal regions at 10p12.31, 11q23.1, 12q23.3, and 22q13.1 (**Fig. 1-2**, **Table 1**). The effect size for the lead variant mapped to 10p12.31 locus was correlated with mean age of the cohort (r = 0.50, p-value = 0.03) (**Supplementary Fig. 4**). No correlation was found for the other lead variants (**Supplementary Fig. 5-10**).

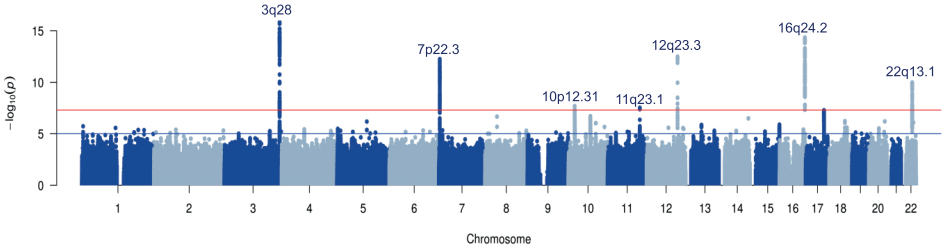


Figure 1. Manhattan plot for stage 3 genome-wide association meta-analysis. Each dot represents a variant. The plot shows –log10 p-values for all variants. Red line represents the genome-wide significance threshold (p-value < 5×10^{-8}), whereas blue line denotes suggestive threshold (p-value < 1×10^{-5}).

Table 1. Genome-wide significant results from the meta-analyses of lateral ventricular volume. Variant that showed the lowest *p*-value in the fixed effect sample-size weighted Z-score meta-analysis for each locus is shown.

SNP	Chr	Annotation	Gene(s)	A1/A2	Stage 1		Stage 2		Stage 3 combined	
					Zscore	P	Zscore	P	Zscore	P
rs34113929*	3q28	intergenic	*SNAR-I,OSTN*	A/G	-6.84	7.70E-12	-5.05	4.44E-07	-8.27	1.37E-16
rs9937293*	16q24.2	intergenic	*FOXL1,C16orf95*	A/G	5.65	1.63E-08	5.61	2.03E-08	7.84	4.45E-15
7:2760334:C_CT*	7p22.3	intergenic	*AMZ1,GNA12*	D/I	-5.88	4.21E-09	-4.48	7.34E-06	-7.21	5.61E-13
rs12146713	12q23.3	intronic	*NUAK1*	T/C	-5.01	5.57E-07	-5.44	5.32E-08	-7.28	3.25E-13
rs4820299	22q13.1	intronic	*TRIOBP*	T/C	-4.79	1.71E-06	-4.49	7.04E-06	-6.46	1.05E-10
rs35587371	10p12.31	intronic	*MLLT10*	A/T	-4.89	1.03E-06	-3.32	9.12E-04	-5.61	2.07E-08
rs7936534	11q23.1	intergenic	*ARHGAP20,C11orf53*	A/G	4.25	2.12E-05	3.71	2.04E-04	5.54	2.96E-08

Abbreviations: SNP - single nucleotide polymorphism; Chr - chromosome; A1/A2 - effect allele/other allele; Freq - Frequency of effect allele; Zscore - Z score from METAL; P - *p*-value;

*Variants that surpassed genome-wide significance threshold in stage 1 meta-analysis; remaining SNPs listed in the table reached genome-wide significance threshold in combined, stage 3, meta-analysis;

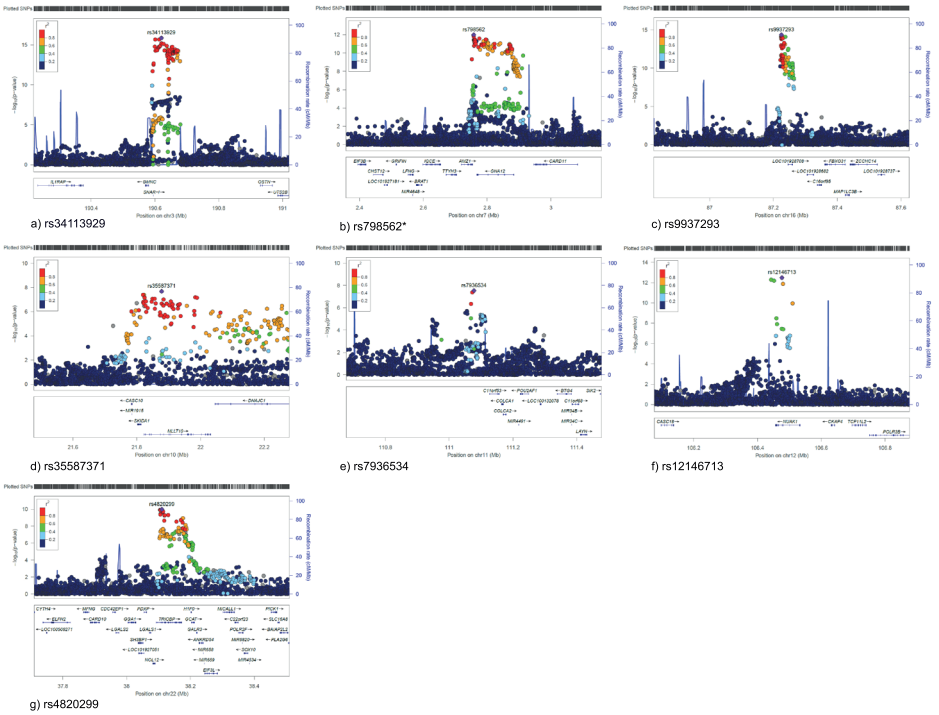


Figure 2. Regional association and recombination plots in combined stage 3 GWA meta-analysis. The left axis represents –log10 *p*-values for association with total later ventricular volume. The right axis represents the recombination rate, and the x-axis represents chromosomal position (hg19 genomic position). The most significant SNPs of the regions are denoted with a purple diamond. Surrounding SNPs are colored according to their pairwise correlation (r^2) with the top-associated SNP of the region. The gene annotations are below the figure.

Even though cohorts of European (EA) and African-American (AA) ancestry were included, all significant associations were mainly driven by EA samples (**Supplementary Fig. 11-12**). The direction of effect size across the EA cohorts for the 7 lead variants was generally concordant and showed no evidence of any single cohort driving the associations (**Supplementary Fig. 11**). Despite the different methods of phenotyping across the cohorts, the cohorts with different phenotyping methods showed evidence of effect suggesting that there is limited heterogeneity in effects (**Supplementary Fig. 12**).

To investigate whether 7 lead variants have an effect in early life, childhood, the analyses were carried out in a children's cohort of 1,141 participants from Generation R study. The percentage of lead variants showing consistent direction of effect with stage 3 was 85.7% (6 out of 7, binomial *p*-value = 0.05) (**Supplementary Data 4**), and a variant mapped to the 12q23.3 region showed nominal association with lateral ventricular volume in the children's cohort (effect = -0.15, *p*-value = 0.01). Additionally, three out of seven lead

variants (or their proxies; $r^2 > 0.7$) showed pleiotropic association (p-value < 5×10^{-8}) with other traits according to the PhenoScanner database (**Supplementary Data 5**).[24]

To capture gender-based differences, sex-stratified GWA analysis was performed (Nmen = 10,358; Nwomen = 12,872). None of the 15,660,719 variants that were tested for heterogeneity between men and women reached genome-wide significance threshold (**Supplementary Fig. 13**). However, an indel located at 4q35.2 showed suggestive evidence of association in men (4:187559262:C_CAA, p-value = 5.43×10^{-8}) but not in women (p-value = 0.88).

Independent signals within loci

The conditional and joint (COJO) analysis using the Genome-wide Complex Trait Analysis (GCTA) identified no other additional variants, after conditioning on the lead variant at the locus 3q28, 7p22.3, 10p12.31, 11q23.1, 12q23.3, 16q24.2, or 22q13.1.

Functional annotation

A large proportion of genome-wide significant variants were intergenic (335/460) (**Supplementary Fig. 14**). Variants with the highest probability of having a regulatory function based on RegulomeDB score (Category 1 RegulomeDB score) were located at 7p22.3 and at 22q13.1 (**Supplementary Data 6**). Of 7 lead variants, 4 were intergenic, 4 were in an active chromatin state and 3 showed expression quantitative trait (eQTL) effects (**Supplementary Data 6**). The lead SNP at 22q13.1 (rs4820299) was associated with differential expression of the largest number of genes (n = 6). In brain tissue, the alternate allele of this SNP was associated with higher expression of *TRIOBP* suggesting that higher expression was associated with smaller lateral ventricles (**Supplementary Fig. 15**).

Partitioned heritability

SNP-based heritability in the sample of European ancestry participants was estimated at 0.20 (SE = 0.02) using LD score regression, and this was higher in women (0.19 (SE = 0.04)) than in men (0.15 (SE = 0.05)). The 7 lead variants explained 1.5% of total variance in lateral ventricular volume. Partitioning of heritability based on functional annotation using LD score regression, revealed significant enrichment of SNPs within 500 bp of highly active enhancers, where 17% of SNPs accounted for 54% of the heritability (p-value = 7.9×10^{-6}, **Supplementary Table 2**). Significant enrichment was also found for histone marks including H3K27ac (which indicates enhancer and promoter regions), H3K9ac (which highlights promoters), H3K4me3 (which indicates promoters/transcription starts), and H3K4me1 (which highlights enhancers) (**Supplementary Table 2**).[25,26]

Functional enrichment analysis

Functional enrichment analysis using regulatory regions from the ENCODE and Roadmap projects using the GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction (GARFIELD) method revealed that SNPs associated with lateral ventricular volume at *p*-value threshold < 10^{-5} were more often located in genomic regions harboring histone marks (H3K9ac (associated with promoters) and H3K36me3 (associated with transcribed regions))[25] and DNaseI hypersensitivity sites (DHS) than a permuted background (**Figure 3**, **Supplementary Data 7**).

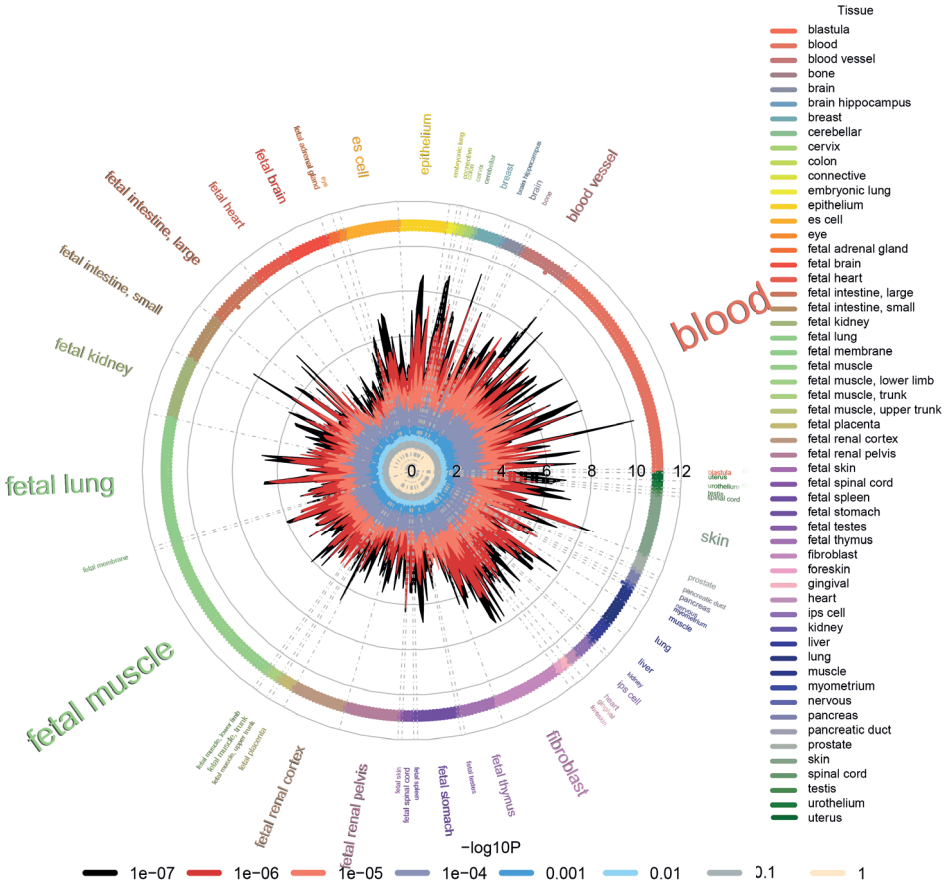


Figure 3. Functional enrichment analysis of lateral ventricular volume loci within DNaseI hypersensitivity spots. The radial lines show fold enrichment (FE) at eight GWA *p*-value thresholds. The results are shown for each of 424 cell types which are sorted by tissue, represented along the outer circle of the plot. The font size is proportional to the number of cell types from the tissue. FE values are plotted with different colors with respect to different GWA thresholds. Significant enrichment for a given cell type is denoted along the outer circle of the plot from a GWA *p*-value threshold < 10^{-5} (outermost) to GWA *p*-value threshold < 10^{-8} (innermost). The results show ubiquitous enrichment.

Integration of gene expression data

Integration of functional data from the Genotype-Tissues Expression (GTEx) project using the MetaXcan method revealed two significant associations between genetically predicted expression in brain tissue and lateral ventricular volume (**Supplementary Fig. 16**). Expression levels of *TRIOBP* at the locus 22q13.1 (p-value = 3.2×10^{-6}) and *MRPS16* at the locus 10q22.2 (p-value = 1.8×10^{-6}) were associated with lateral ventricular volume.

Gene annotation and pathway analysis

The results of gene-based and pathway analyses are illustrated in **Supplementary Table 3** and **Supplementary Data 8**. The pathway analysis identified "regulation of cytoskeleton organization" (GO:0051493) gene-set to be significantly enriched (p-value = 6×10^{-6}). Genes of the "regulation of cytoskeleton organization" pathway have previously been implicated in various neurological or cardiovascular diseases (**Supplementary Data 9**). Furthermore, pathways that pointed towards sphingosine 1 phosphate (S1P) signaling showed suggestive enrichment (**Supplementary Data 8**).

Genetic correlation

Additionally, we examined the genetic overlap between lateral ventricular volume and other traits (**Table 2**). We found that genetically-determined components of thalamus and lateral ventricular volumes appear to be negatively correlated ($\rho_{genetic}$ = -0.59, p-value = 3.14×10^{-6}). This finding was also confirmed at the phenotype level (**Supplementary Table 4**). Weaker genetic overlap was observed with infant head circumference ($\rho_{genetic}$ = 0.28, p-value = 8.7×10^{-3}), intracranial volume ($\rho_{genetic}$ = 0.35, p-value = 9×10^{-3}), height ($\rho_{genetic}$ = -0.14, p-value = 5.7×10^{-3}), and mean pallidum ($\rho_{genetic}$ = -0.29, p-value = 2.5×10^{-2}), whereas no significant genetic overlap was found with neurological diseases, psychiatric diseases, or personality traits.

Genetic risk score

We next examined the association of genetic risk scores (GRS) for Alzheimer's disease, Parkinson's disease, schizophrenia, bipolar disorder, cerebral small vessel disease, and tau-related pathology, including tau and phosphorylated tau levels in cerebrospinal fluid, amyotrophic lateral sclerosis (ALS) and progressive supranuclear palsy (PSP), using the lead SNPs from the largest published GWA study and lateral ventricular volume (**Supplementary Data 10**). We found a suggestive association of GRS for tau levels in cerebrospinal fluid (p-value = 9.59×10^{-3}) and lateral ventricular volume (**Supplementary Table 5**). The association was driven by one SNP (**Supplementary Table 6**). No association was observed with other examined phenotypes (**Supplementary Table 5**).

Table 2. The results of genetic correlation between the lateral ventricular volume and anthropometric traits, brain volumes, neurological and psychiatric diseases and personality traits.

Category	Phenotype	PMID	N	rg	SE	P
Anthropometric						
	Height	20881960	133859	-0.135	0.049	5.70E-03
	Infant head circumference	22504419	10768	0.284	0.108	8.70E-03
	Child birth length	25281659	28459	-0.133	0.089	1.34E-01
	Child birth weight	23202124	26836	-0.118	0.102	2.47E-01
Brain volume						
	Mean Thalamus	25607358	13193	-0.591	0.127	3.14E-06
	Mean Pallidum	25607358	13142	-0.29	0.129	2.47E-02
	ICV	25607358	11373	0.347	0.133	9.00E-03
	Mean Accumbens	25607358	13112	-0.29	0.158	6.64E-02
	Mean Putamen	25607358	13145	-0.15	0.089	9.13E-02
	Mean Hippocampus	25607358	13163	-0.204	0.132	1.20E-01
	Mean Caudate	25607358	13171	0.012	0.105	9.06E-01
Neurological diseases						
	Alzheimer's disease	24162737	54162	0.181	0.11	9.87E-02
	Parkinson's disease	19915575	5691	-0.096	0.084	2.55E-01
	Amyotrophic lateral sclerosis	27455348	36052	-0.032	0.128	8.04E-01
	White matter hyperintensities	25663218	17940	0.100	0.094	2.87E-01
Personality traits						
	Neo-conscientiousness	21173776	17375	-0.359	0.158	2.27E-02
	Neo-openness to experience	21173776	17375	0.088	0.118	4.56E-01
	Neuroticism	27089181	170911	-0.03	0.065	6.45E-01
Psychiatric traits						
	ADHD	20732625	5422	-0.276	0.152	6.90E-02
	PGC cross-disorder analysis	23453885	61220	-0.121	0.071	8.65E-02
	Major depressive disorder	22472876	18759	-0.165	0.102	1.05E-01
	Schizophrenia	25056061	77096	-0.067	0.044	1.30E-01
	Subjective well being	27089181	298420	0.087	0.075	2.50E-01
	ADHD (No GC)	27663945	17666	-0.151	0.149	3.11E-01
	Depressive symptoms	27089181	161460	-0.038	0.071	5.93E-01
	Autism spectrum disorder	0	10263	0.041	0.092	6.53E-01
	Anorexia Nervosa	24514567	17767	0.011	0.056	8.43E-01
	Bipolar disorder	21926972	16731	0.009	0.078	9.12E-01

Abbreviations: rg – genetic correlation; SE - standard error; P - *p*-value;

DISCUSSION

We have performed the first genome-wide association study of lateral ventricular volume including up to 23,533 individuals. We identified statistically significant association between lateral ventricular volume and variants at 7 loci. Additionally, we found that genetically-determined components of thalamus and lateral ventricular volume are correlated.

The strongest association was observed at the intergenic 3q28 locus between non-coding RNA *SNAR-I* and *OSTN*. This region has previously been associated with cerebrospinal fluid tau/ptau levels and Alzheimer's disease risk, tangle pathology and cognitive decline.[27] Similarly, the genome-wide significant locus at 12q23.3 encompasses *NUAK1,* which has also been associated with tau pathology. Nuak1 modulates tau levels in human cells and animal models and associates with tau accumulation in different tauopathies.[28] *NUAK1* is most prominently expressed in the brain where it has a role in mediating axon growth and branching in cortical neurons.[29] The lead SNP of the 12q23.3 locus mapped to an intron of *NUAK1*. This SNP is among the top 1% of most deleterious variants in the human genome based on its Combined Annotation Dependent Depletion (CADD) score of 21.5 and is located in an enhancer region (**Supplementary Data 6**). Interestingly, this variant also showed an effect in early life.

In our data, the significant variants of 7p22.3 region had the highest probability of being regulatory based on the RegulomeDB score (1b). The lead variant at 7p22.3 was in an active chromatin state and was associated with differential expression of *GNA12* (**Supplementary Data 6**). The *GNA12* gene is involved in various transmembrane signaling systems.[30-33] Interestingly, this gene was part of S1P signaling pathways identified to be enriched among genes associated with lateral ventricular volume. S1P, a bioactive sphingolipid metabolite, regulates nervous system development[34] such as neuronal survival, neurite outgrowth, and axon guidance,[35,36] and plays a role in neurotransmitter release.[37] It also plays a role in regulating the development of germinal matrix (GM) vasculature.[38] Disruption of S1P regulation results in defective angiogenesis in GM, hemorrhage, and enlarged ventricles.[38]

The other identified locus, 16q24.2, has previously been connected with small vessel disease and white-matter lesions formation.[39] Further, the alternate allele of the lead SNP at 22q13.1 in *TRIOBP* is associated with higher expression of the same gene in basal ganglia and brain cortex, and the same allele is associated with smaller lateral ventricular volume. Interestingly, predicted expression of this gene in cerebral cortex was significantly associated with lateral ventricular volume, suggesting a causal functional

role of the gene. The same analysis revealed significant association of the expression of *MRPS16* in frontal cortex with lateral ventricular volume. This gene was previously related to agenesis/hypoplasia of corpus callosum and enlarged ventricles.[40]

Finally, the lead intergenic SNP at 11q23.1 maps between *C11orf53* and *ARHGAP20*, whereas the 10p12.31 region encompasses *MLLT10* which has been linked to various leukemias, ovarian cancer, and meningioma.[41,42] The effect size of this variant on lateral ventricular volume was correlated with mean cohort age, with the effect being near zero at younger age and larger at older ages.

The gene-enrichment analysis highlighted "regulation of cytoskeleton organization" (GO:0051493) pathway. Genes that are part of this pathway have previously been implicated in various neurological diseases such as Parkinson's disease (*PARK2*), frontotemporal dementia (*MAPT*), neurofibromatosis 2 (*NF2*), tuberous sclerosis (*TSC1*) (**Supplementary Data 9**). The cytoskeleton is essentially involved in all cellular processes, and therefore crucial for processes in the brain such as cell proliferation, differentiation, migration, and signaling. Dysfunction of cytoskeleton has been associated with neurodevelopmental, psychiatric and neurodegenerative diseases.[43-45]

Previous studies showed significant sex-specific differences in lateral ventricular volume.[46,47] In our study we did not observe sex-specific differences; as for the lead 7 variants, both males and females were contributing to the association signal. However, we observed only one suggestive association at 4q35.2 that showed association in men only. The lead variant (indel) is mapped to *FAT1* which encodes atypical cadherins. Mutation in this gene causes a defect in cranial neural tube closure in a mouse model and an increase in radial precursor proliferation in the cortex.[48] However, the SNP-based heritability estimates were slightly higher in females. This may be explained by the differences in sample size in male and female-specific analyses implying that there is lower precision.

We estimated that 20% of genetic variance in lateral ventricular volume could be explained by common genetic variants, suggesting that common variants represent a substantial fraction of overall genetic component of variance. Moreover, the most statistically significant effect occurred in the regions of highly active enhancers and histone marks, suggesting their involvement in gene expression. Using the LD score regression method, we found a significant negative genetic correlation between lateral ventricular volume and thalamus volume. However, these may not be independent events, but inverse reflections of the same biology. Even though not strictly significant, we also observed trends for genetic correlations with other brain volumetric measures. Fur-

thermore, no genome-wide overlap was found between lateral ventricular volume and various neurological or psychiatric diseases. Given that enlargement of lateral ventricles has been suggested in Alzheimer's disease, we examined the association of *APOE* alleles and found no association between the *APOE* ε4 (*p*-value = 0.86) or *APOE* ε2 (*p*-value = 0.81) and lateral ventricular volume in our study population.

As we identified loci underlying lateral ventricular volume at the genome-wide level, but also genes and common pathways, our results provide various insights into the genetic contribution to lateral ventricular volume variability and a better understanding of the complex genetic architecture of brain structures. The genes with variants that we found to be associated with lateral ventricular volume are relevant to neurological aging given the characteristics of the study population which is relatively free from the disease as participants with stroke, traumatic brain injury and dementia at the time of magnetic resonance imaging (MRI) were excluded. This is in line with the previously published work of Pfefferbaum *et al.* who showed that the stability of lateral ventricles is genetically determined, whereas other factors such as normal aging or trauma and disease play a role in its change.[1,16]

However, while studying genetic overlap of lateral ventricular volume and various neurological or psychiatric disorders at multiple levels (LD score regression/polygenic, GRS/oligogenic, GWA hits/monogenic), we found evidence that some single genetic variants have pleiotropic effect on lateral ventricular volume and biochemical markers for a neurological disease (AD) or meningioma (**Supplementary Data 5**), while no evidence was found for genetic overlap with other neurological or psychiatric disorders (**Table 2**, **Supplementary Table 5**). The pattern of association between lateral ventricular volume and psychiatric disorder i.e. schizophrenia on multiple scales is similar to the findings of Franke *et al.* who evaluated association of various subcortical brain volumes and schizophrenia and reported no evidence of genetic overlap.[49] Even though our study does not provide a definite statement regarding the relationship between lateral ventricular volume and neurological or psychiatric disorders, it lays the foundation for future studies which should disentangle whether lateral ventricular volume is genetically related or unrelated to various neurological and psychiatric disorders (e.g. result from reverse causation). Novel insights may be revealed by improving the power of the studies, studying homogeneous samples with harmonized phenotype assessment methods along with evaluation of common and rare variants.

The strengths of our study are the large sample, population-based design and the use of quantitative MRI. Our study also has several limitations. Despite the effort to harmonize phenotype assessment, the methods used to quantify lateral ventricular volume

differ across cohorts. Because of this phenotypic heterogeneity, association results of participating cohorts were combined using a sample-size weighted meta-analysis, thus limiting discussion on effect sizes. Secondly, phenotypic heterogeneity may have caused the loss of statistical power. However, despite heterogeneity in the phenotype assessment, the association signals were coming from several studies irrespective of the method of phenotype assessment, which suggests robustness of our findings. Furthermore, although we made an effort to include cohorts of EA and AA ancestry, the study comprised predominately of individuals of European origin (22,045 individuals of EA and 1,488 of AA ancestry). Given the disparity in sample size, it is difficult to distinguish whether any inconsistency in results between the 2 groups stems from true genetic differences or from differential power to detect genetic effects. Indeed, this is also exemplified by the plots of the Z-scores (**Supplementary Fig. 11**) showing that direction of effect size in AA cohorts is often inconsistent with the direction of effect size in EA cohorts. However, the same inconsistency can be observed with European cohorts of equally small sample size. This inconsistency may be due to small sample size rather than ethnic background but we cannot rule out that racial-ethnic specific effects may exit. This limitation underscores the need for expanding research studies in non-European populations. Finally, as some loci only reached the genome-wide significance in the combined meta-analysis, they should be considered as highly probable findings and would still require independent replication.

To conclude, we identified genetic associations of lateral ventricular volume with variants mapping to 7 loci and implicating several pathways, including pathway related to tau pathology, cytoskeleton organization, and S1P signaling. These data provide new insights into understanding brain morphology.

METHODS

Study design

The overview of study design is illustrated in **Supplementary Fig. 1**. We performed a GWA meta-analysis of 11,396 participants of mainly European ancestry from 12 studies (stage 1) that contributed summary statistic data before a certain deadline. The deadline was set prior to data inspection and was not influenced by the results of the GWA meta-analysis. Variants that surpassed the genome-wide significance threshold (p-value < 5×10^{-8}) were subsequently evaluated in an independent sample of 12,137 participants of mainly European ancestry from 14 studies (stage 2). Finally, we performed a meta-analysis of all stage 1 and stage 2 studies (stage 3).

Study population

All participating studies are part of the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium.[50] A detailed description of participating studies can be found in **Supplementary Note 1**. General characteristics of study participants are provided in **Supplementary Data 1**. Written informed consent was obtained from all participants. Each study was approved by local ethical committees or the institutional review boards (see **Supplementary Note 1** for details).

Imaging

Each study performed magnetic resonance imaging (MRI) and estimated the volume of the lateral ventricles and intracranial volume (ICV). The field strength of scanners ranged from 0.35 to 3 T. Information on scanner manufacturers and measurement methods is provided in **Supplementary Data 2**. While most of the studies quantified lateral ventricular volume using validated automated segmentation methods, some studies used validated visual grading scales. The visual and volumetric scales were compared previously and showed high agreement for lateral ventricular volume.[2] The assessment of consistency of lateral ventricular volume on volumetric scale across time and different versions of software (freesurfer v4.5, v5.1, and v6.0), revealed high intraclass correlation (ICC > 0.98) in a subset of participants from the Rotterdam Study. Participants with dementia at the time of MRI, traumatic brain injury, prior or current stroke or intracranial tumors were excluded.

Genotyping and imputation

Information on genotyping platforms, quality control procedures and imputations methods for each participating study are provided in **Supplementary Data 3**. All studies used commercially available genotyping arrays, including Illumina or Affymetrix arrays. Similar quality control procedures were applied for each study (**Supplementary Data 3**). Using the validated software (Minimac,[51] IMPUTE,[52] BEAGLE[53]), each study performed genotype imputations using mostly the 1000 Genome phase 1 v3 reference panel.

Genome-wide association (GWA) analysis

Each participating study performed the GWA analysis of total lateral ventricular volume under an additive model using variant allele dosage as predictors and natural logarithm of the total lateral ventricular volume as the dependent variable. Transformation of the lateral ventricular volume was applied to obtain approximately normal distribution (**Supplementary Fig. 17**). The association analyses were adjusted for age, sex, total intracranial volume, age^2 if significant, population stratification, familial relationship (family-based studies) or study site (multi-site studies). Population stratification was controlled for by including principal components derived from genome-wide genotype

data. Study-specific details on covariates and software used are provided in **Supplementary Data 3**. Quality control (QC) was conducted for all participating studies using a standardized protocol provided by Winkler *et al.*[54] Variants with low imputation quality $r^2 < 0.3$ or minor allele count (MAC) ≤ 6 were filtered out. The association results of participating studies were combined using a fixed-effect sample-size weighted Z-score meta-analysis in METAL because of the difference in measurement methods of lateral ventricular volume.[55] Genomic control was applied to account for small amounts of population stratification or unaccounted relatedness. After the meta-analysis, variants with information in less than half the total sample size were excluded. Meta-analyses were performed separately for each of the stages. In the stage 1 meta-analysis, a *p*-value $< 5\times10^{-8}$ was considered significant. Variants that surpassed the threshold were evaluated in the stage 2 meta-analysis. In order to model linkage disequilibrium (LD) between those variants, we first calculated the number of independent tests using the eigenvalues of a correlation matrix using the Matrix Spectral Decomposition (matSpDlite) software.[56] Subsequently, a Bonferroni correction was applied for the effective number of independent tests (0.05/10 independent SNPs = 5×10^{-3}). Additionally, all analyses were stratified by sex. Following the same QC steps as for overall analyses, the sex-stratified association results of participating studies were combined using a fixed-effect sample-size weighted Z-score meta-analysis in METAL while applying genomic control.[55] The variants were assessed only if test statistics (Z-score) were heterogeneous between males and females (*p*-value < 0.1) and if the association in a sex-combined analysis did not reach genome-wide significance threshold.[57]

Conditional analysis

In order to identify variants that were independently associated with lateral ventricular volume, we performed conditional and joint (COJO) GWA analysis using Genome-wide Complex Trait Analysis (GCTA), version 1.26.0.[58] LD pattern was calculated based on 1000 Genome phase 1v3 imputed data of 6,291 individuals from the Rotterdam Study I.

Functional annotation

To annotate genome-wide significant variants with regulatory information, we used HaploReg v4.1[59], RegulomeDB v1.1[60], and Combined Annotation Dependent Depletion (CADD) tools.[61] To determine whether they have an effect on gene expression, we used GTEx data.[62] For the lead variants, we explored 5 chromatin marks assayed in 127 epigenomes (H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3) of RoadMap data.[63] To search for pleiotropic associations between our lead variants and their proxies ($r^2 > 0.7$) with other traits, we used the PhenoScanner database designed to facilitate the cross-referencing of genetic variants with many phenotypes.[9] The association results with genome-wide significance at 5×10^{-8} were extracted.

Variance explained

The proportion of variance in lateral ventricular volume explained by each lead variant was calculated using Pearson's phi coefficient squared as explained in Draisma *et al.*[64] The total proportion of variance in lateral ventricular volume was calculated by adding up the proportions of variance in lateral ventricular volume explained by each lead association signal.

Partitioned heritability

SNP-based heritability and partitioned heritability analyses were performed using LD score regression following the previously described method.[65] Partitioned heritability analysis determines enrichment of heritability in SNPs partitioned into 24 functional classes as reported in Finucane *et al.*[65] To avoid bias, an additional 500 bp window was included around the variants included in the functional classes. Only the HapMap3 variants were included as these seem to be well-imputed across cohorts.

Functional enrichment analysis

We performed functional enrichment analysis using regulatory regions from the ENCODE and Roadmap projects using GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction (GARFIELD) method.[66] The method provides fold enrichment (FE) statistics at various GWA *p*-value thresholds after taking into account LD, minor allele frequency, and local gene density.[66] The FE statistics were calculated at eight GWA *p*-value thresholds (0.1 to 1×10^{-8}). The associations were tested for various regulatory elements including DNase-I hypersensitivity sites, histone modifications, chromatin states and transcription factor binding sites in over 1000 cell and tissue-specific annotations.[66] The significance threshold calculated based on the number of annotations used was set at 4.97×10^{-5}.

Integration of gene expression

To integrate functional data in the context of our meta-analysis results, we used the MetaXcan method, which evaluated the association between lateral ventricular volume and brain-specific gene-expression levels predicted by genetic variants using the data from GTEx project.[62,67] This method is an extension of PrediXcan method modified to use summary statistic data from meta-analysis.[67] Based on a total number of genes tested, the Bonferroni corrected significance threshold was set to $0.05/12{,}379 = 4\times10^{-6}$.

Gene annotation and pathway-based analysis

The gene-based test statistics were computed using VEGAS2 software which tests for enrichment of multiple single variants within the genes while accounting for LD structure.[68] LD structure was computed based on the 1000 Genomes phase 3 population.

Variants within 10 kb of the 5′ and 3′ untranslated regions were included in this analysis in order to maintain regulatory variants.[68] Subsequently, the gene-based scores were used to perform gene-set enrichment analysis using VEGAS2pathway.[69] VEGAS2Pathway approach accounts for LD between variants within a gene, and between neighboring genes, gene size, and pathway size.[69] It uses computationally predicted Gene Ontology pathways and curated gene-sets from the MSigDB, PANTHER, and pathway commons databases.[69] The pathway-based significance threshold was set to the p-value = 1×10^{-5} while taking into account the multiple testing of correlated pathways (0.05/5,000 independent tests).[69]

Genetic correlation

We used the LD score regression method to estimate genetic correlations between lateral ventricular volume and various traits including anthropometric traits, brain volumes, neurological and psychiatric diseases and personality traits. The analyses were performed using a centralized database of summary-level GWA study results and a web interface for LD score regression, the LD-hub.[70] Summary-level GWA study results for white matter hyperintensities were obtained from the CHARGE consortium[71] and the analyses were performed using the ldsc tool (https://github.com/bulik/ldsc).

Genetic risk scores

We generated genetic risk scores (GRS) for Alzheimer's disease, amyotrophic lateral sclerosis (ALS), Parkinson's disease, bipolar disorder, schizophrenia, white matter lesions and tau-related phenotypes. The tau-related phenotypes, including tau and phosphorylated tau levels in cerebrospinal fluid, and progressive supranuclear palsy (PSP), were studied in relatively small sample and are therefore not appropriate for LD score regression. We extracted the lead genome-wide significantly associated SNPs and their effect estimates from the largest published GWA studies (**Supplementary Data 10**). For white matter lesions burden, effect estimate and standard errors were estimated from Z-statistics using the previously published formula.[72] The allele associated with an increased risk in corresponding traits was considered to be the effect allele. The weighted GRS was constructed as the sum of products of effect sizes as weights and respective allele dosages from 1000 Genome imputed data of Rotterdam Study using R software version 3.2.5 (https://www.R-project.org). Variants with low imputation quality ($r^2 < 0.3$) were excluded. Subsequently, the GRS was tested for association with lateral ventricular volume in three cohorts of Rotterdam Study while adjusting for age, sex, total intracranial volume, age^2 and population stratification. The significance threshold for genetic risk score association was set to p-value = 5×10^{-3} (0.05/10) based on the number of genetic risk scores tested.

Data availability

The summary statistics will be made available upon the publication on the CHARGE dbGaP site under the accession number phs000930.v7.p1.

REFERENCES

1. Pfefferbaum, A., Sullivan, E.V. & Carmelli, D. Morphological changes in aging brain structures are differentially affected by time-linked environmental influences despite strong genetic stability. *Neurobiol Aging* **25**, 175-83 (2004).
2. Carmichael, O.T. *et al.* Ventricular volume and dementia progression in the Cardiovascular Health Study. *Neurobiol Aging* **28**, 389-97 (2007).
3. Apostolova, L.G. *et al.* Hippocampal atrophy and ventricular enlargement in normal aging, mild cognitive impairment (MCI), and Alzheimer Disease. *Alzheimer Dis Assoc Disord* **26**, 17-27 (2012).
4. Long, X. *et al.* Healthy aging: an automatic analysis of global and regional morphological alterations of human brain. *Acad Radiol* **19**, 785-93 (2012).
5. Nestor, S.M. *et al.* Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's disease neuroimaging initiative database. *Brain: A Journal of Neurology***.131**, pp (2008).
6. Kuller, L.H. *et al.* Determinants of vascular dementia in the Cardiovascular Health Cognition Study. *Neurology* **64**, 1548-52 (2005).
7. Mak, E. *et al.* Longitudinal whole-brain atrophy and ventricular enlargement in nondemented Parkinson's disease. *Neurobiol Aging* **55**, 78-90 (2017).
8. Kuller, L.H., Lopez, O.L., Becker, J.T., Chang, Y. & Newman, A.B. Risk of dementia and death in the long-term follow-up of the Pittsburgh Cardiovascular Health Study-Cognition Study. *Alzheimers Dement* **12**, 170-183 (2016).
9. Vita, A., De Peri, L., Silenzi, C. & Dieci, M. Brain morphology in first-episode schizophrenia: a meta-analysis of quantitative magnetic resonance imaging studies. *Schizophr Res* **82**, 75-88 (2006).
10. Kempton, M.J., Geddes, J.R., Ettinger, U., Williams, S.C. & Grasby, P.M. Meta-analysis, database, and meta-regression of 98 structural imaging studies in bipolar disorder. *Arch Gen Psychiatry* **65**, 1017-32 (2008).
11. Olabi, B. *et al.* Are there progressive brain changes in schizophrenia? A meta-analysis of structural magnetic resonance imaging studies. *Biological Psychiatry* **70.1**, 88-96 (2011).
12. Mosley, T.H., Jr. *et al.* Cerebral MRI findings and cognitive functioning: the Atherosclerosis Risk in Communities study. *Neurology* **64**, 2056-62 (2005).
13. Appelman, A.P. *et al.* White matter lesions and lacunar infarcts are independently and differently associated with brain atrophy: the SMART-MR study. *Cerebrovasc Dis* **29**, 28-35 (2010).
14. Geerlings, M.I. *et al.* Brain volumes and cerebrovascular lesions on MRI in patients with atherosclerotic disease. The SMART-MR study. *Atherosclerosis* **210**, 130-6 (2010).
15. Horga, G. *et al.* Correlations between ventricular enlargement and gray and white matter volumes of cortex, thalamus, striatum, and internal capsule in schizophrenia. *European Archives of Psychiatry and Clinical Neuroscience* **261.7**, 467-476 (2011).
16. Kremen, W.S. *et al.* Heritability of brain ventricle volume: converging evidence from inconsistent results. *Neurobiol Aging* **33**, 1-8 (2012).
17. Peper, J.S., Brouwer, R.M., Boomsma, D.I., Kahn, R.S. & Hulshoff Pol, H.E. Genetic influences on human brain structure: a review of brain imaging studies in twins. *Hum Brain Mapp* **28**, 464-73 (2007).
18. Schmitt, J.E. *et al.* Review of twin and family studies on neuroanatomic phenotypes and typical neurodevelopment. *Twin Res Hum Genet* **10**, 683-94 (2007).
19. Kremen, W.S. *et al.* Genetic and environmental influences on the size of specific brain regions in midlife: the VETSA MRI study. *Neuroimage* **49**, 1213-23 (2010).

20. Eyler, L.T. *et al.* Genetic patterns of correlation among subcortical volumes in humans: Results from a magnetic resonance imaging twin study. *Human Brain Mapping* **32.4**, 641-653 (2011).
21. Mata, I. *et al.* A neuregulin 1 variant is associated with increased lateral ventricle volume in patients with first-episode schizophrenia. *Biological Psychiatry* **65.6**, 535-540 (2009).
22. Crespo-Facorro, B. *et al.* Low-activity allele of Catechol-O-Methyltransferase (COMTL) is associated with increased lateral ventricles in patients with first episode non-affective psychosis. *Progress in Neuro-Psychopharmacology & Biological Psychiatry* **31**, 1514-1518 (2007).
23. Skol, A.D., Scott, L.J., Abecasis, G.R. & Boehnke, M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature Genetics* **38**, 209-213 (2006).
24. Staley, J.R. *et al.* PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics* **32**, 3207-3209 (2016).
25. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
26. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).
27. Cruchaga, C. *et al.* GWAS of cerebrospinal fluid tau levels identifies risk variants for Alzheimer's disease. *Neuron* **78**, 256-68 (2013).
28. Lasagna-Reeves, C.A. *et al.* Reduction of Nuak1 Decreases Tau and Reverses Phenotypes in a Tauopathy Mouse Model. *Neuron* **92**, 407-418 (2016).
29. Courchet, J. *et al.* Terminal Axon Branching Is Regulated by the LKB1-NUAK1 Kinase Pathway via Presynaptic Mitochondrial Capture. *Cell* **153**, 1510-1525 (2013).
30. Yanamadala, V., Negoro, H., Gunaratnam, L., Kong, T. & Denker, B.M. Galpha12 stimulates apoptosis in epithelial cells through JNK1-mediated Bcl-2 degradation and up-regulation of IkappaBalpha. *J Biol Chem* **282**, 24352-63 (2007).
31. Kelly, P. *et al.* The G12 family of heterotrimeric G proteins promotes breast cancer invasion and metastasis. *Proc Natl Acad Sci U S A* **103**, 8173-8 (2006).
32. Krakstad, B.F., Ardawatia, V.V. & Aragay, A.M. A role for Galpha12/Galpha13 in p120ctn regulation. *Proc Natl Acad Sci U S A* **101**, 10314-9 (2004).
33. Zhu, D., Kosik, K.S., Meigs, T.E., Yanamadala, V. & Denker, B.M. Galpha12 directly interacts with PP2A: evidence FOR Galpha12-stimulated PP2A phosphatase activity and dephosphorylation of microtubule-associated protein, tau. *J Biol Chem* **279**, 54983-6 (2004).
34. Blaho, V.A. & Hla, T. An update on the biology of sphingosine 1-phosphate receptors. *J Lipid Res* **55**, 1596-608 (2014).
35. Strochlic, L., Dwivedy, A., van Horck, F.P., Falk, J. & Holt, C.E. A role for S1P signalling in axon guidance in the Xenopus visual system. *Development* **135**, 333-42 (2008).
36. Herr, D.R. *et al.* Sphingosine 1-phosphate (S1P) signaling is required for maintenance of hair cells mainly via activation of S1P2. *J Neurosci* **27**, 1474-8 (2007).
37. Shen, H. *et al.* Coupling between endocytosis and sphingosine kinase 1 recruitment. *Nat Cell Biol* **16**, 652-62 (2014).
38. Ma, S., Santhosh, D., Kumar, T.P. & Huang, Z. A Brain-Region-Specific Neural Pathway Regulating Germinal Matrix Angiogenesis. *Dev Cell* **41**, 366-381 e4 (2017).
39. Traylor, M. *et al.* Genetic variation at 16q24.2 is associated with small vessel stroke. *Annals of Neurology* **81.3**, 383-394 (2017).
40. Miller, C. *et al.* Defective mitochondrial translation caused by a ribosomal protein (MRPS16) mutation. *Ann Neurol* **56**, 734-8 (2004).

41. Pharoah, P.D.P. *et al.* GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. *Nature Genetics* **45**, 362-370 (2013).
42. Egan, K.M. *et al.* Brain tumor risk according to germ-line variation in the MLLT10 locus. *European Journal of Human Genetics* **23**, 132-134 (2015).
43. Paus, T., Pesaresi, M. & French, L. White matter as a transport system. *Neuroscience* **276**, 117-25 (2014).
44. McMurray, C.T. Neurodegeneration: diseases of the cytoskeleton? *Cell Death Differ* **7**, 861-5 (2000).
45. Cairns, N.J., Lee, V.M. & Trojanowski, J.Q. The cytoskeleton in neurodegenerative diseases. *J Pathol* **204**, 438-49 (2004).
46. Hasan, K.M., Moeller, F.G. & Narayana, P.A. DTI-based segmentation and quantification of human brain lateral ventricular CSF volumetry and mean diffusivity: validation, age, gender effects and biophysical implications. *Magn Reson Imaging* **32**, 405-12 (2014).
47. Pfefferbaum, A. *et al.* Variation in longitudinal trajectories of regional brain volumes of healthy men and women (ages 10 to 85 years) measured with atlas-based parcellation of MRI. *Neuroimage* **65**, 176-93 (2013).
48. Badouel, C. *et al.* Fat1 interacts with Fat4 to regulate neural tube closure, neural progenitor proliferation and apical constriction during mouse brain development. *Development* **142**, 2781-91 (2015).
49. Franke, B. *et al.* Genetic influences on schizophrenia and subcortical brain volumes: large-scale proof of concept. *Nat Neurosci* **19**, 420-431 (2016).
50. Psaty, B.M. *et al.* Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circ Cardiovasc Genet* **2**, 73-80 (2009).
51. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* **44**, 955-9 (2012).
52. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
53. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**, 1084-97 (2007).
54. Winkler, T.W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc* **9**, 1192-212 (2014).
55. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).
56. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* **95**, 221-227 (2005).
57. Zeggini, E. & Ioannidis, J.P. Meta-analysis in genome-wide association studies. *Pharmacogenomics* **10**, 191-201 (2009).
58. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
59. Ward, L.D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* **40**, D930-4 (2012).
60. Boyle, A.P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* **22**, 1790-7 (2012).

61. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).
62. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-5 (2013).
63. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**, 215-6 (2012).
64. Draisma, H.H.M. *et al.* Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nat Commun* **6**, 7208 (2015).
65. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47**, 1228 (2015).
66. Iotchkova, V. *et al.* GARFIELD - GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction. *bioRxiv* (2016).
67. Barbeira, A. *et al.* MetaXcan: Summary Statistics Based Gene-Level Association Method Infers Accurate PrediXcan Results. *bioRxiv* (2016).
68. Mishra, A. & Macgregor, S. VEGAS2: Software for More Flexible Gene-Based Testing. *Twin Res Hum Genet* **18**, 86-91 (2015).
69. Mishra, A. & MacGregor, S. A Novel Approach for Pathway Analysis of GWAS Data Highlights Role of BMP Signaling and Muscle Cell Differentiation in Colorectal Cancer Susceptibility. *Twin Res Hum Genet* **20**, 1-9 (2017).
70. Zheng, J. *et al.* LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272-279 (2017).
71. Verhaaren, B.F. *et al.* Multiethnic genome-wide association study of cerebral white matter hyperintensities on MRI. *Circ Cardiovasc Genet* **8**, 398-409 (2015).
72. Chauhan, G. *et al.* Association of Alzheimer's disease GWAS loci with MRI markers of brain aging. *Neurobiol Aging* **36**, 1765 e7-1765 e16 (2015).

Chapter 2.2

Genetic determinants of general cognitive function and their association to circulating metabolites: a cross-omics study

Dina Vojinovic, Caroline Hayward, Jennifer A. Smith, Wei Zhao, Jan Bressler, Stella Trompet, Chloé Sarnowski, Murali Sargurupremraj, Jingyun Yang, Paul R.H.J. Timmers, Narelle K. Hansell, Ari Ahola-Olli, Eva Krapohl, Joshua C. Bis, Daniel E. Gustavson, Teemu Palviainen, Yasaman Saba, Anbu Thalamuthu, Sudheer Giddaluru, Leonie Weinhold, Najaf Amin, Nicola Armstrong, Lawrence F. Bielak, Anne C. Böhmer, Patricia A. Boyle, Henry Brodaty, Harry Campbell, David W. Clark, Baptiste Couvy-Duchesne, Philip L De Jager, Jeremy A. Elman, Thomas Espeseth, Jessica D. Faul, Annette Fitzpatrick, Scott D. Gordon, Thomas Hankemeier, Edith Hofer, M. Arfan Ikram, Peter K. Joshi, Rima Kaddurah-Daouk, Jaakko Kaprio, Sharon LR Kardia, Katherine A. Kentistou, Luca Kleineidam, Nicole Kochan, John Kwok, Markus Leber, Teresa Lee, Terho Lehtimäki, Anu Loukola, Anders Lundquist, Leo-Pekka Lyytikäinen, Karen Mather, Grant W. Montgomery, Simone Reppermund, Richard J. Rose, Suvi Rovio, Perminder Sachdev, Matthias Schmid, Helena Schmidt, Andre G. Uitterlinden, Eero Vuoksimaa, Michael Wagner, Holger Wagner, David R. Weir, Margaret J. Wright, Miao Yu, Lars Nyberg, Alfredo Ramirez, Stephanie Le Hellard, David Ames, Peter Schofield, Reinhold Schmidt, Danielle Dick, David Porteous, William S. Kremen, Bruce M. Psaty, Olli Raitakari, Nicholas G. Martin, James F. Wilson, David A. Bennett, Stephanie Debette, J. Wouter Jukema, Thomas H Mosley, Jr, Sudha Seshadri, Cornelia M. van Dujin

This chapter is in preparation.

The supplemental information for this paper is available at https://drive.google.com/drive/folders/1yQGEPmTmY4kcnrsx8EiosNPL1w9rUx57?usp=sharing

ABSTRACT

General cognitive function is a heritable predictor of health outcomes. Here, we performed a genome-wide association study of general cognitive function in 245,117 participants of European (EA) and African American (AA) ancestry from Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium and UK Biobank. We reported 32 novel genetic loci in individuals of EA which have previously been associated with various disorders including psychiatric illnesses such as schizophrenia, autistic disorder, bipolar disorder, depression, mood and anxiety disorders. The risk score based on our findings and previously identified genetic risk loci underlying general cognitive function in EA was associated with general cognitive function in individuals of AA ancestry (N = 2,117). Genes associated with general cognitive function could be linked to circulating metabolites including tyrosine, creatinine, 22:6 docosahexaenoic acid (DHA), glycoprotein acetyl, acetate, and citrate. Using Mendelian randomization, we examined whether these metabolites were cause or rather a consequence of biological pathways underlying cognitive function. Genes determining glycoprotein acetyls and tyrosine also determine general cognitive function, suggesting that these metabolites are in the causal pathway, whereas DHA is rather a consequence of the physiological process determining cognitive function. These results provide new insights into general cognitive function.

INTRODUCTION

General cognitive function is an important predictor of health outcomes, including mortality and morbidity varying from dementia to depression and other psychiatric diseases.[1-4] Differences in cognitive function are determined by various factors including lifestyle and genetic factors.[5] Morbidities such as cardiometabolic diseases and cancer also contribute to cognitive performance and cognitive decline in later life.[6] For long, the relationship between metabolic factors and cognitive function was poorly understood. We recently identified circulating metabolites to be associated with the general cognitive function in healthy individuals.[7] The metabolic profile included the subfractions of high-density lipoprotein particles, fatty acids, amino acids, and acute phase reaction markers.[7,8] We successfully associated these metabolites to environmental factors such as lifestyle and diet.[7] A question to answer is whether these metabolites are in the causal pathway and a target of cognitive function or rather a consequence of physiological processes underlying general cognitive function and associated lifestyle and pathology.

Our human genome is another major driver of the circulating metabolites and general cognitive function.[9] General cognitive function has heritability of 50% and over 140 genomic regions have been identified in the genome-wide association studies (GWASs) performed to date.[10-13] Despite the overwhelming progress, the polygenic profile score capturing the joint effects of those variants explained only up to 4.3% of trait variance.[11] Yet, common genetic variants underlying general cognitive function were associated with various neurological and psychiatric disorders when checking for genome-wide genetic overlap using LD score regression.[14] Based on this method, there is also evidence for a shared genetic origin with body mass index, waist to hip ratio, high-density lipoprotein cholesterol, and cardiovascular diseases which are key drivers of the human metabolism.[11] Up until now, we have not linked these genetic determinants to the metabolites in the circulation, which may bring to surface new insights in metabolic pathways that play a key role in general cognitive function. An omission of previous studies is that only participants of European ancestry were included. A question to answer is whether the findings are generalizable to other ethnic groups. Finally, the studies of general cognitive function conducted to date mainly focused on the imputations generated as part of the 1000 Genomes Project. Recently, the Haplotype Reference Consortium (HRC) made available a large haplotype reference panel which increased imputation accuracy.[15,16]

Here, we performed a GWAS of general cognitive function in 243,000 participants of European ancestry and 2,117 participants of African-American from Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium and UK biobank

ancestry using HRC imputation panel.[11] We examined the association of genes implicated in general cognitive function and circulating metabolites and evaluated their causal relationship using Mendelian randomization approach.

METHODS

Study population

Our study population encompassed 243,000 participants of European ancestry (EA) from 32 studies from CHARGE consortium and UK biobank and 2,117 participants of African-American (AA) ancestry from 3 studies that were part of the CHARGE consortium.[11,17] Participating studies are described in detail in the **Supplementary Note**. General characteristics of study populations are provided in **Supplementary Table 1**. Local ethical committees or the institutional review boards approved each of the studies and written informed consent was obtained from all participants.

Phenotype assessment

The general cognitive function was constructed from a number of cognitive tasks for each of the CHARGE cohorts.[18] Each participating study performed principal component analysis using at least three cognitive tests that assess different cognitive domains. Only one score was used from each of the cognitive tests. The general cognitive function was the first unrotated principal component. The phenotype was constructed in such a way that higher score indicated higher cognitive function. Participants with dementia and stroke were excluded. Information on cognitive tests used to create general cognitive function score in each participating study and correlation between the general cognitive function and each cognitive test is provided in **Supplementary Table 2.** General cognitive function explained between 34.7% and 59.3% of the total test variance.

The cognitive test from UK Biobank was a verbal and numerical reasoning score assessed by 13 multiple choice questions which has a high genetic correlation with general cognitive function.[11,19] A detailed information regarding the samples of UK Biobank participants with verbal-numerical reasoning scores is provided elsewhere.[11] In the current analysis, four samples of UK Biobank participants were used.

Genotyping and imputation

Description of genotyping platforms, calling method and quality control procedures in each of the CHARGE cohorts is provided in **Supplementary Table 3**. The study participants were genotyped using commercially available genotyping arrays. Each study used free imputation servers (Michigan or Sanger) to perform genotype imputation using

Haplotype Reference Consortium (HRC) reference panel.[16] Description of genotyping platforms and quality control in UK Biobank is provided in Bycroft *et al.* (http://www.biorxiv.org/content/early/2017/07/20/166298).

Genome-wide association analysis

Each of the CHARGE cohorts performed genome-wide association analysis of general cognitive function while adjusting for age, gender, principal components if needed, familial relationship if appropriate and study center if needed. Details on analysis methods for each cohort are provided in **Supplementary Table 3**. The quality control (QC) was performed using EasyQC.[20] Genetic variants with low imputation quality ($r^2 < 0.5$) or minor allele count below 5 were removed. The genome-wide association summary results of verbal and numerical reasoning score in UK Biobank were obtained from http://www.ccace.ed.ac.uk/node/335.[11] The summary statistic results of CHARGE participating studies and UK biobank were combined using sample size weighted meta-analysis in METAL.[21] The meta-analysis was performed separately for each ethnic group. For each genome-wide association analysis, LD score regression method was used to estimate intercept which can distinguish between the inflation due to a polygenic signal and the inflation due to population stratification or cryptic relatedness.[22]

Conditional association analysis

Approximate conditional genome-wide analysis was performed using Genome-wide Complex Trait Analysis (GCTA), version 1.26.0, in order to identify genetic variants conditionally independent on genetic signals previously reported in the largest GWAS of general cognitive function to date.[11,23] Genetic variants with high collinearity (0.9) were ignored. Complete linkage equilibrium was assumed for genetic variants located more than 10Mb away from each other. The linkage disequilibrium pattern (LD) between the genetic variants was calculated based on data of 11,496 individuals from the Rotterdam Study imputed with HRC reference panel.

Characterization of genomic loci

Genomic loci were characterized using Functional Mapping and Annotation of genetic associations (FUMA).[24] First, independent genetic variants were defined as genome-wide significant variants that are not in linkage disequilibrium with each other ($r^2 < 0.6$).[24] Independent significant variants with $r^2 \geq 0.1$ were assigned to the same genomic risk locus and were merged into a single locus if they were 250 bp or closer.[24] Each genomic risk locus was represented by the top lead genetic variant defined as an independent significant variant ($r^2 < 0.1$).

Functional annotation

The genome-wide significant variants were annotated using the Combined Annotation Dependent Depletion (CADD), HaploReg v4.1, and RegulomeDB tools.[25,26] Furthermore, GTEx data was used to determine whether these genetic variants have an effect on expression.[27]

Correlating genetic determinants of cognition and circulating metabolites

Association of individual genetic variants underlying general cognitive function was explored in a GWAS of circulating metabolites including ~25,000 individuals.[11,28] The metabolites were measured by nuclear magnetic resonance (^{1}H-NMR) on Nightingale Health platform. To model correlation between metabolites and linkage disequilibrium between the genetic variants, we first calculated the number of independent tests using the method of Li and Ji.[29] The Bonferroni corrected *p*-value was calculated based on the number of independent tests and set at 0.05/(32 independent metabolites × 342 independent genetic variants) = 4.57×10^{-6}.

Next, more global test was used to link general cognitive function and metabolites using the LD score regression method. Genetic correlation was estimated between general cognitive function and metabolites measured by ^{1}H-NMR on Nightingale Health platform.[30] The analyses were performed using a web interface, LD-hub.[31] The significance threshold was determined based on a number of traits tested and was set at 0.05/111 = 4.5×10^{-4}.

Mendelian randomization

To evaluate whether the association of the metabolites to general cognitive function is a cause or consequence of the physiological processes underlying general cognitive function bidirectional Mendelian randomization was performed for each metabolite associated with genetic variants underlying general cognitive function. The associations were estimated based on the present GWAS and that of circulating metabolites including ~25,000 individuals.[28] The effect of genetic risk score was constructed using the summary statistic level data and method implemented in gtx package.[32] Genetic risk scores based on more than 5 genetic variants that explain more than 1% of variance in exposure were taken forward.

RESULTS

Genome-wide association study of general cognitive function in individuals of EA

A detailed description of the genome-wide association analysis in individuals of EA is given in the supplementary material. The quantile-quantile plot suggested inflation (λ = 1.62, mean χ^2 = 1.9) (**Supplementary Figure 1, Supplementary Table 4**). However, LD score regression revealed intercept of 1.049 (SE = 0.012) and a ratio of 0.0568 suggesting that inflation is mainly due to polygenicity and only 5.68% of the inflation is due to other causes. General cognitive function was associated with 9,521 genetic variants distributed across all autosomal chromosomes at genome-wide significance level (*p*-value < 5×10^{-8}), of which 358 independent genetic variants mapped to 139 genomic loci (**Supplementary Table 5-7**). After conditioning on genetic signals previously reported in the largest GWAS of general cognitive function to date,[11] 311 genetic variants surpassed genome-wide significance threshold, including 33 novel independent genetic variants mapped to 32 genomic loci (**Figure 1, Supplementary Table 8-10**). The list of pleiotropic associations for these variants and tagged variants is provided in **Supplementary Table 11**, whereas the genes to which these independent variants were mapped to and disease they have been implicated in are listed in **Supplementary Table 12**.

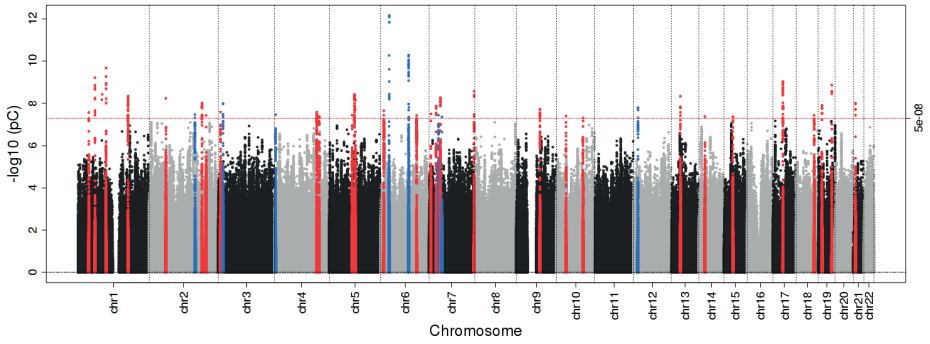


Figure 1. The results of genome-wide association meta-analysis including participants of EA in CHARGE cohorts and UK biobank after conditioning on genetic variants identified in the largest GWAS of general cognitive function to date. The novel loci defined as > 1Mb from previously reported genome-wide variants are depicted in red, whereas the known loci are depicted in blue.

Genome-wide association study of general cognitive function in individuals of AA

There was no variant that surpassed the genome-wide significant threshold in the sample of AA. Among 9,521 genome-wide significant variants associated with general cognitive function in EA cohorts, 66.7% genetic variants had the same direction of effect size in AA individuals and 10.6% variants showed at least nominal evidence of significance (*p*-value < 0.05) (**Supplementary Table 5** and **8** for all loci and independent loci,

respectively). When combining the 148 loci described previously and the 32 discovered in EA in our study into a genetic risk score, the EA risk score was significantly associated with the general cognitive function in AA (p-value = 1.88×10^{-7}).

Correlating genetic determinants of cognition and circulating metabolites

To explore association of the genes involved in general cognitive function and metabolites, we first examined the association of individual genetic variants underlying general cognitive function in the GWAS of circulating metabolites (**Supplementary Table 13**). Two associations surpassed the threshold for multiple testing (**Figure 2**). A genetic variant in *PKD1L3* was associated with lower level of tyrosine (p-value = 5.7×10^{-7}), whereas a variant near *ITIH1* was associated with lower levels of creatinine (p-value = 2.3×10^{-6}). Next, we performed a global genetic test to link general cognitive function using LD score regression. Nominally significant genome-wide genetic overlap was observed between general cognitive function and circulating metabolites including acetate ($\rho_{genetic}$ = 0.21, p-value = 3.9×10^{-3}), citrate ($\rho_{genetic}$ = 0.18, p-value = 8.3×10^{-3}), glycoprotein acetyls ($\rho_{genetic}$ = -0.12, p-value = 2.8×10^{-2}), and 22:6 docosahexaenoic acid (DHA) ($\rho_{genetic}$ = 0.12, p-value = 4.8×10^{-2}) (**Figure 3**).

Mendelian randomization

To evaluate whether the association of the metabolites that were associated with general cognitive function in the single variant or global evaluation are a cause or consequence of the physiological processes underlying general cognitive function, we performed a Mendelian randomization experiment. When testing the hypothesis that the genes determining cognition are also implicated in circulating metabolites, we found evidence for such mechanism for DHA (p-value = 1.3×10^{-5}) when adjusting for multiple testing (**Figure 4**). When testing the hypothesis that the genes determining circulating metabolites also determine general cognitive function, we found evidence for such mechanism for tyrosine (p-value = 5.8×10^{-5}) and glycoprotein acetyls (p-value = 8.99×10^{-3}) (**Figure 4**).

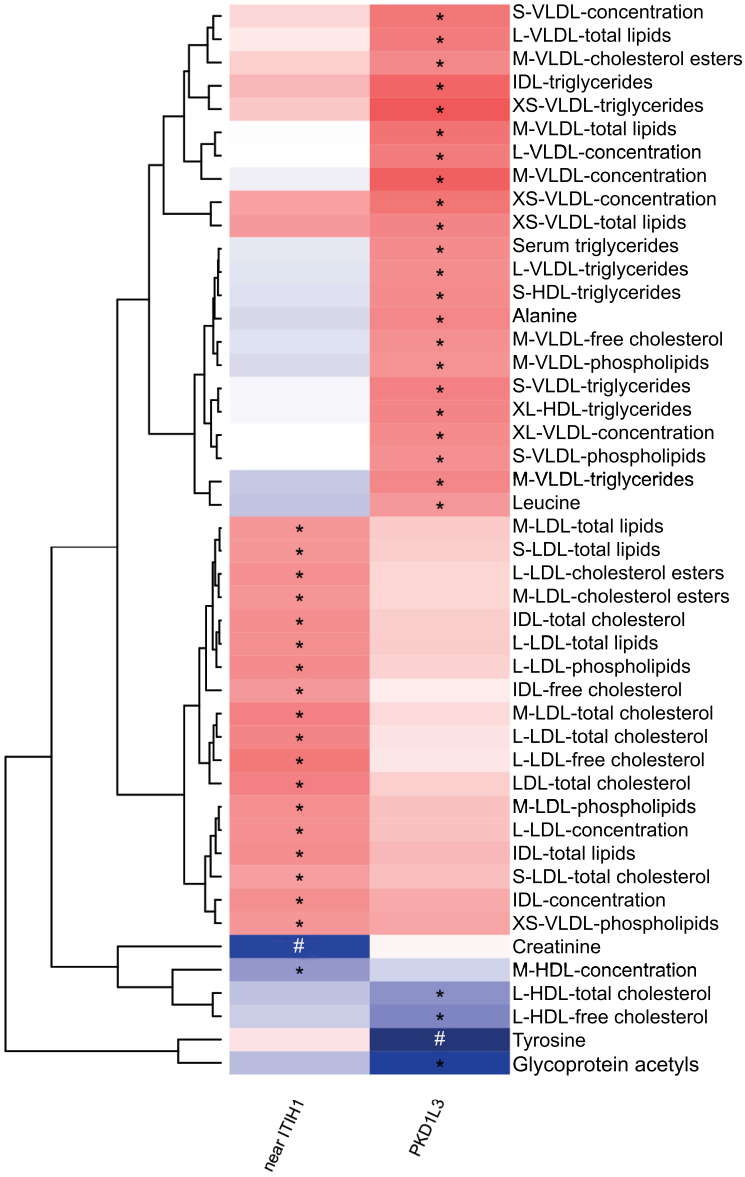


Figure 2. Genetic determinants of general cognitive function and metabolites. Red color denotes positive association and blue color depicts inverse association. Associations that surpassed threshold for multiple testing (p-value < 4.6×10^{-6}) are indicated by hash symbol, whereas nominal associations (p-value < 0.05) are labeled with star.

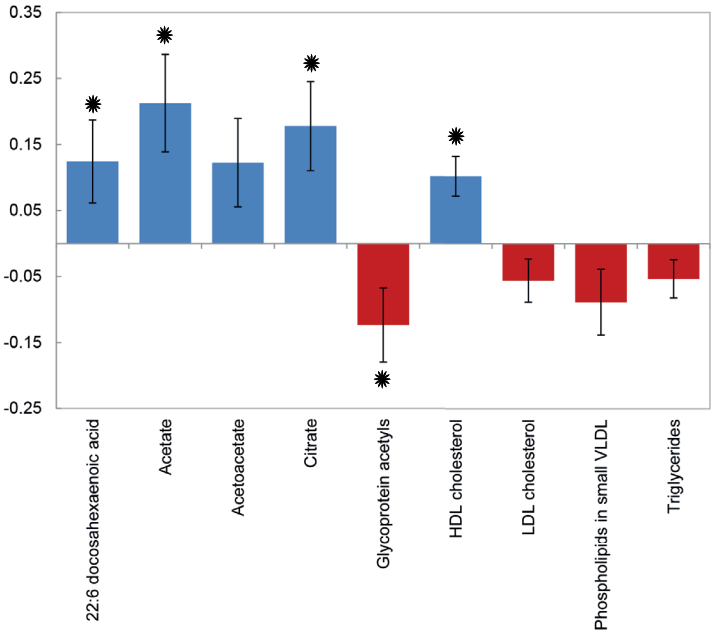


Figure 3. The top results of genetic correlation between the general cognitive function (summary statistic generated in our project) and metabolites for which summary statistic data was available on LD-hub (*p*-value < 0.2). The nominally significant associations (*p*-value < 0.05) are denoted with a star.

DISCUSSION

In a GWAS of general cognitive ability in 243,000 individuals of EA we detected 32 novel findings, bringing the total number of independent loci implicated in general cognitive function up to 180. The risk score based on 180 loci was significantly associated with general cognitive function in AA. Two genes implicated in general cognitive function could directly be linked to circulating levels of tyrosine and creatine, whereas more global genome-wide genetic overlap was found for DHA, glycoprotein acetyl, acetate, and citrate. Mendelian randomization suggests that genes determining glycoprotein acetyl and tyrosine also determine general cognitive function while DHA is rather a consequence of the physiological process determining cognitive function.

Among the 32 novel loci, the variants with the highest probability of having regulatory function based on RegulomeDB score (1f, 1d) were mapped to 17q12 (**Supplementary Table 9**). Deletion of this region has been associated with a syndrome in which about half of the people have delayed development, intellectual disability, or psychiatric disorders such as autism spectrum disorder, schizophrenia, anxiety, and bipolar disorder.[33] The lead variant at 17q12 was mapped to *DHRS11* gene that metabolizes steroid hor-

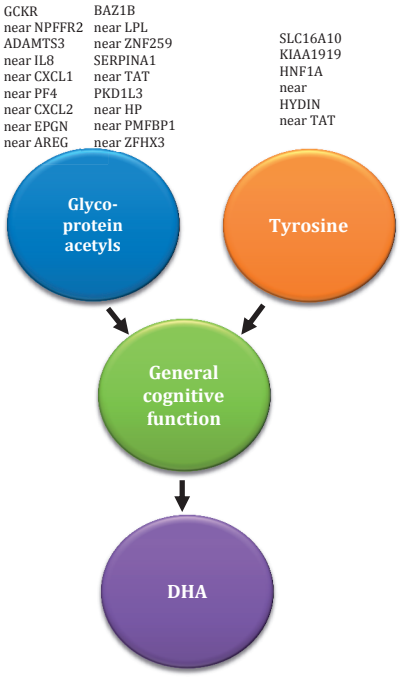


Figure 4. Suggested paths for general cognitive function and metabolites. The genes used in the genetic risk score are located above the metabolite name.

mones, prostaglandins, retinoids, lipids, and xenobiotics (**Supplementary Table 9**).[34] Additionally, other novel locus 6p21.33 had also RegulomeDB score of 1f and is located near *IER3* gene previously related to inflammatory diseases and hypertension.[35,36] Other independent significant variants at novel loci were mapped to genes previously implicated in bipolar disorder, autistic disorder, or near the genes previously implicated in depression, schizophrenia, mood and anxiety disorders (**Supplementary Table 12**).[37-41] Our findings also overlap with those of Savage *et al*. which targeted intelligence rather than general cognitive findings.[42] When performing LD score regression of the two traits the genetic correlation was 0.98. Yet there are differences observed in lead variants in the same regions which imply heterogeneity: statistical or genetic.

This study is unique in two aspects. First, it included a study of genomic variants implicated in general cognitive function in AA. Although this study was too small to yield genome-wide significant findings, the study showed that the variants implicated in EA are also highly significantly determining general cognitive function in AA. Second, our study has successfully examined the role of the genes implicated in general cognitive function in metabolic changes in the circulation. Our findings showed association with circulating metabolites including tyrosine and creatinine. A missense genetic variant in *PKD1L3* at 16q22.2 was associated with general cognitive function (p-value = 1.13×10^{-11}).

This variant, predicted to be damaging (PolyPhen = 0.99), showed association with tyrosine. Tyrosine is an amino acid that plays an important role in synthesis of dopamine, a key neurotransmitter in the brain, and is known to modulate cognitive functions in healthy population by modulating dopamine function.[43,44] Several diseases that involve dopamine dysfunction, such as Parkinson's disease, schizophrenia, and attention deficit hyperactivity, also show alteration of cognitive function.[45] Of note is that there is also a genetic correlation of general cognitive function and schizophrenia and attention deficit hyperactivity disorder, suggesting common pathogenesis. This finding not only highlights the potential mechanism through which the established genetic variants of general cognitive function may act but also opens opportunities to prevent cognitive decline by targeting tyrosine, which is according to our Mendelian randomization experiments most likely in the causal pathway. Also, the finding that the established genetic variant of general cognitive function (rs3755799, 1f RegulomeDB score) modulates levels of creatinine is interesting from a preventive perspective. Creatinine is an organic molecule and a breakdown product of muscle creatine phosphate and it is widely used as a measure of renal function. Higher levels of creatinine and renal impairment were previously associated with lower cognitive performance and increased risk of dementia, asking for further follow-up research in view of potential preventive interventions.[46,47]

In the LD regression, we identified at nominal significance genetic correlation between general cognitive function and several circulating metabolites including DHA, acetate, citrate, and glycoprotein acetyls. These were also included in a formal Mendelian randomization to evaluate whether the metabolites are more likely in the causal pathway or rather a consequence of the various biological processed determining general cognitive function. Unfortunately, the Mendelian randomization experiments for citrate and acetate were not possible because reliable data on genetics was missing. Mendelian randomization suggests that glycoprotein acetyl is causally related to general cognitive function while the association with DHA is rather a consequence of the physiological process determining cognitive function. Circulating levels of DHA, a long-chain omega-3 polyunsaturated fatty acid, have been associated with cognitive function and risk of Alzheimer's disease and dementia.[7] These are strongly associated with fish consumption.[7] However, circulating levels of DHA are also genetically determined and subject of enzymatic processes.[28] Our Mendelian randomization experiment suggests that endogenous processes related to cognitive function are related to levels of DHA rather than that DHA is a driver of general cognitive function. Glycoprotein acetyls appear to be more likely in the causal pathway. Circulating glycoprotein acetyl levels, a marker of acute phase reaction, have been implicated in chronic inflammatory disease and cancer.[48] The levels of this protein were also associated with future risk of all-cause mortality.[49] The fact

that at middle age, these proteins also associate to lower cognitive ability asks for more research on what influences circulating levels of glycoprotein acetyls.[7]

The strengths of our study are large sample size, population-based design, use of large imputation reference panel, and integration of genetic and metabolomics data. However, our study also has limitations. The association results of participating cohorts were combined using a sample-size weighted meta-analysis due to phenotypic heterogeneity limiting discussion on effect sizes. Despite concordance in the direction of effects observed between EA and AA participants, limited sample size of AA sample yielded low statistical power and influenced our ability to explore genetic determinants of general cognitive function in other ethnic groups. Future research studies should focus on non-Europeans. When exploring metabolic pathophysiology underlying genetic variants associated with the general cognitive function, we focused on circulating metabolites measured by NMR technology using Nightingale Health platform.[7] This platform detects various metabolites including amino acids, ketone bodies, fatty acids, and a large proportion of metabolites are lipoproteins and lipid subclasses which provides an excellent opportunity to study cognition.[50] However, the studied metabolites represent only small proportion of circulating metabolites, therefore, future studies focusing on a wider spectrum of metabolites are needed.[51] By improving the power of GWASs of metabolites, novel associations may be revealed.

We have reported association of general cognitive function with 32 novel genetic loci in EA sample and showed that there is genetic overlap of the loci determining general cognitive function in EA and AA. We have also found association of established and novel genetic determinants of general cognitive function and circulating metabolites, providing a starting point for new preventive studies.

REFERENCES

1. Deary, I.J., Weiss, A. & Batty, G.D. Intelligence and personality as predictors of illness and death: How researchers in differential psychology and chronic disease epidemiology are collaborating to understand and address health inequalities. *Psychological Science in the Public Interest***.11**, pp (2010).
2. Batty, G.D., Deary, I.J. & Gottfredson, L.S. Premorbid (early life) IQ and later mortality risk: systematic review. *Ann Epidemiol* **17**, 278-88 (2007).
3. Wraw, C., Deary, I.J., Gale, C.R. & Der, G. Intelligence in youth and health at age 50. *Intelligence. Vol.53 2015, pp. 23-32. Nov-Dec Intelligence* (2015).
4. Hagenaars, S.P., Gale, C.R., Deary, I.J. & Harris, S.E. Cognitive ability and physical health: a Mendelian randomization study. *Sci Rep* **7**, 2651 (2017).
5. Yaffe, K. *et al.* Predictors of maintaining cognitive function in older adults: the Health ABC study. *Neurology* **72**, 2029-35 (2009).
6. Calvin, C.M. *et al.* Intelligence in youth and all-cause-mortality: systematic review with meta-analysis. *Int J Epidemiol* **40**, 626-44 (2011).
7. van der Lee, S.J. *et al.* Circulating metabolites and general cognitive ability and dementia: Evidence from 11 cohort studies. *Alzheimers Dement* (2018).
8. Toledo, J.B. *et al.* Metabolic network failures in Alzheimer's disease: A biochemical road map. *Alzheimers Dement* **13**, 965-984 (2017).
9. Draisma, H.H.M. *et al.* Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nat Commun* **6**, 7208 (2015).
10. Davies, G. *et al.* Genetic contributions to variation in general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE consortium (N=53 949). *Molecular Psychiatry* **20**, 183-192 (2015).
11. Davies, G. *et al.* Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nat Commun* **9**, 2098 (2018).
12. Haworth, C.M. *et al.* The heritability of general cognitive ability increases linearly from childhood to young adulthood. *Mol Psychiatry* **15**, 1112-20 (2010).
13. Plomin, R. & Deary, I.J. Genetics and intelligence differences: five special findings. *Mol Psychiatry* **20**, 98-108 (2015).
14. Davies, G. *et al.* Ninety-nine independent genetic loci influencing general cognitive function include genes associated with brain health and structure (N = 280,360). *biorxiv* (2017).
15. Iglesias, A.I. *et al.* Haplotype reference consortium panel: Practical implications of imputations with large reference panels. *Hum Mutat* **38**, 1025-1032 (2017).
16. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-83 (2016).
17. Psaty, B.M. *et al.* Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circ Cardiovasc Genet* **2**, 73-80 (2009).
18. Davies, G. *et al.* Genetic contributions to variation in general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE consortium (N=53949). *Mol Psychiatry* **20**, 183-92 (2015).
19. Hill, W.D. *et al.* Molecular genetic aetiology of general cognitive function is enriched in evolutionarily conserved regions. *Translational Psychiatry* **6**(2016).

20. Winkler, T.W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc* **9**, 1192-212 (2014).
21. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).
22. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).
23. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
24. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).
25. Boyle, A.P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research* **22**, 1790-1797 (2012).
26. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310 (2014).
27. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **45**, 580-585 (2013).
28. Kettunen, J. *et al.* Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat Commun* **7**, 11122 (2016).
29. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* **95**, 221-227 (2005).
30. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**, 1236-41 (2015).
31. Zheng, J. *et al.* LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272-279 (2017).
32. Liu, J. *et al.* A Mendelian Randomization Study of Metabolite Profiles, Fasting Glucose, and Type 2 Diabetes. *Diabetes* **66**, 2915-2926 (2017).
33. Mitchel, M.W. *et al.* 17q12 Recurrent Deletion Syndrome. (1993).
34. Persson, B. *et al.* The SDR (short-chain dehydrogenase/reductase and related enzymes) nomenclature initiative. *Chem Biol Interact* **178**, 94-8 (2009).
35. Arlt, A. & Schafer, H. Role of the immediate early response 3 (IER3) gene in cellular stress response, inflammation and tumorigenesis. *Eur J Cell Biol* **90**, 545-52 (2011).
36. Shahid, M. *et al.* Impaired 3',5'-cyclic adenosine monophosphate-mediated signaling in immediate early responsive gene X-1-deficient vascular smooth muscle cells. *Hypertension* **56**, 705-12 (2010).
37. Ikeda, M. *et al.* A genome-wide association study identifies two novel susceptibility loci and trans population polygenicity associated with bipolar disorder. *Mol Psychiatry* **23**, 639-647 (2018).
38. Li, X. *et al.* Common variants in the CDH7 gene are associated with major depressive disorder in the Han Chinese population. *Behav Genet* **44**, 97-101 (2014).
39. Hawi, Z. *et al.* The role of cadherin genes in five major psychiatric disorders: A literature update. *Am J Med Genet B Neuropsychiatr Genet* **177**, 168-180 (2018).
40. Sacchetti, E. *et al.* The GRM7 gene, early response to risperidone, and schizophrenia: a genome-wide association study and a confirmatory pharmacogenetic analysis. *Pharmacogenomics J* **17**, 146-154 (2017).
41. Wu, G. *et al.* Central functions of neuropeptide Y in mood and anxiety disorders. *Expert Opin Ther Targets* **15**, 1317-31 (2011).

42. Savage, J.E. *et al.* Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat Genet* **50**, 912-919 (2018).
43. Jongkees, B.J., Hommel, B. & Colzato, L.S. People are different: tyrosine's modulating effect on cognitive control in healthy humans may depend on individual differences related to dopamine function. *Frontiers in Psychology* **5** (2014).
44. Jongkees, B.J., Hommel, B., Kuhn, S. & Colzato, L.S. Effect of tyrosine supplementation on clinical and healthy populations under stress or cognitive demands--A review. *J Psychiatr Res* **70**, 50-7 (2015).
45. Nieoullon, A. Dopamine and the regulation of cognition and attention. *Prog Neurobiol* **67**, 53-83 (2002).
46. Elias, M.F. *et al.* Chronic kidney disease, creatinine and cognitive functioning. *Nephrology Dialysis Transplantation* **24**, 2446-2452 (2009).
47. Seliger, S.L. *et al.* Moderate renal impairment and risk of dementia among older adults: The cardiovascular health cognition study. *Journal of the American Society of Nephrology* **15**, 1904-1911 (2004).
48. Connelly, M.A., Gruppen, E.G., Otvos, J.D. & Dullaart, R.P.F. Inflammatory glycoproteins in cardiometabolic disorders, autoimmune diseases and cancer. *Clin Chim Acta* **459**, 177-186 (2016).
49. Lawler, P.R. *et al.* Circulating N-Linked Glycoprotein Acetyls and Longitudinal Mortality Risk. *Circ Res* **118**, 1106-15 (2016).
50. Hottman, D.A., Chernick, D., Cheng, S., Wang, Z. & Li, L. HDL and cognition in neurodegenerative disorders. *Neurobiol Dis* **72 Pt A**, 22-36 (2014).
51. Wishart, D.S. *et al.* HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic Acids Res* **41**, D801-7 (2013).

SUPPLEMENTARY MATERIAL

Supplementary Note

Supplementary Table 1. Descriptive statistics of study population.

Supplementary Table 2. Information on phenotype assessment.

Supplementary Table 3. Information on genotyping platforms and quality control.

Supplementary Table 4. Genomic inflation factor for each study.

Supplementary Table 5. The variants associated with general cognitive factor at genome-wide significance level.

Supplementary Table 6. Annotation of genome-wide significant variants in CHARGE (EA) + UKB meta-analysis.

Supplementary Table 7. Independent significant SNPs and their distribution across the genomic loci.

Supplementary Table 8. Independent significant SNPs after conditioning on previously reported genetic signal.

Supplementary Table 9. Annotation of genome-wide significant variants after conditioning on previously reported signals.

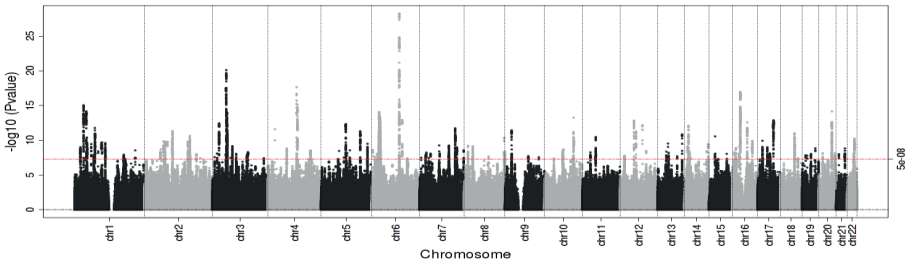
Supplementary Table 10. Independent significant SNPs and their distribution across the genomic loci after conditioning on previously reported genetic signal.

Supplementary Table 11. Overview of pleiotropic associations for novel independent significant genetic variants after conditional analysis and tagged SNPs. Only associations that passed genome-wide significance threshold were listed.

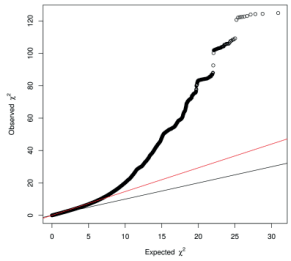
Supplementary Table 12. Genes to which independent genetic variants are mapped to and diseases they have been implicated in according to the DisGeNET database. Only association supported by at least two curated databases (Score > 0.2) are shown.

Supplementary Table 13. Independent genetic variants associated with general cognitive factor in the largest study of general cognitive ability to date and current study that were extracted from GWAS of circulating metabolites.

A)



B)



Supplementary Figure 1. (A) The results of genome-wide association meta-analysis including participants of EA in CHARGE cohorts and UK biobank. The x-axis represents chromosomes and y-axis –log10 *p*-values. Variants are represented by dots. The genome-wide significance threshold (*p*-value < 5×10^{-8}) is depicted by the red dashed line. (B) Quantile-quantile plot of the genetic variants associated with general cognitive function.

Chapter 2.3

Meta-analysis of epigenome-wide association studies of cognitive abilities

Riccardo E. Marioni*, Allan F. McRae*, Jan Bressler*, Elena Colicino*, Eilis Hannon*, Shuo Li*, Diddier Prada*, Jennifer A. Smith*, Letizia Trevisi*, Pei-Chien Tsai*, Dina Vojinovic*, Jeannette Simino, Daniel Levy, Chunyu Liu, Michael Mendelson, Claudia L. Satizabal, Qiong Yang, Min A. Jhun, Sharon L.R. Kardia, Wei Zhao, Stefania Bandinelli, Luigi Ferrucci, Dena G. Hernandez, Andrew B. Singleton, Sarah E. Harris, John M. Starr, Douglas P. Kiel, Robert R. McLean, Allan C. Just, Joel Schwartz, Avron Spiro III, Pantel Vokonas, Najaf Amin, M. Arfan Ikram, Andre G. Uitterlinden, Joyce B.J. van Meurs, Tim D. Spector, Claire Steves, Andrea A. Baccarelli*, Jordana T. Bell*, Cornelia M. van Duijn*, Myriam Fornage*, Yi-Hsiang Hsu*, Jonathan Mill*, Thomas H. Mosley*, Sudha Seshadri* & Ian J. Deary*

* These authors contributed equally to this work

This chapter was published in Mol Psychiatry. 2018 Jan 8. [Epub ahead of print]

The supplemental information for this paper is available online at https://doi.org/10.1038/s41380-017-0008-y

ABSTRACT

Cognitive functions are important correlates of health outcomes across the life-course. Individual differences in cognitive functions are partly heritable. Epigenetic modifications, such as DNA methylation, are susceptible to both genetic and environmental factors and may provide insights into individual differences in cognitive functions. Epigenome-wide meta-analyses for blood-based DNA methylation levels at ~420,000 CpG sites were performed for seven measures of cognitive functioning using data from 11 cohorts. CpGs that passed a Bonferroni correction, adjusting for the number of CpGs and cognitive tests, were assessed for: longitudinal change; being under genetic control (methylation QTLs); and associations with brain health (structural MRI), brain methylation and Alzheimer's disease pathology. Across the seven measures of cognitive functioning (meta-analysis n range: 2,557–6,809), there were epigenome-wide significant (p-value $< 1.7\times10^{-8}$) associations for global cognitive function (cg21450381, p-value = 1.6×10^{-8}), and phonemic verbal fluency (cg12507869, p-value = 2.5×10^{-9}). The CpGs are located in an intergenic region on chromosome 12 and the *INPP5A* gene on chromosome 10, respectively. Both probes have moderate correlations (~0.4) with brain methylation in Brodmann area 20 (ventral temporal cortex). Neither probe showed evidence of longitudinal change in late-life or associations with white matter brain MRI measures in one cohort with these data. A methylation QTL analysis suggested that rs113565688 was a cis methylation QTL for cg12507869 (p-value = 5×10^{-5} and 4×10^{-13} in two lookup cohorts). We demonstrate a link between blood-based DNA methylation and measures of phonemic verbal fluency and global cognitive ability. Further research is warranted to understand the mechanisms linking genomic regulatory changes with cognitive function to health and disease.

BACKGROUND

Cognitive function is an important predictor of health outcomes and mortality.[1,2,3,4] Whether this is due to differences in health literacy and lifestyle choices or if there is a biological predisposition is not clear.[5] The complex balance between genetic and environmental contributions to cognitive function is poorly understood.[6] Epigenetic modifications may provide insight into the link between cognitive function, perturbed biological pathways and relevance for lifelong health.

Molecular genetic studies of unrelated individuals show that around 30% of the variance in general cognitive function can be explained by common genetic polymorphisms (single-nucleotide polymorphisms: SNPs) and variants in linkage disequilibrium with them.[7,8,9] However, there are relatively few well-established individual SNP predictors of cognitive function and those that have been identified explain a very small proportion of the variance in cognitive test scores.[8]

Epigenetic marks may help us better understand the interaction between genes, the environment, and health-related quantitative traits, such as cognitive function, and common disease outcomes.[10, 11] The epigenome helps to regulate genes via, for example, chemical modifications to DNA. DNA methylation typically refers to the addition of a methyl group to a cytosine nucleotide placed next to a guanine in the DNA sequence. The addition or removal of the methyl group is a dynamic process and can be tissue specific with, for example, different epigenetic signatures in blood and brain. The proportion of cytosines methylated at a specific CpG site can be partly explained by both genetics and lifestyle/environment or a combination of these.[12] Studies have examined the association between DNA methylation with genotype,[13, 14] metabolic factors, such as body mass index,[15, 16] and environmental factors, such as smoking.[17] However, no large-scale population-based studies have examined the association of cognitive function with DNA methylation in circulating leucocytes.

One aspect of note for epigenetic epidemiology studies of brain-related traits (cognitive functions, schizophrenia, depression, dementia, etc.) is tissue (and cellular) specificity. As brain samples are not likely to be available until post-mortem, a proxy tissue is an attractive possibility to be explored for building relevant epigenetic signatures. In epidemiological studies, the most likely candidate is blood, which, although its methylation patterns are often dissimilar to those in the brain,[18, 19] they have still been linked to mental health traits.[20,21,22] Identifying robust methylomic differences in relation to cognitive traits may improve our ability to predict cognitive decline and better understand the mechanistic link between cognitive function and deleterious health outcomes.

Here, we examine, using a meta-analytic approach, the associations between blood-based DNA methylation and several individual tests of cognitive functions in up to 6809 healthy, older-aged adults. First, we test which, if any, CpG probes are associated with individual cognitive functions at an epigenome-wide level. Then we investigate these probes to see if they are (1) under genetic control (methQTLs), (2) stable over time, (3) associated with structural brain-imaging measures, (4) associated with Alzheimer's disease case-control status or neuropathology, (5) associated with DNA methylation levels in different brain regions and (6) associated with blood-based gene expression.

METHODS

Overview

Epigenome-wide association studies were performed in 11 independent cohorts for seven cognitive function phenotypes. The number of cohorts contributing to each of the seven tests of cognitive function ranged from 3 to 10 (**Table S1**). A sample-size-based meta-analysis of Z-scores was performed on the overlapping cohort summary output for each cognitive test.

Cohorts

Nine of the eleven cohorts that contributed to the analysis included participants of European ancestry: Framingham Heart Study, InCHIANTI, Lothian Birth Cohort 1921, Lothian Birth Cohort 1936, MOBILIZE Boston, Normative Aging Study, Rotterdam Study (Rotterdam Bios and Rotterdam III) and Twins UK. The Atherosclerosis Risk in the Community (ARIC) and Genetic Epidemiology Network of Arteriopathy's (GENOA) cohorts included participants of African American ancestry. Details of each cohort are presented in **Appendix 1**.

Cognitive measures

Scores from seven different cognitive tests were assessed:

1. Wechsler Logical Memory[23, 24] as a measure of verbal declarative memory. The sum of the immediate and delayed tasks was used.
2. Wechsler Digit Symbol Test[25] or Symbol Digit Modalities Test[26] or Letter Digit Substitution Test[27] as a measure of processing speed, hereafter referred to as Digit Test. The total number of correct answers in the allocated time period was used. The three tests listed above are highly correlated.[28]
3. Semantic Verbal Fluency[29] as a measure of an aspect of executive function (animal naming - total score).

4. Phonemic Verbal Fluency[29] as a measure of an aspect of executive function (letter fluency - total score).
5. Trail Making Test Part B[30] as a measure of an aspect of executive function (Natural log (ln) of the time taken in seconds).
6. Boston Naming Test[31] or National Adult Reading Test[32] or any other measure of vocabulary. The total number of correct answers was assessed.
7. Mini-Mental State Examination (MMSE)[33] as a measure of general cognitive function. Individuals with a score of less than 24 out of 30 were excluded from the analysis.

With the exception of the MMSE scores, any cognitive score that fell above or below 3.5 standard deviations from the mean was set to the mean plus or minus 3.5 standard deviations, respectively. These analyses were performed within each cohort independently for each cognitive test. Full details of the tests available within each cohort are provided in **Appendix 1**.

DNA methylation

Whole-blood DNA methylation was assessed in each cohort using the Illumina HumanMethylation450 BeadChips.[34] Quality control was performed according to cohort-specific thresholds, described in **Appendix 1**. The blood samples for DNA methylation and cognitive ability were measured concurrently.

Structural brain imaging

1.5 T structural brain imaging was assessed in one of the participating epigenome-wide association study (EWAS) cohorts: The Lothian Birth Cohort 1936. Full details have been reported previously.[35] Here, we considered two measures of white matter connectivity -fractional anisotropy (directional coherence of water diffusion) and mean diffusivity (average magnitude of water diffusion) - that have been previously associated with cognitive function.[36, 37]

Gene expression

The association between DNA methylation and gene expression was assessed using the Affymetrix Human Exon 1.0 ST Array in one of the participating cohorts: The Framingham Heart Study. Methodological details are provided in **Appendix 1**.

Ethics

Ethical permission for each cohort is described in **Appendix 1**. Written informed consent was obtained from all subjects.

Statistical analysis

Epigenome-wide association testing

For each cognitive test, two linear regression models were considered - a basic-adjustment model and a full-adjustment model. Both models treated methylation at the CpG sites (untransformed methylation beta value) as the dependent variable with the cognitive test score as the independent predictor of interest. In the basic-adjustment model, covariates included age, sex, white-blood cell counts (either measured or imputed),[38] technical covariates such as plate, chip, array and hybridization date, and, where required, genetic principal components to account for population stratification. In the fully adjusted model, the following additional covariate terms were included: a quadratic term for age, an age × sex interaction; smoking status (current, ever, never) and body mass index. The findings from the fully adjusted model were considered as the primary output. Measurement details for all variables are presented in **Appendix 1**. Age was standardized within cohort to mean 0, variance 1, to avoid potential model convergence issues. Individuals with prevalent dementia or clinical stroke (including self-reported) were excluded.

Quality control filtering

Prior to the meta-analysis, all probes on sex chromosomes were removed along with non-CpG probes, and any cross-reactive probes as reported by Chen *et al*.[39] Genomic correction was applied to any cohort-specific results file with an empirical lambda of more than 1. The total number of probes included in the meta-analysis for each cognitive trait ranged between 421,335 and 421,633.

Trait-specific meta-analysis

The primary analyses were conducted in R;[40] sample-size weighted meta-analyses were conducted in METAL.[41] Several significance thresholds were considered. The most liberal threshold was a within meta-analysis Benjamini-Hochberg false discovery rate of $Q < 0.05$. Next was a within meta-analysis Bonferroni corrected *p*-value threshold: $0.05/n_{probes\_max} = 0.05/421{,}633 = 1.2\times10^{-7}$. Finally, the most conservative threshold applied was a Bonferroni corrected *p*-value that also adjusted for the seven meta-analyses: $0.05/(n_{probes}\times n_{meta\text{-}analyses}) = 0.05/(421{,}633\times7) = 1.7\times10^{-8}$.

Summary meta-analysis combining all cognitive traits

Finally, a meta-analysis of the summary output from the seven meta-analyses was conducted for the fully adjusted models using the CPASSOC software[42] in R. As the cohorts contributed to multiple EWAS, and as the as cognitive test scores are positively correlated,[43] a correlation matrix of the CpG Z-scores for the seven cognitive traits was

included to reduce the false-positive rate.[42] A test assuming heterogeneity was assumed and default input arguments were set.

Methylation quantitative trait loci

To determine if the significant EWAS findings (at the most conservative threshold of p-value < 1.7×10^{-8}) were partly under genetic control, a methylation QTL analysis lookup was performed using data from the Lothian Birth Cohorts of 1921 and 1936 (combined n = 1,366), and the Brisbane Systems Genetics Study (n = 614).[44] The discovery and replication thresholds set in that study were p-value < 1×10^{-11} and p-value < 1×10^{-6}, respectively, with the combined LBC cohorts acting as a discovery dataset (p-value < 1×10^{-11}) with BSGS as the replication study (p-value < 1×10^{-6}) and vice versa. SNPs within 2 Mbp of a CpG site were labeled cis methylation QTLs, and only the most significant SNP for each CpG were considered.

Longitudinal change in methylation

For the significantly associated CpG probes identified in the meta-analyses, longitudinal data from the Lothian Birth Cohort 1936 were used to chart change in methylation at these CpGs between ages 70 and 76 years. Stability in methylation levels might be indicative of potential genetic control or a long-term fixed effect of differential cognitive function on the probe. Variability in methylation levels may be a by-product or cause of cognitive change over time. Methylation data were available on participants at ages 70 (n = 920), 73 (n = 800) and 76 (n = 618) years. Linear mixed models with random intercept terms, adjusting for sex, imputed white-blood cell counts and technical variables, were used to determine the rate of change over time (the coefficient for the fixed effect age variable in the model) for each probe.

Structural brain-imaging associations with methylation

As cognitive function is a brain-related phenotype, it was of interest to see if blood-based methylation signatures for cognitive function were related to brain-imaging measures. Structural MRI data and covariate information were also available in 552 participants at the second wave of the Lothian Birth Cohort 1936 - data from only the first wave of the cohort were included in the EWAS. The top associations from the EWAS meta-analyses were assessed at the second wave of the Lothian Birth Cohort 1936 in relation to age- and sex-adjusted brain structural fractional anisotropy and mean diffusivity using linear regression models, adjusting for age, sex, imputed white cell counts and technical covariates.

Blood-brain methylation correlations

Lookup analyses of significant CpG sites were performed in published data sets for both blood and brain (prefrontal cortex, entorhinal cortex, superior temporal gyrus, and cerebellum) based EWAS findings for Braak staging and Alzheimer's disease status.[21] A second lookup was performed using results from blood and Brodmann areas 7, 10 and 20 from post-mortem samples of 16 individuals.[45]

Gene expression associations

Transcriptome-wide association studies (TWAS) were conducted in the Framingham Heart Study for any significant probes from the cognitive EWAS. Linear mixed effects models with expression of each gene as the dependent variable, methylation as exposure and identical covariates to the EWAS were considered. A Bonferroni correction was applied (p-value $< 0.05/n_{probes} = 0.05/17{,}873 = 2.8\times10^{-6}$) to identify statistically significant associations.

RESULTS

Study sample characteristics

Participants came from 11 cohorts - ranging in size from 219 to 2,307 individuals (Q1–Q3: 435–920), with between 0 and 100% female participants (Q1–Q3: 52–65%), mean age ranged from 56 to 79 years (Q1–Q3: 60–73). Two of the cohorts (ARIC and GENOA) included participants of African American ancestry; all other cohorts included participants of European ancestry. The cohort-specific summary details for each cognitive test are presented in **Supplementary Table 1**. The basic-adjustment meta-analytic sample-size ranged from 2,557 individuals for the Trail Making Test to 6,809 individuals for the MMSE. Similar sample-sizes were observed for the fully adjusted models with the meta-analytic results presented in **Fig. 1** and **Table 1**.

Epigenome-wide association study model diagnostics

Heterogeneity was observed in the EWAS inflation statistics, both within and across cohorts (**Supplementary Table 2**). For example, the minimum and maximum lambda values in LBC1936 were 1.05 and 1.25, respectively. Prior to meta-analysis, within-cohort genomic correction was applied where lambda exceeded 1. The meta-analysis genomic inflation statistics for the basic and fully adjusted models ranged from 0.93 to 1.30, and 0.92 to 1.26, respectively (**Table 1**).

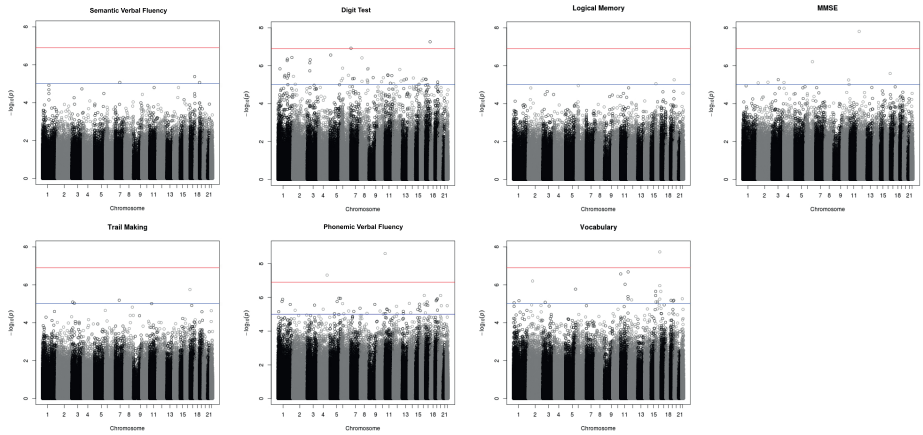


Fig. 1. Meta-analysis EWAS Manhattan Plots for the seven cognitive tests - full adjustment models*. *Models adjusted for age, sex, age × sex, age^2, self-reported smoking status, body mass index, white-blood cell counts, technical covariates and principal components (population stratification).

Table 1. Summary of meta-analysis results for the seven cognitive tests

	Cognitive test						
	Phonemic Verbal Fluency	**MMSE**[a]	**Trail Making**	**Logical Memory**	**Vocabulary**	**Digit Test**	**Semantic Verbal Fluency**
$N_{participants\ m1}$	6405	6809	2557	2988	3013	4794	3678
λ_{m1}	1.3	1.21	0.95	0.97	1.08	1.06	0.93
$N\ loci_{m1}$	4	1	0	0	3	29	0
$N_{participants\ m2}$	6390	6780	2549	2983	3007	4780	3658
λ_{m2}	1.26	1.16	0.97	0.99	1.1	1.03	0.92
$N\ loci_{m2}$	2	1	0	0	1	2	0

[a]MMSE Mini-Mental State Examination;
m1 model 1, adjusted for age, sex, cell counts, technical covariates and population stratification (genetic principal components-cohort specific);
m2 model 2, adjusted for model 1 covariates, smoking, body mass index, age^2 and an age × sex interaction term;

Epigenome-wide association study of seven cognitive traits

A list of the within-test epigenome-wide significant associations within a given cognitive test across both models are presented in **Supplementary Table 3**. Significant associations (p-value $< 1.2\times10^{-7}$) were observed in the basic and full adjustment models for Phonemic Verbal Fluency (n = 4 and n = 2), MMSE (n = 1 for both models), Vocabulary (n = 3 and n = 1), and Digit Test (n = 29 and n = 2). From the basic-adjustment model, significant CpGs were located in genes associated with, for example: alcohol metabolism

(*ALDH2*, Digit Test, cg12142865),[46] smoking (*AHRR*, Digit Test, cg05575921),[17] inflammation (*CCR9* and *PRRC2A*, cg10475172 and cg14943908, respectively)[47, 48] and neurodegeneration through the beta-amyloid precursor protein interactor GAPDH (Digit Test, cg00252813).[49] In the fully adjusted model, significant CpGs were located in genes associated with, for example: inflammation (*SOCS3*, Digit Test, cg18181703),[50] epithelial cell splicing (*ESRP2*, Vocabulary, cg04513006)[51] and transcription activation of NOTCH proteins (*MAML3*, Phonemic Verbal Fluency, cg16201957).[52] No CpGs were significantly associated with the Trail Making, Logical Memory or Semantic Verbal Fluency tests. Methylation at cg21450381 was not associated with any of the six other cognitive traits in the fully adjusted meta-analytic results at a nominal significance threshold of *p*-value < 0.05 (**Table 2**). However, cg12507869 was associated with lower scores for both Logical Memory (*p*-value = 0.043) and Vocabulary (*p*-value = 9.4×10^{-5}).

Table 2. Lookup of top EWAS associations across all cognitive tests in the fully adjusted models. The *p*-values for the initial EWAS associations at *p*-value < 1.7×10^{-8} are highlighted in bold

Cognitive Test	***N***	***Z***	***P***
Digit Test			
cg21450381	4780	0.51	0.61
cg12507869	4780	-1.44	0.15
Vocabulary			
cg21450381	3007	-0.61	0.54
cg12507869	3007	-3.91	9.4×10^{-5}
Semantic Verbal Fluency			
cg21450381	3658	-1.11	0.27
cg12507869	3658	-1.18	0.24
Logical Memory			
cg21450381	2983	-1.65	0.099
cg12507869	2983	-2.03	0.043
MMSE			
cg21450381	6780	-5.66	**1.6×10^{-8}**
cg12507869	6780	-1.26	0.21
Trail-making Test			
cg21450381	2549	1.06	0.29
cg12507869	2549	0.87	0.38
Phonemic Verbal Fluency			
cg21450381	6390	-0.61	0.54
cg12507869	6390	-5.96	**2.5×10^{-9}**

MMSE Mini-mental state examination;

Variation in results when modifying the significance threshold

Using a less conservative FDR correction for multiple testing identified associations at a q-value threshold of 0.05 in both the basic and fully adjusted models for Phonemic Verbal Fluency (n = 49 and n = 2), MMSE (n = 1 for both models), Vocabulary (n = 7 and n = 3) and Digit Test (n = 309 and n = 14). The FDR-significant probes are presented in **Supplementary Table 4**.

After Bonferroni correction for CpG sites and the seven cognitive traits - *p*-value < 0.05/(420,000×7) - two remaining differentially methylated CpGs were cg21450381 (R^2 = 0.47%, *p*-value = 1.6×10^{-8}) with MMSE scores, and cg12507869 (R^2 = 0.55%, *p*-value = 2.5×10^{-9}) with Phonemic Verbal Fluency. In both cases, higher methylation was associated with lower cognitive scores across all of the contributing cohorts. cg21450281 is located in an intergenic region of chromosome 12; cg12507869 is located in the inositol polyphosphate-5-phosphatase, 40 kDa (*INPP5A*) gene on chromosome 10. Both probes were approximately normally distributed in the Lothian Birth Cohort 1936 (**Fig. 2**). A forest plot of the Z-scores by cohort sample-size is presented in **Fig. 3** and shows no evidence of ethnic outliers or single cohorts driving the associations.

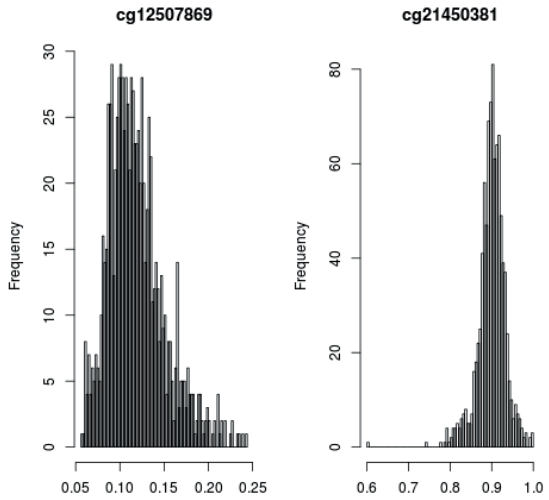


Fig. 2. Histogram showing the distribution of beta values for the two significant CpGs in the Lothian Birth Cohort 1936 (n = 920).

Combined meta-analysis of all seven cognitive traits

There was no evidence from the combined meta-analytic results of the seven tests for a globally significant CpG across all tests in the fully adjusted model (minimum Benjamini-Hochberg FDR q-value of 0.057 for cg12507869).

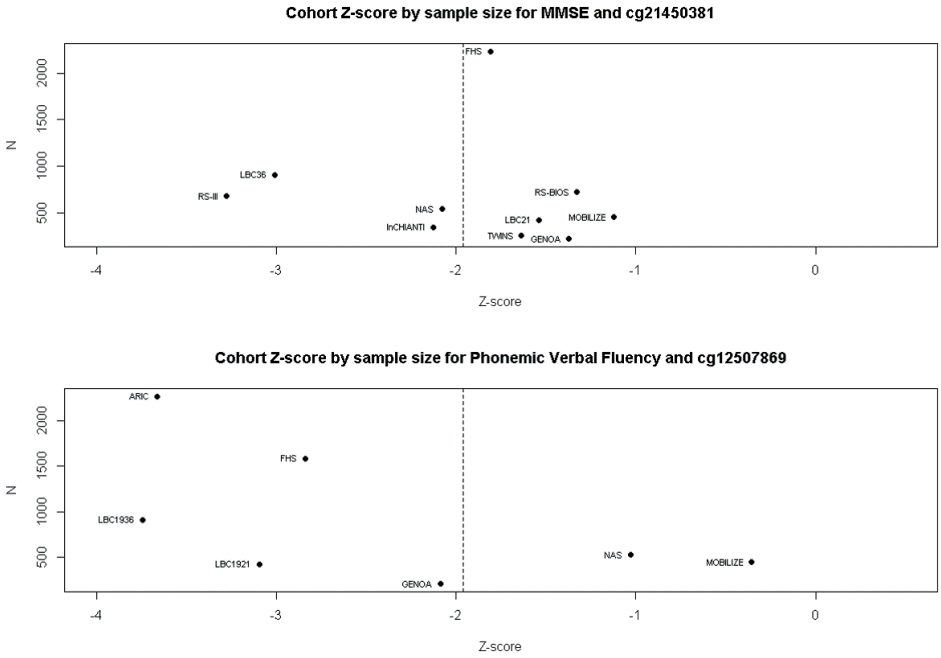


Fig. 3. Forest plots of the Z-scores by cohort sample size for the two significant CpGs. ARIC Atherosclerosis Risk in the Community, FHS Framingham Heart Study Offspring Cohort, GENOA Genetic Epidemiology Network of Arteriopathy, InCHIANTI Invecchiare in Chianti, LBC Lothian Birth Cohort, MOBILIZE Maintenance of Balance, Independent Living, Intellect and Zest in the Elderly of Boston, NAS Normative Aging Study, RS Rotterdam Study, RS-Bios Rotterdam Study-Biobank-based Integrative Omics Studies.

Genetic contributions to cognitive-related differential methylation

A methylation QTL lookup[44] analyses identified no SNPs to be associated with cg21450381. The top SNP for cg12507869 (rs113565688 in the *INPP5A* gene on chromosome 10) explained around 1.2% of the variance in methylation (p-values of 3.6×10^{-13} and 5.4×10^{-5} in the Australian and Scottish cohorts, respectively). There is no overlap of this SNP with cognitive traits based on a recent GWAS conducted in the UK Biobank cohort: rs113565688 association with memory (p-value = 0.55), reaction time (p-value = 0.42), verbal-numerical reasoning (p-value = 0.17) and educational attainment (p-value = 0.13).[8]

Longitudinal changes in methylation at cognitive-related differential methylation sites

Longitudinal analyses over three waves of data (ages 70, 73 and 76 years) from LBC1936, adjusting for sex, imputed white-blood cell counts and technical variables, found no evidence for a linear change in the methylation of either probe over a relatively narrow

period in later-life. The mixed model standardized effect size for change in cg12507869 was 0.02 standard deviations per year, *p*-value = 0.13; the standardized effect size for cg21450381 was -0.02, *p*-value = 0.40. Without adjustment for covariates, the across wave correlations for cg12507869 were 0.62 (age 70: age 73), 0.63 (age 70: age 76) and 0.68 (age 73: age 76). The corresponding correlations for cg21450381 were 0.04, 0.10 and 0.30, respectively.

Association of brain MRI features with cognitive-related differential methylation

There were no significant associations between the top two CpGs and either of the brain MRI measures of white matter connectivity (mean diffusivity minimum *p*-value = 0.56; fractional anisotropy minimum *p*-value = 0.28) at age 73 in the LBC1936 (n = 552).

Correlation of blood and brain methylation at the cognitive-related differential methylation

Two blood-brain comparisons were conducted. The first, using a blood-brain DNA methylation comparison tool[18] [http://epigenetics.essex.ac.uk/bloodbrain/], provided no evidence for a significant correlation between blood-methylation at either probe with methylation in four brain regions (prefrontal cortex, entorhinal cortex, superior temporal gyrus and cerebellum, **Supplementary Figs. 1** and **2**). Whereas the mean of the cg21450381 probe was similar to the means for the four brain regions, the mean of the cg12507869 probe in blood was markedly different (hypomethylated) to the means for the prefrontal cortex, entorhinal cortex, superior temporal gyrus (**Supplementary Figs. 1** and **2**). It was, however, similar to the mean of the cerebellum. The second comparison, using BECon[45] [https://redgar598.shinyapps.io/BECon/] showed the same mean methylation levels for cg21450381 between blood and Brodmann areas 7, 10 and 20; cg12507869 was again hypomethylated in blood compared to the three brain regions. There were moderate correlations between blood-methylation and Brodmann area 20 for both CpGs (r = 0.43 for cg12507869 and r = -0.46 for cg21450381) and between Brodmann area 7 and cg12507869 (r = 0.31).

Association of cognitive-related differential methylation with Braak staging and Alzheimer's disease

None of the six CpGs that were epigenome-wide significant in the fully adjusted EWASs at *p*-value $< 1.2\times10^{-7}$ were associated with Braak staging or Alzheimer's case-control status in either blood or brain-based methylation (minimum FDR q-value 0.51, **Supplementary Table 5**).

Transcriptome-wide association study

There were no significant TWAS results for cg21450381. The minimum *p*-value observed was 0.00013 (Q = 0.51). There were nine significant TWAS results for cg12507869 at *p*-value < 2.8×10^{-6} and 41 at Q < 0.05. There was a nominal inverse association between the INPP5A transcript and CpG (*p*-value = 0.049, Q = 0.65). The full TWAS output for the two CpGs is shown in **Supplementary Tables 6** and **7**.

DISCUSSION

This study presents a meta-analysis of the relationship between blood-based DNA methylation and cognitive function. We analyzed seven different cognitive tests and found two epigenome-wide methylation correlations: cg21450381, located in an intergenic region of chromosome 12, with global cognitive function (as measured by the MMSE); and cg12507869, located in the *INPP5A* gene on chromosome 10, with phonemic verbal fluency. Methylation at the latter CpG was also associated with two other cognitive tests (logical memory and vocabulary) at a nominal *p*-value < 0.05 threshold. Genetic analyses of the top two CpGs showed a modest cis regulation for one of the probes, suggesting that the vast majority of the methylation variation at the cognitive-related differentially methylated sites are due to environmental influences. Blood-based methylation levels at both of the CpGs correlated with methylation levels in Brodmann area 20 (cerebral cortex).

INPP5A is a member of the inositol polyphosphate-5-phosphatase (*INPP5*) family of genes that encode enzymes that hydrolyze inositol 1,4,5 triphosphate (IP3). It is involved in the mobilization of intracellular calcium and has been implicated in cerebellar degeneration in mice.[53] A second *INPP5* family member, *INPP5D*, has been associated with Alzheimer's Disease and cognitive decline,[54, 55] further implicating this gene family in cognitive functions. cg21450381 is located in an intergenic region of chromosome 12, that contains a histone modification mark (H3K27Ac), DNAaseI hypersensitivity clusters and evidence of transcription factor-binding sites, which indicates that the region may be involved in gene regulation.[56]

In a TWAS analysis of the top two probes in the Framingham Heart Study (n > 1,900), there was no evidence for an association between cg21450381 and blood-based gene expression. Of the nine Bonferroni-significant transcripts in the TWAS of cg12507869, eight were trans associations, with the cis association occurring in *ADAM12*, which is more than 6 Mb from *INPP5A*. There was no evidence of a cis effect of the CpG on the *INPP5A* expression levels.

Disentangling correlation from causation is particularly tricky when studying epigenetic marks in a non-target tissue. By increasing the sample sizes of the meta-analytic EWAS and replicating any findings across different cognitive domains will reduce the chances of false-positive associations. It is, of course, possible that a reliable blood-based epigenetic marker of cognitive function may be several degrees of separation away from the biological processes that drive cognitive skills. For example, the signal could be in response to neurotoxic events, such as inflammation, oxidative stress or small vessel disease. However, the discrimination of cause from consequence is something that affects many epigenetic epidemiology studies. Approaches that may overcome this include Mendelian randomization studies where a methQTL can be used as an instrument or the use of mouse models to dissect functional consequences of DNA methylation on gene regulation.

There are additional limitations of this study: a varying number of participants with cognitive data available for each test; heterogeneity in relation to the ethnicity and geographical location of the participants across cohorts; and relating a blood-based methylation signature to a brain-based outcome. We attempted to counter these limitations by: plotting cohort sample-size by Z-score to see if there was bias due to outliers or clustering by ethnicity; adjusting for population stratification in the cohorts with admixture; correlating the blood-based CpG associations with methylation levels in several brain regions; looking at the association between brain region-specific methylation and Alzheimer's disease phenotypes for the blood-based CpG associations. It is possible that bias may have been introduced in the secondary analyses that focused on the MRI, gene expression and longitudinal methylation data, as both the LBC1936 and Framingham studies contributed to the discovery meta-analyses. Re-running the meta-analyses without these cohorts yielded: *p*-values of 1.3×10^{-7} and 7.1×10^{-6} for the phonemic verbal fluency finding (cg12507869), excluding Framingham and LBC1936, respectively; and *p*-values of 1.7×10^{-8} and 3.3×10^{-6} for the MMSE finding (cg21450381), again excluding Framingham and LBC1936, respectively. Whereas the longitudinal methylation and MRI findings were null, the cis and trans expression-methylation associations warrant confirmation in an independent sample. The methQTL findings were based on highly stringent discovery and replication *p*-value thresholds in both LBC and an independent cohort, BSGS.

Neither of the top two CpGs showed signs of linear change in methylation levels between the ages of 70 and 76 years in one of the participating studies (LBC1936) that had three waves of longitudinal data. It is possible that non-linear changes may be present although additional waves of data collection would be required to test this robustly. In

addition, a 6-year window is possibly too narrow to observe substantial changes in the CpG levels.

It is notable that the two significant CpG associations were found for the cognitive tests that were completed by the largest number of participants (n > 6,000). The study provided results for a list of cognitive tests that cover several major cognitive domains: memory, processing speed, executive function, vocabulary and global ability. The heterogeneity with respect to ethnicity and geographic location can allow us to generalize our findings to multiple populations.

Blood is the most feasible tissue for epigenetic epidemiology analyses of cognitive function. Brain would be the ideal target tissue although this would make it impossible to have simultaneous cognitive function data. Moreover, epigenome-wide studies of other brain-related outcomes, such as schizophrenia, have identified putative blood-based methylation signatures.[22]

In conclusion, we have presented evidence for blood-based epigenetic correlates of cognitive function. Specifically, we identified methylation sites that are linked to an aspect of executive function and global cognitive ability. The latter finding relied on a relatively crude cognitive test (the MMSE), which is commonly used to identify individuals at risk of dementia. One of the two CpG sites identified was under modest genetic control, with a cis SNP explaining over 1% of its variance. Unlike other traits, such as smoking and body mass index,[15, 17] there are relatively modest methylation signatures for cognitive function. However, our analyses concur with other recent studies to suggest that blood-based methylation signatures may be useful tools to interrogate differences in brain-related outcomes.

REFERENCES

1. Calvin, C.M. *et al.* Intelligence in youth and all-cause-mortality: systematic review with meta-analysis. *Int J Epidemiol* **40**, 626-44 (2011).
2. Deary, I.J., Weiss, A. & Batty, G.D. Intelligence and personality as predictors of illness and death: How researchers in differential psychology and chronic disease epidemiology are collaborating to understand and address health inequalities. *Psychological Science in the Public Interest* **11.2**, 53-79 (2010).
3. Henderson, M., Richards, M., Stansfeld, S. & Hotopf, M. The association between childhood cognitive ability and adult long-term sickness absence in three British birth cohorts: :a cohort study. *Bmj Open* **2** (2012).
4. Sorberg, A. *et al.* Cognitive Ability in Late Adolescence and Disability Pension in Middle Age: Follow-Up of a National Cohort of Swedish Males. *Plos One* **8** (2013).
5. Hagenaars, S.P. *et al.* Shared genetic aetiology between cognitive functions and physical and mental health in UK Biobank (N=112151) and 24 GWAS consortia. *Molecular Psychiatry* **21**, 1624-1632 (2016).
6. Deary, I.J., Johnson, W. & Houlihan, L.M. Genetic foundations of human intelligence. *Human Genetics* **126**, 215-232 (2009).
7. Davies, G. *et al.* Genetic contributions to variation in general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE consortium (N=53 949). *Molecular Psychiatry* **20**, 183-192 (2015).
8. Davies, G. *et al.* Genome-wide association study of cognitive functions and educational attainment in UK Biobank (N=112151). *Molecular Psychiatry* **21**, 758-767 (2016).
9. Marioni, R.E. *et al.* Molecular genetic contributions to socioeconomic status and intelligence. *Intelligence* **44**, 26-32 (2014).
10. Bakulski, K.M. & Fallin, M.D. Epigenetic epidemiology: Promises for public health research. *Environmental and Molecular Mutagenesis* **55**, 171-183 (2014).
11. Rakyan, V.K., Down, T.A., Balding, D.J. & Beck, S. Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics* **12**, 529-541 (2011).
12. Shah, S. *et al.* Genetic and environmental exposures constrain epigenetic drift over the human life course. *Genome Research* **24**, 1725-1733 (2014).
13. Lemire, M. *et al.* Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. *Nature Communications* **6** (2015).
14. Gaunt, T.R. *et al.* Systematic identification of genetic influences on methylation across the human life course. *Genome Biology* **17** (2016).
15. Dick, K.J. *et al.* DNA methylation and body-mass index: a genome-wide analysis. *Lancet* **383**, 1990-1998 (2014).
16. Mendelson, M.M. *et al.* Association of Body Mass Index with DNA Methylation and Gene Expression in Blood Cells and Relations to Cardiometabolic Disease: A Mendelian Randomization Approach. *Plos Medicine* **14** (2017).
17. Joehanes, R. *et al.* Epigenetic Signatures of Cigarette Smoking. *Circulation-Cardiovascular Genetics* **9**, 436-447 (2016).
18. Hannon, E., Lunnon, K., Schalkwyk, L. & Mill, J. Interindividual methylomic variation across blood, cortex, and cerebellum: implications for epigenetic studies of neurological and neuropsychiatric phenotypes. *Epigenetics* **10**, 1024-1032 (2015).

19. Walton, E. *et al.* Correspondence of DNA Methylation Between Blood and Brain Tissue and Its Application to Schizophrenia Research. *Schizophrenia Bulletin* **42**, 406-414 (2016).
20. Hannon, E. *et al.* An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation. *Genome Biology* **17** (2016).
21. Lunnon, K. *et al.* Methylomic profiling implicates cortical deregulation of ANK1 in Alzheimer's disease. *Nature Neuroscience* **17**, 1164-1170 (2014).
22. Montano, C. *et al.* Association of DNA Methylation Differences With Schizophrenia in an Epigenome-Wide Association Study. *Jama Psychiatry* **73**, 506-514 (2016).
23. Wechsler, D. WMS-IIIUK administration and scoring manual, (Psychological Corporation, London, UK, 1998).
24. Wechsler, D. Wechsler Memory Scale - Revised, (Psychological Corporation, New York, NY, New York, 1987).
25. Wechsler, D. WAIS-IIIUK administration and scoring manual, (Psychological Corporation, London, UK, 1998).
26. Smith, A. Symbol Digit Modalities Test manual - revised, (Western Psychological Services, Los Angeles, CA, USA, 1992).
27. van der Elst, W., van Boxtel, M.P.J., van Breukelen, G.J.P. & Jolles, J. The Letter Digit Substitution Test: Normative data for 1,858 healthy participants aged 24-81 from the Maastricht Aging Study (MAAS): Influence of age, education, and sex. *Journal of Clinical and Experimental Neuropsychology* **28**, 998-1009 (2006).
28. Ibrahim-Verbaas, C.A. *et al.* GWAS for executive function and processing speed suggests involvement of the CADM2 gene. *Molecular Psychiatry* **21**, 189-197 (2016).
29. Lezak, M. Neuropsychological testing, (Oxford University Press, Oxford, UK, 2004).
30. Bowie, C.R. & Harvey, P.D. Administration and interpretation of the Trail Making Test. *Nat Protoc* **1**, 2277-81 (2006).
31. Kaplan, E., Goodglass, H. & Weintraub, S. Boston Naming Test, (Lea & Febiger, Philadelphia, 1983).
32. Nelson, H.E. & Willison, J.R. National Adult Reading Test (NART) Test Manual (Part II), (NFER-Nelson, Windsor, UK, 1991).
33. Folstein, M.F., Folstein, S.E. & McHugh, P.R. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res* **12**, 189-98 (1975).
34. Bibikova, M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288-95 (2011).
35. Wardlaw, J.M. *et al.* Brain aging, cognition in youth and old age and vascular disease in the Lothian Birth Cohort 1936: rationale, design and methodology of the imaging protocol. *International Journal of Stroke* **6**, 547-559 (2011).
36. Booth, T. *et al.* Brain White Matter Tract Integrity and Cognitive Abilities in Community-Dwelling Older People: The Lothian Birth Cohort, 1936. *Neuropsychology* **27**, 595-607 (2013).
37. Penke, L. *et al.* Brain white matter tract integrity as a neural foundation for general intelligence. *Molecular Psychiatry* **17**, 1026-1030 (2012).
38. Houseman, E.A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *Bmc Bioinformatics* **13** (2012).
39. Chen, Y.A. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8**, 203-209 (2013).
40. R Core Team R: A language and environment for statistical computing. (R Foundation for Statistical Computing, Vienna, Austria, 2017).

41. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).
42. Zhu, X. *et al.* Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *Am J Hum Genet* **96**, 21-36 (2015).
43. Spearman, C. "General Intelligence," objectively determined and measured. *Am J Psychol*, 201-92 (1904).
44. McRae, A.F. *et al.* Identification of 55,000 Replicated DNA Methylation QTL. *bioRxiv* (2017).
45. Edgar, R.D., Jones, M.J., Meaney, M.J., Turecki, G. & Kobor, M.S. BECon: A tool for interpreting DNA methylation findings from blood in the context of brain. *bioRxiv* (2017).
46. Edenberg, H.J. The genetics of alcohol metabolism: role of alcohol dehydrogenase and aldehyde dehydrogenase variants. *Alcohol Res Health* **30**, 5-13 (2007).
47. Hashimoto, M. *et al.* Genetic contribution of the BAT2 gene microsatellite polymorphism to the age-at-onset of insulin-dependent diabetes mellitus. *Hum Genet* **105**, 197-9 (1999).
48. Zabel, B.A. *et al.* Human G protein-coupled receptor GPR-9-6/CC chemokine receptor 9 is selectively expressed on intestinal homing T lymphocytes, mucosal lymphocytes, and thymocytes and is required for thymus-expressed chemokine-mediated chemotaxis. *J Exp Med* **190**, 1241-56 (1999).
49. Butterfield, D.A., Hardas, S.S. & Lange, M.L. Oxidatively modified glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and Alzheimer's disease: many pathways to neurodegeneration. *J Alzheimers Dis* **20**, 369-93 (2010).
50. Carow, B. & Rottenberg, M.E. SOCS3, a Major Regulator of Infection and Inflammation. *Front Immunol* **5**, 58 (2014).
51. Warzecha, C.C., Sato, T.K., Nabet, B., Hogenesch, J.B. & Carstens, R.P. ESRP1 and ESRP2 Are Epithelial Cell-Type-Specific Regulators of FGFR2 Splicing. *Molecular Cell* **33**, 591-601 (2009).
52. Wu, L., Sun, T., Kobayashi, K., Gao, P. & Griffin, J.D. Identification of a family of mastermind-like transcriptional coactivators for mammalian notch receptors. *Molecular and Cellular Biology* **22**, 7688-7700 (2002).
53. Yang, A.W., Sachs, A.J. & Nystuen, A.M. Deletion of Inpp5a causes ataxia and cerebellar degeneration in mice. *Neurogenetics* **16**, 277-285 (2015).
54. Lambert, J.C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics* **45**, 1452-U206 (2013).
55. Andrews, S.J., Das, D., Anstey, K.J. & Easteal, S. Late Onset Alzheimer's Disease Risk Variants in Cognitive Decline: The PATH Through Life Study. *Journal of Alzheimers Disease* **57**, 423-436 (2017).
56. Rosenbloom, K.R. *et al.* ENCODE Data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Research* **41**, D56-D63 (2013).

Chapter 2.4

The dystrophin gene and cognitive function in the general population

Dina Vojinovic, Hieab H.H. Adams, Sven J. van der Lee, Carla A. Ibrahim-Verbaas, Rutger Brouwer, Mirjam C.G.N. van den Hout, Edwin Oole, Jeroen van Rooij, Andre Uitterlinden, Albert Hofman, Wilfred F.J. van IJcken, Annemieke Aartsma-Rus, GertJan B. van Ommen, M. Arfan Ikram, Cornelia M. van Duijn, Najaf Amin

This chapter was published in Eur J Hum Genet. 2015 Jun;23(6):837-43.

The supplemental information for this paper is available online on European Journal of Human Genetics website (http://www.nature.com/ejhg)

ABSTRACT

The aim of our study is to investigate whether single-nucleotide dystrophin gene (*DMD*) variants associate with variability in cognitive functions in healthy populations. The study included 1,240 participants from the Erasmus Rucphen Family (ERF) study and 1,464 individuals from the Rotterdam Study (RS). The participants whose exomes were sequenced and who were assessed for various cognitive traits were included in the analysis. To determine the association between *DMD* variants and cognitive ability, linear (mixed) modeling with adjustment for age, sex and education was used. Moreover, Sequence Kernel Association Test (SKAT) was used to test the overall association of the rare genetic variants present in the *DMD* with cognitive traits. Although no *DMD* variant surpassed the prespecified significance threshold (p-value < 1×10^{-4}), rs147546024:A>G showed strong association (β = 1.786, p-value = 2.56×10^{-4}) with block-design test in the ERF study, while another variant rs1800273:G>A showed suggestive association (β = −0.465, p-value = 0.002) with Mini-Mental State Examination test in the RS. Both variants are highly conserved, although rs147546024:A>G is an intronic variant, whereas rs1800273:G>A is a missense variant in the *DMD* which has a predicted damaging effect on the protein. Further gene-based analysis of *DMD* revealed suggestive association (p-values = 0.087 and 0.074) with general cognitive ability in both cohorts. In conclusion, both single variant and gene-based analyses suggest the existence of variants in the *DMD* which may affect cognitive functioning in the general populations.

INTRODUCTION

The dystrophin gene (*DMD*) is localized on the X chromosome. Variants in *DMD* have been recognized as a cause of the most common form of muscular dystrophy during childhood, Duchenne muscular dystrophy (DMD).[1] This fatal, X-linked disorder leads to progressive muscle weakness and less well-described non-progressive central nervous system (CNS) manifestations.[2]

A consistent finding among patients with DMD is the reduction in full-scale intelligence quotient. Although most individuals are not intellectually disabled, risk for cognitive impairment is increased among affected males and up to 30% of patients have intellectual disability.[3, 4, 5] Apart from intellectual abilities, frequently reported neurocognitive function impairment has been published.[6] Deficits in short-term memory, executive functions, visuospatial ability, as well as deficits in some aspect of attention, problems with narrative, linguistic and reading skills have been described, irrespective of general intelligence.[7-12] Moreover, a higher incidence of different neuropsychiatric disorders, such as autism spectrum, attention deficit hyperactivity disorder, obsessive-compulsive disorders, and social behavior problems has been revealed among affected males.[13-17]

The impact of *DMD* on cognitive ability in cognitively healthy populations has not been studied to the best of our knowledge; therefore, in the current study we aim to investigate whether single-nucleotide *DMD* variants associate with variability in cognitive functions in general populations, suggesting loci in the *DMD* contributing to cognition, besides genuine *DMD* variants.

MATERIAL AND METHODS

Study populations

Our study population consisted of subjects from Erasmus Rucphen Family (ERF) and Rotterdam Study (RS). ERF is a family-based study that includes inhabitants of a genetically isolated community in the South-West of the Netherlands, studied as part of the Genetic Research in Isolated Population (GRIP) program.[18] Study population includes ~3,000 individuals who are living descendants of 22 couples who had at least six children baptized in the community church. All data were collected between 2002 and 2005. The population shows minimal immigration and high inbreeding; therefore, frequency of rare alleles is increased in this population. All participants gave informed consent, and the Medical Ethics Committee of the Erasmus University Medical Centre approved the study.

The RS is a prospective, population study from a well-defined Ommoord district in the Rotterdam city that investigates the occurrence and determinants of diseases in the elderly.[19] The cohort was initially defined in 1990 among ~7,900 persons who underwent a home interview and extensive physical examination at the baseline and during follow-up rounds every 3–4 years. Cohort was extended in 2000 and 2005.[19] RS is an outbred population, predominantly of Dutch origin. The Medical Ethics Committee of the Erasmus Medical Center, Rotterdam, approved the study. Written informed consent was obtained from all participants.

Data collection procedure

Participants from both cohorts underwent extensive neuropsychological examination. In ERF study, different cognitive domains were assessed using Dutch validated battery of neuropsychological tests.[20,21] We focused on neurocognitive domains which are known to be affected in patients with DMD.[8-12] General cognitive ability was assessed with the Dutch Adult Reading Test (DART). Memory function was measured with a word learning test from which immediate recall and learning scores were derived while executive function was assessed with the Trail Making Test (TMT) parts A and B[22] and verbal fluency tests.[22] Visuospatial ability was assessed with the WAIS-III block-design subtest.

In the RS, global cognitive function was assessed with the Mini-Mental State Examination (MMSE) test, while executive function and information processing speed were assessed with the Letter-Digit Substitution Task (LDST),[23] the Word Fluency Test (WFT)[24] and the abbreviated Stroop test.[25] Examination was performed at baseline (MMSE) and during follow-up rounds (MMSE, LDST, and WTF).

Participants from both cohorts who had dementia or clinical stroke were excluded from the analysis as these conditions can influence neuropsychological assessment.

Genotyping/sequencing

The exomes of 1,336 individuals from the ERF population were sequenced 'in-house' at the Center for Biomics of the Cell Biology Department of the Erasmus MC, The Netherlands, using the Agilent version V4 capture kit (Agilent Technologies, Santa Clara, CA, USA) on an Illumina Hiseq2000 sequencer (Illumina, San Diego, CA, USA) using the TruSeq Version 3 protocol (Illumina). The sequence reads were aligned to the human genome build 19 (hg19) using BWA and the NARWHAL pipeline.[26,27] The aligned reads were processed further using the IndelRealigner, MarkDuplicates and TableRecalibration tools from the Genome Analysis Toolkit (GATK) and Picard (http://picard.sourceforge.net). Genetic variants were called using the Unified Genotyper tool of the GATK. About 1.4 million single-nucleotide variants (SNVs) were called and after removing the low-

quality variants (QUAL < 150) we retrieved 577,703 SNVs in 1,309 individuals. Further, for prediction of the functionality of the variants, annotations were performed using the SeattleSeq database (http://snp.gs.washington. edu/SeattleSeqAnnotation131).

In the RS, exomes of 1,764 individuals from the RS-I population were sequenced using the Nimblegen SeqCap EZ V2 capture kit (Roche NimbleGen, Madison, WI, USA) on an Illumina Hiseq2000 sequencer and the TruSeq Version 3 protocol. The sequences reads were aligned to the hg19 using Burrows-Wheeler Aligner.[27] Subsequently, the aligned reads were processed further using Picard (http://picard.sourceforge.net), SAMtools[28] and GATK.[29] Genetic variants were called using Unified Genotyper Tool from GATK. Samples with low concordance to genotyping array (< 95%), low transition/transversion ratio (< 2.3) and high heterozygote to homozygote ratio (> 2.0) were removed from the data. The final data set consisted of 903,316 SNVs in 1,524 individuals.

Statistical analysis

Baseline descriptive analysis was performed with SPSS version 17 (IBM, New York, NY, USA). Deviation from normality of cognitive functions was assessed by histograms and P-P plots. As the ERF study includes related individuals, all single variants in *DMD* were tested for association applying additive linear-mixed modeling with the 'mmscore' function adjusting for age, sex and education in the GenABEL library of the R software.[30] The 'mmscore' function uses the relationship matrix estimated from genomic data in the linear mixed model to correct for relatedness among the samples. Additionally, for the most interesting results gender-stratified analysis was also performed. As most of these cognitive tests are correlated (the Pearson correlation coefficient ranged from 0.219 to 0.670), to adjust for multiple testing we first calculated the effective number of independent tests using the eigenvalues of a correlation matrix using the Matrix Spectral Decomposition (matSpDlite) software,[31] finally Bonferroni correction was applied for the effective number of independent tests. The same strategy was also adopted for modeling linkage disequilibrium between the SNVs of the *DMD*. Considering the number of independent cognitive tests and independent variants, the significance threshold was set to 0.05/(4 independent cognitive tests×124 independent variants) = 1×10^{-4}, whereas suggestive threshold was set to 1/(4 independent cognitive tests×124 independent variants) = 2×10^{-3}. SNVs were coded 0, 1, 2 for genotypes AA, AB, BB in females, respectively, and 0, 2 for genotypes A, B in males.

Since sequencing is likely to reveal several variants that may be population specific, we also performed the gene-based Sequence Kernel Association Test (SKAT), a test specifically designed to analyze rare sequence variation in a specific gene/region.[32] Assessing the joint effect of multiple variants within the gene/region, the SKAT is proposed as a

more powerful approach for rare variants than a classical single variant analysis and several burden tests.[32] The significance threshold for gene-wise analysis was set to 0.05/4 independent cognitive tests = 0.0125, while the suggestive threshold was set to 1/4 independent test = 0.25.

To assess the relationship between the SNVs outside the protein-coding regions with gene expression in the tissue, we used the Genotype-Tissue Expression (GTEx) project database.[33]

The data were deposited in the GWAS Central database, under the accession number HGVST1824 (http://www.gwascentral.org/study/HGVST1824).

RESULTS

General characteristics of the studied populations are shown in **Table 1**. The mean age in ERF was 48 years and 39% of the participants were males while mean age in RS was around 68 years and 44% of the participants were males. Around 30% of participants in the ERF study had only primary education compared with around 36% subjects in the RS.

Table 1. Descriptive statistics of the study populations

	ERF	**RS baseline**	**RS follow up**
N	1241	1464	902
Age	47.9 (14.4)	68.1 (9.4)	72.0 (7.1)
Gender (% of males)	39.3%	44.3%	44.8%
Education (% of only primary education)	29.8%	35.6%	29.3%
Cognitive tests			
Dutch Adult Reading Test, mean (sd)	58.56 (20.31)		
AVLT - Immediate recall, mean (sd)	4.37 (1.69)		
AVLT - Learning, mean (sd)	33.55 (9.01)		
Ratio TMT-B / TMT-A, mean (sd)	2.68 (1.02)		
Verbal fluency, mean (sd)	61.66 (18.21)		
Block design test, mean (sd)	8.24 (2.77)		
Mini-mental state examination, mean (sd)		27.7 (1.8)	27.7 (2.0)
Letter-Digit Substitution Task, mean (sd)			27.0 (7.2)
Word Fluency Test, mean (sd)			21.3 (5.5)

Abbreviations: AVLT, Auditory Verbal Learning Test; ERF, Erasmus Rucphen Family; N, number of participants; RS, Rotterdam Study; TMT-A, TMT-B, Trail Making Test parts A and B.

Number of SNVs in the *DMD* discovered by exome sequencing was 165 in the ERF and 482 in the RS (**Supplementary Table 1**). Around 70% of variants in the *DMD* had minor allele frequency (MAF) lower than 0.05 in ERF compared with around 98% of variants in the RS.

The results of the association analysis between SNVs in the *DMD* and cognitive functions with nominal level of significance in ERF study are presented in **Table 2**. Although none of the findings surpassed multiple testing correction using a Bonferroni threshold of 1×10^{-4}, strong association was observed between rs147546024:A>G (β = 1.786, *p*-value = 2.56×10^{-4}) and the block-design test. Gender-stratified analysis showed nominally significant association in both genders (β = 1.796, *p*-value = 0.009 in males and β = 1.623, *p*-value = 0.018 in females). This rare (A→G) variant with MAF of 0.011 was localized in the intron 1 of the *DMD* (chrX.hg19:g.33146086A>G) and although being highly conserved over species (conservation score GERP = 4.08) has an unknown effect on the protein. On the basis of localization, we studied the relationship of this variant with gene expression in human tissues GTEx database but no significant eQTLs were found for this variant. The family-based design of the ERF study allowed us to check whether all the carriers (n = 24) of this variant were closely related. All carriers were connected to each other in 10 generations (**Figure 1**).

Next, we explored the association of rs147546024:A>G in the population-based study (RS). Even though rs147546024:A>G is a previously identified genetic variation in dbSNP database (present in 6 copies in 1000 Genomes with an MAF of 0.004) it was not present in RS and was not in linkage disequilibrium with any of the other SNVs of *DMD*. This prompted us to look for overlapping variants between the two studies. Among 34 overlapping variants we identified the most interesting overlapping finding that is shown in **Table 3**. Among these variants, rs1800273 (chrX.hg19:g.31986607G>A) had similar MAF in both studies (0.038 in the ERF and 0.033 in the RS), similar effect size and same direction of the effect in both cohorts and was suggestively associated with block-design test in the ERF study (β = −0.424, *p*-value = 0.066) and with MMSE in RS (β = −0.465, *p*-value = 0.002) (**Table 3**). This G→A variant is localized in exon 45 of the *DMD* and is classified as a missense variant with a predicted damaging effect on the protein (PolyPhen score = 0.99, conservation score GERP = 2.52). This variant is present in 23 copies in 1000 Genomes with an MAF of 0.014. All carriers of the variant in the ERF were connected to each other (**Figure 2**).

In the gene-based analysis using SKAT suggestive associations (*p*-values 0.087 and 0.074) were also observed both in ERF and in RS for DART and MMSE, respectively.

Table 2. Association of *DMD* variants with cognitive abilities in ERF study

Cognitive test	Name	Genomic position[a]	Reference allele	Variant allele	N	Effect	SE	Nominal p-value	MAF	HWE p-value	PolyPhen prediction	GERP conservation score
General cognitive ability												
Dutch Adult Reading Test	rs72470515	32716133	G	C	1222	3.839	1.456	8.59E-03	0.042	0.392	unknown	0.018
	rs72470514	32716132	G	T	1225	3.226	1.419	2.35E-02	0.043	0.392	unknown	-1.75
	rs1800278	31496426	T	C	1225	-3.448	1.528	2.45E-02	0.035	1	0.281	1.66
	rs41305353	31496431	T	A	1225	-3.448	1.528	2.45E-02	0.035	1	0.981	5.4
	rs183429765	31838024	C	T	1225	-9.496	4.213	2.47E-02	0.004	1	unknown	-1.47
	rs17338590	31497369	T	C	1146	-3.246	1.530	3.44E-02	0.034	0.006	unknown	-0.067
	rs16989970	31950056	G	A	1215	-3.053	1.460	3.72E-02	0.038	0.161	unknown	4.25
	rs17309542	32614065	A	G	1225	-1.815	0.882	4.03E-02	0.124	0.081	unknown	2.76
	rs5927082	32591811	A	G	1225	-1.639	0.798	4.07E-02	0.160	0.499	unknown	2.12
	rs5927083	32591931	T	C	1225	-1.639	0.798	4.07E-02	0.160	0.499	unknown	-1.32
	rs72468656	32459449	A	G	1221	-8.105	4.089	4.82E-02	0.006	1	unknown	3.9
	rs72466537	31165350	G	C	1225	-2.814	1.428	4.96E-02	0.042	0.105	unknown	2.83
Memory												
AVLT - Immediate recall	rs1800279	31496398	T	C	1228	0.311	0.138	3.04E-02	0.035	0.282	0.01	2.92
	23:32715801	32715801	G	A	1221	0.836	0.408	4.85E-02	0.003	1	unknown	2.84
AVLT - Learning	23:32715801	32715801	G	A	1221	6.139	2.015	2.93E-03	0.003	1	unknown	2.84
	rs2293667	31224881	A	G	1228	1.161	0.472	1.62E-02	0.076	0.467	unknown	1.29
	rs2293668	31224684	G	A	1228	1.161	0.472	1.62E-02	0.076	0.467	unknown	3.96
	rs2293666	31224994	G	A	1194	1.145	0.468	1.70E-02	0.077	0.141	unknown	3.65

Table 2. Association of *DMD* variants with cognitive abilities in ERF study (*continued*)

Cognitive test	Name	Genomic position[a]	Reference allele	Variant allele	N	Effect	SE	Nominal p-value	MAF	HWE p-value	PolyPhen prediction	GERP conservation score
	23:31838262	31838262	A	G	1228	-7.458	3.541	3.97E-02	0.001	1.000	unknown	-1.35
	rs1800279	31496398	T	C	1228	1.419	0.685	4.30E-02	0.035	0.282	0.01	2.92
Executive												
Ratio TMT-B / TMT-A	rs7891425	32361033	C	T	1223	-0.101	0.048	3.99E-02	0.140	0.072	unknown	5.36
	rs56094071	32430503	A	T	1202	-0.098	0.048	4.55E-02	0.149	0.570	unknown	5.15
Verbal fluency	rs72468668	32486917	T	G	1225	7.426	3.124	2.12E-02	0.007	1	unknown	2.06
	rs72470511	32663417	G	A	1229	-8.246	3.758	3.34E-02	0.004	1	unknown	2.15
	rs12837503	32404249	A	G	1229	-4.038	1.993	4.95E-02	0.018	1	unknown	-5.13
Visuospatial												
Block design test	rs147546024	33146086	A	G	1211	1.786	0.470	**2.56E-04**	0.011	1	unknown	4.08
	rs72470511	32663417	G	A	1218	-2.144	0.629	1.01E-03	0.004	1	unknown	2.15
	rs183429765	31838024	C	T	1220	-1.673	0.650	1.32E-02	0.004	1	unknown	-1.47
	23:32834523	32834523	A	G	1220	-2.513	1.043	2.03E-02	0.002	1	unknown	0.531

Abbreviations; AVLT, Auditory Verbal Learning Test; DMD, dystrophin gene; GERP, the program that generates the conservation score; HWE, Hardy-Weinberg equilibrium; MAF, minor allele frequency; N, number of individuals; SE, standard error; TMT-A, TMT-B, Trail Making Test parts A and B. The most significant finding is printed in bold.
[a]Genomic positions are according to hg19 assembly.

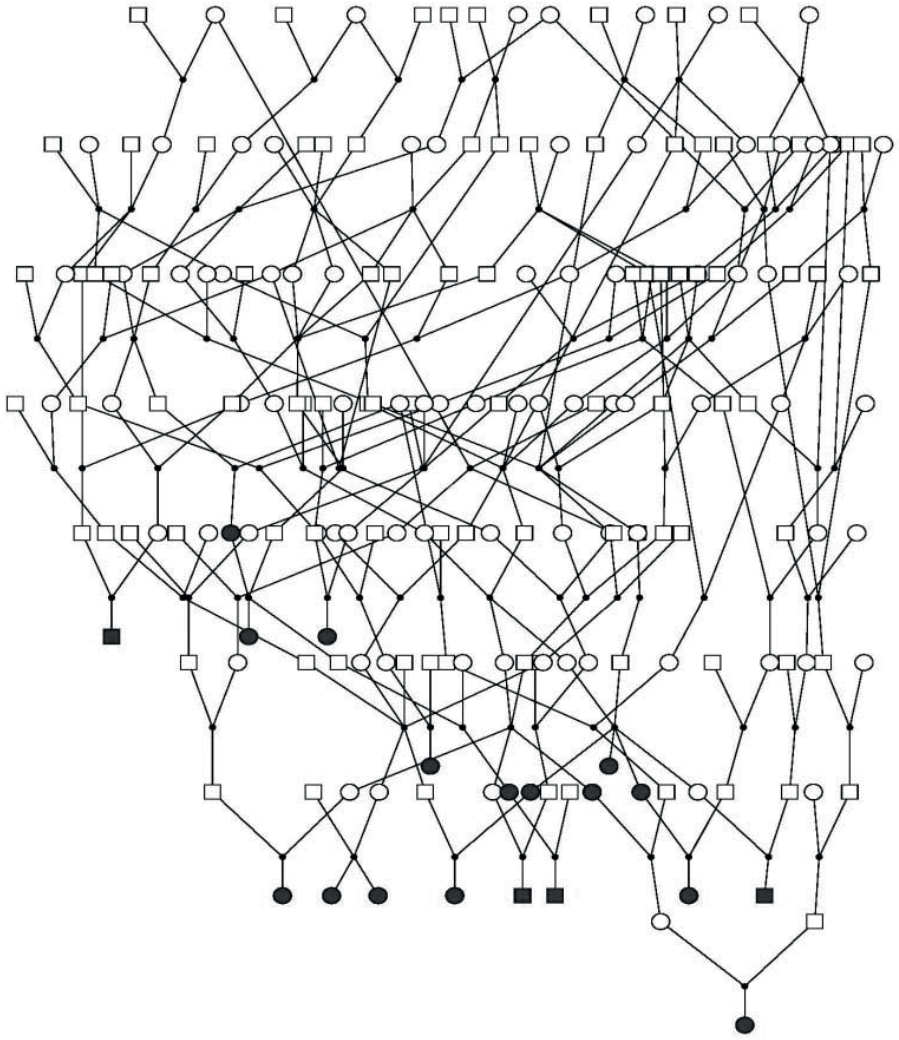


Figure 1. Carriers of the SNV that achieved the strongest association in the ERF. Carriers are indicated in black.

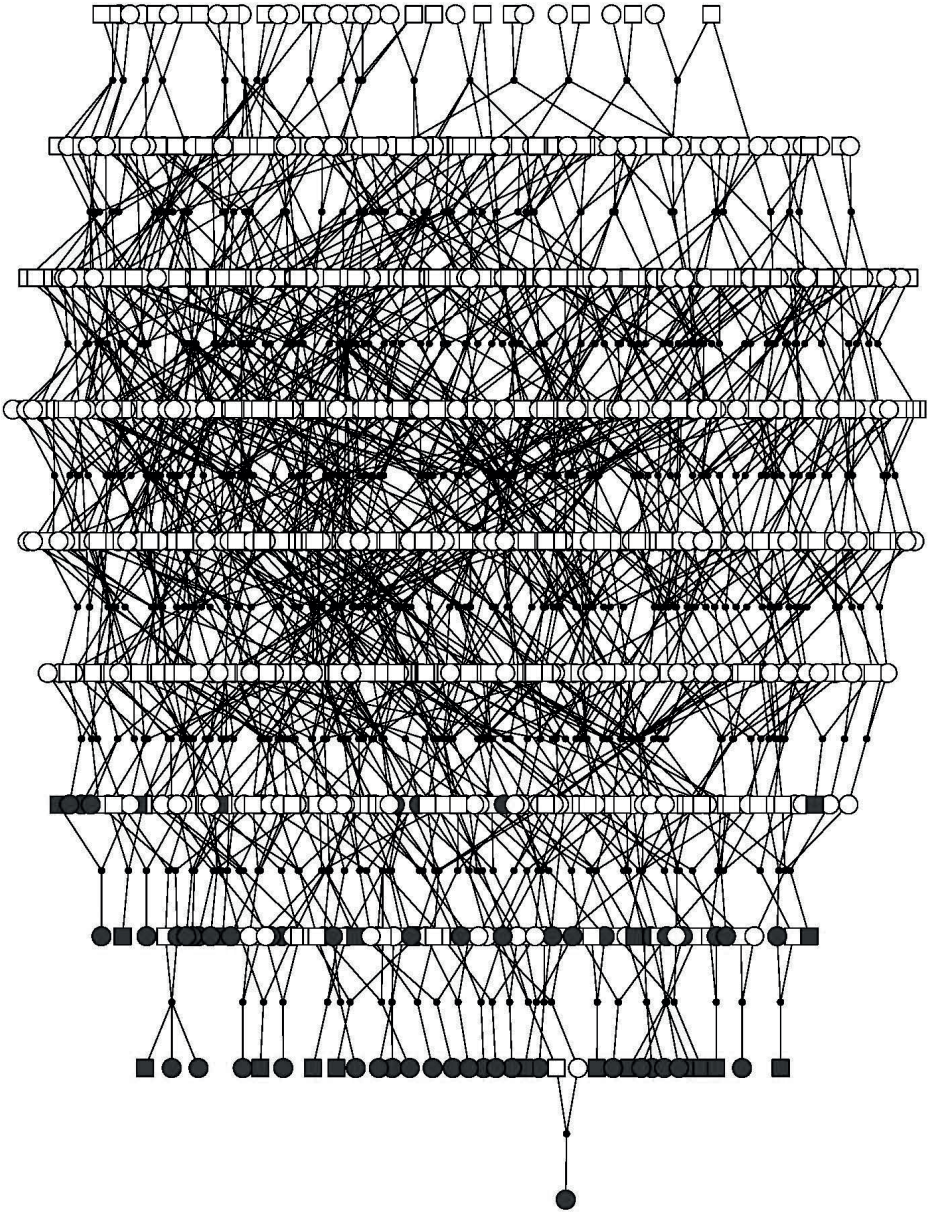


Figure 2. Carriers of the overlapping SNV in the ERF. Carriers are indicated in black.

Table 3. Overlapping variant in both cohorts

	Name	Genomic position[a]	N	Reference allele	Variant allele	Effect	SE	*p-value*	MAF	PolyPhen prediction	GERP conservation score
ERF											
Block design test	rs1800273	31986607	1220	G	A	-0.424	0.222	**0.066**	0.038	0.999	2.52
RS											
MMSE	rs1800273	31986607	1418	G	A	-0.465	0.151	**0.002**	0.033	0.999	2.52

Abbreviations: ERF, Erasmus Rucphen Family; GERP, the program that generates the conservation score; MAF, minor allele frequency; MMSE, mini-mental state examination; N, number of individuals; RS, Rotterdam study; SE, standard error.

[a]Genomic positions are according to hg19 assembly.

DISCUSSION

The aim of this study was to investigate possible impact of genetic variants in the *DMD* on cognitive ability in the general population. Even though none of the *DMD* variants surpassed the prespecified significance threshold, rs147546024:A>G was suggestively associated with block-design test in ERF, whereas rs1800273:G>A was nominally associated with MMSE test in the RS and marginally associated with block-design test in ERF.

rs147546024:A>G is localized in the intron 1,196 bp far from the promoter of full-length protein isoform (Dp427p), which is expressed predominantly in the Purkinje cells of the hippocampus. The frequency of this variant in 1000 Genomes was observed to be 0.005 in individuals of European origin compared with ERF where the frequency was 0.011. This enrichment is expected due to genetic drift and isolation of the ERF population.[18] Functional prediction of this variant showed high conservation score and unknown effect on the protein while gene expression analysis found no significant eQTLs in various human tissues. Interestingly, the rare allele of rs147546024:A>G was associated with better cognitive performance on block-design test which is designed to assess visuo-spatial ability. Similar to some studies which have described a sex difference in cognitive ability with a male advantage on the spatial domains,[34] our study confirmed slight, but not significant, higher scoring of males on block-design test. It is known that better performance on block-design test is associated with autistic spectrum disorder[35-37] and *DMD* is recognized as one of susceptibility genes for autism disorder.[38,39] Suppression of the global configuration to process the information in a detailed manner, essential for this test, is described as a main characteristic of autistic patients.[40-43]

Another biologically interesting finding while searching for overlapping variants in both studies was the missense G→A variant, rs1800273:G>A, which we found associated with block-design test in ERF and the test of global cognitive ability (MMSE) in RS. This variant was observed at a frequency of 0.033 in the individuals of European origin and absent in those of African and Asian origin. Localized in exon 45 of the *DMD*, this variant was classified as a missense variant with a predicted damaging effect on the protein. Since the *DMD* has three upstream and four intragenic promoters that control expression of full-length (Dp427c, Dp427m and Dp427p) and short protein isoforms (Dp260, Dp140, Dp116 and Dp71), exon 45 is present in the four different isoforms (Dp427c, Dp427m, Dp427p and Dp260) among which Dp427c and Dp427p are expressed in the brain.[44] The Dp427c is expressed predominantly in neurons of the cortex and the CA regions of the hippocampus. It has been shown that this form of protein dystrophin colocalizes with inhibitory GABA receptor clusters at the postsynaptic membranes of hippocampal and neocortical pyramidal neurons where the synapse function is modulated.[45, 46, 47, 48]

According to various studies this dystrophin isoform has a stabilizing effect on the GABA receptors by limiting their lateral diffusion outside the synapse.[49,50] Importance of GABA receptors for the regulation of cognition, emotion and memory is increasingly being recognized.[51,52] The Dp427p is expressed in the cerebellar and hippocampal Purkinje cells and in the cortical brain.[53,54] However, exon 45 does not affect three shorter *DMD* isoforms (Dp140, Dp116 and Dp71) which are known to be associated with cognitive function in DMD.[55,56] rs1800273:G>A was detected earlier in DMD patients and is present in the Leiden Muscular dystrophy database.[57] Since majority of DMD patients have cognitive impairment, the association of rs1800273:G>A with DMD may represent association with cognitive impairment. However, the presence of this variant and lack of the dystrophin protein - which can by itself lead to cognitive impairment - would make it difficult to study the separate effect of this variant in DMD patients.

One of the difficulties that our study had to deal with is heterogeneity in classification of phenotypes. Even though various cognitive tests are used in the studied populations, different cognitive domains can be compared since they are correlated. Therefore, moderate correlation (the Pearson correlation coefficient of 0.429, p-value < 0.0001) between visuospatial ability and global cognition ability in the ERF, as well as correlation (the Pearson correlation coefficient of 0.460, p-value < 0.0001) between visuospatial ability and executive function which is recognized as a central domain of cognitive functioning[58,59] allow us to compare association of the most interesting overlapping variant with block-design test in the ERF and MMSE test in the RS.

The majority of variants called in our study were rare variants. Even though there is growing evidence that rare variants contribute to etiology of different complex traits, the search for rare variants is very difficult and challenging. Standard methods used to test for association with single common genetic variants are not powerful enough for the analysis of rare variants.[60, 61, 62] Therefore with the available sample size, our study had limited power to detect association. This we attempted to overcome using the recently proposed gene-based analysis (SKAT) design for rare variant analysis.[32] Assessing the cumulative effect of multiple variants in *DMD* implied only suggestive p-value for both cohorts. Still like other approaches that deal with rare variants this approach also has limitations in terms of power but suggestive p-values generated by SKAT pointed out that variants in the *DMD* may affect cognitive functioning in healthy populations.

In conclusion, analyzing the sequence variants in the exon of *DMD* in two cognitively healthy cohorts we find evidence of association of *DMD* with cognitive functioning in healthy individuals. Larger studies are required for confirmation.

REFERENCES

1. Hoffman, E.P., Brown, R.H., Jr. & Kunkel, L.M. Dystrophin: the protein product of the Duchenne muscular dystrophy locus. *Cell* **51**, 919-28 (1987).
2. Anderson, J.L., Head, S.I., Rae, C. & Morley, J.W. Brain function in Duchenne muscular dystrophy. *Brain* **125**, 4-13 (2002).
3. Cotton, S., Voudouris, N.J. & Greenwood, K.M. Intelligence and Duchenne muscular dystrophy: full-scale, verbal, and performance intelligence quotients. *Dev Med Child Neurol* **43**, 497-501 (2001).
4. Emery, A.E.H. Duchenne muscular dystrophy, x, 270 p. (Oxford University Press, Oxford ; New York, 2003).
5. Cotton, S.M., Voudouris, N.J. & Greenwood, K.M. Association between intellectual functioning and age in children and young adults with Duchenne muscular dystrophy: further results from a meta-analysis. *Dev Med Child Neurol* **47**, 257-65 (2005).
6. Sollee, N.D., Latham, E.E., Kindlon, D.J. & Bresnan, M.J. Neuropsychological impairment in Duchenne muscular dystrophy. *J Clin Exp Neuropsychol* **7**, 486-96 (1985).
7. Cyrulnik, S.E., Fee, R.J., De Vivo, D.C., Goldstein, E. & Hinton, V.J. Delayed developmental language milestones in children with Duchenne's muscular dystrophy. *J Pediatr* **150**, 474-8 (2007).
8. Wicksell, R.K., Kihlgren, M., Melin, L. & Eeg-Olofsson, O. Specific cognitive deficits are common in children with Duchenne muscular dystrophy. *Dev Med Child Neurol* **46**, 154-9 (2004).
9. Hinton, V.J., De Vivo, D.C., Nereo, N.E., Goldstein, E. & Stern, Y. Poor verbal working memory across intellectual level in boys with Duchenne dystrophy. *Neurology* **54**, 2127-32 (2000).
10. Hinton, V.J., Fee, R.J., Goldstein, E.M. & De Vivo, D.C. Verbal and memory skills in males with Duchenne muscular dystrophy. *Dev Med Child Neurol* **49**, 123-8 (2007).
11. Mento, G., Tarantino, V. & Bisiacchi, P.S. The neuropsychological profile of infantile Duchenne muscular dystrophy. *Clin Neuropsychol* **25**, 1359-77 (2011).
12. D'Angelo, M.G. *et al.* Neurocognitive profiles in Duchenne muscular dystrophy and gene mutation site. *Pediatr Neurol* **45**, 292-9 (2011).
13. Perronnet, C. & Vaillend, C. Dystrophins, utrophins, and associated scaffolding complexes: role in mammalian brain and implications for therapeutic strategies. *J Biomed Biotechnol* **2010**, 849426 (2010).
14. Hendriksen, J.G. & Vles, J.S. Neuropsychiatric disorders in males with duchenne muscular dystrophy: frequency rate of attention-deficit hyperactivity disorder (ADHD), autism spectrum disorder, and obsessive--compulsive disorder. *J Child Neurol* **23**, 477-81 (2008).
15. Wu, J.Y., Kuban, K.C., Allred, E., Shapiro, F. & Darras, B.T. Association of Duchenne muscular dystrophy with autism spectrum disorder. *J Child Neurol* **20**, 790-5 (2005).
16. Kohane, I.S. *et al.* The co-morbidity burden of children and young adults with autism spectrum disorders. *PLoS One* **7**, e33224 (2012).
17. Nakamura A, M.Y., Kumagai T, Suzuki Y, Miura K. Various central nervous system involvements in dystrophinopathy: clinical and genetic considerations. *No To Hattatsu* **40**, 10-4 (2008).
18. Pardo, L.M., MacKay, I., Oostra, B., van Duijn, C.M. & Aulchenko, Y.S. The effect of genetic drift in a young genetically isolated population. *Ann Hum Genet* **69**, 288-95 (2005).
19. Hofman, A. *et al.* The Rotterdam Study: 2014 objectives and design update. *Eur J Epidemiol* **28**, 889-926 (2013).
20. Sleegers, K. *et al.* Familial clustering and genetic risk for dementia in a genetically isolated Dutch population. *Brain* **127**, 1641-9 (2004).

21. Liu, F. *et al.* The apolipoprotein E gene and its age-specific effects on cognitive function. *Neurobiol Aging* **31**, 1831-3 (2010).
22. Reitan, R.M. The relation of the trail making test to organic brain damage. *J Consult Psychol* **19**, 393-4 (1955).
23. Lezak MD, H.D., Loring DW. Neuropsychological assessment. (2004).
24. Welsh, K.A. *et al.* The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part V. A normative study of the neuropsychological battery. *Neurology* **44**, 609-14 (1994).
25. Golden, C.J. Identification of brain disorders by the Stroop Color and Word Test. *J Clin Psychol* **32**, 654-8 (1976).
26. Brouwer, R.W., van den Hout, M.C., Grosveld, F.G. & van Ijcken, W.F. NARWHAL, a primary analysis pipeline for NGS data. *Bioinformatics* **28**, 284-5 (2012).
27. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).
28. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
29. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-303 (2010).
30. Aulchenko, Y.S., Ripke, S., Isaacs, A. & van Duijn, C.M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294-6 (2007).
31. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity (Edinb)* **95**, 221-7 (2005).
32. Wu, M.C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**, 82-93 (2011).
33. Consortium, T.G. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-5 (2013).
34. Voyer, D., Voyer, S. & Bryden, M.P. Magnitude of sex differences in spatial abilities: a meta-analysis and consideration of critical variables. *Psychol Bull* **117**, 250-70 (1995).
35. Lord, C. *et al.* Autism diagnostic observation schedule: a standardized observation of communicative and social behavior. *J Autism Dev Disord* **19**, 185-212 (1989).
36. Caron, M.J., Mottron, L., Berthiaume, C. & Dawson, M. Cognitive mechanisms, specificity and neural underpinnings of visuospatial peaks in autism. *Brain* **129**, 1789-802 (2006).
37. Shah A, F.U. Why do autistic individuals show superior performance on the block design task? . *J Child Psychol Psychiatry* **34**, 1351–1364 (1993).
38. Pagnamenta, A.T. *et al.* A family with autism and rare copy number variants disrupting the Duchenne/Becker muscular dystrophy gene DMD and TRPM3. *J Neurodev Disord* **3**, 124-31 (2011).
39. Chung, R.H. *et al.* An X chromosome-wide association study in autism families identifies TBL1X as a novel autism spectrum disorder candidate gene in males. *Mol Autism* **2**, 18 (2011).
40. Pellicano, E., Maybery, M., Durkin, K. & Maley, A. Multiple cognitive capabilities/deficits in children with an autism spectrum disorder: "weak" central coherence and its relationship to theory of mind and executive control. *Dev Psychopathol* **18**, 77-98 (2006).
41. Ropar, D. & Mitchell, P. Susceptibility to illusions and performance on visuospatial tasks in individuals with autism. *J Child Psychol Psychiatry* **42**, 539-49 (2001).
42. Rumsey, J.M. & Hamburger, S.D. Neuropsychological findings in high-functioning men with infantile autism, residual state. *J Clin Exp Neuropsychol* **10**, 201-21 (1988).
43. Happe, F. & Frith, U. The weak coherence account: detail-focused cognitive style in autism spectrum disorders. *J Autism Dev Disord* **36**, 5-25 (2006).
44. Muntoni, F., Torelli, S. & Ferlini, A. Dystrophin and mutations: one gene, several proteins, multiple phenotypes. *Lancet Neurol* **2**, 731-40 (2003).

45. Lidov, H.G., Byers, T.J., Watkins, S.C. & Kunkel, L.M. Localization of dystrophin to postsynaptic regions of central nervous system cortical neurons. *Nature* **348**, 725-8 (1990).
46. Sekiguchi, M. *et al.* A deficit of brain dystrophin impairs specific amygdala GABAergic transmission and enhances defensive behaviour in mice. *Brain* **132**, 124-35 (2009).
47. Kueh, S.L., Head, S.I. & Morley, J.W. GABA(A) receptor expression and inhibitory post-synaptic currents in cerebellar Purkinje cells in dystrophin-deficient mdx mice. *Clin Exp Pharmacol Physiol* **35**, 207-10 (2008).
48. Vaillend, C. & Billard, J.M. Facilitated CA1 hippocampal synaptic plasticity in dystrophin-deficient mice: role for GABAA receptors? *Hippocampus* **12**, 713-7 (2002).
49. Fritschy, J.M., Schweizer, C., Brunig, I. & Luscher, B. Pre- and post-synaptic mechanisms regulating the clustering of type A gamma-aminobutyric acid receptors (GABAA receptors). *Biochem Soc Trans* **31**, 889-92 (2003).
50. Craig, A.M. & Kang, Y. Neurexin-neuroligin signaling in synapse development. *Curr Opin Neurobiol* **17**, 43-52 (2007).
51. Mohler, H. Role of GABAA receptors in cognition. *Biochem Soc Trans* **37**, 1328-33 (2009).
52. Millan, M.J. *et al.* Cognitive dysfunction in psychiatric disorders: characteristics, causes and the quest for improved therapy. *Nat Rev Drug Discov* **11**, 141-68 (2012).
53. Holder, E., Maeda, M. & Bies, R.D. Expression and regulation of the dystrophin Purkinje promoter in human skeletal muscle, heart, and brain. *Hum Genet* **97**, 232-9 (1996).
54. Gorecki, D.C. *et al.* Expression of four alternative dystrophin transcripts in brain regions regulated by different promoters. *Hum Mol Genet* **1**, 505-10 (1992).
55. Daoud, F. *et al.* Analysis of Dp71 contribution in the severity of mental retardation through comparison of Duchenne and Becker patients differing by mutation consequences on Dp71 expression. *Hum Mol Genet* **18**, 3779-94 (2009).
56. Taylor, P.J. *et al.* Dystrophin gene mutation location and the risk of cognitive impairment in Duchenne muscular dystrophy. *PLoS One* **5**, e8803 (2010).
57. Aartsma-Rus, A., Van Deutekom, J.C., Fokkema, I.F., Van Ommen, G.J. & Den Dunnen, J.T. Entries in the Leiden Duchenne muscular dystrophy mutation database: an overview of mutation types and paradoxical cases that confirm the reading-frame rule. *Muscle Nerve* **34**, 135-44 (2006).
58. Miyake A, F.N., Rettinger DA, Shah P, Ph.D., Hegarty M. How are Visuospatial Working Memory, Executive Functioning, and Spatial Abilities Related? A Latent-Variable Analysis. *Journal of Experimental Psychology - General* **130**, 532–545 (2005).
59. Salthouse, T. Relations Between Cognitive Abilities and Measures of Executive Functioning. *Neuropsychology* **19**, 532-545 (2005).
60. Ladouceur, M., Dastani, Z., Aulchenko, Y.S., Greenwood, C.M. & Richards, J.B. The empirical power of rare variant association methods: results from sanger sequencing in 1,998 individuals. *PLoS Genet* **8**, e1002496 (2012).
61. Li, B. & Leal, S.M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 83, 311-21 (2008).
62. Madsen, B.E. & Browning, S.R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* **5**, e1000384 (2009).

Chapter 2.5

Intellectual ability in the Duchenne muscular dystrophy and dystrophin gene mutation location

Vedrana Milic Rasic, Dina Vojinovic, Jovan Pesovic, Gordana Mijalkovic, Vera Lukic, Jelena Mladenovic, Ana Kosac, Ivana Novakovic, Nela Maksimovic, Stanka Romac, Slobodanka Todorovic, Dusanka Savic Pavicevic

This chapter was published in Balkan J Med Genet. 2015 Apr 10; 17(2): 25–35.

ABSTRACT

Duchenne muscular dystrophy (DMD) is the most common form of muscular dystrophy during childhood. Mutations in dystrophin (*DMD*) gene are also recognized as a cause of cognitive impairment. We aimed to determine the association between intelligence level and mutation location in *DMD* genes in Serbian patients with DMD. Forty-one male patients with DMD, aged 3 to 16 years, were recruited at the Clinic for Neurology and Psychiatry for Children and Youth in Belgrade, Serbia. All patients had defined *DMD* gene deletions or duplications [multiplex ligation-dependent probe amplification (MLPA), polymerase chain reaction (PCR)] and cognitive status assessment (Wechsler Intelligence Scale for Children, Brunet-Lezine scale, Vineland-Doll scale). In 37 patients with an estimated full-scale intelligence quotient (FSIQ), six (16.22%) had borderline intelligence ($70 < FSIQ \leq 85$), while seven (18.92%) were intellectually impaired ($FSIQ < 70$). The FSIQ was not associated with proximal and distal mutations when boundaries were set at exons 30 and 45. However, FSIQ was statistically significantly associated with mutation location when we assumed their functional consequence on dystrophin isoforms and when mutations in the 5′-untranslated region (5′UTR) of Dp140 (exons 45–50) were assigned to affect only Dp427 and Dp260. Mutations affecting Dp140 and Dp71/Dp40 have been associated with more frequent and more severe cognitive impairment. Finally, the same classification of mutations explained the greater proportion of FSIQ variability associated with cumulative loss of dystrophin isoforms. In conclusion, cumulative loss of dystrophin isoforms increases the risk of intellectual impairment in DMD and characterizing the genotype can define necessity of early cognitive interventions in DMD patients.

INTRODUCTION

Duchenne muscular dystrophy (DMD) is the most common form of muscular dystrophy during childhood, affecting 1 in 3,500 live born males.[1] This fatal, X linked disease, leads to progressive muscular weakness and less well described non progressive central nervous system manifestations.

The consistent finding in patients with DMD is reduction in a full-scale intelligence quotient (FSIQ) by one standard deviation (SD) from the population mean.[1,2] Although most affected boys are not intellectually disabled, the risk of cognitive impairment is increased in DMD patients. Therefore, up to 30% of patients have intellectual disability with a FSIQ of less than 70, including around 3% of them with severe impairment and FSIQ of less than 50.[2,3] Duchenne muscular dystrophy is caused by various types of mutations within the dystrophin gene (*DMD*),[4] which changes the reading frame of coding transcripts affecting the production of protein dystrophin.[5] Expression of the dystrophin is controlled by three upstream promoters, which produce full-length dystrophin isoform (Dp427) and four internal promoters that regulate production of shorter dystrophin isoforms (Dp260, Dp140, Dp116 and Dp71).[6–9] Dp427 is expressed in skeletal and cardiac muscle, brain and Purkinje cells, Dp260 is expressed in retina,[10] Dp140 in brain, retina and kidney,[11] while Dp116 is present in peripheral nerves.[12] Dp71 is the most abundant isoform in non muscular tissues and represents the major product in the adult brain.[13] The dystrophin is a part of dystrophin-associated glycoprotein complex, and in the brain, it is involved in the clustering of ion channels and postsynaptic membrane receptors during synaptogenesis,[9] suggesting that loss of its function may be responsible for intellectual impairment and cognitive deficits in DMD patients. Cognitive deficit is likely the result of cumulative loss of Dp427, Dp140 and Dp71,[9,14] whereas loss of Dp71 contributes to the severity of cognitive impairment.[14,15] Recently, one more dystrophin isoform (Dp40), produced from the same promoter as Dp71 but by the use of an alternative polyadenylation site, has been implicated in presynaptic function.[16]

Although cognitive impairment has frequently been reported, systematized data on the cognitive profile of patients with DMD in Serbia is lacking. Therefore, the aim of this study was to determine frequency of intellectual impairment and to examine association of intelligence level with mutation location and affected dystrophin isoforms among our patients with DMD.

MATERIALS AND METHODS

Patient Data

Forty-one patients with DMD were recruited retrospectively at the Clinic for Neurology and Psychiatry for Children and Youth in Belgrade, Serbia, during the period between 1992 and 2013. The diagnosis of DMD was based on the clinical onset of the disease before 5 years of age, initial or clear neurological signs of decline of motor function at the age of 6 years, decline of motor function or positive family history of DMD for boys younger than 6 years, elevated serum creatine kinase levels and confirmed mutation in the *DMD* gene. All recruited patients were unrelated except of one sibling pair.

All patients and/or their parents gave informed consent concerning the use of the data for research. This study was approved by the Ethics Committee of the Clinic for Neurology and Psychiatry for Children and Youth, Belgrade, Serbia.

Methods: Genetic Analysis

Deletions and duplications were detected via multiplex ligation-dependent probe amplification (MLPA) using two probe mixes, P034 and P035 (MRC Holland, Amsterdam, The Netherlands)[17] at the Center for Human Molecular Genetics at the Faculty of Biology in Belgrade, Serbia. A few samples were analyzed via multiplex polymerase chain reaction (PCR)[18–20] at the Institute of Human Genetics at the Faculty of Medicine in Belgrade, Serbia, were included in the analyses as the mutation location allowed us to unequivocally assign the altered dystrophin isoforms.

Mutations were described using Human Genome Variation Society (HGVS) nomenclature.[21] The positions of the mutations were determined in relation to reference sequence NM_004006 (GenBank) at cDNA level and in relation to reference sequence UniProtKB:11532 (Uni ProtKB/Swiss-Prot) at protein level. The mutation effect on the reading frame was determined using software *DMD* gene reading frame checker (available at http://www.dmd.nl/).

Since the effect of mutation location on FSIQ was expected to be different, all mutations were divided into two structural groups according to previously applied classifications.[22,23] The mutations localized upstream from exon 30 (1–30) and the mutations localized upstream from exon 45 (1–45) were defined as proximal mutations, while the mutations downstream from exon 30 (31–79) or exon 45 (46–79) were considered as distal mutations.

Considering the complex organization of *DMD* gene and dystrophin isoforms produced from the inner promoters, mutations were assigned to altered expression of dystrophin isoforms (Dp427, Dp260, Dp140, Dp116, Dp71, and Dp40), and were divided into various groups. We took into consideration that the Dp140 transcript has a long 5′ untranslated region (5′UTR), consisting of exons 45–50, so that the effect of mutations in this region on Dp140 expression could not be confidently predicted. Therefore, we also analyzed clustering, in which mutations within this region (Dp140utr) were assumed to be coding exon mutations affecting only expression of Dp427 and Dp260 but not Dp140, while mutations in the promoter and protein coding region of Dp140 (Dp140pc) were assigned to affect expression of Dp140.[14] As the promoter that regulates Dp140 expression lies within intron 44 of the *DMD* gene, all patients who had a deletion breakpoint within intron 44 were tested by PCR for the presence of the Dp140 promoter using the following primers: IN44F (5′-GCC CTA AGT GCT TCC AGA AA-3′) and IN44R (5′-CTC ACA GCT CCT GCA TCA GA-3′). This original approach allowed us to accurately group patients with the affected Dp140 expression. To assess the cumulative effect of dystrophin isoforms on FSIQ, the patients were divided into three groups with respect to the preservation or loss of Dp140 and Dp71/Dp40.

Cognitive Assessment

All DMD patients were psychologically tested. Taking into account the patients' age, different psychological instruments were used for cognitive status assessment.

In order to assess intelligence level in children younger than 16 years and 11 months, the Wechsler Intelligence Scale for Children (WISC) was used.[24] The test generates FSIQ, which represents overall cognitive ability. Patients with FSIQ ≤ 70 were considered mentally disabled, while patients with FSIQ ranging between 70 and 85 (70 < FSIQ ≤ 85) were defined as borderline.

The Brunet-Lezine scale, statistically adapted for our population, was applied in order to assess psychomotor developmental quotient (DQ) in children from 0 to 30 months.[25] With complementary tests, this scale can be used until 6 years of age.

The Vineland-Doll scale was used to measure social maturity in individuals with mental retardation or individuals who had difficulty performing in testing situations.[26] Based on literature findings, an estimate of social quotient (SQ) provided by this test, was highly correlated (0.80) with intelligence.

Statistical Analysis

Exploration of normality, assessed with the Shapiro-Wilk test, revealed normal distribution of age and FSIQ in our sample. Therefore, parametric tests were used for further analysis. To investigate whether the mean FSIQ was statistically different from normative values (100 ± 15) one sample t-test was applied. To evaluate whether age had an impact on cognitive status, the Pearson correlation coefficient was calculated. The results were considered significant when probability was less than 0.05.

The cognitive abilities of patients with mutations localized in proximal and distal parts of the *DMD* gene were compared using the parametric one-way t-test for independent samples. The same test was used to assess an effect of loss of different dystrophin isoforms on FSIQ. In order to adjust for multiple testing, we made an adjustment of *p*-value calculating Bonferroni correction (0.05/3 tests = 0.017). To determine whether Dp140 and Dp71/Dp40 were associated with cognitive abilities, patients with presumably intact or absent Dp140 and/or Dp71/Dp40 were compared using one-way ANOVA analysis of variance. Statistical analyses were performed using commercially available software (Statistical Package for the Social Sciences, IBM Corporation, Armonk, NY, USA; Statistics for Windows, SPSS Version 20.0).

RESULTS

Clinical and Genetic Characteristics

Forty-one patients with genetically confirmed deletion or duplication in the *DMD* gene and cognitive status assessment were recruited at the Clinic for Neurology and Psychiatry for Children and Youth in Belgrade, Serbia. All affected patients were males, aged 3 to 16 (8.34 ± 2.56) years. Deletions were confirmed in 37 patients (90.24%), while duplications were identified in four patients (9.75%). The identified deletions and duplications, description of the mutations at cDNA and protein level, mutation effect on reading frame, as well as dystrophin isoforms impaired by mutation, the age at the onset of disease, the age at psychological testing and data of the psychological exploration are shown in **Table 1**. Although patients with in-frame mutations could express milder phenotypes, all our patients with in-frame mutations developed a DMD phenotype: one patient showed delayed psychomotor development and was wheelchair-bound at the age of 10 (ID 28), the second had the onset of the disease at the age of 3.5 years and developed positive Gowers sign at the age of 7 (ID 5), and two boys, the youngest one (IDs 2 and 38), expressed initial signs of motor decline at the age of 6.

Table 1. Observed deletions and duplications for all examined patients with Duchenne muscular dystrophy and corresponding cognitive abilities.

ID	Sex-Age at Onset (years)	Detected *DMD* Gene	Mutation at cDNA Level	Mutation at Protein Level	Mutation Effect	Affected Dystrophin Isoforms	FSIQ	Age at FSIQ Test (years)
1	5	deletion of exon 1 and Dp427c	c.(?_-128297)_31+?del	N/A	N/A	Dp427	72	9
2	5	deletion of exons 3_4	c.94-?_264+?del	p.F32_N88del	in frame	Dp427	100	6
3	3	duplication of exons 3_7	c.94-?_649+?dup	p.D217V fs*7	out of frame	Dp427	96	11
4	2	deletion of exons 3_17	c.94-?_2168+?del	p.L724V fs*4	out of frame	Dp427	75	9
5	3.5	deletion of exons 3_18	c.94-?_2292+?del	p.F32_N764del	in frame	Dp427	106	7
6	2	duplication of exons 8_12	c.650-?_1482+?dup	p.V495M fs*13	out of frame	Dp427	110	10
7	2	duplication of exons 16_17	c.1813-?_2168+?dup	p.L724F fs*2	out of frame	Dp427	88	10
8	5	deletion of exons 20_23	c.2381-?_3162+?del	p.N1055E fs*12	out of frame	Dp427	111	8
9	3	deletion of exons 8_34	c.650-?_4845+?del	p.A1616G fs*7	out of frame	Dp427, Dp260	100	6
10	2	deletion of exon 45[a]	c.6439-?_6614+?del	p.L2206A fs*17	out of frame	Dp427, Dp260, Dp140utr	109	9
11	2	deletion of exons 45_50[a]	c.6439-?_7309+?del	p.S2437L fs*9	out of frame	Dp427, Dp260, Dp140utr	96	5
12	5	deletion of exons 45_50[a]	c.6439-?_7309+?del	p.S2437L fs*9	out of frame	Dp427, Dp260, Dp140utr	118	9
13	2	deletion of exons 46_47	c.6615-?_6912+?del	p.V2305F fs*16	out of frame	Dp427, Dp260, Dp140utr	105	10
14	4	deletion of exons 46_48	c.6615-?_7098+?del	p.E2367K fs*4	out of frame	Dp427, Dp260, Dp140utr	68	6
15	pdd	deletion of exons 46_50	c.6615-?_7309+?del	p.R2205S fs*16	out of frame	Dp427, Dp260, Dp140utr	94	7
16	2	deletion of exons 46_50	c.6615-?_7309+?del	p.R2205S fs*16	out of frame	Dp427, Dp260, Dp140utr	114	9
17	?-3	deletion of exons 48_50	c.6913-?_7309+?del	p.S2437L fs*9	out of frame	Dp427, Dp260, Dp140utr	89	10
18	<2	deletion of exons 48_50	c.6913-?_7309+?del	p.S2437L fs*9	out of frame	Dp427, Dp260, Dp140utr	94	7
19	4	deletion of exons 49_50	c.7099-?_7309+?del	p.S2437L fs*9	out of frame	Dp427, Dp260, Dp140utr	63	7
20	5	deletion of exons 49_50	c.7099-?_7309+?del	p.S2437L fs*9	out of frame	Dp427, Dp260, Dp140utr	99	9
21	<2	deletion of exon 50	c.7201-?_7309+?del	p.S2437L fs*9	out of frame	Dp427, Dp260, Dp140utr	63	12

Table 1. Observed deletions and duplications for all examined patients with Duchenne muscular dystrophy and corresponding cognitive abilities. (*continued*)

ID	Sex-Age at Onset (years)	Detected *DMD* Gene	Mutation at cDNA Level	Mutation at Protein Level	Mutation Effect	Affected Dystrophin Isoforms	FSIQ	Age at FSIQ Test (years)
22	pdd	deletion of exon 50	c.7201-?_7309+?del	p.S2437L fs*9	out of frame	Dp427, Dp260, Dp140utr	95	6
23	4.5	deletion of exon 44[a]	c.6291-?_6438+?del	p.E2147N fs*16	out of frame	Dp427, Dp260, Dp140pc	71	8
24	4	deletion of exon 44[a]	c.6291-?_6438+?del	p.E2147N fs*16	out of frame	Dp427, Dp260, Dp140pc	94	11
25	3.5	deletion of exons 45_52[a]	c.6439-?_7660+?del	p.I2554L fs*22	out of frame	Dp427, Dp260, Dp140pc	50	9
26	4	deletion of exons 46_51	c.6615-?_7542+?del	p.A2515Q fs*23	out of frame	Dp427, Dp260, Dp140pc	95	14
27	2.5	deletion of exons 46_52	c.6763-?_7542+?del	p.R2205S fs*2	out of frame	Dp427, Dp260, Dp140pc	88	9
28	pdd	deletion of exons 47_51	c.6763-?_7542+?del	p.I2255_K2514del	in frame	Dp427, Dp260, Dp140pc	96	10
29	N/A	deletion of exons 50_53[b]	N/A	N/A	N/A	Dp427, Dp260, Dp140pc	75	7
30	5	deletion of exons 50_53[b]	N/A	N/A	N/A	Dp427, Dp260, Dp140pc	79	7
31	3.5	deletion of exon 51	c.7310-?_7542+?del	p.A2515C fs*33	out of frame	Dp427, Dp260, Dp140pc	87	11
32	5	deletion of exon 52	c.7543-?_7660+?del	p.I2554L fs*22	out of frame	Dp427, Dp260, Dp140pc	101	11
33	3	deletion of exon 53[b]	N/A	N/A	N/A	Dp427, Dp260, Dp140pc	98	7
34	<2	deletion of exon 53	c.76610?_7872+?del	p.Q2625T fs*18	out of frame	Dp427, Dp260, Dp140pc	84	6
35	4	deletion of exon 61	c.90834-?_9163+?del	p.T3055R fs*34	out of frame	Dp427, Dp260, Dp140pc, Dp116	55	8
36	2.5	deletion of exons 45_73[b]	c.6439-?_10394+?del	p.D3466R fs*2	out of frame	Dp427, Dp260, Dp140pc, Dp116, Dp40	44	16
37	pdd	deletion of exons 45_76[a]	c.6439-?_10921+?del	p.G3641V fs*16	out of frame	Dp427, Dp260, Dp140pc, Dp116, Dp40	58	7
38	5	deletion of exons 10_23	c.961-?_3162+?del	p.H321_Q1054del	in frame	Dp427	SQ 41	6
39	3	deletion of exons 48_50	c.6913-?_7309+?del	p.S2437L fs*9	out of frame	Dp427, Dp260, Dp140utr	DQ 104	3
40	4	deletion of exons 45_76[a]	c.6439-?_10921+?del	p.G3641V fs*16	out of frame	Dp427, Dp260, Dp140pc, Dp116, Dp40	DQ 82	4

Table 1. Observed deletions and duplications for all examined patients with Duchenne muscular dystrophy and corresponding cognitive abilities. (*continued*)

ID	Sex-Age at Onset (years)	Detected *DMD* Gene	Mutation at cDNA Level	Mutation at Protein Level	Mutation Effect	Affected Dystrophin Isoforms	FSIQ	Age at FSIQ Test (years)
41	pdd	duplication of exons 52_55, 63_67, triplication of exons 68_79	c.7543-?_8217+?dup, c.9225-?_9807+?dup, c.9808-?_(*2691_?)trip	p.A3270P fs*22	out of frame	Dp427, Dp260, Dp140pc, Dp116, Dp40	SQ 58	6

ID: identification number; FSIQ: full-scale intelligence quotient; N/A: not available; ppd: psychomotor development delay; SQ: social quotient; DQ: developmental quotient. Mutations at the cDNA and protein levels were described according to nomenclature suggested by the Human Genome Variation Society (HGVS), (15); mutations at the cDNA level were determined in relation to the reference sequence NM_004006 (GenBank), and at the protein level in relation to reference sequence UniProtKB:P11532 (Uni-ProtKB/Swiss-Prot); mutation effect on reading frame was determined using software *DMD* gene reading frame checker (available at http://www.dmd.nl/).

[a]The patient was tested for the presence of the Dp140 promoter region in intron 44.

[b]The mutation was detected with multiplex PCR; therefore, it was not possible to identify its precise nomenclature on cDNA and protein levels or its effect on the open reading frame.

The FSIQ was estimated for 37 participants, DQ was estimated for two patients, while SQ was assessed for two patients. The patients with estimated DQ and SQ were excluded from statistical analysis and they are described separately (**Table 2**).

Table 2. Scores according to different psychological tests.

Psychological Exploration	*n*	Mean Value (range)
FSIQ	37	87.57±18.79 (44, 118)
DQ	2	(82, 105)
SQ	2	(41, 58)

n: number of patients; FSIQ: full scale intelligence quotient; DQ: development quotient; SQ: social quotient.

General intelligence evaluation of the analyzed population demonstrated statistically significant difference from population normative values ($t = -4.024$, p-value = 0.00015). The FSIQ with mean of 87.57 (SD 18.79) had a broad range of values between 44 and 118. Thirteen patients (35.14%) had FSIQ lower than 85, of whom six patients (16.22%) had borderline intelligence levels ($70 < \text{FSIQ} \leq 85$), while seven patients (18.92%) were considered intellectually disabled ($\text{FSIQ} \leq 70$). Two intellectually disabled patients were severely impaired ($\text{FSIQ} < 50$). Above average FSIQ was assessed in four patients ($\text{FSIQ} \geq 110$). Assessed correlation coefficient showed no significant association between age and FSIQ (Pearson's $r = -0.129$).

Association of Intelligence Level with Mutation Location and Affected Dystrophin Isoforms

The analysis of the intellectual ability with respect to the structural classification of mutation locations did not show statistically significant difference in the mean FSIQ between the patients with mutation proximal and distal to exon 30 ($t = 1.23$, p-value = 0.114) (**Table 3**). The patients with mutation proximal to exon 30 had a mean FSIQ of 94.75 (SD 15.14), while for the patients with mutations distal to exon 30, the mean FSIQ was 88.59 (SD 19.44). A similar result was observed when exon 45 was used as a boundary ($t = 1.45$, p-value = 0.130) (**Table 3**). The patients with the mutation proximal

Table 3. Association between proximal and distal mutations and full scale intelligence quotient.

Location of Mutation in the *DMD* Gene	*n*	Mean FSIQ (SD)	*t*-Test
Proximal to exon 30 (1-30)	8	94.75 (15.14)	$t = 1.23$; $df = 35$; $p = 0.114$
Distal to exon 30 (31-79)	29	88.59 (19.44)	
Proximal to exon 45 (1-45)	11	93.00 (14.71)	$t = 1.45$; $df = 35$; $p = 0.130$
Distal to exon 45 (46-79)	26	85.27 (20.09)	

FSIQ: full scale intelligence quotient; *n*: number of patients; SD: standard deviation.

to exon 45 had a mean FSIQ of 93.0 (SD 14.71), while for the patients with mutation in the distal part, the mean FSIQ was 85.27 (SD 20.09).

However, the analysis of intellectual ability with respect to mutation location assigned to the impaired dystrophin isoforms indicated statistically significantly different scoring between patients whose mutations affected expression of Dp427 and Dp260, and patients whose mutations additionally affected expression of Dp140, Dp116, Dp71 and Dp40 (t = 2.67, p-value = 0.0057) (**Table 4**). This result was obtained after the group of patients with a mutation in the Dp140utr region was clustered together with the patients whose mutations altered expression of Dp427 and Dp260. The frequency of borderline intellectual ability in the group of patients with altered Dp427, Dp260, and Dp140utr, was 9.09%, while in the group of patients with altered Dp140pc, Dp116, Dp71 and Dp40, it was 26.67%. The frequency of intellectual disability in the two groups was

Table 4. Association between mutation locations assigned to altered expression of dystrophin isoforms and full-scale intelligence quotient.

Location of the Mutation in the *DMD* Gene	**Affected Dystrophin Isoforms**	***n***	**Mean FSIQ (SD)**	***t*-Test**
Proximal to intron 29	Dp427	8	94.75 (15.14)	
Distal to intron 29	+Dp260, Dp140, Dp115, Dp71, Dp40	29	85.59 (19.44)	$t = 1.23$, $df = 35$, $p = 0.113$
Proximal to intron 44	Dp427, Dp260	9	95.33 (14.27)	
Dp140 promoter and distal to intron 44	+Dp140, Dp116, Dp71, Dp40	28	85.07 (19.59)	$t = 1.45$, $df = 35$, $p = 0.078$
Proximal to exon 51 excluding the Dp140 promoter	Dp427, Dp260, Dp140utr	22	93.86 (16.36)	
Dp140 promoter and distal to exon 51 including exon 51	+Dp140pc, Dp115, Dp71, Dp40	15	78.33 (18.79)	$t = 2.67$, $df = 35$, $p = 0.0057$[a]
Proximal to intron 55	Dp427, Dp260, Dp140	34	90.68 (16.10)	
Distal to intron 55	+Dp116, Dp71, Dp40	3	52.33 (7.37)[b]	–
Proximal to intron 62	Dp427, Dp260, Dp140, Dp116	35	89.66 (16.97)	
Distal to intron 62	+Dp71, Dp40	2	51.00 (9.90)[c]	–

n: number of patients; FSIQ: full scale intelligence quotient; SD: standard deviation; Dp140: Dp140utr+Dp140pc; Dp140utr: 5′ untranslated region of Dp140; Dp140pc: promoter and protein coding region of Dp140; +: additionally affected dystrophin isoforms; –: only few observation in one group, statistical test was not reliable.

[a]$p < 0.01$ statistical significance (Bonferroni correction 0.05/3).

[b]Three patients with FSIQ of 55, 44 and 58.

[c]Two patients with FSIQ of 44 and 58.

13.64 vs. 26.67%. Furthermore, three patients had affected Dp116, Dp71, and Dp40 and they were intellectually disabled (IDs 36, 37 and 41). Considering the few observations in this group of patients, a statistical test was not performed.

Additionally, 28 patients with the mutation affecting Dp140, were classified into groups based on the mutation localization in Dp140utr or Dp140pc. The patients with mutation within the Dp140utr region had a mean FSIQ of 92.85 (SD 18.16), while for patients with a mutation in the Dp140pc region, the mean FSIQ was 78.33 (SD 18.79). The difference in the mean FSIQ was statistically significantly different ($t = 2.07$, p-value = 0.024) between those two groups.

Finally, the distribution of FSIQ with respect to functional consequence on the Dp140 and Dp71/Dp40 isoforms indicated that cognitive impairment in DMD patients was associated with the cumulative loss of dystrophin isoforms (**Figure 1**). A greater proportion of variability in FSIQ was explained after assuming that mutations within the Dp140utr affect the functional loss of Dp427 and Dp260, but not Dp140 ($F = 7.454$, p-value = 0.002 vs. $F = 5.76$, p-value = 0.007).

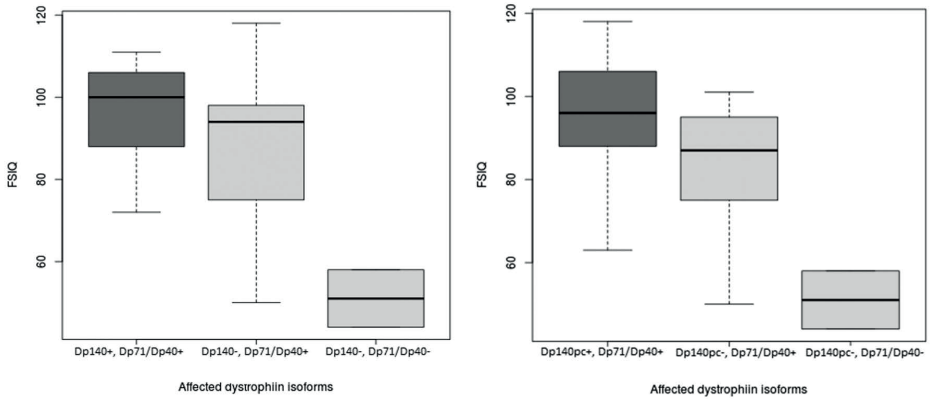


Figure 1. Distribution of FSIQ with respect to the preservation or loss of Dp140 and Dp71/Dp40 indicates that cognitive impairment in DMD patients is associated with the cumulative loss of dystrophin isoforms. A greater proportion of variability on FSIQ was explained after clustering mutations within the Dp140utr together with mutations affecting expression of Dp427 and Dp260 isoforms, but no expression of the Dp140 isoform. FSIQ: full scale intelligence quotient; +/−: preservation/absence of appropriate dystrophin isoform; Dp140-Dp140utr+Dp140pc: Dp140utr: 5′ untranslated region of Dp140; Dp140pc: promoter and protein coding region of Dp140.

For the youngest two patients, psychomotor development was measured with a scale for early infancy. Estimated DQ suggested normal psychomotor development (DQ = 105) for a boy (ID 39) whose mutation affected expression of Dp427, Dp260, and Dp140utr,

while slightly delayed psychomotor development (DQ = 82) was reported for the other boy (ID 40) whose mutation affected expression of all dystrophin isoforms.

Moreover, for two DMD patients with difficulties in a test situation, SQ was estimated. One patient (ID 38), with autistic like-behavior and undeveloped speech, had SQ 41, while the other (ID 41) had SQ 58. The patient with autistic like-behavior had a deletion that altered expression of Dp427 only, while duplication/triplication in the distal part of the *DMD* gene caused loss of all dystrophin isoforms in the other patient. Cognitive assessment for one sibling pair showed intellectual impairment in one sibling (ID 37), while slightly delayed psychomotor development was noted in the other (ID 40).

DISCUSSION

The aim of this study was to determine the frequency of intellectual impairment and relationship between intelligence level and dystrophin mutations in the Serbian group of patients with DMD. Consistent with previously published studies, general intellectual level in our study group was statistically significantly different from normative values.[27] The FSIQ was reduced for almost 1 SD (15 FSIQ points) from population average, which is in agreement with the results from other DMD cohorts[28–30], or DMD plus intermediate muscular dystrophy cohorts.[31] Seven boys in our sample (18.92%) had intellectual impairment with FSIQ < 70, which is slightly lower but still in agreement with the results reported in other studies (19–35%).[2,3,28–30]

The majority of patients in our study had confirmed causal deletion in the *DMD* gene, so it was difficult to analyze the effect of mutation type on intellectual level. Even though only four patients had duplication in the *DMD* gene, three of them (IDs 3, 6 and 7) had normal intellectual ability, while for one patient (ID 41), the estimated SQ, which highly correlates with FSIQ, was low. The patients with normal intellectual ability had a duplication in the proximal part of the *DMD* gene that affected expression of Dp427, while the patient with low SQ had a duplication/triplication that altered the expression of all dystrophin isoforms. Our study did not include patients with DMD clinical presentation without deletion or duplication and with possible point mutations. However, Taylor *et al*.[14] published a study in which there was no significant correlation between mutation type and FSIQ.

Despite the fact that some of the previously published studies found association between the structural location of mutations and FSIQ,[22,23,28] we were not able to replicate this association in our study when boundaries for proximal and distal mutations were

set at exon 30 and exon 45. However, association between the intellectual ability of DMD patients and mutation location in regard to their functional consequence, loss of expression of different dystrophin isoforms,[14] was confirmed with our results.

The loss of the Dp427 isoform is a common feature among all DMD patients, which may result in cognitive impairment. Dp427 is expressed in the neocortex, cerebellum and amygdala,[6,32,33] where it plays not only a structural role in central synapses but likely regulates $GABA_A$ receptor clustering at inhibitory synapses.[9] In eight patients whose mutations abolished the expression of Dp427, two boys (IDs 1 and 4) (25%) had borderline FSIQ, while none had an intellectual impairment (FSIQ < 70). However, the mutations affecting the expression of Dp140 and Dp71/Dp40, in addition to Dp427, have been associated with higher frequency and severe cognitive impairment in our DMD patients, suggesting the effect of cumulative loss of dystrophin isoforms and the important role of Dp140 and Dp71/Dp40 on intellectual ability. This finding is in agreement with the results of previously published studies.[14,15,34–36]

The role of the Dp140 isoform on intellectual functioning was recognized by Felisariet *et al.*,[37] who described the association between the mutations affecting the expression of Dp140 and intellectual ability in DMD patients, has been replicated in other studies.[14,28,29] Dp140 is detected throughout the brain (cerebral cortex, cerebellum, hippocampus, brain stem and olfactory bulb) and in the spinal cord,[11] but its function is still unclear. Our results suggest statistically significantly lower FSIQ in patients with altered Dp140, Dp116, Dp71 and Dp40 compared to patients with the mutations affecting only the expression of Dp427 and Dp260. However, a statistically significant difference has been obtained after clustering mutations in the Dp140utr region together with the mutations affecting the expression of Dp427 and Dp260. Additionally, statistically significant difference in the mean FSIQ was obtained when patients with the mutation affecting Dp140 were classified into groups with the mutation localization in 5'UTR (Dp140utr) or in promoter and protein-coding region (Dp140pc). These results underline the importance of assuming that the expression of the Dp140 isoform is not mainly affected by mutations located in its long 5'UTR, which includes frequently deleted exons 45–50. All the same, in our group of patients, three boys (IDs 14, 19 and 21) with mutations in the Dp140utr coding region were intellectually impaired, suggesting that some regulatory elements in the 5'UTR might be affected, influencing the alteration in the expression of Dp140. Our results are in agreement with a previously published study suggesting that mutations in the Dp140utr have a lesser effect on FSIQ when compared to the mutations affecting Dp140pc.[14]

The Dp71 isoform is a major product of the *DMD* gene in the brain. It has been confirmed that Dp71 is abundant in the fetal as well as adult brain, particularly in the cerebral cortex and hippocampus.[38] Although the function of this isoform remains unknown, it has been reported that Dp71 has a role in the stabilization and/or formation of the synaptic membrane.[39] The dysfunctions of proteins involved in the regulation of synaptic structure and function influence neuronal connectivity and the ability of the brain to process information, and may be related to the cognitive impairment.[36] Additionally, it was shown that Dp71 has a regulatory role in excitatory synapse organization and function, by clustering glutamate receptors and organizing signaling in postsynaptic densities.[15] The systematic occurrence of mild-to-severe mental retardation was noticed in more than 50 patients with dystrophinopathies and the mutation located in the Dp71.[15] Although the number of patients with altered Dp71 in our study was small (n = 2), which is in accordance with the low frequency of deletions and duplications in the most distal part of the *DMD* gene, both patients (IDs 36 and 37) were intellectually disabled, whereas one (ID 36) had the lowest FSIQ (44) within the entire study group. The mutations affecting the expression of Dp71 also effect the expression of all dystrophin isoforms, supporting the importance of cumulative loss of dystrophin isoforms apart from the loss of Dp71 only, reported in this and other studies.[14,15,34]

Until recently, the role of the shortest isoform Dp40 in the brain was unknown. A finding that Dp40 is enriched in the synaptic vesicle fraction where it assembles a group of presynaptic proteins involved in the exocytosis of synaptic vesicles, indicates that Dp40 might have an important role in presynaptic function.[16] Even though other studies did not analyze the effect of Dp40, the patients with intellectual impairment had mutations that influenced the expression of both Dp71 and Dp40,[15,16,34] implying that Dp40 might have a function relevant to cognitive processes.

Interestingly, one pair of siblings who were observed in our study had the mutation that affects Dp71, but the use of different psychological instruments made it difficult to correlate genetic data with cognitive ability assessment and to analyze variation in DMD expression between siblings (IDs 37 and 40) with the same mutations. Unlike the FSIQ, the DQ is a ratio reflecting the child's overall development without precisely defined correlation with FSIQ in later life. Still, the infant who scores low often turns out to be intellectually disabled.[40]

In addition, the association of DMD with neuropsychiatric disorders has also been recognized. Wu *et al.*[41] published a study in which they confirmed a previously unrecognized relation between DMD and an autistic spectrum disorder. Therefore, it is not surprising that one boy (ID 38) with autism like-behavior, qualitatively different from the behavior

of DMD boys with mental retardation, was described in our sample. His SQ, which was estimated to be 41, indicates below average achievement and impairment in adaptability, including communication, daily living, and socialization. The mutation in this boy affected only the expression of Dp427.

The limitations of our study were the retrospective design and limited sample size, but since psychological testing is a standard procedure in the care of DMD patients at the Clinic for Neurology and Psychiatry for Children and Youth, Belgrade, Serbia, we enabled an unbiased selection of recruited patients, overcoming selection issues discussed in other studies.[2] The non-longitudinal design of our study excluded the possibilities of defining subtypes within DMD[23] and to define the clinical severity of two very young DMD patients (IDs 2 and 38) with in-frame mutations. In general, in-frame mutations are associated with milder form of dystrophinopathy, but exceptions to the reading frame hypothesis exist.[42]

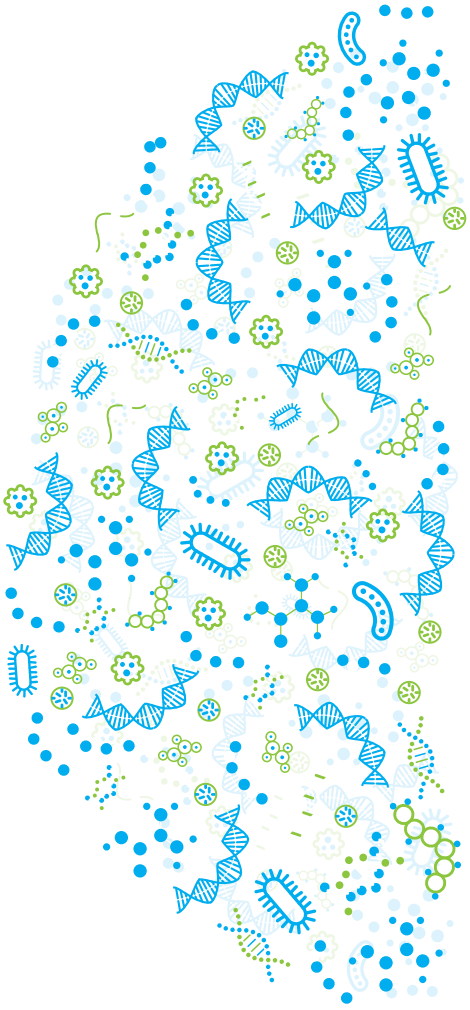
In summary, the classification of the mutations based on their functional consequence on dystrophin isoforms, with the assumption that the expression of Dp140 is not mainly affected by the mutations in its 5'UTR, explained the genetic influence on variability of FSIQ with the effect of cumulative loss of dystrophin isoforms, suggesting an important role of Dp140, Dp71 and Dp40 isoforms on intellectual ability. Defining the functional loss of dystrophin isoforms allows the recognition of the subgroup of DMD boys with greater risk for cognitive problems. Early interventions and the support in cognitive, emotional and behavioral development could be very useful and more effective than interventions in the older period of childhood or adolescence.

REFERENCES

1. Emery, A. & Muntoni, F. Duchenne muscular dystrophy. New York, NY: *Oxford University Press* **3rd ed**(2003).
2. Cotton, S., Voudouris, N.J. & Greenwood, K.M. Intelligence and Duchenne muscular dystrophy: full-scale, verbal, and performance intelligence quotients. *Dev Med Child Neurol* **43**, 497-501 (2001).
3. Cotton, S., Voudouris, N. & Douglas, J. Key findings from a meta-analytical study on intellectual functions in Duchenne muscular dystrophy. *Journal of Intellectual Disability Research* **44**, 248-248 (2000).
4. Koenig, M. *et al.* Complete Cloning of the Duchenne Muscular-Dystrophy (Dmd) Cdna and Preliminary Genomic Organization of the Dmd Gene in Normal and Affected Individuals. *Cell* **50**, 509-517 (1987).
5. Monaco, A.P., Bertelson, C.J., Liechti-Gallati, S., Moser, H. & Kunkel, L.M. An Explanation for the Phenotypic Differences between Patients Bearing Partial Deletions of the DMD Locus. *Genomics* **2**, 90-95 (1988).
6. Lidov, H.G.W., Byers, T.J., Watkins, S.C. & Kunkel, L.M. Localization of Dystrophin to Postsynaptic Regions of Central-Nervous-System Cortical-Neurons. *Nature* **348**, 725-727 (1990).
7. Tokarz, S.A. *et al.* Redefinition of dystrophin isoform distribution in mouse tissue by RT-PCR implies role in nonmuscle manifestations of duchenne muscular dystrophy. *Mol Genet Metab* **65**, 272-81 (1998).
8. Anderson, J.L., Head, S.I., Rae, C. & Morley, J.W. Brain function in Duchenne muscular dystrophy. *Brain* **125**, 4-13 (2002).
9. Perronnet, C. & Vaillend, C. Dystrophins, utrophins, and associated scaffolding complexes: role in mammalian brain and implications for therapeutic strategies. *J Biomed Biotechnol* **2010**, 849426 (2010).
10. Pillers, D.A.M. *et al.* Dystrophin Expression in the Human Retina Is Required for Normal Function as Defined by Electroretinography. *Nature Genetics* **4**, 82-86 (1993).
11. Lidov, H.G.W., Selig, S. & Kunkel, L.M. Dp140 - a Novel 140-Kda Cns Transcript from the Dystrophin Locus. *Human Molecular Genetics* **4**, 329-335 (1995).
12. Byers, T.J., Lidov, H.G.W. & Kunkel, L.M. An Alternative Dystrophin Transcript Specific to Peripheral-Nerve. *Nature Genetics* **4**, 77-81 (1993).
13. Bar, S. *et al.* A Novel Product of the Duchenne Muscular-Dystrophy Gene Which Greatly Differs from the Known Isoforms in Its Structure and Tissue Distribution. *Biochemical Journal* **272**, 557-560 (1990).
14. Taylor, P.J. *et al.* Dystrophin Gene Mutation Location and the Risk of Cognitive Impairment in Duchenne Muscular Dystrophy. *Plos One* **5**(2010).
15. Daoud, F. *et al.* Analysis of Dp71 contribution in the severity of mental retardation through comparison of Duchenne and Becker patients differing by mutation consequences on Dp71 expression. *Human Molecular Genetics* **18**, 3779-3794 (2009).
16. Tozawa, T. *et al.* The Shortest Isoform of Dystrophin (Dp40) Interacts with a Group of Presynaptic Proteins to Form a Presumptive Novel Complex in the Mouse Brain. *Molecular Neurobiology* **45**, 287-297 (2012).
17. Schouten, J.P. *et al.* Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Research* **30** (2002).

18. Chamberlain, J.S., Gibbs, R.A., Ranier, J.E., Nguyen, P.N. & Caskey, C.T. Deletion Screening of the Duchenne Muscular-Dystrophy Locus Via Multiplex DNA Amplification. *Nucleic Acids Research* **16**, 11141-11156 (1988).
19. Beggs, A.H., Koenig, M., Boyce, F.M. & Kunkel, L.M. Detection of 98-Percent of Dmd/Bmd Gene Deletions by Polymerase Chain-Reaction. *Human Genetics* **86**, 45-48 (1990).
20. Kunkel, L.M., Snyder, J.R., Beggs, A.H., Boyce, F.M. & Feener, C.A. Searching for Dystrophin Gene Deletions in Patients with Atypical Presentations. *Etiology of Human Disease at the DNA Level* **80**, 51-60 (1991).
21. den Dunnen, J.T. & Antonarakis, S.E. Mutation nomenclature extensions and suggestions to describe complex mutations: A discussion. *Human Mutation* **15**, 7-12 (2000).
22. Bushby, K.M.D. *et al.* Deletion Status and Intellectual Impairment in Duchenne Muscular-Dystrophy. *Developmental Medicine and Child Neurology* **37**, 260-269 (1995).
23. Desguerre, I. *et al.* Clinical Heterogeneity of Duchenne Muscular Dystrophy (DMD): Definition of Sub-Phenotypes and Predictive Criteria by Long-Term Follow-Up. *Plos One* **4** (2009).
24. Wechsler, D. Manual for the Wechsler Intelligence Scale for Children. *3rd ed. New York, NY: Psychological Corporation* (1991).
25. Brunet, O. & Lezine, P. Le Developpement Psychologique del al Premiere Enfance Issy-les-Moulineaux. *France: Editions Scientifiques et Psychotechniques* (1951).
26. Doll, E.A. Vineland Social Maturity Scale. Oxford UK: Educational Test Bureau (1947).
27. Hinton, V.J., Fee, R.J., Goldstein, E.M. & De Vivo, D.C. Verbal and memory skills in males with Duchenne muscular dystrophy. *Dev Med Child Neurol* **49**, 123-8 (2007).
28. D'Angelo, M.G. *et al.* Neurocognitive profiles in Duchenne muscular dystrophy and gene mutation site. *Pediatr Neurol* **45**, 292-9 (2011).
29. Wingeier, K. *et al.* Neuropsychological impairments and the impact of dystrophin mutations on general cognitive functioning of patients with Duchenne muscular dystrophy. *J Clin Neurosci* **18**, 90-5 (2011).
30. Lorusso, M.L. *et al.* Specific profiles of neurocognitive and reading functions in a sample of 42 Italian boys with Duchenne Muscular Dystrophy. *Child Neuropsychol* **19**, 350-69 (2013).
31. Chamova, T. *et al.* ASSOCIATION BETWEEN LOSS OF Dp140 AND COGNITIVE IMPAIRMENT IN DUCHENNE AND BECKER DYSTROPHIES. *Balkan Journal of Medical Genetics* **16**, 21-29 (2013).
32. Knuesel, I. *et al.* Differential expression of utrophin and dystrophin in CNS neurons: An in situ hybridization and immunohistochemical study. *Journal of Comparative Neurology* **422**, 594-611 (2000).
33. Sekiguchi, M. *et al.* A deficit of brain dystrophin impairs specific amygdala GABAergic transmission and enhances defensive behaviour in mice. *Brain* **132**, 124-135 (2009).
34. Moizard, M.P. *et al.* Severe cognitive impairment in DMD: obvious clinical indication for Dp71 isoform point mutation screening. *European Journal of Human Genetics* **8**, 552-556 (2000).
35. Moizard, M.P. *et al.* Are Dp71 and Dp140 brain dystrophin isoforms related to cognitive impairment in Duchenne muscular dystrophy? *American Journal of Medical Genetics* **80**, 32-41 (1998).
36. Chelly, J., Khelfaoui, M., Francis, F., Cherif, B. & Bienvenu, T. Genetics and pathophysiology of mental retardation. *European Journal of Human Genetics* **14**, 701-713 (2006).
37. Felisari, G. *et al.* Loss of Dp140 dystrophin isoform and intellectual impairment in Duchenne dystrophy. *Neurology* **55**, 559-564 (2000).
38. Haenggi, T., Soontornmalai, A., Schaub, M.C. & Fritschy, J.M. The role of utrophin and Dp71 for assembly of different dystrophin-associated protein complexes (DPCs) in the choroid plexus and microvasculature of the brain. *Neuroscience* **129**, 403-413 (2004).

39. Jung, D., Yang, B., Meyer, J., Chamberlain, J.S. & Campbell, K.P. Identification and characterization of the dystrophin anchoring site on beta-dystroglycan. *J Biol Chem* **270**, 27305-10 (1995).
40. Sigelman, C. & Rider, E.A. Life-Span Human Development. *7th ed. Belmont, CA: Wadsworth Cengage Learning* (2012).
41. Wu, J.Y., Kuban, K.C., Allred, E., Shapiro, F. & Darras, B.T. Association of Duchenne muscular dystrophy with autism spectrum disorder. *J Child Neurol* **20**, 790-5 (2005).
42. Ferlini, A., Neri, M. & Gualandi, F. The medical genetics of dystrophinopathies: molecular genetic diagnosis and its impact on clinical practice. *Neuromuscul Disord* **23**, 4-14 (2013).



Chapter 3

Omics of neurovascular pathology

Chapter 3.1

Whole-genome linkage scan combined with exome sequencing identifies novel candidate genes for carotid intima-media thickness

Dina Vojinovic, Maryam Kavousi, Mohsen Ghanbari, Rutger W.W. Brouwer, Jeroen G.J. van Rooij, Mirjam C.G.N. van den Hout, Robert Kraaij, Wilfred F.J. van IJcken, Andre G. Uitterlinden, Cornelia M. van Duijn, Najaf Amin

This chapter is accepted for publication in Frontiers in Genetics.

The supplemental information for this paper is available at https://drive.google.com/drive/folders/13m2uhJ5MJ2kjsvaH5CqoyNHCV-IIAcWu?usp=sharing

ABSTRACT

Carotid intima-media thickness (cIMT) is an established heritable marker for subclinical atherosclerosis. In this study, we aim to identify rare variants with large effects driving differences in cIMT by performing genome-wide linkage analysis of individuals in the extremes of cIMT trait distribution (> 90th percentile) in a large family-based study from a genetically isolated population in the Netherlands. Linked regions were subsequently explored by fine-mapping using exome sequencing. We observed significant evidence of linkage on chromosomes 2p16.3 (rs1017418, heterogeneity LOD (HLOD) = 3.35), 19q13.43 (rs3499, HLOD = 9.09), 20p13 (rs1434789, HLOD = 4.10) and 21q22.12 (rs2834949, HLOD = 3.59). Fine-mapping using exome sequencing data identified a non-coding variant (rs62165235) in *PNPT1* gene under the linkage peak at chromosome 2 that is likely to have a regulatory function. The variant was associated with quantitative cIMT in the family-based study population (effect = 0.27, *p*-value = 0.013). Furthermore, we identified several genes under the 21q22 linkage peak highly expressed in tissues relevant for atherosclerosis. To conclude, our linkage analysis identified four genomic regions significantly linked to cIMT. Further analyses are needed to demonstrate involvement of identified candidate genes in development of atherosclerosis.

INTRODUCTION

Cardiovascular diseases, including heart and cerebrovascular diseases, are listed among the leading causes of death in developed countries. [1] The underlying pathology in the majority of cases is atherosclerosis.[2] Carotid intima-media thickness (cIMT), a quantitative measure of carotid artery wall thickening, is a marker for subclinical atherosclerosis that has been shown to predict future cardiovascular events in large epidemiological studies.[3-5] cIMT is determined by both traditional cardiovascular risk factors, such as aging, blood pressure, body mass index, plasma lipid levels, diabetes mellitus or smoking, and genetic factors.[6] Genetic factors play a key role in the etiology of cIMT with heritability estimates ranging from 30-60%.[7,8] Several genome-wide linkage studies of quantitative cIMT published up to date, reported significant and suggestive evidence of linkage on chromosomes 2q33-q35, 6p12-p22, 7p, 11q23, 12q24, 13q32-q33, and 14q31.[8-11] The largest genome-wide association study (GWAS) of cIMT, including 42,484 individuals, identified only three genomic regions of common non-coding genetic variation on 8q24 (near *ZHX2*), 19q13 (near *APOC1*) and 8q23.1 (*PINX1*) and an additional suggestive region on 6p22 (near *SLC17A4*).[12] In addition, an exome-wide association study in 52,869 individuals identified the association of protein-coding variants in *APOE* with cIMT.[13] The identified variants provide valuable insights into the genetic architecture of cIMT but explain a small proportion of the trait variance.[12] A previous sequencing study of cIMT candidate regions in population-based cohorts yielded inconclusive results due to limited power.[14] A more powerful approach for uncovering the role of rare variants is a family-based study design due to the higher frequency of the rare variants.[15] The chances of success for family-based studies are even higher in genetic isolates since rare variants become more frequent due to founder effect, genetic drift and inbreeding.[15-17]

In this study, we hypothesized that there may be rare variants with large effects driving differences in cIMT independently of traditional cardiovascular risk factors and that these variants are enriched in the extremes of the cIMT distribution. To the best of our knowledge, no study to date explored extremes of quantitative cIMT. However, this approach has been demonstrated as successful for some other quantitative traits. Following the same approach as described in our study, Amin *et al.* successfully identified a rare variant of large effect in large extended families.[18] To discover such variants in the extremes of cIMT distribution, we performed affected-only genome-wide linkage analysis of cIMT followed by fine-mapping using exome sequencing in a large family-based study from a genetically isolated population in the Netherlands.

MATERIAL AND METHODS

Study population

Our discovery population consisted of participants from Erasmus Rucphen Family (ERF) study. ERF is a family-based cohort that includes around 3,000 inhabitants of a genetically isolated community in the South-West of the Netherlands.[19] The community was constituted as a religious isolate at the middle of the 18th century by a limited number of founders.[19] The population has remained in isolation with minimal immigration rate and high inbreeding.[19,20] All ERF participants are living descendants of a limited number of founders living in the 19th century. The Medical Ethical Committee of the Erasmus University Medical Center, Rotterdam approved the study. Written informed consent was obtained from all participants.

Phenotypes

Participants from ERF underwent extensive clinical examination between 2002 and 2005. cIMT was measured using high-resolution B-mode ultrasonography with a 7.5-MHz linear array transducer (ATL UltraMark IV). Maximum cIMT was measured on the 3 still, longitudinal, two-dimensional ultrasound images of the near and far wall from both left and right arteries, as described previously.[21] The mean value of these measurements was used for the analyses.

Information on covariates for both studies included age, sex, and smoking status. Body mass index (BMI) was defined as weight divided by the square of height (kg/m^2) and waist-hip ratio (WHR) was computed by dividing the waist and hip circumferences with each other. Hypertension was defined as systolic blood pressure above 140 mmHg, diastolic blood pressure above 90 mmHg or use of medication for treatment of hypertension. Dyslipidemia was defined as total cholesterol above 6.2 mmol/L or use of lipid-lowering medication, whereas diabetes was defined as fasting plasma glucose levels above 7 mmol/L, random plasma glucose above 11.1 mmol/L or use of medication indicated for treatment of diabetes.

Genotyping

Genotyping on the Illumina 6K Array

Genomic DNA was extracted from peripheral venous blood of all study participants using the salting out procedure.[22] Genotyping was performed using the 6K Illumina Linkage IV Panels (Illumina, San Diego, CA, USA) at the Centre National de Genotypage in France. Markers with a minor allele frequency (MAF) < 5%, call rate < 98% or which failed an exact test of Hardy-Weinberg equilibrium (HWE) (p-value < 10^{-8}) were removed

during the quality control process. In total 5,250 autosomal variants were available for analysis.

Exome sequencing

The exomes of randomly selected participants from the ERF study were sequenced at the Cell Biology Department of the Erasmus MC, The Netherlands. Sequencing was performed at a median depth of 57× using the Agilent version V4 capture kit on an Illumina Hiseq2000 sequencer using the TruSeq Version 3 protocol.[23,24] After quality control, we retrieved 528,617 single nucleotide variants (SNVs) in 1,308 individuals, of which 1,046 had cIMT data available. Annotation of the SNVs was performed using the SeattleSeq annotation database (http://snp.gs.washington.edu/SeattleSeqAnnotation138/). To further assess the functionality of the variants, we used RegulomeDB database that annotates SNVs with known and predicted regulatory elements and Combined Annotation Dependent Depletion (CADD) tool for scoring the deleteriousness of variants.[25,26] The ERF data is available in the European Genome-phenome Archive (EGA) public repository with ID number EGAS00001001134.

Statistical analysis

Linkage analysis

We performed affected only genome-wide multipoint non-parametric linkage (NPL) analysis in MERLIN 1.1.2 using individuals from the ERF study.[27] Individuals that scored above the 90th percentile of the distribution of the residuals from the regression of cIMT onto age, age^2, sex, smoking status, BMI, WHR, diabetes, dyslipidemia, and hypertension were set as affected (N = 103). Descriptive characteristics of the selected individuals are presented in **Table 1**. The selected individuals were older and higher cIMT measurements compared to all ERF study participants (**Table 1**). They also had a higher prevalence of hypertension, dyslipidemia, and diabetes than all ERF study participants, whereas the body mass index and waist to hip ratio were comparable (**Table 1**). These 103 affected individuals were connected to each other in a large pedigree consisting of 5,083 individuals. To facilitate linkage analysis, the 103 affected individuals were clustered into 21 smaller non-overlapping sub-pedigrees with a maximum bit size of 24 using the PEDCUT software version 1.19.[28] Bit size value is used to characterize the maximal number of subjects of interest who share a common ancestor.[28] The number of affected subjects of interest in the sub-pedigrees ranged from two to eight. MEGA2 software tool version 4.4[29] was used to create input files for MERLIN. Mendelian inconsistencies were set to missing within the whole sub-pedigree. There were 543 Mendelian inconsistencies observed among 5,250 autosomal variants. After they were set to missing, 4,707 autosomal variants were used in the linkage analysis. We also performed affected only parametric

linkage analysis under the dominant and recessive models assuming incomplete penetrance of 0.5 and a disease allele frequency of 0.01 using MERLIN. Marker allele frequencies were calculated from all genotyped individuals in the pedigrees. Subsequently, we carried out per family analyses in order to identify families that were contributing predominantly to the linkage signals, henceforth referred to as 'contributing families'. Additionally, we performed variance component linkage analysis in MERLIN using quantitative cIMT in the total study population. To facilitate analysis PEDCUT software was used to cluster individuals into 116 non-overlapping sub-pedigrees. The number of subjects of interest in the sub-pedigrees ranged from two to eighteen. To determine the significance of each test, the logarithm of the odds (LOD) score was calculated as the log10 of the likelihood ratio. The LOD score of 3.3 or higher was considered to represent genome-wide significance threshold, whereas the LOD score of 1.9 was used to declare genome-wide suggestive threshold.[30]

Identification of variants under the linkage peaks using exome sequencing

We used exome sequence data to identify variants that could explain observed linkage peaks. To this end, we looked for variants that were shared among the majority of affected individuals from the contributing families within the respective linkage peak. We only considered variants with MAF < 5% or absent in 1000 Genome Project (1kG) and MAF < 5% in the ERF controls which were defined as individuals who scored below the mean of the distribution of the residuals from the regression of cIMT onto age, age^2, sex, smoking status, BMI, WHR, diabetes, dyslipidemia, and hypertension. The

Table 1. Descriptive statistics of study populations including ERF cases (N=103) selected for the linkage analysis and ERF overall.

Characteristics	ERF cases	ERF overall
Age, mean (sd)	53.6 (13.6)	48.3 (14.2)
Gender, % of males	45.6%	40.2%
IMT (mm), mean (sd)	1.1 (0.2)	0.8 (0.2)
Smoking, % of ever smokers	44.7%	41.9%
BMI (kg/m^2), mean (sd)	26.9 (3.7)	26.7 (4.4)
WHR, mean (sd)	0.9 (0.1)	0.9 (0.1)
Hypertension, % of cases with hypertension	63.1%	48.7%
Dyslipidemia, % of cases with dyslipidemia	51.5%	36.2%
Diabetes, % of patients with diabetes	6.8%	4.5%

Abbreviations: IMT - intima-media thickness, BMI - body mass index, WHR - waist to hip ratio;
Hypertension: systolic blood pressure above 140 mmHg, diastolic blood pressure above 90 mmHg or use of medication for treatment of hypertension; Dyslipidemia - total cholesterol above 6.2 mmol/L or use of lipid-lowering medication; Diabetes - fasting plasma glucose levels above 7 mmol/L, random plasma glucose above 11.1 mmol/L or use of medication indicated for the treatment of diabetes;

MAF of variants absent in 1kG project was checked in NHLBI Exome Sequencing Project (http://evs.gs.washington.edu/EVS/). Candidate variants were subjected to quantitative trait association analysis with cIMT in the ERF under the same model as in the sharing analysis (additive, dominant, recessive) using the RVtests software.[31] Inverse normalized residuals from the regression of cIMT onto age, age^2, sex, smoking status, BMI, WHR, diabetes, dyslipidemia, and hypertension were used in the association analysis. To take into account multiple tests, we first calculated a number of independent tests using the method of Li and Ji.[32] Subsequently, Bonferroni corrected *p*-value was calculated based on number of independent tests. GTEx portal (https://www.gtexportal.org/home/) was used to check for gene expression.

RESULTS

The results of affected only genome-wide non-parametric and parametric linkage scans are illustrated in **Figure 1**. Regions with significant (LOD > 3.3) evidence of linkage in either the non-parametric or the parametric analyses are shown in **Table 2**. Significant evidence of linkage for cIMT was observed to chromosomes 2p16.3, 19q13.43, 20p13, and 21q22.12 in the parametric linkage analysis under the dominant model, and to chromosomes 19q13.43 and 20p13 in the parametric linkage analysis under the recessive model. The families contributing predominantly to these linkage peaks and the distribution of their per-family heterogeneity LOD (HLOD) scores are shown in **Supplementary Figure 1-4**.

Table 2. Genome-wide significant results of linkage analyses for cIMT. Start and end SNV are reported for base to base linkage regions.

Region*	Start SNV	End SNV	Start position**	End position**	SNV with max LOD	Dominant (HLOD)	Recessive (HLOD)	Non-parametric (LOD)
2p16.3	rs1447107	rs1017267	45272197	56785785	rs1017418	**3.35**	2.88	1.40
19q13.43	rs897783	rs3499	52031162	59093484	rs3499	**7.17**	**9.09**	**3.73**
20p13	rs1434789	rs241605	137900	3915064	rs1434789	**4.10**	**3.87**	**3.34**
21q22.12	rs762173	rs2836301	33832675	40351780	rs2834949	**3.59**	1.86	2.14

Abbreviations: SNV - Single Nucleotide Variant; HLOD - heterogeneity LOD;
*Regions with significant evidence of linkage are in bold;
**Start and end positions correspond to genetic position of SNV at the start or end of the base to base linkage region according to hg19 assembly;

We next determined to what extent the affected members in these families shared rare variants under the linkage peaks. Sharing analyses under the base to base linkage peak at 2p16 (family specific HLOD = 3.63) identified intronic and coding-synonymous vari-

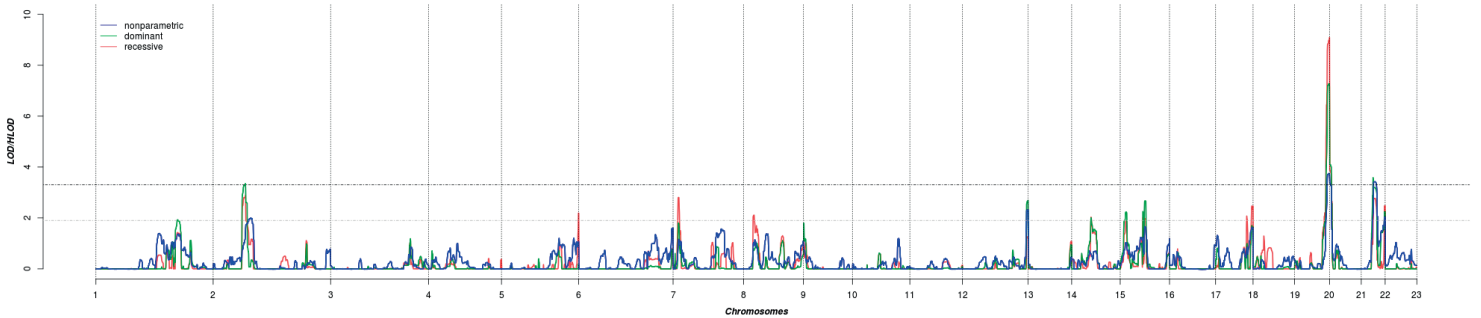


Figure 1. The results of genome-wide linkage scan for non-parametric (blue) and parametric analyses under the dominant (green) and recessive (red) models. The x-axis shows 22 autosomal chromosomes, whereas the y-axis shows the LOD scores for non-parametric model and heterogeneity LOD (HLOD) scores for dominant and recessive model. Black dotted line depicts the genome-wide significant threshold, whereas gray dotted line shows the suggestive threshold.

Table 3. Variants shared among the affected family members of the family that predominantly contributed to the LOD score at 2p16

Name*	Function	Gene	MAF ERF controls**	MAF 1kG***	CADD	Regulome DB	Association analysis in ERF			
							Beta	$Beta_{untransformed}$	SE	P
rs375801385	intron	*FBXO11*	0.031	0.056	17.18	6	0.023	0.014	0.145	0.876
2:48848294	intron	*GTF2A1L*	0.031	NA	5.82	-	0.023	0.014	0.145	0.876
rs149304214	coding-synonymous	*SPTBN1*	0.016	0.002	15.95	-	0.311	0.059	0.182	0.087
rs62165235	intron	*PNPT1*	0.044	0.038	4.413	2b	0.265	0.037	0.107	0.013
rs114706375	intron	*USP34*	0.026	0.010	0.122	5	0.124	0.036	0.145	0.393
rs144629927	intron	*XPO1*	0.021	0.003	4.595	3a	0.131	0.041	0.159	0.409

Abbreviations: MAF - minor allele frequency; 1kG - 1000 Genomes; CADD - Combined Annotation Dependent Depletion score; Regulome DB - Regulome DB score; Beta - effect estimate; Betauntransformed - effect estimate from association analysis in which untransformed cIMT was used; SE - standard error of Beta; P - p-value; ERF - Erasmus Rucphen Family study;

*The variants are ordered based on their genomic position (hg19 assembly). First four variants are shared by 6 of 8 affected family members in a family contributing predominately to the linkage peak, and last two variants are shared by 5 of 8 affected family members;

**ERF controls were defined as individuals who scored below the mean of the distribution of the residuals from the regression of cIMT onto age, age^2, sex, smoking status, BMI, WHR, diabetes, dyslipidemia, and hypertension;

***If MAF was unknown in 1kG, MAF reported from Exome Sequencing Project if available;

Table 4. Functional annotation of rs62165235 variant and variants that are in LD ($r^2 > 0.6$) using HaploReg 4.1[69]

Variant	LD (r^2)	Ref	Alt	GERP cons*	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	Motifs changed	Selected eQTL hits	RefSeq genes	function
rs7591128	0.65	T	G	No	-	BLD	-	-	6 altered motifs	1 hit	39kb 5′ of *CCDC88A*	intergenic
rs78928997	0.75	T	A	Yes	-	-	-	-	7 altered motifs	-	*SMEK2*	3′-UTR
rs62165193	0.77	C	T	No	-	2 tissues	-	-	GR	-	*SMEK2*	intronic
rs62165227	0.61	G	A	No	-	-	-	-	-	-	*PNPT1*	intronic
rs62165231	0.95	C	T	No	-	-	HRT	-	Rad21,Tgif1	1 hit	*PNPT1*	intronic
rs62165235	1	T	C	No	24 tissues	-	15 tissues	E2F6	7 altered motifs	-	*PNPT1*	intronic
rs62165236	0.96	T	C	No	-	-	-	-	ZID	-	2.1kb 5′ of *PNPT1*	intergenic
rs79873145	0.76	T	C	No	-	GI	-	-	4 altered motifs	1 hit	30kb 5′ of *PNPT1*	intergenic

Abbreviations: Ref - reference allele; Alt - Alternative allele; GERP - conservation score;*GERP conservation score indicates whether the element is conserved or not according to the algorithm; The variant highlighted in red was identified in the sharing and association analysis in the ERF;

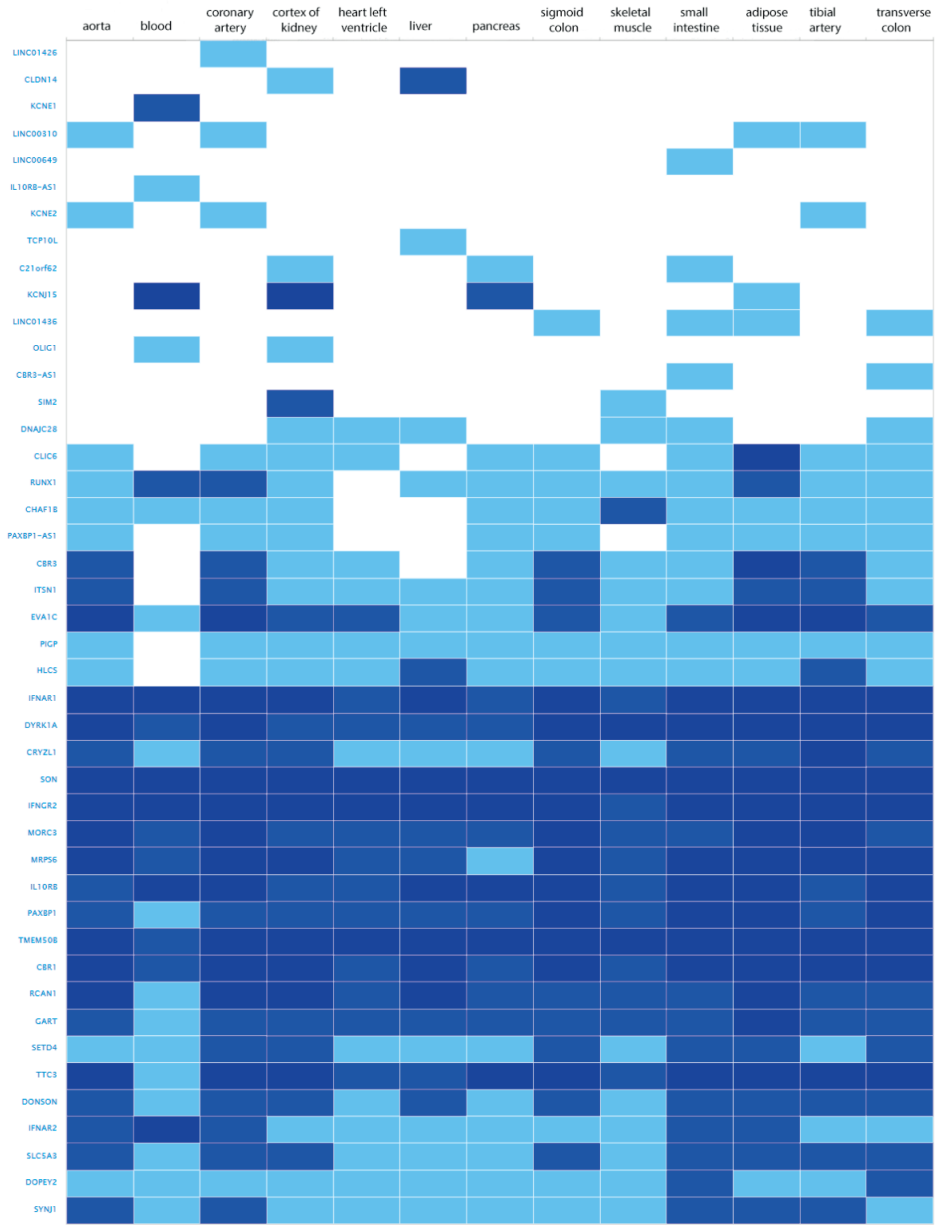


Figure 2. Genes under the base to base linkage peak at 21q22 and their expression levels in tissues relevant for atherosclerosis according to GTEx database. Gene names are shown on y-axis and tissues on x-axis. Colour depicts expression level estimated as fragments per kilobase of exon model per million reads mapped (FPKM). The grey box stands for expression level below cutoff (0.5 FPKM), light blue box stands for low expression level (between 0.5 to 10), medium blue for medium expression level (between 11 to 1000 FPKM), and dark blue for high expression level (more than 1000 FPKM or more than 1000 TPM) and white for no data available.

ants (**Table 3**). The most interesting finding is a variant (rs62165235) with MAF 0.038 in 1kG mapping to *PNPT1*. The variant, shared by 6 out of 8 affected relatives, is likely having a regulatory function and affecting transcription factor binding and matched DNase Footprinting and DNase sensitivity (Category 2b Regulome DB score; **Table 3-4**). The variant was sequenced at a read depth of 37x and it showed association with quantitative cIMT in the ERF (effect = 0.27, *p*-value = 0.013, **Table 3**) after applying Bonferroni correction (*p*-value = 0.05/3 independent tests = 0.017). The effect estimate of the minor allele C on untransformed cIMT suggested a mean increase of 0.04 mm for each minor allele (0.04 mm for heterozygote C/T carriers and 0.08 for homozygous C/C carriers) (**Table 3**). This variant explained 0.3% of variation in the ERF.

The search for shared variants within the linkage region 19q13.43, 20p13 and 21q22.12 identified several variants to be shared among the affected family members, however, none of the variants showed significant association with quantitative cIMT (**Supplementary Table 1-3**). There are, however, several potentially interesting candidate genes for atherosclerosis in each of these linked regions, for instance, among the genes under the base to base 19q13 and 20p13 linkage peaks, several genes have been implicated in the pathogenesis of cardiovascular disease, including *FCAR, TNNT1, OSCAR, FPR2* under the 19q13 and *ADAM33, TRIB3, HSPA12B* under the 20p13 peak. The 21q22 region harbors several genes that are highly expressed in tissues relevant for atherosclerosis, including *IFNAR1, DYRK1A, SON, IFNGR2, MORC3, MRPS6, IL10RB, TMEM50B, CBR1, RCAN1,* and *TTC3* (**Figure 2**). According to the Ingenuity Pathway Analysis (IPA) tool (QIAGEN Inc., https://www.qiagenbioinformatics.com/products/ingenuitypathway-analysis), which exposes possible functional relationship between the genes by expanding upstream analysis to include regulators that are not directly connected to targets in the dataset, these genes connected to a network illustrated in **Supplementary Figure 5**.

There were several regions that showed suggestive evidence of linkage, including 1q31.1, 5q35.3, 7p21.3, 8p22, 12q24.33, 14q22.2, 15q21.3 and 17q25.3. As the 12q24 region has previously been linked to cIMT, we have explored it further (**Figure 1**). Search for shared variants within 12q24 identified no variants that can explain linkage signal.

The results of linkage analysis when using cIMT as a quantitative outcome are shown in **Supplementary Figure 6**.

DISCUSSION

In this study, we have identified genomic regions at 2p16.3, 19q13.43, 20p13 and 21q22.12 with significant evidence of linkage to cIMT. These regions have not been reported before. Identification of variants under the linkage peaks using exome sequencing revealed a variant with likely regulatory function mapping to *PNPT1* gene at chromosome 2 and several candidate genes at 21q22.

As the present study targets genes with relatively large effects, we studied the extremes of cIMT distribution in the discovery population. Even though extreme trait approach neglects much of the overall distribution of the trait and some rare variants with the moderate effects may be missed, it has been shown that the power to detect rare variants can be increased due to an excess of rare variants in the upper tails of the distribution.[15,34,35] Comparison of the results obtained in the linkage analysis of the extremes of cIMT distribution and those using cIMT as a quantitative trait revealed no overlap, highlighting the power of the approach we have followed.

When comparing the results of several genome-wide linkage studies of cIMT that have been conducted so far, we noticed that a region with suggestive evidence of linkage in our study, 12q24, has previously been linked to cIMT through the linkage scan.[10] Similarly to the previous study which identified 12q24, we did not identify variants by sharing analysis that could explain the linkage signal. Even though linkage findings from prior studies already showed limited generalizability across the reported linkage peaks due to selected nature of the cohorts, the overlap of our finding with the literature suggests that our study population is representative of the general population. However, the linkage single at chromosome 12q24 is relatively weak in our population.

Among the regions with significant evidence of linkage to cIMT in our study, we identified 2p16.3 which gave significant linkage signal under the dominant model and suggestive signal under the recessive model. This region has previously been associated with polycystic ovary syndrome (PCOS) and primary open angle glaucoma (POAG).[36,37]Interestingly, women with PCOS are at a greater risk of premature atherosclerosis,[38] whereas atherosclerosis associated with vascular conditions that are correlated with POAG.[39] The region has further been implicated in body mass index[40] and glycated hemoglobin levels.[41] Both obesity and poor glycemic control are risk factors for variety of diseases including atherosclerosis and cardiovascular diseases.

Identification of variants under the base to base peak at 2p16 using exome sequence data revealed a variant that lies in DNase sites, promote histone marks and protein

binding regions and changes regulatory motifs based on the variant allele change. The variant is mapped to intron 1 of Polyribonucleotide Nucleotidyltransferase 1 (*PNPT1*) gene which encodes a protein predominantly localized in the mitochondrial intermembrane space and is involved in import of RNA to mitochondria (http://www.genecards.org/cgi-bin/carddisp.pl?gene=PNPT1). *PNPT1* has been characterized as a type I interferon-inducible early response gene.[42,43] Type I interferons promote atherosclerosis by enhancing macrophage-endothelial cell adhesion and promoting leukocyte attraction to atherosclerosis-prone sites in animal models.[44] Even though our finding supports a role of *PNPT1* as a candidate gene in atherosclerosis, we acknowledge that *PNPT1* variant is unlikely to be causal and cannot explain the linkage signal at 2p16.3 to cIMT. Furthermore, we attempted to replicate the association of this variant with cIMT in the Rotterdam Study, a population-based cohort study (detailed information provided in **Supplementary Methods** and **Supplementary Table 4**). However, the variant was not available in the exome sequencing data of the Rotterdam Study.

The other interesting region includes 21q22, which is also known as a Down critical region. Interestingly, persons with Down syndrome are protected against atherosclerosis, in spite of increases in metabolic disturbances and obesity in Down syndrome.[45] Even though identification of variants using exome sequencing did not identify a causal variant, this region contains several plausible candidate genes which are highly expressed in relevant tissues[46], including *IFNAR1, DYRK1A, SON, IFNGR2, MORC3, MRPS6, IL10RB, TMEM50B, CBR1, RCAN1,* and *TTC3*. *DYRK1A* signaling pathway is linked to homocysteine cycle which is associated with an increased risk of atherosclerosis.[47,48] *IFNGR2* and *RCAN1* also play a role in atherosclerosis.[49,50] Notably, IPA analysis revealed that those genes are connected in one network, and directly or indirectly linked to *TP53*. *TP53* encodes a tumor suppressor gene p53 involved in regulation of cell proliferation and apoptosis. Numerous studies implicated p53 in development of atherosclerosis and vascular smooth muscle cell apoptosis.[51-55] Higher plasma p53 levels were also associated with an increased cIMT.[56] However, it is important to note that network analysis is based on the knowledge databases that are always evolving and new discoveries happen all time.

Furthermore, we identified 19q13.43 and 20p13 regions with significant evidence of linkage to cIMT. Several genes under the linkage peak have previously been implicated in the pathogenesis of cardiovascular disease. The base to base peak at 19q13 encompassed *FCAR* and *TNNT1* genes associated with coronary heart disease[57,58] and *OSCAR* and *FPR2* genes associated with atherosclerosis plaque phenotype,[59,60] whereas the base to base peak at 20p13 encompassed *ADAM33* and *TRIB3* associated with extent and promotion of atherosclerosis[61-65] and *HSPA12B* which is found to be enriched in atherosclerotic lesions.[66]

Our study presents the linkage analysis using extreme phenotype approach that was designed to capture region with genetic variants that have large effects on cIMT. Combination of linkage analysis in a large family-based study and exome sequence data provide a unique opportunity to explore the variants in the linkage regions. However, despite these distinct advantages, we were able to identify a genetic variant for only one of the several linked genomic regions, for which, there may be several reasons including structural variants, and intronic or intergenic single-nucleotide variants that were not evaluated in the current study. Interestingly, the 19q13.43, 20p13 and 21q22.12 linkage peaks were previously associated with various phenotypes in our study population including personality traits and depressive symptoms.[67,68]

To conclude, our linkage analysis identified four genomic regions at 2p16.3, 19q13.43, 20p13 and 21q22.12 for cIMT. The significant linkage regions contain several plausible candidate genes. Further analyses are needed to demonstrate their involvement in atherosclerosis.

REFERENCES

1. Xu, J., Murphy, S.L., Kochanek, K.D. & Bastian, B.A. Deaths: Final Data for 2013. *Natl Vital Stat Rep* **64**, 1-119 (2016).
2. Falk, E. Pathogenesis of atherosclerosis. *J Am Coll Cardiol* **47**, C7-12 (2006).
3. Lorenz, M.W., Markus, H.S., Bots, M.L., Rosvall, M. & Sitzer, M. Prediction of clinical cardiovascular events with carotid intima-media thickness: a systematic review and meta-analysis. *Circulation* **115**, 459-67 (2007).
4. Polak, J.F. *et al.* Carotid-wall intima-media thickness and cardiovascular events. *N Engl J Med* **365**, 213-21 (2011).
5. Den Ruijter, H.M. *et al.* Common carotid intima-media thickness measurements in cardiovascular risk prediction: a meta-analysis. *JAMA* **308**, 796-803 (2012).
6. Lusis, A.J. Genetics of atherosclerosis. *Trends Genet* **28**, 267-75 (2012).
7. Fox, C.S. *et al.* Genetic and environmental contributions to atherosclerosis phenotypes in men and women: heritability of carotid intima-media thickness in the Framingham Heart Study. *Stroke* **34**, 397-401 (2003).
8. Sacco, R.L. *et al.* Heritability and linkage analysis for carotid intima-media thickness: the family study of stroke risk and carotid atherosclerosis. *Stroke* **40**, 2307-12 (2009).
9. Wang, D. *et al.* A genome-wide scan for carotid artery intima-media thickness: the Mexican-American Coronary Artery Disease family study. *Stroke* **36**, 540-5 (2005).
10. Fox, C.S. *et al.* Genomewide linkage analysis for internal carotid artery intimal medial thickness: evidence for linkage to chromosome 12. *Am J Hum Genet* **74**, 253-61 (2004).
11. Kuipers, A.L. *et al.* Genetic epidemiology and genome-wide linkage analysis of carotid artery ultrasound traits in multigenerational African ancestry families. *Atherosclerosis* **231**, 120-3 (2013).
12. Bis, J.C. *et al.* Meta-analysis of genome-wide association studies from the CHARGE consortium identifies common variants associated with carotid intima media thickness and plaque. *Nat Genet* **43**, 940-7 (2011).
13. Natarajan, P. *et al.* Multiethnic Exome-Wide Association Study of Subclinical Atherosclerosis. *Circ Cardiovasc Genet* **9**, 511-520 (2016).
14. Bis, J.C. *et al.* Sequencing of 2 subclinical atherosclerosis candidate regions in 3669 individuals: Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium Targeted Sequencing Study. *Circ Cardiovasc Genet* **7**, 359-64 (2014).
15. Auer, P.L. & Lettre, G. Rare variant association studies: considerations, challenges and opportunities. *Genome Med* **7**, 16 (2015).
16. Stacey, S.N. *et al.* A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. *Nat Genet* **43**, 1098-103 (2011).
17. Gudmundsson, J. *et al.* A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nature Genetics* **44**, 1326-1329 (2012).
18. Amin, N. *et al.* A rare missense variant in RCL1 segregates with depression in extended families. *Molecular Psychiatry* **23,** 1120-1126 (2018).
19. Pardo, L.M., MacKay, I., Oostra, B., van Duijn, C.M. & Aulchenko, Y.S. The effect of genetic drift in a young genetically isolated population. *Ann Hum Genet* **69**, 288-95 (2005).
20. Aulchenko, Y.S. *et al.* Linkage disequilibrium in young genetically isolated Dutch population. *Eur J Hum Genet* **12**, 527-34 (2004).
21. Sayed-Tabatabaei, F.A. *et al.* Heritability of the function and structure of the arterial wall: findings of the Erasmus Rucphen Family (ERF) study. *Stroke* **36**, 2351-6 (2005).

22. Miller, S.A., Dykes, D.D. & Polesky, H.F. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* **16**, 1215 (1988).
23. Amin, N. *et al.* Exome-sequencing in a large population-based study reveals a rare Asn396Ser variant in the LIPG gene associated with depressive symptoms. *Mol Psychiatry* (2016).
24. Amin, N. *et al.* Nonsynonymous Variation in NKPD1 Increases Depressive Symptoms in European Populations. *Biol Psychiatry* **81**, 702-707 (2017).
25. Boyle, A.P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research* **22**, 1790-1797 (2012).
26. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* **46**, 310 (2014).
27. Abecasis, G.R., Cherny, S.S., Cookson, W.O. & Cardon, L.R. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30**, 97-101 (2002).
28. Liu, F., Kirichenko, A., Axenovich, T.I., van Duijn, C.M. & Aulchenko, Y.S. An approach for cutting large and complex pedigrees for linkage analysis. *Eur J Hum Genet* **16**, 854-60 (2008).
29. Baron, R.V., Kollar, C., Mukhopadhyay, N. & Weeks, D.E. Mega2: validated data-reformatting for linkage and association analyses. *Source Code Biol Med* **9**, 26 (2014).
30. Ott, J., Wang, J. & Leal, S.M. Genetic linkage analysis in the age of whole-genome sequencing. *Nat Rev Genet* **16**, 275-84 (2015).
31. Zhan, X.W., Hu, Y.N., Li, B.S., Abecasis, G.R. & Liu, D.J.J. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics* **32**, 1423-1426 (2016).
32. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* **95,** 221-227 (2005).
33. Ramsey, S.A., Gold, E.S. & Aderem, A. A systems biology approach to understanding atherosclerosis. *Embo Molecular Medicine* **2,** 79-89 (2010).
34. Lamina, C. Digging into the extremes: a useful approach for the analysis of rare variants with continuous traits? *BMC Proc* **5** Suppl 9, S105 (2011).
35. Coassin, S. *et al.* Investigation and functional characterization of rare genetic variants in the adipose triglyceride lipase in a large healthy working population. *PLoS Genet* **6**, e1001239 (2010).
36. Mutharasan, P. *et al.* Evidence for chromosome 2p16.3 polycystic ovary syndrome susceptibility locus in affected women of European ancestry. J *Clin Endocrinol Metab* **98**, E185-90 (2013).
37. Liu, Y.T., Qin, X.J., Schmidt, S., Allingham, R.R. & Hauser, M.A. Association between chromosome 2p16.3 variants and glaucoma in populations of African descent. *Proceedings of the National Academy of Sciences of the United States of America* **107**, E61-E61 (2010).
38. Meyer, M.L., Malek, A.M., Wild, R.A., Korytkowski, M.T. & Talbott, E.O. Carotid artery intima-media thickness in polycystic ovary syndrome: a systematic review and meta-analysis. *Human Reproduction Update* **18**, 112-126 (2012).
39. Belzunce, A. & Casellas, M. [Vascular risk factors in primary open angle glaucoma] Factortes de riesgo vascular en el glaucoma primario de angulo abierto. *An Sist Sanit Navar* **27**, 335-44 (2004).
40. Akiyama, M. *et al.* Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. *Nat Genet* **49**, 1458-1467 (2017).
41. Wheeler, E. *et al.* Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis. *PLoS Med* **14**, e1002383 (2017).

42. Leszczyniecka, M. *et al.* Identification and cloning of human polynucleotide phosphorylase, hPNPase old-35, in the context of terminal differentiation and cellular senescence. *Proc Natl Acad Sci U S A* **99**, 16636-41 (2002).
43. Leszczyniecka, M., Su, Z.Z., Kang, D.C., Sarkar, D. & Fisher, P.B. Expression regulation and genomic organization of human polynucleotide phosphorylase, hPNPase(old-35), a Type I interferon inducible early response gene. *Gene* **316**, 143-56 (2003).
44. Goossens, P. *et al.* Myeloid type I interferon signaling promotes atherosclerosis by stimulating macrophage recruitment to lesions. *Cell Metab* **12**, 142-53 (2010).
45. Colvin, K.L. & Yeager, M.E. What people with Down Syndrome can teach us about cardiopulmonary disease. *European Respiratory Review* **26** (2017).
46. Ramsey, S.A., Gold, E.S. & Aderem, A. A systems biology approach to understanding atherosclerosis. *EMBO Mol Med* **2**, 79-89 (2010).
47. Noll, C. *et al.* DYRK1A, a novel determinant of the methionine-homocysteine cycle in different mouse models overexpressing this Down-syndrome-associated kinase. *PLoS ONE* **4**, e7540 (2009).
48. Tlili, A. *et al.* Hepatocyte-specific Dyrk1a gene transfer rescues plasma apolipoprotein A-I levels and aortic Akt/GSK3 pathways in hyperhomocysteinemic mice. *Biochim Biophys Acta* **1832**, 718-28 (2013).
49. Mendez-Barbero, N. *et al.* A major role for RCAN1 in atherosclerosis progression. *EMBO Mol Med* **5**, 1901-17 (2013).
50. Voloshyna, I., Littlefield, M.J. & Reiss, A.B. Atherosclerosis and interferon-gamma: new insights and therapeutic targets. *Trends Cardiovasc Med* **24**, 45-51 (2014).
51. Boesten, L.S. *et al.* Macrophage p53 controls macrophage death in atherosclerotic lesions of apolipoprotein E deficient mice. *Atherosclerosis* **207**, 399-404 (2009).
52. Tabas, I. p53 and atherosclerosis. *Circ Res* **88**, 747-9 (2001).
53. Mayr, M., Hu, Y., Hainaut, H. & Xu, Q. Mechanical stress-induced DNA damage and rac-p38MAPK signal pathways mediate p53-dependent apoptosis in vascular smooth muscle cells. *FASEB J* **16**, 1423-5 (2002).
54. Mercer, J., Figg, N., Stoneman, V., Braganza, D. & Bennett, M.R. Endogenous p53 protects vascular smooth muscle cells from apoptosis and reduces atherosclerosis in ApoE knockout mice. *Circ Res* **96**, 667-74 (2005).
55. Varela, A. *et al.* Elevated expression of mechanosensory polycystins in human carotid atherosclerotic plaques: association with p53 activation and disease severity. *Sci Rep* **5**, 13461 (2015).
56. Chen, W. *et al.* p53 Levels positively correlate with carotid intima-media thickness in patients with subclinical atherosclerosis. *Clin Cardiol* **32**, 705-10 (2009).
57. Iakoubova, O.A. *et al.* Asp92Asn polymorphism in the myeloid IgA Fc receptor is associated with myocardial infarction in two disparate populations - CARE and WOSCOPS. *Arteriosclerosis Thrombosis and Vascular Biology* **26**, 2763-2768 (2006).
58. Guay, S.P. *et al.* Epigenetic and genetic variations at the TNNT1 gene locus are associated with HDL-C levels and coronary artery disease. *Epigenomics* **8**, 359-371 (2016).
59. Goettsch, C. *et al.* The Osteoclast-Associated Receptor (OSCAR) Is a Novel Receptor Regulated by Oxidized Low-Density Lipoprotein in Human Endothelial Cells. *Endocrinology* **152**, 4915-4926 (2011).
60. Petri, M.H. *et al.* The role of the FPR2/ALX receptor in atherosclerosis development and plaque stability. *Cardiovascular Research* **105**, 65-74 (2015).
61. Figarska, S.M., Vonk, J.M., van Diemen, C.C., Postma, D.S. & Boezen, H.M. ADAM33 gene polymorphisms and mortality. A prospective cohort study. *PLoS One* **8**, e67768 (2013).

62. Holloway, J.W. *et al.* ADAM33 expression in atherosclerotic lesions and relationship of ADAM33 gene variation with atherosclerosis. *Atherosclerosis* 211, 224-30 (2010).
63. Formoso, G. *et al.* The TRIB3 R84 variant is associated with increased carotid intima-media thickness in vivo and with enhanced MAPK signalling in human endothelial cells. *Cardiovascular Research* **89**, 184-192 (2011).
64. Wang, Z.H. *et al.* Silence of TRIB3 Suppresses Atherosclerosis and Stabilizes Plaques in Diabetic ApoE(-/-)/LDL Receptor(-/-) Mice. *Diabetes* **61**, 463-473 (2012).
65. Prudente, S. *et al.* Infrequent TRIB3 coding variants and coronary artery disease in type 2 diabetes. *Atherosclerosis* **242**, 334-339 (2015).
66. Han, Z.H., Truong, Q.A., Park, S. & Breslow, J.L. Two Hsp70 family members expressed in atherosclerotic lesions. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 1256-1261 (2003).
67. Amin, N. *et al.* A rare missense variant in RCL1 segregates with depression in extended families. *Mol Psychiatry (*2017).
68. Amin, N. *et al.* A genome-wide linkage study of individuals with high scores on NEO personality traits. *Mol Psychiatry* **17**, 1031-41 (2012).
69. Ward, L.D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Research* **40,** D930-D934 (2012).

SUPPLEMENTARY METHODS AND TABLES

Supplementary Methods

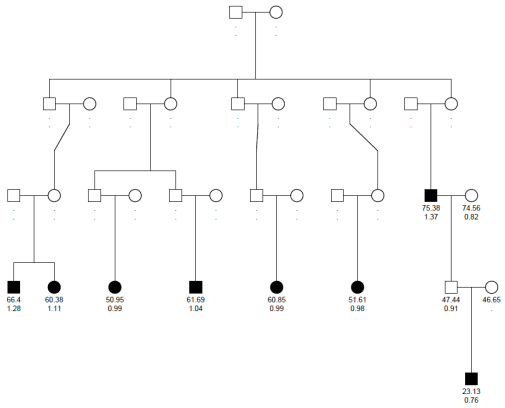
Supplementary Table 1. Variants shared by majority of affected family members in the family contributing the most to the linkage peak at chromosome 19 and their association with cIMT in the ERF.

Supplementary Table 2. Variants shared by majority of affected family members in the family contributing the most to the linkage peak at chromosome 20 and their association with cIMT in the ERF.

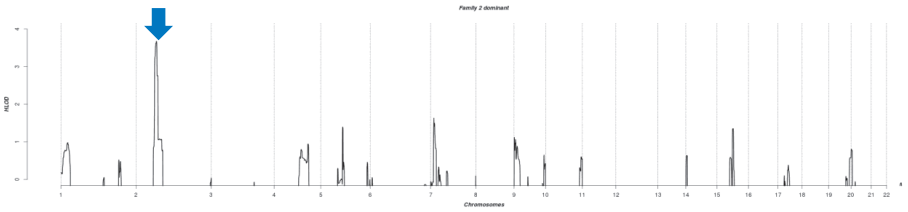
Supplementary Table 3. Variants shared by majority of affected family members in the family contributing the most to the linkage peak at chromosome 21 and their association with cIMT in the ERF.

Supplementary Table 4. Descriptive statistics of the Rotterdam Study.

(A)

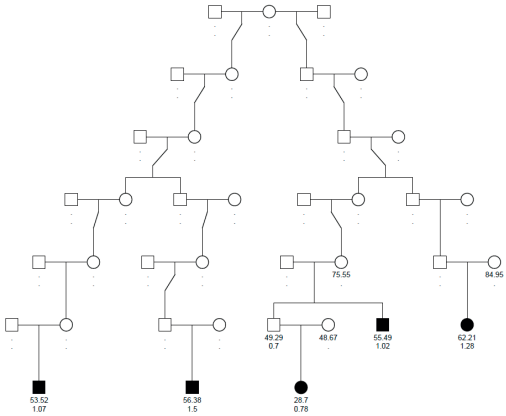


(B)

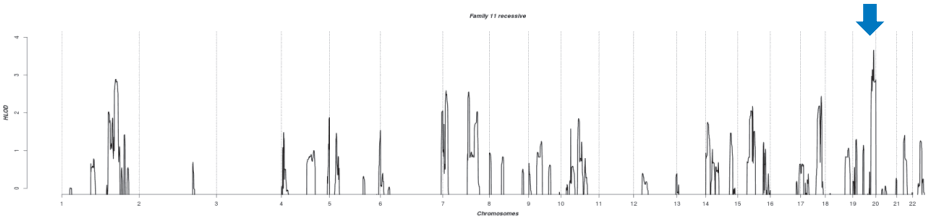


Supplementary Figure 1. The pedigree of highest HLOD score contributing family and the results of parametric per-family linkage analysis under the dominant model for the chromosome 2p16.3. (A) Squares represent males and circles females. Solid symbols depict affected family members. These family members were used in the linkage analysis. Age of individual and carotid intima-media thickness are displayed on the pedigree. Open symbols denote unaffected individuals (carotid intima-media thickness is displayed underneath the symbol) or individuals with no data available for analysis (carotid intima-media thickness is missing). (B) The x-axis shows 22 autosomal chromosomes, and the y-axis shows the heterogeneity LOD (HLOD) scores for the dominant model.

(A)

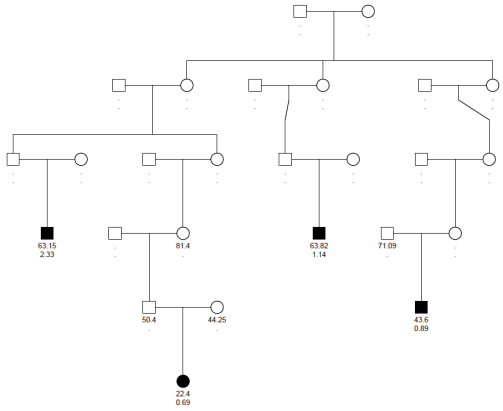


(B)

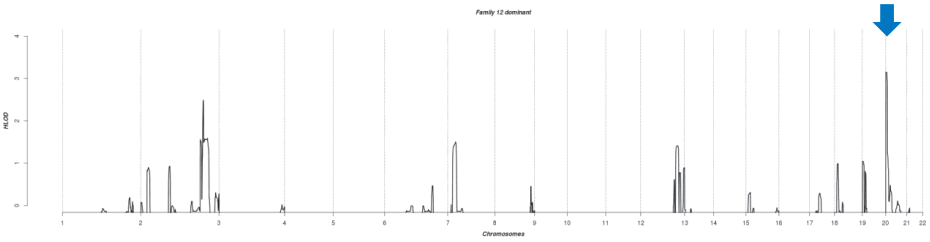


Supplementary Figure 2. The pedigree of highest HLOD score contributing family and the results of parametric per-family linkage analysis under the recessive model for the chromosome 19q13.43. (A) Squares represent males and circles females. Solid symbols depict affected family members. These family members were used in the linkage analysis. Age of individual and carotid intima-media thickness are displayed on the pedigree. Open symbols denote unaffected individuals (carotid intima-media thickness is displayed underneath the symbol) or individuals with no data available for analysis (carotid intima-media thickness is missing). (B) The x-axis shows 22 autosomal chromosomes, and the y-axis shows the heterogeneity LOD (HLOD) scores for recessive model.

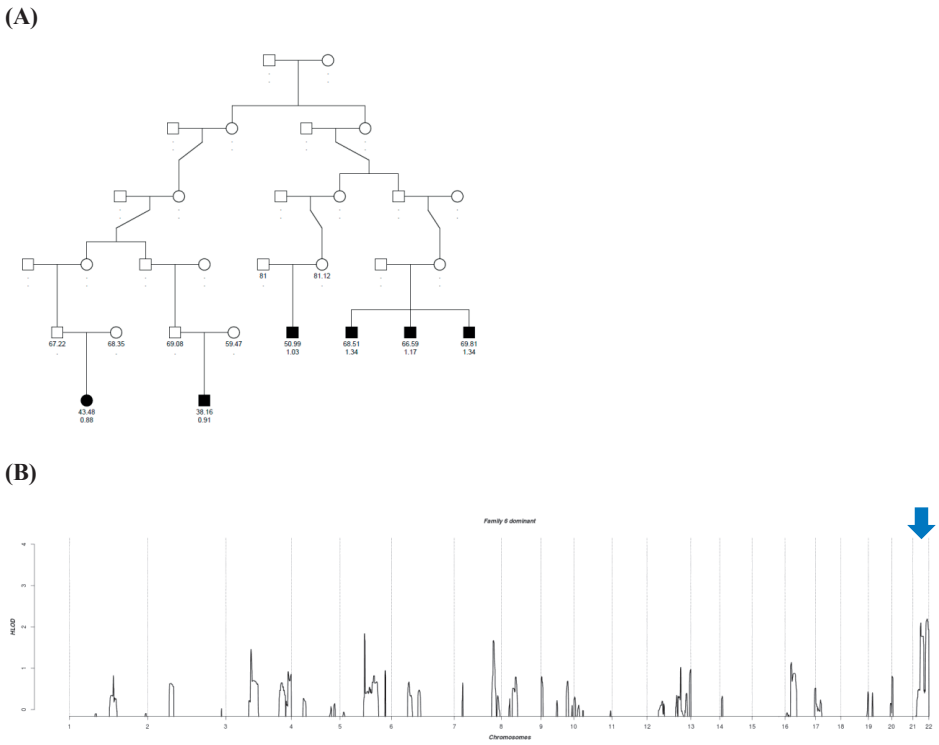
(A)



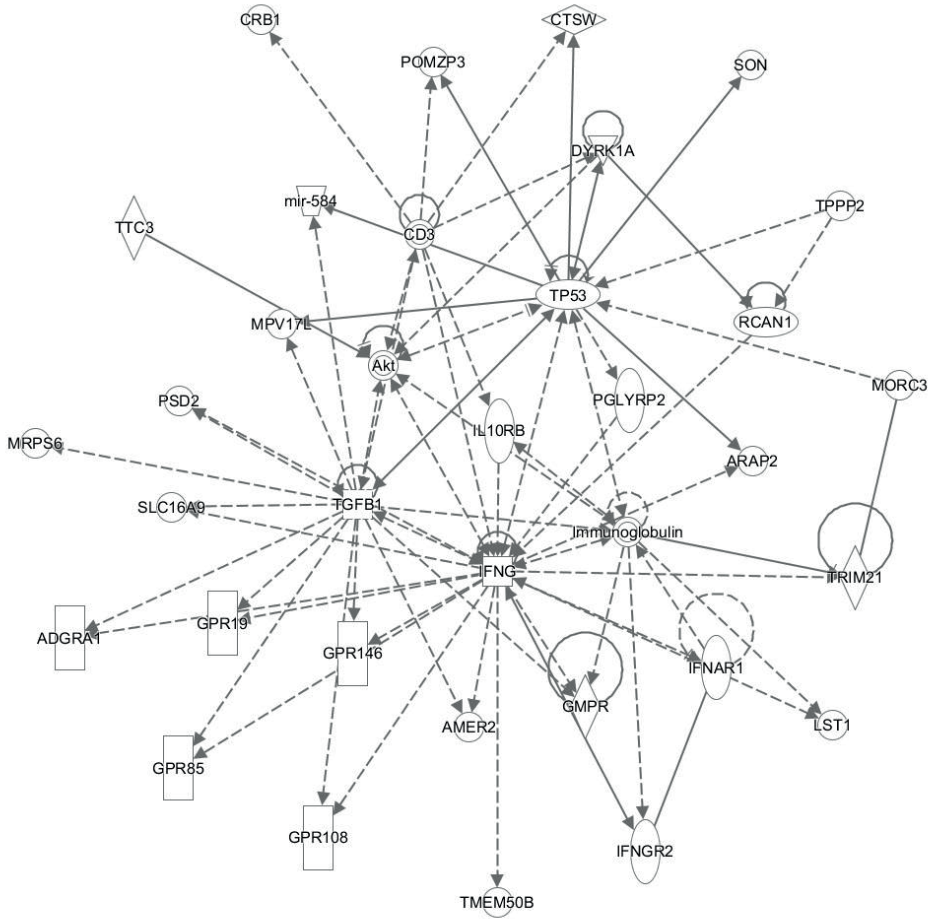
(B)



Supplementary Figure 3. The pedigree of highest HLOD score contributing family and the results of parametric per-family linkage analysis under the dominant model for the chromosome 20p13. (A) Squares represent males and circles females. Solid symbols depict affected family members. These family members were used in the linkage analysis. Age of individual and carotid intima-media thickness are displayed on the pedigree. Open symbols denote unaffected individuals (carotid intima-media thickness is displayed underneath the symbol) or individuals with no data available for analysis (carotid intima-media thickness is missing). (B) The x-axis shows 22 autosomal chromosomes, and the y-axis shows the heterogeneity LOD (HLOD) scores for the dominant model.



Supplementary Figure 4. The pedigree of highest HLOD score contributing family and the results of parametric per-family linkage analysis under the dominant model for the chromosome 21q22.12. (A) Squares represent males and circles females. Solid symbols depict affected family members. These family members were used in the linkage analysis. Age of individual and carotid intima-media thickness are displayed on the pedigree. Open symbols denote unaffected individuals (carotid intima-media thickness is displayed underneath the symbol) or individuals with no data available for analysis (carotid intima-media thickness is missing). (B) The x-axis shows 22 autosomal chromosomes, and the y-axis shows the heterogeneity LOD (HLOD) scores for the dominant model.



Supplementary Figure 5. Ingenuity Pathway Analysis tool network of genes under the linkage peak at 21q22 which are highly expressed in tissues relevant for atherosclerosis. Lines without arrow indicate interactions (chemical-chemical, protein-protein, chemical-protein, RNA-RNA, correlation) while lines with an arrow indicate activation, causation, expression, localization, transcription, molecular cleavage, membership, modification, phosphorylation, protein-DNA and/or protein-RNA interactions. Solid lines indicate direct interaction while dashed lines indicate indirect interactions. Diamond molecule shape denotes enzyme, rhombus - peptidase, inverted triangle - kinase, inverted trapezium - microRNA, circle in a circle - complex/group, vertical rectangle - G-protein coupled receptor, horizontal rectangle - ligand-dependent nuclear receptor, vertical ellipse - transmembrane receptor, horizontal ellipse - transcription regulator, square - cytokine, and circle - other.

Supplementary Figure 6. The results of quantitative trait linkage analysis using variance component linkage.

Chapter 3.2

Metabolic profiling of intra- and extracranial carotid artery atherosclerosis

Dina Vojinovic*, Sven J. van der Lee*, Cornelia M. van Duijn, Meike W. Vernooij, Maryam Kavousi, Najaf Amin, Ayşe Demirkan, M. Arfan Ikram, Aad van der Lugt, Daniel Bos

* These authors contributed equally to this work

This chapter was published in Atherosclerosis. 2018 Mar 8;272:60-65.

The supplemental information for this paper is available online at https://doi.org/10.1016/j.atherosclerosis.2018.03.015

ABSTRACT

Background and aims

Increasing evidence shows that intracranial carotid artery atherosclerosis may develop under the influence of a differential metabolic risk factor profile than atherosclerosis in the extracranial part of the carotid artery. To further elucidate these differences, we investigated associations of a wide range of circulating metabolites with intracranial and extracranial carotid artery atherosclerosis.

Methods

From the population-based Rotterdam Study, blood samples from 1,111 participants were used to determine a wide range of metabolites by proton nuclear magnetic resonance (NMR). Moreover, these participants underwent non-contrast computed tomography of the neck and head to quantify the amount of extra- and intracranial carotid artery calcification (ECAC and ICAC), as a proxy of atherosclerosis. We assessed associations of the metabolites with ICAC and ECAC and compared the metabolic association patterns of the two.

Results

We found that one standard deviation (SD) increase in concentration of 3-hydroxybutyrate, a ketone body, was significantly associated with a 0.11 SD increase in ICAC volume (p-value = 1.8×10^{-4}). When we compared the metabolic association pattern of ICAC with that of ECAC, we observed differences in glycolysis-related metabolite measures, lipoprotein subfractions, and amino acids. Interestingly, glycoprotein acetyls were associated with calcification in both studied vessel beds. These associations were most prominent in men.

Conclusions

We found that a higher circulating level of 3-hydroxybutyrate was associated with an increase in ICAC. Furthermore, we found differences in metabolic association patterns of ICAC and ECAC, providing further evidence for location-specific differences in the etiology of atherosclerosis.

INTRODUCTION

Carotid artery atherosclerosis is established as the single most important cause of stroke worldwide.[1-4] Importantly, increasing evidence suggests that the specific location of carotid atherosclerosis, i.e. extracranial versus intracranial, harbors unique, differential information with regard to the risk of subsequent stroke.[3-4] In addition, it was also found that the contribution of traditional cardiovascular risk factors to intracranial carotid artery atherosclerosis is different from that to extracranial carotid artery atherosclerosis.[5-7] In particular, diabetes mellitus and insulin resistance, i.e. expressions of disrupted glucose and insulin metabolism, seem to play a more prominent role in the development of intracranial carotid artery atherosclerosis.[6,8,9] This apparent location-specific susceptibility to metabolic disturbances warrants further in-depth investigation of the metabolic underpinnings of carotid artery atherosclerosis. Interestingly, methods to perform such an in-depth investigation of large spectra of active metabolites in relation to disease have only recently become available.[10,11] With the use of nuclear magnetic resonance (NMR), metabolites can now be inexpensively and reproducibly quantified on a large-scale, which enables metabolomics studies in large population-based cohorts. Successful examples include metabolic profiling of type 2 diabetes,[12,13] and cardiovascular events.[14–17]

Applying a similar approach to carotid artery atherosclerosis may expose important metabolites contributing to the disease. To date, several inflammatory markers have been associated with different stages and manifestation of carotid artery atherosclerosis, such as interleukin-6 and tumor necrosis factor-α.[18,19] Ultimately, this knowledge may provide opportunities for the development of specific therapeutic and preventive strategies.

Hence, the aim of this study was to investigate associations of a broad range of metabolites with intracranial and extracranial carotid artery calcification (ICAC and ECAC), as a proxy of atherosclerosis, and to compare the metabolic association profile of ICAC with that of ECAC.

MATERIALS AND METHODS

Study population

Our study population consisted of participants from the Rotterdam Study, a prospective population-based cohort study among individuals aged 45 years and over, who are living in the well-defined Ommoord district in Rotterdam, the Netherlands.[20] The study started in 1990, with 7,983 participants (first Rotterdam Study cohort, RS-I), and was extended in 2000/2001 (RS-II, 3,011 participants) and 2006/2008 (RS-III, 3,932 participants).[20] All participants were invited for extensive re-examinations every 3–4 years. At each visit,

blood was drawn after overnight fasting. The Rotterdam Study has been approved by the Medical Ethics Committee of the Erasmus MC and by the Ministry of Health, Welfare and Sport of the Netherlands, implementing the Wet Bevolkingsonderzoek: ERGO (Population Studies Act: Rotterdam Study).[20] All participants provided written informed consent to participate in the study and to obtain information from their treating physicians.

Population for analysis

Metabolites were available for two independent datasets of the Rotterdam Study. The first set encompassed all individuals from the RS-I cohort that participated in the fourth examination round at the study center (N = 2,975). Of these, 730 underwent a computed tomography (CT) scan to visualize calcification in the carotid arteries. The second dataset consisted of 768 participants from the RS-I, RS-II and RS-III cohorts of whom 381 also underwent a CT scan. This dataset was the subset of samples previously included in the Biobank-based Integrative Omics Studies Consortium (BIOS Consortium).[20,21] The CT scan was performed on average 4 months (interquartile range (IQR) 2–4 months) after metabolite measuring for the first Rotterdam Study dataset, and 6 years (IQR 5.9–6.2 years) before metabolite measuring for the second Rotterdam Study dataset.

Metabolite quantification

The metabolites were quantified from EDTA plasma samples using high-throughput proton Nuclear Magnetic Resonance (NMR) metabolomics (Nightingale Health, Helsinki, Finland). This method provides simultaneous quantification of metabolic measures, i.e. routine lipids, lipoprotein subclass profiling with lipid concentrations within 14 subclasses, fatty acid composition, and various low-molecular-weight metabolites including amino acids, ketone bodies and gluconeogenesis-related metabolites in molar concentration units. The lipoprotein subclasses include very low-density lipoprotein (VLDL), intermediate-density lipoprotein (IDL), low-density lipoprotein (LDL) and high-density lipoprotein (HDL). In these subclasses, the concentration is measured as well as the subfraction of lipids, triglycerides, cholesterol esters, free cholesterol, and phospholipids. Details of the experimentation and applications of this NMR metabolomics platform have been described previously.[22,23] For this study we analyzed in total 166 non-derived metabolites that were measured across both cohorts.

Assessment of atherosclerosis

A 16-slice (n = 785) or 64-slice (n = 1,739) multidetector CT scanner (Somatom Sensation 16 or 64; Siemens, Forchheim, Germany) was used to perform non-enhanced scanning of intracranial and extracranial carotid arteries to visualize calcification as a proxy of atherosclerosis. Detailed information regarding the protocol and imaging settings is provided elsewhere.[4] ICAC was semi-automatically quantified from the horizontal seg-

ment of the petrous internal carotid artery to the top of the internal carotid artery.[8] Details of this quantification method were described previously.[4] Briefly, regions of interest were drawn in the course of the intracranial internal carotid arteries in consecutive CT sections. Next, calcification volumes were calculated by multiplying the number of pixels in excess of 130 Hounsfield units by the pixel size and the increment.[8] Calcification volumes in the extracranial internal carotid arteries were quantified using dedicated commercially available software (Syngo Calcium Scoring; Siemens).[4] All calcification volumes are expressed in cubic millimeters.

Other measurements

Information on cardiovascular risk factors was obtained by means of interview, physical examination or blood sampling. Hypertension was defined as a systolic blood pressure ≥ 140 mmHg, diastolic blood pressure ≥ 90 mmHg, or use of medication for the treatment of hypertension.[24] Diabetes was defined as fasting plasma glucose levels above 7 mmol/L or use of medication indicated for the treatment of diabetes.[24] Hypercholesterolemia was defined as a total cholesterol ≥ 6.2 mmol/L or use of lipid-lowering medication.[24] BMI was calculated as weight in kilograms divided by square of height in meters. A history of cardiovascular disease was defined as previous myocardial infarction, percutaneous transluminal coronary angioplasty, coronary artery bypass graft or stroke.[8,24]

Statistical analysis

The distributions of metabolic measures were visually inspected for non-normality and were, if necessary, natural logarithmic transformed to obtain approximately normal distributions (**Supplementary Table 1**). The metabolites were scaled to standard deviation (SD) units to enable direct comparisons of effect estimates across the different measures. Because ICAC and ECAC volumes were non-normally distributed, we used natural logarithmic transformed values. To deal with calcium volumes of zero we added 1.0 mm^3 to the non-transformed values. Subsequently, we scaled these new values to SD units to unify reporting and interpretation of the results. To assess the relation of metabolites with ICAC and ECAC per dataset, we performed linear regression analysis while adjusting for age, gender, and lipid-lowering medication (Model 1). The associations were further adjusted for hypertension, diabetes, hypercholesterolemia, smoking, and BMI (Model 2). Finally, we additionally adjusted for history of cardiovascular disease (Model 3). The summary statistic results of the two datasets were meta-analyzed using inverse variance-weighted fixed-effect meta-analysis. Additionally, all analyses were performed in males and females separately.

As metabolic measures are highly correlated (median absolute correlation coefficient = 0.24, IQR = 0.11–0.50), we used the method of Li and Ji[25] to correct for multiple

testing. The method calculates the number of independent variables (and thus tests) in correlated measures. The 166 metabolites corresponded to 33 independent variables. Bonferroni correction was applied for the number of independent variables tested (*p*-value threshold for significance: $0.05/33 = 1.5 \times 10^{-3}$). All analyses were performed with R (R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (http://www.R-project.org/)).

RESULTS

The characteristics of the study population are shown in **Table 1** and the summary statistics of metabolites are shown in **Supplementary Table 1**. Participants from the first dataset of Rotterdam Study (N = 730, 51.2% women, mean age 73.8 ± 5.5 years) were older than participants from the second dataset of Rotterdam Study (N = 381, 53%

Table 1. Descriptive characteristic of study population

	Rotterdam Study dataset 1[a]	**Rotterdam Study dataset 2**[a]
Number of Participants	730	381
Age at CT scan, years	73.8 ± 5.5	64.9 ± 3.2
Women	374 (51.2%)	202 (53.0%)
Diastolic blood pressure, mmHg	79.7 ± 11.4	81.3 ± 10.5
Systolic blood pressure, mmHg	151.2 ± 21.2	142.3 ± 18.0
Hypertension	406 (56.2%)	162 (43.7%)
Glucose, mmol/l	5.8 ± 1.4	5.6 ± 1.2
Participants at CT with diabetes	105 (14.4%)	29 (7.6%)
Total cholesterol, mmol/l	5.6 ± 1.0	5.8 ± 1.0
HDL-Cholesterol, mmol/l	1.4 ± 0.4	1.5 ± 0.4
Hypercholesterolemia	351 (48.5%)	186 (49.7%)
Smoking (never/past/current) (%)	206/403/91 (28.2/55.2/12.5)	110/200/59 (28.9/52.5/15.5)
BMI, kg/m^2	27.3 ± 4	27.8 ± 3.8
Participants at CT with cardiovascular disease	95 (13.1%)	27 (7.1%)
Participants at CT with coronary heart disease	72 (9.9%)	17 (4.5%)
Participants at CT with stroke	32 (4.4%)	11 (2.9%)
ICAC volume, median (IQR), cm^3	64.8 (13.0-205.6)	22.1 (3.8-75.4)
ECAC volume, median (IQR), cm^3	48 (3.1-176.7)	10.4 (0-60.4)

Abbreviations: BMI - body mass index; HDL - high-density lipoprotein; ICAC - intracranial carotid artery calcification; ECAC - extracranial carotid artery calcification; IQR - interquartile range.

[a] Values are means ± standard deviation for continuous variables and number (percentages) for dichotomous variables.

women, mean age 64.9 ± 3.2 years), resulting in differences in age-related clinical characteristics and average volume of calcifications (**Table 1**). However, the prevalence of ICAC was comparable being 83.0% and 80.6% in the first and second group respectively. The prevalence of ECAC was 79.9% in the first and 65.6% in the second dataset.

Table 2. Association of 3-hydroxybutyrate with ICAC volume.

Models	Effect (±SE)[a]	P	N
Model 1 Age, sex and, lipid-lowering medication	0.107(±0.029)	1.76×10^{-4}	1095
Model 2 Model 1 + hypertension, diabetes, hypercholesterolemia, smoking, and BMI	0.092(±0.030)	2.10×10^{-3}	1059
Model 3 Model 2 + history of cardiovascular disease	0.092(±0.030)	2.02×10^{-3}	1054

[a] Effect estimates are SD change in ICAC per 1-SD 3-hydroxybutyrate concentration.

We found a significant association of 3-hydroxybutyrate, a glycolysis-related metabolite, with ICAC (one SD increase in the concentration of 3-hydroxybutyrate was related to a 0.11 SD increase in ICAC; p-value = 1.8×10^{-4}, **Table 2**). The effect estimates were consistent across the two datasets (first Rotterdam Study dataset: effect = 0.10, p-value = 6.5×10^{-3}; second Rotterdam study dataset; effect: 0.13, p-value = 9.06×10^{-3}). Further adjustments for traditional cardiovascular risk factors or history of cardiovascular disease did not influence the effect estimate (**Table 2**). We found no statistically significant association of any of the metabolites with ECAC (**Supplementary Table 2**).

When comparing the metabolic association pattern between ICAC and ECAC we found specific differences (**Fig. 1**, **Supplementary Table 2**). Among the glycolysis-related metabolic measures, 3-hydroxybutyrate which was significantly associated with ICAC, showed nominally significant association with ECAC (effect = 0.07, p-value = 0.02). Glucose was nominally significant associated with both ICAC (effect = 0.07, p-value = 0.01), and ECAC (effect = 0.06, p-value = 0.03), whereas citrate was nominally associated with ECAC (effect = 0.06, p-value = 0.03) (**Supplementary Table 2**). Interestingly, among lipoprotein subfractions, only triglycerides in medium-sized LDL were nominally associated with ICAC (effect = 0.06, p-value = 0.03, **Fig. 1A**, **Supplementary Table 2**), whereas total and free cholesterol and cholesterol esters in extra-large HDL showed nominally significant association with ECAC (**Fig. 1A**, **Supplementary Table 2**). Among the amino-acids, isoleucine was nominally associated with ICAC, and histidine was nominally associated with ECAC. Glycoprotein acetyls were associated with calcification volume in both studied vessel beds.

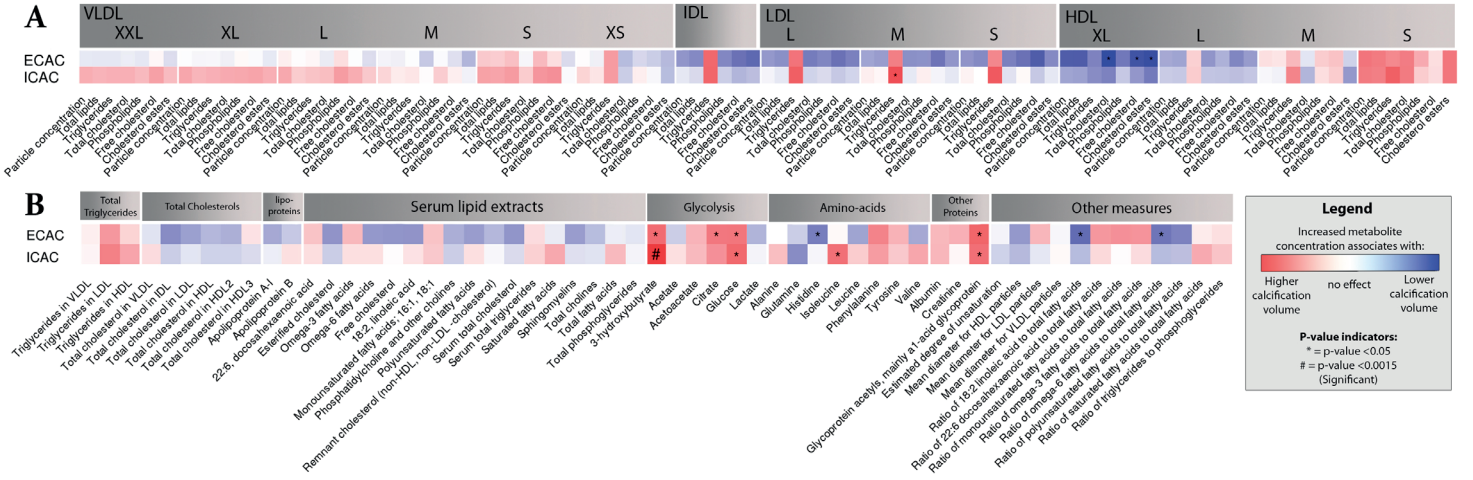


Figure 1. Metabolic association profiles of ICAC and ECAC. The colors represent standardized effect estimates of the metabolites with the calcification volume. The effect estimates of lipoprotein subfractions with calcification in all studied vessel beds are shown in Figure 1A grouped by classes of lipoproteins. The effect estimates of other metabolite measures are shown in Figure 1B grouped by type of metabolic measure. The effect estimates are adjusted for age, gender, and lipid-lowering medication.

When we stratified the analysis by sex the association of 3-hydroxybutyrate with ICAC was nominally significant in both men (effect = 0.13, *p*-value = 2.8×10^{-3}) and women (effect = 0.08, *p*-value = 0.04) (**Supplementary Figs. 1** and **2**, **Supplementary Tables 3** and **4**). Interestingly, the association of glycoprotein acetyls with ICAC and ECAC was mainly driven by men (**Supplementary Figs. 1** and **2**). However, after further adjustments for traditional cardiovascular risk factors or history of cardiovascular disease, glycoprotein acetyls in men were not associated with ICAC and ECAC (*p*-value > 0.05, **Supplementary Table 3**). Other metabolites that were significantly associated with ECAC in men were: the ratio of 18:2 linoleic acid to total fatty acids (effect = −0.17, *p*-value = 4.7×10^{-5}) and the ratio of omega-6 fatty acids to total fatty acids (effect = −0.15, *p*-value = 3.4×10^{-4}). These associations were not modified by the additional adjustments made in model 2 and 3 (**Supplementary Table 3**). No statistically significant associations were observed of metabolites with ECAC in women (**Supplementary Table 4**).

DISCUSSION

In this population-based study, we found that glycolysis-related metabolite measures were associated with a larger volume of ICAC. In particular, higher levels of 3-hydroxybutyrate substantially contributed to larger ICAC volumes. When comparing the metabolic association profile of ICAC with that of ECAC, we found specific differences in glycolysis-related metabolite measures, lipoprotein subfractions, and amino acids.

To our knowledge, this is the first study investigating associations of metabolomics with ICAC and ECAC. The most intriguing finding was the association of 3-hydroxybutyrate (also called beta-hydroxybutyric acid) with ICAC. The ketone 3-hydroxybutyrate is the most abundant of the three known ketone bodies (acetoacetate, 3-hydroxybutyrate, and acetone) that is produced by the liver during fasting and represents an alternative energy source for the brain.[26] In addition, fasting-induced 3-hydroxybutyrate has been found to enhance expression of the glucose transporter Glut1 in brain endothelial cells, which plays an important role in glucose transport across the blood-brain barrier.[27] In general, ketone bodies are considered to exert beneficial effects on brain functioning.[28] In this light, our finding that higher concentrations of 3-hydroxybutyrate relate to larger volumes of ICAC seems to contrast these beneficial effects, especially because ICAC is a risk factor for (subclinical) stroke, cognitive decline, and dementia.[29,30] Yet, a potential mechanism underlying this association may be found in the property of 3-hydroxybutyrate to form polymers known as Poly-(R)-3-hydroxybutyrates (PHB)s. These short-chain PHBs reside in the lipid core of lipoprotein(a) (Lp(a)), a lipoprotein with profound atherogenic effects and also causally related to coronary heart disease.[31-34]

Another explanation for the relation of 3-hydroxybutyrate with ICAC may be impaired glucose tolerance. Impaired glucose tolerance is the (pre-) clinical state of diabetes mellitus type 2 (DM2) and is associated with an elevated risk of and a poor prognosis after cardiovascular events.[35,36] 3-hydroxybutyrate levels were found to be increased in individuals with impaired glucose tolerance and in patients with DM2, in whom it predicted worsening of hyperglycemia and incident DM2 in the next 5 years.[37] These data could hypothetically place 3-hydroxybutyrate in the pathway that leads from an impaired glucose tolerance to increased ICAC and eventually cardiovascular events. Another explanation may be that higher levels of 3-hydroxybutyrate compensate for defective transport of 3-hydroxybutyrate across the blood-brain barrier due to intracranial arteriosclerosis, i.e. reverse causation. Finally, a partial common genetic background might explain the relation between 3-hydroxybutyrate and ICAC.

We also compared the metabolic association patterns of ICAC with that of ECAC. The association between glycoprotein acetyls was observed with both ICAC and ECAC, and glycoprotein acetyls associated with calcifications in men. Attenuation of these associations in model 2, suggests that glycoprotein acetyls might in part reflect pathology related to cardiovascular risk. Levels of this protein are strongly associated with smoking and physical activity and glycoprotein acetyl concentration has been shown to be a strong predictor of 10-year mortality.[38,39] The protein is a marker of acute-phase reactions and may be implicated in this way in depression,[40] diabetes,[41] cardiovascular disease,[42] and cancer[43]. Furthermore, we observed specific differences that further underline the location-specific properties of atherosclerosis.[5,44,45] In addition to differences in glycolysis-related metabolic measures between ICAC and ECAC which are discussed above, another interesting difference we found was that higher concentrations of HDL subfractions were associated with lower volumes of ECAC, but not with ICAC.

The strength of our study includes the large sample with standardized assessments of metabolic measures and arterial calcification in intracranial and extracranial carotid artery, enabling comparisons of the metabolic association patterns of calcification in these two vessel beds. The metabolomics platform that we used contains a large proportion of lipoprotein or other lipid measures which provides an excellent opportunity to study atherosclerosis.[15,22,23,46] However, it should be acknowledged that many other metabolites can be measured using more detailed techniques,[47] which may also be of importance to atherosclerosis. There are also other limitations of our study that should be noted. First, even though calcification is a validated marker of atherosclerosis, non-calcified atherosclerotic disease was not taken into account. Especially, these non-calcified components of the atherosclerotic plaque may also be influenced by the studied metabolites.[48] Another limitation of the current study is that metabolites and CT scan

measures have not been taken at the same time. However, the results were concordant in the two datasets despite the time difference in the metabolites and CT scan measures. Finally, although we adjusted our analyses for various known potential confounders, residual confounding by unknown factors remains possible. We urge future replication efforts of our findings in independent datasets.

CONCLUSIONS

We found a prominent association between 3-hydroxybutyrate and the amount of ICAC. Investigation of the underlying biological mechanisms for the identified association should be the subject of future biological studies. When comparing the metabolic association profile of ICAC with that of ECAC, we found specific differences in glycolysis-related metabolite measures, lipoprotein subfractions, and amino acids, further corroborating the evidence for the existence of location-specific differences in the etiology of carotid artery atherosclerosis.

REFERENCES

1. Gorelick, P.B., Wong, K.S., Bae, H.J. & Pandey, D.K. Large artery intracranial occlusive disease: a large worldwide burden but a relatively neglected frontier. *Stroke* **39**, 2396-9 (2008).
2. Arenillas, J.F. Intracranial atherosclerosis: current concepts. *Stroke* **42**, S20-3 (2011).
3. Qureshi, A.I. & Caplan, L.R. Intracranial atherosclerosis. *Lancet* **383**, 984-98 (2014).
4. Bos, D. *et al.* Intracranial carotid artery atherosclerosis and the risk of stroke in whites: the Rotterdam Study. *JAMA Neurol* **71**, 405-11 (2014).
5. Bos, D. *et al.* Genetic loci for coronary calcification and serum lipids relate to aortic and carotid calcification. *Circ Cardiovasc Genet* **6**, 47-53 (2013).
6. Lopez-Cancio, E. *et al.* Biological signatures of asymptomatic extra- and intracranial atherosclerosis: the Barcelona-AsIA (Asymptomatic Intracranial Atherosclerosis) study. *Stroke* **43**, 2712-9 (2012).
7. Allison, M.A., Criqui, M.H. & Wright, C.M. Patterns and risk factors for systemic calcified atherosclerosis. *Arterioscler Thromb Vasc Biol* **24**, 331-6 (2004).
8. Bos, D. *et al.* Intracranial carotid artery atherosclerosis: prevalence and risk factors in the general population. *Stroke* **43**, 1878-84 (2012).
9. Mazighi, M. *et al.* Autopsy prevalence of intracranial atherosclerosis in patients with fatal stroke. *Stroke* **39**, 1142-7 (2008).
10. Quehenberger, O. & Dennis, E.A. The human plasma lipidome. *N Engl J Med* **365**, 1812-23 (2011).
11. Inouye, M. *et al.* Metabonomic, transcriptomic, and genomic variation of a population cohort. *Mol Syst Biol* **6**, 441 (2010).
12. Wang, T.J. *et al.* Metabolite profiles and the risk of developing diabetes. *Nat Med* **17**, 448-53 (2011).
13. Mahendran, Y. *et al.* Glycerol and fatty acids in serum predict the development of hyperglycemia and type 2 diabetes in Finnish men. *Diabetes Care* **36**, 3732-8 (2013).
14. Shah, S.H., Kraus, W.E. & Newgard, C.B. Metabolomic profiling for the identification of novel biomarkers and mechanisms related to common cardiovascular diseases: form and function. *Circulation* **126**, 1110-20 (2012).
15. Wurtz, P. *et al.* High-throughput quantification of circulating metabolites improves prediction of subclinical atherosclerosis. *Eur Heart J* **33**, 2307-16 (2012).
16. Stegemann, C. *et al.* Lipidomics profiling and risk of cardiovascular disease in the prospective population-based Bruneck study. *Circulation* **129**, 1821-31 (2014).
17. Roberts, L.D. & Gerszten, R.E. Toward new biomarkers of cardiometabolic diseases. *Cell Metab* **18**, 43-50 (2013).
18. Ammirati, E and Fogacci, F, Clinical relevance of biomarkers for the identification of patients with carotid atherosclerotic plaque: Potential role and limitations of cysteine protease legumain. *Atherosclerosis* **257**:248-249 (2017).
19. Ammirati, E, Moroni, F, Norata, GD, *et al.* Markers of inflammation associated with plaque progression and instability in patients with carotid atherosclerosis, *Mediators Inflamm* **718329** (2015).
20. Hofman, A. *et al.* The Rotterdam Study: 2016 objectives and design update. *Eur J Epidemiol* **30**, 661-708 (2015).
21. Huan, T. *et al.* A meta-analysis of gene expression signatures of blood pressure and hypertension. *PLoS Genet* **11**, e1005035 (2015).
22. Soininen, P. *et al.* High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. *Analyst* **134**, 1781-5 (2009).

23. Soininen, P., Kangas, A.J., Wurtz, P., Suna, T. & Ala-Korpela, M. Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circ Cardiovasc Genet* **8**, 192-206 (2015).
24. Odink, A.E. *et al.* Risk factors for coronary, aortic arch and carotid calcification; The Rotterdam Study. *J Hum Hypertens* **24**, 86-92 (2010).
25. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* **95**, 221-227 (2005).
26. Owen, O.E. *et al.* Brain metabolism during fasting. *J Clin Invest* **46**, 1589-95 (1967).
27. Tanegashima, K. *et al.* Epigenetic regulation of the glucose transporter gene Slc2a1 by -hydroxybutyrate underlies preferential glucose supply to the brain of fasted mice. *Genes to Cells* **22**, 71-83 (2017).
28. Rahman, M. *et al.* The beta-hydroxybutyrate receptor HCA2 activates a neuroprotective subset of macrophages. *Nat Commun* **5**, 3944 (2014).
29. Bos, D. *et al.* Intracranial Carotid Artery Atherosclerosis and the Risk of Stroke in Whites The Rotterdam Study. *Jama Neurology* **71**, 405-411 (2014).
30. Bos, D. *et al.* Atherosclerotic calcification is related to a higher risk of dementia and cognitive decline. *Alzheimers & Dementia* **11**, 639-647 (2015).
31. Kettunen, J. *et al.* Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat Commun* **7**, 11122 (2016).
32. Tregouet, D.A. *et al.* Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. *Nature Genetics* **41**, 283-285 (2009).
33. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics* **43**, 333-U153 (2011).
34. Reusch, R.N. Poly-(R)-3-hydroxybutyrates (PHB) are atherogenic components of lipoprotein Lp(a). *Medical Hypotheses* **85**, 1041-1043 (2015).
35. Barr, E.L.M. *et al.* Risk of cardiovascular and all-cause mortality in individuals with diabetes mellitus, impaired fasting glucose, and impaired glucose tolerance - The Australian diabetes, obesity, and lifestyle study (AusDiab). *Circulation* **116**, 151-157 (2007).
36. Kurihara, O. *et al.* Impact of Prediabetic Status on Coronary Atherosclerosis A multivessel angioscopic study. *Diabetes Care* **36**, 729-733 (2013).
37. Mahendran, Y. *et al.* Association of Ketone Body Levels With Hyperglycemia and Type 2 Diabetes in 9,398 Finnish Men. *Diabetes* **62**, 3618-3626 (2013).
38. Fischer, K. *et al.* Biomarker profiling by nuclear magnetic resonance spectroscopy for the prediction of all-cause mortality: an observational study of 17,345 persons. *PLoS Med* **11**, e1001606 (2014).
39. Singh-Manoux, A. *et al.* Association between inflammatory biomarkers and all-cause, cardiovascular and cancer-related mortality. *CMAJ* (2016).
40. Harley, J. *et al.* Orosomucoid influences the response to antidepressants in major depressive disorder. *J Psychopharmacol* **24**, 531-5 (2010).
41. El-Beblawy, N.M.S. *et al.* Serum and Urinary Orosomucoid in Young Patients With Type 1 Diabetes: A Link Between Inflammation, Microvascular Complications, and Subclinical Atherosclerosis. *Clinical and Applied Thrombosis-Hemostasis* **22**, 718-726 (2016).
42. Carriere, I. *et al.* Biomarkers of inflammation and malnutrition associated with early death in healthy elderly people. *Journal of the American Geriatrics Society* **56**, 840-846 (2008).

43. Bruno, R. *et al.* alpha-1-acid glycoprotein as an independent predictor for treatment effects and a prognostic factor of survival in patients with non-small cell lung cancer treated with docetaxel. *Clinical Cancer Research* **9**, 1077-1082 (2003).
44. Bos, D. *et al.* Calcification in Major Vessel Beds Relates to Vascular Brain Disease. *Arteriosclerosis Thrombosis and Vascular Biology* **31**, 2331-2337 (2011).
45. Bos, D. *et al.* Comparison of Atherosclerotic Calcification in Major Vessel Beds on the Risk of All-Cause and Cause-Specific Mortality: The Rotterdam Study. *Circ Cardiovasc Imaging* **8** (2015).
46. Wurtz, P. *et al.* Metabolite profiling and cardiovascular event risk: a prospective study of 3 population-based cohorts. *Circulation* **131**, 774-85 (2015).
47. Wishart, D.S. *et al.* HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic Acids Res* **41**, D801-7 (2013).
48. Stary, H.C. *et al.* A definition of advanced types of atherosclerotic lesions and a histological classification of atherosclerosis. A report from the Committee on Vascular Lesions of the Council on Arteriosclerosis, American Heart Association. *Arterioscler Thromb Vasc Biol* **15**, 1512-31 (1995).

Chapter 3.3

Circulating metabolites and risk of stroke in seven population-based cohorts

Dina Vojinovic, Marita Kalaoja, Stella Trompet, Krista Fischer, Martin J. Shipley, Shuo Li, Aki S. Havulinna, Markus Perola, Veikko Salomaa, Qiong Yang, Naveed Sattar, Pekka Jousilahti, Najaf Amin, Ramachandran S. Vasan, M. Arfan Ikram, Mika Ala-Korpela, J. Wouter Jukema, Sudha Seshadri, Johannes Kettunen, Mika Kivimaki, Tonu Esko, Cornelia M. van Duijn

This chapter is in preparation.

The supplemental information for this paper is available at https://drive.google.com/file/d/1XXBT5NWkjSvq_MDHaWDlh_mXe6MqtqgV/view?usp=sharing

ABSTRACT

Stroke is a leading cause of death and long-term disability worldwide. Detailed profiling of metabolic status can provide insights into metabolic changes and lead to identification of individuals with higher risk of stroke. We investigated association of a wide range of metabolites with risk of stroke in seven prospective population-based cohorts including 1,791 incident stroke events among 38,797 participants. The analyses were performed considering all incident stroke events and ischemic and hemorrhagic events separately. The analysis revealed ten significant metabolite associations. Amino acid histidine (hazard ratio (HR) per SD= 0.9, 95% confidence interval (CI): 0.85, 0.94), glycolysis-related metabolite pyruvate (HR per SD = 1.09, 95% CI: 1.04, 1.14), acute phase reaction marker glycoprotein acetyls (HR per SD = 1.09, 95% CI: 1.03, 1.15), cholesterol in high-density lipoprotein (HDL) 2 and several other lipoprotein particles including cholesterol in medium HDL and triglycerides in medium and large low-density lipoprotein (LDL) particles were associated with risk of stroke. When focusing on incident ischemic stroke, a significant association was observed with phenylalanine (HR per SD = 1.12, 95% CI: 1.05, 1.19) and total and free cholesterol in large HDL particles. When comparing our findings to those of a study in the China Kadoorie Biobank, glycoprotein acetyls are replicated both in Caucasians and Chinese. However, we also observed very significant associations that were specific for Western societies. To conclude, we found association of amino acids, glycolysis-related metabolites, acute phase reaction markers, and several lipoprotein subfractions with the risk of stroke. The biological mechanisms underlying these associations should be subject of further studies.

INTRODUCTION

Stroke is a leading cause of death and serious long-term disability worldwide.[1] The majority of strokes are of the ischemic type, while the hemorrhagic type occurs less often but is associated with a higher mortality risk.[1,2] Stroke risk is determined by various modifiable risk factors such as hypertension, diabetes mellitus, cardiovascular disease, smoking, and obesity, whereas association of stroke with cholesterol and its subfractions has shown inconsistent results.[1-6] Opportunities for therapeutic interventions in stroke patients depend on the type of stroke and relay on brain imaging techniques.[7] Despite advances in brain imaging techniques, costs are still high, availability is limited and not all patients show a relevant lesion on neuroimaging.[7,8] New technology is needed to identify high-risk patients, to understand the etiology of stroke and develop future prevention strategies. Detailed profiling of metabolic status can provide insights into metabolic changes that lead to a higher risk of stroke. As the metabolome reflects both genome and exposome including exposures to risk factors that determine the risk of stroke, this new –omics technology may open new avenues towards stroke prevention. To date, only few studies have analyzed metabolic disturbances in stroke and identified various metabolites to be associated with stroke.[9-11] However, these studies are based on a relatively small sample or on participants of non-European ancestry.[12] The most comprehensive study to date was conducted by Holmes *et al.* within the China Kadoorie Biobank including patients with ischemic stroke (N = 1,146) and intracerebral hemorrhage (N = 1,138).[12] The study reported an association between lipids and lipoprotein particles of various sizes with ischemic stroke but not with hemorrhage.[12] Furthermore, the study identified glycoprotein acetyls, ketone bodies, glucose and docosahexaenoic acid to be associated with both ischemic and hemorrhagic stroke.[12]

As large metabolomic studies of stroke in persons of European origin are lacking, the aim of our study is to conduct a comprehensive analysis of circulating metabolites and incident stroke in large prospective population-based setting including 1,791 incident stroke events among 38,797 participants of European origin.

METHODS

Study population

Our study population included 38,797 participants from seven cohorts including Rotterdam Study, the Whitehall II study (Whitehall II), the national FINRISK studies (FINRISK97 and FINRISK07), PROspective Study of Pravastatin in the Elderly at Risk (PROSPER), Estonian biobank (EGCUT), and Framingham Heart Study (FHS). Description of participating

studies is provided in the **Supplementary Note 1**. Each of the participating studies was approved by local ethical committees or institutional review boards. All participants provided written informed consent.

Stroke assessment

Details on stroke assessment are provided in the **Supplementary Note 2.** The incident stroke events were assessed through follow-up of health records, while in some studies additional periodic visits to research centers were used (e.g. Rotterdam Study, FHS). Participants of the Rotterdam Study were monitored for incident stroke using automated linkage of medical records from general practitioners with the study database.[13] Incident stroke events in the Whitehall II study were ascertained through linkage to electronic records from hospitalizations due to stroke and national statistics death registries,[14,15] whereas in the FINRISK cohorts linkage to national health registries was used (https://www.biorxiv.org/content/early/2018/03/12/280677). Ascertainment of incident stroke events in EGCUT was also performed through linkage to electronic records from multiple databases (https://thl.fi/publications/morgam/cohorts/ full/estonia/est-esta.htm), while information regarding domiciliary visits or hospitalizations associated with possible cardiovascular events including stroke, and information on all deaths was used for classification of study endpoints in PROSPER.[16] In the FHS, incident clinical stroke was identified as part of ongoing clinic and hospital surveillance, and additional stroke surveillance by annual phone health updates and collaboration with primary care physicians and local emergency departments.[17,18] Participants with a history of stroke at baseline were excluded from the analyses.

Other measurements

The baseline measurements included measures of blood pressure, plasma glucose levels, weight, and height. Hypertension was defined as a systolic blood pressure ≥ 140 mmHg, diastolic blood pressure ≥ 90 mmHg, or use of antihypertensive medication. Diabetes was defined as fasting plasma glucose levels above 7 mmol/L or use of medication indicated for the treatment of diabetes. Body mass index (BMI) was calculated as weight in kilograms divided by square of heights in meters.

Metabolite quantification

Circulating metabolites were quantified using a high-throughput Nuclear Magnetic Resonance (^{1}H-NMR) technology. In all participating studies except the FHS, the Nightingale Health metabolomics platform (Helsinki, Finland) was used for simultaneous quantification of a wide range of metabolites, including routine lipids, 14 lipoprotein subclasses and their lipids (esterified cholesterol, free cholesterol, total cholesterol, triglycerides, phospholipids, and total lipids), fatty acids, amino acids, ketone bodies,

and various glycolysis precursors. A detailed description of the methodology has been provided previously.[19,20] In the FHS, lipoprotein subclasses were measured by proton NMR spectroscopic assay (LipoScience, Raleigh, NC).[21,22] Blood samples were collected after overnight fasting in all studied except for FINRISK97 and FINRISK07 in which the samples were collected after 4 hours of fasting (semi-fasting state).[23,24] The sample material was EDTA-plasma in the Rotterdam Study, FHS, and EGCUT, whereas the serum was used in FINRISK, PROSPER and Whitehall II.[23-26] There were 147 primary non-derived metabolite measurements quantified in absolute concentration units that were further analyzed in this study.

Statistical analyses

To obtain an approximately normal distribution, all metabolites were natural logarithmic transformed prior the analyses. To deal with zero values, one was added to all values of the metabolites prior to the transformation. The metabolite measurements were subsequently scaled to standard deviation (SD) units. The relationship between metabolites and stroke was assessed by Cox proportional hazards regression models. The analyses were performed while adjusting for age, gender, BMI, lipid-lowering medication, and study-specific covariates if needed (Model 1). The associations were further adjusted for smoking status, diabetes, and hypertension (Model 2). The summary statistics results of participating studies were combined using inverse variance-weighted fixed-effect meta-analysis. The analyses were performed considering all incident stroke events and ischemic and hemorrhagic events separately.

As metabolite measures are highly correlated, we calculated the number of independent tests using the previously described method of Li and Ji.[27] Subsequently, the number of independent tests was used for calculation of Bonferroni corrected *p*-value (0.05/30 independent metabolites = 1.7×10^{-3}).

To determine the discrimination power of metabolite measures discovered in our study, we used the Rotterdam Study to calculate the area under the receiver-operating characteristic curve (AUC). We also determined the discrimination of metabolite measures discovered in the China Kadoorie Biobank and furthermore, their discrimination power when combined together with the metabolites discovered in our study and Framingham Stroke Risk Score. The analyses were performed in R version 3.2.5 (http://www.R-project.org/).

RESULTS

The baseline descriptive characteristics of study participants are shown in **Table 1**. In total there were 1,791 incident stroke events observed among 38,797 participants across the seven cohorts. The mean follow-up time ranged from 2 years in PROSPER, 6 years in the Rotterdam Study, and 7 years in EGCUT and FINRISK07 to 13 years in Whitehall II and 15 years in FINRISK97 and FHS.

The results of association analysis between circulating metabolites and incident stroke are shown in **Table 2**. The analysis revealed 27 significant metabolite associations in model 1. After further adjustment for hypertension status, diabetes and smoking, 7 metabolite associations survived correction for multiple testing. These included the amino

Table 1. Descriptive statistics of study population.

	Rotterdam Study		Whitehall II**		Finrisk97	
Variable*	Incident cases	Controls	Incident cases	Controls	Incident cases	Controls
N	257	2308	197	5792	474	6384
Age (years)	76.9 (6.2)	75.0 (6.1)	59.4 (5.9)	55.6 (6)	59.6 (10.4)	47.0 (12.9)
Women	54.1%	58%	25.4%	29.1%	38.6%	52.6%
Current Smoking	15.2%	13%	15.70%	9.40%	24.5%	23.7%
Diabetes	17.9%	14.3%	8.6%	4.4%	16.9%	4.9%
Hypertension	85.6%	81.0%	41.1%	28.0%	48.9%	21%
Systolic blood pressure (mmHg)	156.8 (23.9)	151.4 (20.1)	127.4 (16.3)	122.9 (16.5)	147.7 (22.3)	134.7 (19.2)
Diastolic blood pressure (mmHg)	79.7 (12.4)	79.2 (11.1)	78.3 (10.2)	77.5 (10.5)	86.1 (11.9)	81.9 (11.2)
Antihypertensive medication	51%	47.10%	21.8%	12.3%	27.4%	11.5%
BMI (kg/m2)	27.2 (3.5)	27.4 (4.2)	26.2 (4.2)	26 (3.9)	28.4 (4.8)	26.5 (4.5)
Follow-up time (years)	5.7 (3.5)	9.8 (3.5)	12.5 (4.9)	18.2 (3)	15.03 (4.2)	16.9 (3)
Total cholesterol	5.5 (0.98)	5.6 (0.98)	5.8 (1.1)	5.9 (1.1)	5.79 (1.1)	5.52 (1.1)
HDL cholesterol	1.4 (0.4)	1.5 (0.4)	1.5 (0.4)	1.5 (0.4)	1.29 (0.3)	1.41 (0.4)
LDL cholesterol	NA	NA	3.8 (1.01)	3.9 (0.9)	3.71 (0.9)	3.46 (0.9)
Triglycerides	NA	NA	1.3 (0.8)	1.4 (0.9)	1.79 (1.1)	1.46 (1.0)
Lipid lowering medication	21.4%	20.6%	5.1%	3.0%	7.8%	3.1%
Coronary Heart Disease	13.6%	10.8%	11.7%	5.9%	9.1%	1.9%
Stroke						
Hemorrhagic	32 (12.5%)	-	48 (24.4%)	-	69 (14.6%)	-
Ischemic	183 (71.2%)	-	126 (64.0%)	-	405 (85.4%)	-
Not defined	42 (16.3%)	-	23 (11.6%)	-	-	-

* Values are means ± standard deviation for continuous variables and percentages for dichotomous variables.
** While all other cohorts included participants of European ancestry, 87.3% of Whitehall II study cases were of European ancestry, 6.6% of Asian, 5.1% of African American and 1% of other.
***Lipid levels are expressed in mmol/l for all cohorts except for FHS (mg/dl).

acid histidine (hazard ratio (HR) per SD = 0.9, 95 % confidence interval (CI): 0.85, 0.94) and cholesterol in high-density lipoprotein (HDL) 2 (HR per SD = 0.91, 95% CI: 0.87, 0.97) which were associated with a decreased risk of stroke, glycolysis-related metabolite pyruvate (HR per SD = 1.1, 95% CI: 1.04, 1.14) and acute phase reaction markers glycoprotein acetyls (HR per SD = 1.09, 95% CI: 1.03, 1.15) which were associated with an increased risk of stroke, and several lipoprotein particles including HDL and low-density lipoprotein (LDL subfractions) (**Table 2, Figure 1**). Cholesterol in medium HDL was associated with decreased risk (HR per SD = 0.92, 95% CI: 0.87, 0.97), whereas triglycerides in medium and large LDL were associated with an increased risk of stroke (HR per SD = 1.09, 95% CI: 1.03, 1.14 and HR per SD = 1.09, 95% CI: 1.03, 1.14, respectively) (**Table 2, Figure 1**). The direction of effect across the cohorts showed no evidence of single cohort driving the associations (**Figure 2**). Whereas the Whitehall II study showed opposite

Finrisk07		PROSPER		EGCUT		FHS***	
Incident cases	Controls	Incident cases	Controls	Incident cases	Controls	Incident cases	Controls
107	4424	197	4627	308	10268	251	3203
62.0 (10.4)	51.9 (13.5)	75.9 (3.7)	75.2 (3.3)	66.3 (12.5)	44.5 (17.1)	58.1 (8.98)	51.7 (10.1)
42.1%	53.7%	54%	52.2%	54.9%	63.3%	47.4%	51.4%
21.5%	17.4%	28.90%	27.10%	17.5%	29.9%	24.9%	24.6%
15.9%	8.9%	18.3%	10.6%	35.4%	7.7%	15.9%	5.0%
43%	16.5%	58.9%	62.5%	66.2%	24.4%	60.2%	34.2%
149.8 (23.7)	136.4 (20.2)	157.1 (21.9)	154.5 (21.8)	142.8 (18.8)	125.7 (16.9)	137.8 (20.8)	126.2 (18.5)
83.1 (13.6)	79.3 (11.0)	84.6 (11.8)	83.7 (11.4)	83.4 (10.9)	77.6 (10.7)	81.6 (10.4)	78.9 (9.9)
34.6%	22.1%	70.6%	74.4%	69.5%	24.3%	36.3%	16.4%
28.0 (5)	27.16 (4.8)	26.5 (4.1)	26.9 (4.2)	29.1 (5.7)	26.4 (5.4)	27.6 (5.1)	26.7 (4.8)
7.25 (1.5)	7.75 (0.7)	1.9 (1.0)	3.3 (0.5)	6.9 (3.1)	8.9 (1.8)	14.7 (7)	22.4 (6.0)
5.23 (0.96)	5.28 (1)	5.64 (0.85)	5.68 (0.90)	6.0 (1.2)	5.7 (1.2)	215.4 (42.0)	205.1 (38.5)
1.44 (0.4)	1.44 (0.4)	1.25 (0.32)	1.28 (0.35)	1.5 (0.4)	1.6 (0.5)	47.4 (15.5)	49.7 (14.9)
3.11 (0.8)	3.20 (0.9)	3.77 (0.76)	3.79 (0.80)	2.5 (0.7)	2.3 (0.6)	138.2 (36.8)	131.4 (35.1)
1.49 (0.8)	1.42 (0.9)	1.57 (0.70)	1.54 (0.69)	1.9 (1.0)	1.6 (0.9)	150.9 (104.7)	123.5 (101.8)
25.2%	14.7%	52.3%	49.5%	13.6%	4.7%	6.0%	3.7%
5.6%	2.9%	16.8%	13.1%	35.1%	9.2%	12.4%	5.7%
23 (21.5%)	-	-	-	45 (14.6%)	-	30 (12%)	-
84 (78.5%)	-	-	-	261 (84.7%)	-	219 (87.3%)	-
-	-	-	-	11 (3.6%)	-	2 (0.8%)	-

direction of effect for apolipoprotein A, HDL, and HDL2 cholesterol, the findings showed a general spread for most HDL subfractions.

Table 2. Results of association analysis between incident stroke and metabolites.

	Model 1					Model 2				
Metabolite	**N**	**Ncases**	**HR**	**CI**	**P**	**N**	**Ncases**	**HR**	**CI**	**P**
Phenylalanine	35091	1527	1.11	1.06;1.17	4.88E-05	35036	1524	1.08	1.03;1.14	3.36E-03
Histidine*	35017	1526	0.89	0.84;0.93	7.94E-06	34962	1523	0.9	0.85;0.94	4.45E-05
plasma-ApoA1	35107	1529	0.91	0.86;0.96	7.14E-04	35052	1526	0.94	0.88;0.99	1.79E-02
HDL-cholesterol	35107	1529	0.89	0.84;0.94	2.89E-05	35052	1526	0.92	0.87;0.97	3.20E-03
HDL2-cholesterol*	35107	1529	0.88	0.84;0.93	9.13E-06	35052	1526	0.91	0.87;0.97	1.41E-03
IDL-triglycerides	38561	1780	1.1	1.05;1.16	6.06E-05	38494	1775	1.07	1.02;1.12	9.91E-03
LDL-triglycerides	35107	1529	1.12	1.06;1.18	3.93E-05	35052	1526	1.08	1.03;1.14	2.47E-03
Glucose	34980	1524	1.15	1.1;1.2	7.81E-11	34925	1521	1.06	1.01;1.11	1.87E-02
Lactate	35100	1529	1.12	1.07;1.18	1.11E-05	35045	1526	1.08	1.02;1.13	5.09E-03
Pyruvate*	24423	1205	1.13	1.08;1.18	1.37E-07	24368	1202	1.09	1.04;1.14	7.45E-04
Glycoprotein acetyls*	35101	1529	1.15	1.09;1.21	1.25E-07	35046	1526	1.09	1.03;1.15	1.27E-03
HDL-diametar	35107	1529	0.89	0.84;0.94	3.05E-05	35052	1526	0.92	0.87;0.98	6.73E-03
S-HDL-triglycerides	35108	1529	1.11	1.06;1.17	6.80E-05	35053	1526	1.07	1.01;1.12	1.97E-02
M-HDL-cholesterol*	38560	1780	0.89	0.85;0.94	2.07E-05	38493	1775	0.92	0.87;0.97	1.35E-03
M-HDL-cholesterol esters	35106	1529	0.9	0.85;0.95	2.05E-04	35051	1526	0.92	0.87;0.97	3.73E-03
M-HDL-free cholesterol	35106	1529	0.91	0.86;0.96	7.33E-04	35051	1526	0.93	0.88;0.98	8.24E-03
L-HDL-cholesterol	38555	1780	0.89	0.84;0.94	2.13E-05	38488	1775	0.92	0.88;0.98	5.50E-03
L-HDL-cholesterol esters	35101	1529	0.9	0.84;0.95	2.03E-04	35046	1526	0.93	0.88;0.99	1.37E-02
L-HDL-free cholesterol	35101	1529	0.89	0.84;0.94	1.25E-04	35046	1526	0.92	0.87;0.98	9.96E-03
L-HDL-total lipids	35101	1529	0.9	0.85;0.95	2.12E-04	35046	1526	0.93	0.88;0.99	1.70E-02
L-HDL-phospholipids	35101	1529	0.9	0.85;0.96	6.29E-04	35046	1526	0.94	0.89;1	3.49E-02
L-HDL concentration	35101	1529	0.9	0.85;0.96	8.53E-04	35046	1526	0.94	0.89;1	4.21E-02
XL-HDL-free cholesterol	35099	1527	0.91	0.86;0.96	8.31E-04	35044	1524	0.94	0.89;1	3.55E-02
S-LDL-triglycerides	29120	1332	1.12	1.06;1.18	2.95E-05	29065	1329	1.09	1.03;1.15	2.81E-03
L-LDL-triglycerides*	35107	1529	1.12	1.06;1.17	3.00E-05	35052	1526	1.09	1.03;1.14	1.67E-03
M-LDL-triglycerides*	35106	1529	1.12	1.06;1.18	1.68E-05	35051	1526	1.09	1.03;1.14	1.19E-03
XL-VLDL-triglycerides	38284	1769	1.09	1.04;1.14	1.56E-04	38217	1764	1.05	1;1.1	4.66E-02

Abbreviations: N - Total samples size; Ncases - Number of cases; HR - Hazard Ratio; 95% CI - 95% confidence interval; P - *p*-value; Model 1 - adjustment for age, gender, BMI, lipid-lowering medication and study-specific covariates if needed; Model 2- additional adjustment for smoking status, diabetes, and hypertension; *Associations that surpassed significance threshold in model 2.

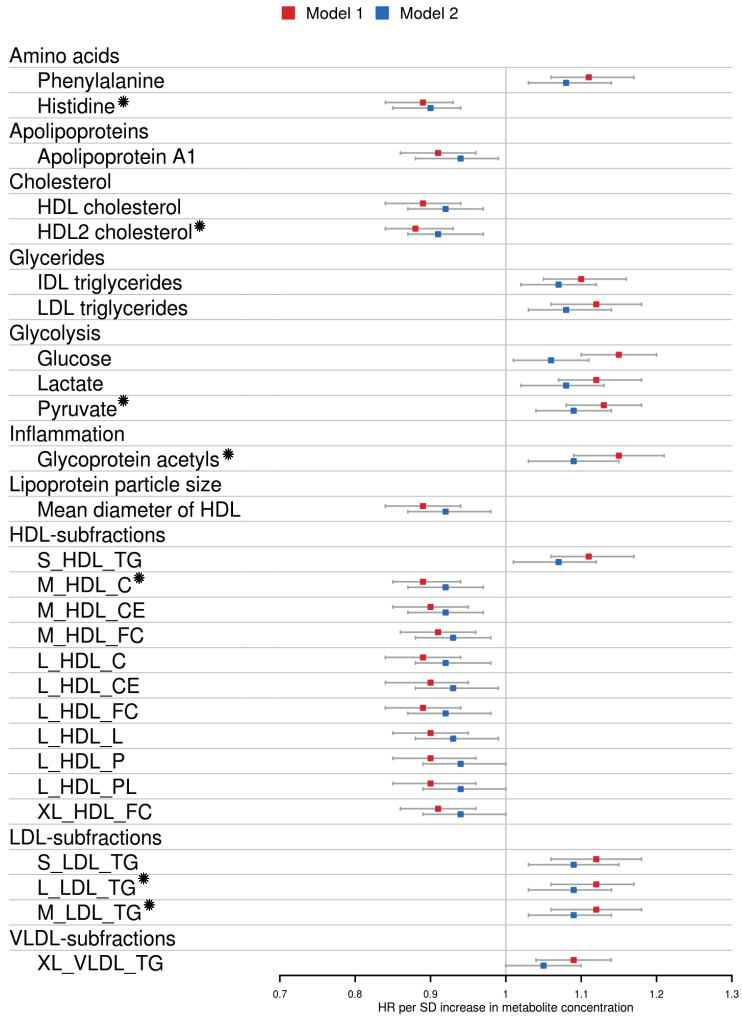


Figure 1. The results of association analysis between incident stroke and metabolites across two different models. Only associations that surpassed significance threshold are illustrated. The results are shown in red for model 1 and in blue for model 2. Hazard Ratios (HR) are denoted with boxes, while corresponding 95% confidence intervals of effect estimates are represented with whiskers. The associations that remained significant in model 2 are denoted by *.

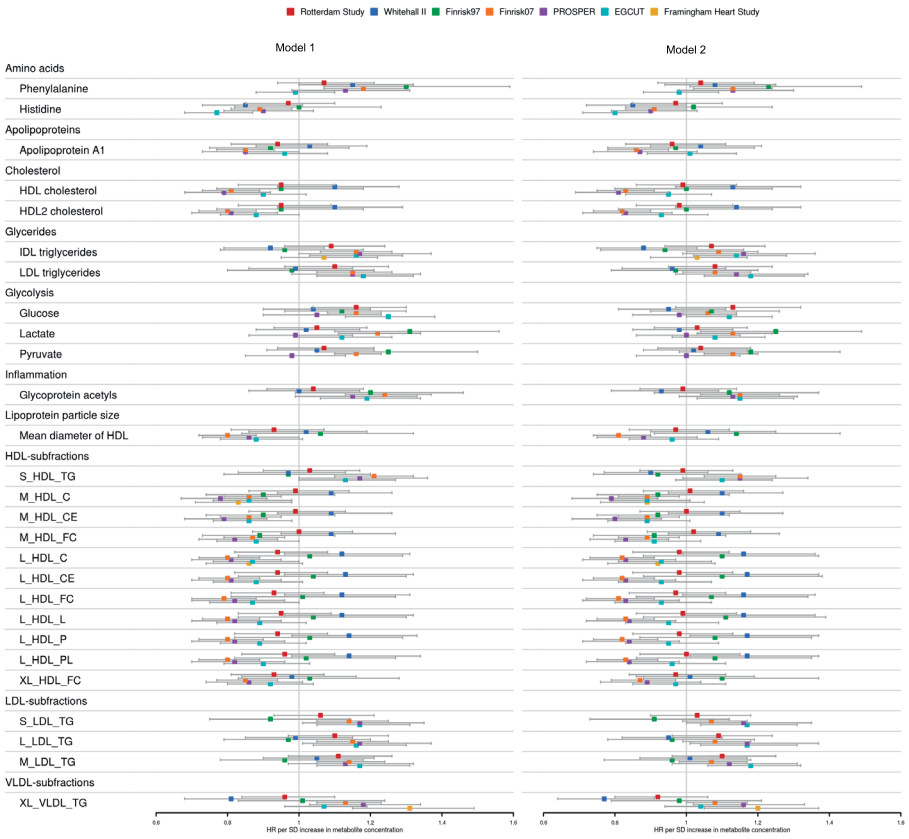


Figure 2. Associations that surpassed the threshold for multiple testing across two different models per study. The results for Rotterdam Study are shown in red, for Whitehall II in blue, for FINRISK97 in green, for FINRISK07 in orange, for PROSPER in purple, for EGCUT in turquoise and for FHS in golden. Hazard Ratios (HR) are denoted with boxes, while corresponding 95% confidence intervals of effect estimates are represented with whiskers.

When we stratified the analysis by stroke type, we observed differences between ischemic and hemorrhagic stroke events (**Table 3**). Amino acid histidine and cholesterol in HDL2 were associated with decreased risk of ischemic but not hemorrhagic incident stroke (**Table 3**). The differences were also observed for glycolysis-related metabolite pyruvate and acute phase reaction markers glycoprotein acetyls which were associated with increased risk of ischemic but not hemorrhagic stroke (**Table 3)**. Association between incident stroke events and LDL and HDL particles of various sizes was observed only in the overall analysis, suggesting contribution of both stroke subtypes (**Table 3**).

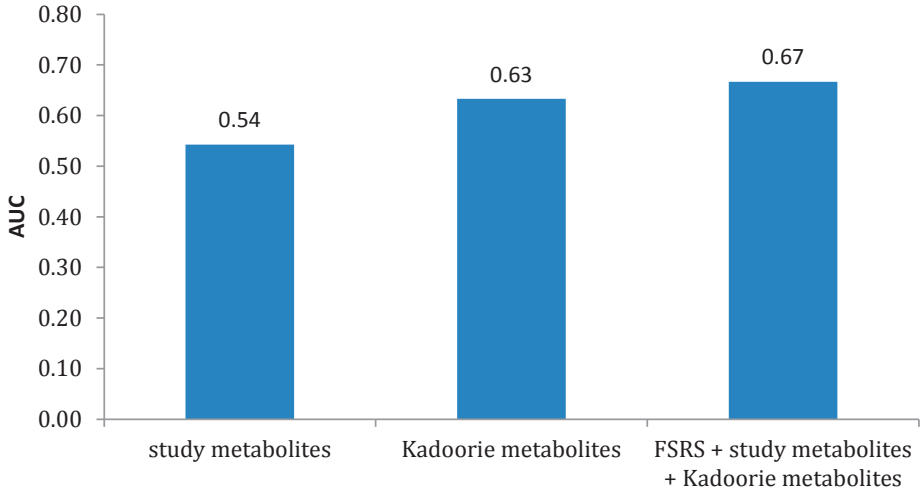


Figure 3. The area under the receiver-operating characteristic curve (AUC) in different prediction models. Study metabolites refer to circulating metabolites discovered in our study, while Kadoorie metabolites refer to metabolites discovered in the China Kadoorie Biobank. FSRS stands for Framingham Stroke Risk Score.

Furthermore, a significant association was observed between phenylalanine levels and increased risk of incident ischemic stroke (HR per SD = 1.12, 95% CI: 1.05, 1.19) and decreased risk of ischemic stroke and level of cholesterol in large HDL (HR per SD = 0.89, 95% CI: 0.84, 0.95) and of free cholesterol in the same particles (HR per SD = 0.89, 95% CI: 0.82, 0.95). There was no metabolite that surpassed the significant threshold in the analysis for hemorrhagic stroke.

The metabolites discovered in our study discriminate future stroke with the AUC of 0.54 (95% CI: 0.50, 0.58) (**Figure 3**). When analyses were repeated using the metabolites discovered in the China Kadoorie Biobank, the AUC was 0.63 (95% CI: 0.60, 0.67). Combining the metabolites discovered in our study and China Kadoorie Biobank, which is using the same metabolomics platform in Chinese population, together with Framingham Stroke Risk Score, lead to further improvement of the discrimination of future patients (AUC: 0.67, 95% CI: 0.63, 0.70) (**Figure 3**).

Table 3. Significant associations for overall incident stroke events and when classified by stroke type.

		Model 1					Model 2				
Metabolite	**Type**	N	Ncases	HR	CI	**P**	**N**	**Ncases**	**HR**	**CI**	**P**
Phenylalanine	all	35091	1527	1.11	1.06;1.17	**4.88E-05**	35036	1524	1.08	1.03;1.14	3.36E-03
	hemorrhagic	30144	214	0.94	0.81;1.09	3.90E-01	30092	214	0.91	0.78;1.05	2.00E-01
	ischemic*	30290	1051	1.16	1.09;1.23	**3.15E-06**	30236	1049	1.12	1.05;1.19	**4.13E-04**
Histidine	all*	35017	1526	0.89	0.84;0.93	**7.94E-06**	34962	1523	0.9	0.85;0.94	**4.45E-05**
	hemorrhagic	30070	214	0.95	0.82;1.1	4.78E-01	30018	214	0.96	0.83;1.11	6.16E-01
	ischemic*	30216	1050	0.88	0.82;0.94	**1.18E-04**	30162	1048	0.89	0.84;0.95	**4.94E-04**
Apolipoprotein A1	all	35107	1529	0.91	0.86;0.96	**7.14E-04**	35052	1526	0.94	0.88;0.99	1.79E-02
	hemorrhagic	30155	216	1.03	0.89;1.19	7.32E-01	30103	216	1.04	0.9;1.2	6.17E-01
	ischemic	30301	1051	0.89	0.83;0.95	**5.87E-04**	30247	1049	0.92	0.86;0.98	1.43E-02
HDL-cholesterol	all	35107	1529	0.89	0.84;0.94	**2.89E-05**	35052	1526	0.92	0.87;0.97	3.20E-03
	hemorrhagic	30155	216	1.04	0.9;1.21	5.63E-01	30103	216	1.07	0.92;1.24	3.97E-01
	ischemic	30301	1051	0.86	0.81;0.92	**1.82E-05**	30247	1049	0.9	0.84;0.96	1.89E-03
HDL2-cholesterol	all*	35107	1529	0.88	0.84;0.93	**9.13E-06**	35052	1526	0.91	0.87;0.97	**1.41E-03**
	hemorrhagic	30155	216	1.04	0.9;1.21	5.75E-01	30103	216	1.07	0.92;1.24	3.90E-01
	ischemic*	30301	1051	0.85	0.8;0.91	**2.85E-06**	30247	1049	0.89	0.83;0.95	**5.29E-04**
IDL-triglycerides	all	38561	1780	1.1	1.05;1.16	**6.06E-05**	38494	1775	1.07	1.02;1.12	9.91E-03
	hemorrhagic	33609	246	0.92	0.81;1.06	2.63E-01	33545	246	0.89	0.78;1.02	1.02E-01
	ischemic	33755	1270	1.13	1.07;1.2	**6.91E-06**	33689	1266	1.09	1.03;1.15	2.01E-03
LDL-triglycerides	all	35107	1529	1.12	1.06;1.18	**3.93E-05**	35052	1526	1.08	1.03;1.14	2.47E-03
	hemorrhagic	30155	216	1.02	0.88;1.18	8.27E-01	30103	216	0.99	0.85;1.14	8.62E-01
	ischemic	30301	1051	1.14	1.07;1.21	**2.89E-05**	30247	1049	1.1	1.04;1.17	1.82E-03
Glucose	all	34980	1524	1.15	1.1;1.2	**7.81E-11**	34925	1521	1.06	1.01;1.11	1.87E-02
	hemorrhagic	30033	214	1.13	0.99;1.28	7.07E-02	29981	214	1.09	0.96;1.24	2.20E-01

Table 3. Significant associations for overall incident stroke events and when classified by stroke type. (*continued*)

		Model 1					Model 2				
Metabolite	**Type**	N	Ncases	HR	CI	P	N	Ncases	HR	CI	P
	ischemic	30179	1048	1.17	1.12;1.23	**5.37E-11**	30125	1046	1.07	1.01;1.13	2.13E-02
Lactate	all	35100	1529	1.12	1.07;1.18	**1.11E-05**	35045	1526	1.08	1.02;1.13	5.09E-03
	hemorrhagic	30153	216	1.06	0.92;1.22	4.04E-01	30101	216	1.04	0.9;1.19	6.13E-01
	ischemic	30299	1051	1.16	1.09;1.24	**1.11E-06**	30245	1049	1.1	1.04;1.17	1.98E-03
Pyruvate	all*	24423	1205	1.13	1.08;1.18	**1.37E-07**	24368	1202	1.09	1.04;1.14	**7.45E-04**
	hemorrhagic	19481	167	0.99	0.84;1.16	8.70E-01	19429	167	0.96	0.81;1.12	5.94E-01
	ischemic*	19627	778	1.17	1.11;1.23	**2.86E-10**	19573	776	1.13	1.07;1.19	**1.93E-05**
Glycoprotein acetyls	all*	35101	1529	1.15	1.09;1.21	**1.25E-07**	35046	1526	1.09	1.03;1.15	**1.27E-03**
	hemorrhagic	30154	216	1.02	0.88;1.18	8.06E-01	30102	216	0.96	0.83;1.11	6.28E-01
	ischemic*	30300	1051	1.2	1.13;1.28	**8.55E-09**	30246	1049	1.13	1.06;1.2	**2.17E-04**
Mean diameter of HDL	all	35107	1529	0.89	0.84;0.94	**3.05E-05**	35052	1526	0.92	0.87;0.98	6.73E-03
	hemorrhagic	30155	216	1.03	0.89;1.2	6.98E-01	30103	216	1.07	0.92;1.24	4.08E-01
	ischemic	30301	1051	0.86	0.8;0.92	**1.28E-05**	30247	1049	0.9	0.84;0.96	2.93E-03
S-HDL-triglycerides	all	35108	1529	1.11	1.06;1.17	**6.80E-05**	35053	1526	1.07	1.01;1.12	1.97E-02
	hemorrhagic	30156	216	1	0.86;1.15	9.71E-01	30104	216	0.96	0.83;1.11	5.84E-01
	ischemic	30302	1051	1.14	1.08;1.22	**1.99E-05**	30248	1049	1.09	1.02;1.16	8.32E-03
M-HDL-cholesterol	all*	38560	1780	0.89	0.85;0.94	**2.07E-05**	38493	1775	0.92	0.87;0.97	**1.35E-03**
	hemorrhagic	33608	246	1	0.87;1.15	9.88E-01	33544	246	1.02	0.88;1.17	8.27E-01
	ischemic	33754	1270	0.88	0.83;0.93	**3.11E-05**	33688	1266	0.91	0.85;0.97	1.95E-03
M-HDL-cholesterol esters	all	35106	1529	0.9	0.85;0.95	**2.05E-04**	35051	1526	0.92	0.87;0.97	3.73E-03
	hemorrhagic	30154	216	0.99	0.86;1.14	8.84E-01	30102	216	1	0.87;1.16	9.72E-01
	ischemic	30300	1051	0.89	0.83;0.95	**4.60E-04**	30246	1049	0.91	0.86;0.98	6.99E-03
M-HDL-free cholesterol	all	35106	1529	0.91	0.86;0.96	**7.33E-04**	35051	1526	0.93	0.88;0.98	8.24E-03

Table 3. Significant associations for overall incident stroke events and when classified by stroke type. (*continued*)

		Model 1					Model 2				
Metabolite	**Type**	N	Ncases	HR	CI	P	**N**	**Ncases**	**HR**	**CI**	**P**
	hemorrhagic	30154	216	1.02	0.88;1.18	8.25E-01	30102	216	1.02	0.88;1.18	7.79E-01
	ischemic	30300	1051	0.9	0.84;0.96	1.75E-03	30246	1049	0.92	0.86;0.98	1.56E-02
L-HDL-cholesterol	all	38555	1780	0.89	0.84;0.94	**2.13E-05**	38488	1775	0.92	0.88;0.98	5.50E-03
	hemorrhagic	33603	246	1.07	0.93;1.24	3.19E-01	33539	246	1.11	0.96;1.28	1.56E-01
	ischemic*	33749	1270	0.85	0.8;0.91	**2.03E-06**	33683	1266	0.89	0.84;0.95	**9.00E-04**
L-HDL-cholesterol esters	all	35101	1529	0.9	0.84;0.95	**2.03E-04**	35046	1526	0.93	0.88;0.99	1.37E-02
	hemorrhagic	30149	216	1.07	0.92;1.24	3.88E-01	30097	216	1.1	0.95;1.28	2.13E-01
	ischemic	30295	1051	0.86	0.8;0.92	**4.28E-05**	30241	1049	0.9	0.84;0.97	3.59E-03
L-HDL-free cholesterol	all	35101	1529	0.89	0.84;0.94	**1.25E-04**	35046	1526	0.92	0.87;0.98	9.96E-03
	hemorrhagic	30149	216	1.09	0.93;1.26	2.81E-01	30097	216	1.12	0.96;1.3	1.44E-01
	ischemic*	30295	1051	0.85	0.79;0.91	**1.04E-05**	30241	1049	0.89	0.82;0.95	**1.33E-03**
L-HDL-total lipids	all	35101	1529	0.9	0.85;0.95	**2.12E-04**	35046	1526	0.93	0.88;0.99	1.70E-02
	hemorrhagic	30149	216	1.06	0.91;1.23	4.71E-01	30097	216	1.09	0.93;1.27	2.77E-01
	ischemic	30295	1051	0.87	0.81;0.93	**5.91E-05**	30241	1049	0.91	0.84;0.97	6.00E-03
L-HDL-concentration	all	35101	1529	0.9	0.85;0.96	**8.53E-04**	35046	1526	0.94	0.89;1	4.21E-02
	hemorrhagic	30149	216	1.09	0.94;1.27	2.73E-01	30097	216	1.12	0.96;1.3	1.44E-01
	ischemic	30295	1051	0.87	0.81;0.93	**1.30E-04**	30241	1049	0.91	0.85;0.98	1.04E-02
L-HDL-phospholipids	all	35101	1529	0.9	0.85;0.96	**6.29E-04**	35046	1526	0.94	0.89;1	3.49E-02
	hemorrhagic	30149	216	1.08	0.93;1.26	3.21E-01	30097	216	1.11	0.95;1.29	1.85E-01
	ischemic	30295	1051	0.87	0.81;0.93	**9.62E-05**	30241	1049	0.91	0.85;0.98	9.32E-03
XL-HDL-free cholesterol	all	35099	1527	0.91	0.86;0.96	**8.31E-04**	35044	1524	0.94	0.89;1	3.55E-02
	hemorrhagic	30147	216	1.07	0.93;1.24	3.52E-01	30095	216	1.09	0.94;1.26	2.38E-01

Table 3. Significant associations for overall incident stroke events and when classified by stroke type. (*continued*)

		Model 1					Model 2				
Metabolite	**Type**	N	Ncases	HR	CI	**P**	**N**	**Ncases**	**HR**	**CI**	**P**
	ischemic	30293	1049	0.88	0.82;0.94	**3.46E-04**	30239	1047	0.92	0.86;0.99	1.75E-02
S-LDL-triglycerides	all	29120	1332	1.12	1.06;1.18	**2.95E-05**	29065	1329	1.09	1.03;1.15	2.81E-03
	hemorrhagic	24168	168	1.04	0.88;1.22	6.44E-01	24116	168	1.01	0.86;1.19	8.73E-01
	ischemic	24314	925	1.14	1.07;1.21	**7.12E-05**	24260	923	1.09	1.03;1.17	5.79E-03
L-LDL-triglycerides	all*	35107	1529	1.12	1.06;1.17	**3.00E-05**	35052	1526	1.09	1.03;1.14	**1.67E-03**
	hemorrhagic	30155	216	1.01	0.87;1.17	9.12E-01	30103	216	0.98	0.85;1.13	7.87E-01
	ischemic	30301	1051	1.13	1.07;1.2	**4.39E-05**	30247	1049	1.1	1.03;1.17	2.20E-03
M-LDL-triglycerides	all*	35106	1529	1.12	1.06;1.18	**1.68E-05**	35051	1526	1.09	1.03;1.14	**1.19E-03**
	hemorrhagic	30154	216	1.04	0.9;1.2	5.70E-01	30102	216	1.02	0.88;1.17	8.35E-01
	ischemic	30300	1051	1.14	1.07;1.21	**2.84E-05**	30246	1049	1.1	1.04;1.17	1.80E-03
XL-VLDL-triglycerides	all	38284	1769	1.09	1.04;1.14	**1.56E-04**	38217	1764	1.05	1;1.1	4.66E-02
	hemorrhagic	33352	242	0.98	0.85;1.12	7.66E-01	33288	242	0.96	0.83;1.1	5.23E-01
	ischemic	33499	1263	1.12	1.06;1.18	**2.00E-05**	33433	1259	1.07	1.01;1.13	1.41E-02

Abbreviations: N - Total samples size; Ncases - Number of cases; HR - Hazard Ratio; 95% CI - 95% confidence interval; P - *p*-value; Model 1 - adjustment for age, gender, BMI, lipid-lowering medication and study-specific covariates if needed; Model 2- additional adjustment for smoking status, diabetes, and hypertension;
*Associations that surpassed significance threshold in model 2.

DISCUSSION

In this study, we identified ten metabolites associated with the risk of stroke. These include amino acid histidine, glycolysis-related metabolite pyruvate, acute phase reaction markers glycoprotein acetyls, cholesterol in HDL2, and lipoprotein subfractions such as cholesterol in medium HDL and triglycerides in medium and large LDL particles which showed association with incident stroke events. Amino acid phenylalanine and HDL subfractions including cholesterol and free cholesterol in large HDL were associated with ischemic incident stroke. This metabolite profile was independent of traditional risk factors including hypertension, diabetes, smoking, and BMI.

The strongest association was observed between amino acid histidine and risk of stroke. One SD increase in concentration of histidine was associated with 10% lower risk of stroke. The effect was very similar across studies, with only the Finrisk97 study showing no effect. Even though the same direction of effect was observed in both ischemic and hemorrhagic stroke subtype, the association was mainly driven by ischemic stroke. Histidine is a semi-essential amino acid as adults generally produce it while children may not. Histidine can be converted to histamine which shows a strong effect on vasodilatation and functions as a neurotransmitter in the brain.[28,29] Previous studies reported that oral administration of histidine can reduce blood pressure.[30-32] Plasma concentrations of histidine have been inversely associated with inflammation and oxidative stress in patients with chronic kidney disease and obese women with metabolic syndrome.[33-35] Furthermore, histidine has also been studied in relation to cerebral ischemia. Recent animal studies reported that histidine treatment remarkably alleviated the infarction induced by middle cerebral artery occlusion[36] and showed long term-neuroprotection after cerebral ischemia with decreased infarct volume and improved neurological function.[37] Even though our findings support the results of previous studies, in the most comprehensive study of stroke to date, histidine was not associated with ischemic and hemorrhagic stroke in individuals within the China Kadoorie Biobank. However, in the China Kadoorie Biobank, nominal association was found with myocardial infarction.[12] This could be explained either by environmental and ethnic differences of studied populations or difference in confounder adjusted for, in the present study we adjusted for more potential confounders including BMI, lipid-lowering medication, diabetes, and hypertension.

We also found the glycolysis-related metabolite, pyruvate, to be associated with increased risk of stroke. The analyses of stroke subtypes suggested that this association was driven by ischemic incident stroke events. Our findings suggested that 1 SD increase in pyruvate concentration was associated with 12% higher risk of ischemic stroke. Pyruvate is

the end-product of glycolysis and it is critical for supplying energy to the cell.[38] Pyruvate has previously been shown to protect against experimental stroke possibly by blocking inflammation.[39,40] In this light, our finding seems to contrast previously described effects of pyruvate. However, in a combined study of myocardial infarction and stroke using the same metabolomics platform as the present study, higher levels of pyruvate were also associated with a higher risk of cardiovascular disease.[41] The mechanism through which circulating level of pyruvate relates to stroke and cardiovascular disease is still to be elucidated.

Furthermore, acute phase marker glycoprotein acetyls mainly alpha-1 glycoprotein was associated with higher risk of stroke. The analyses revealed that the association was strongest for ischemic subtype, for which we found that an increase of 1 SD in the circulating compound was associated with 13% higher risk of ischemic stroke. Our results confirmed association of glycoprotein acetyl and ischemic stroke that was observed in individuals within the China Kadoorie Biobank.[12] Circulating levels of glycoprotein acetyls have previously been associated with cardiovascular diseases and dementia but also inflammatory disease, cancer, and mortality.[41-43]

Analyses focused on stroke subtypes revealed the association of essential amino acid phenylalanine with increased risk of ischemic stroke. One SD increase in concentration of phenylalanine was associated with 15% higher risk of ischemic stroke. Phenylalanine is a precursor for tyrosine and catecholamines including dopamine, epinephrine, and norepinephrine. Phenylalanine has previously been associated with risk of diabetes and cardiovascular disease.[41,44,45] As the association with phenylalanine remains after adjustment for diabetes, the association with stroke cannot be explained by impaired glucose tolerance. Phenylalanine did not associate with risk of hemorrhagic stroke.

Majority of circulating biomarkers measured by NMR metabolomics technology belong to lipid concentrations and composition of 14 lipoprotein subparticles. This provides an excellent opportunity for comprehensive investigation of lipoprotein particles in stroke, as the analyses of cholesterol and cholesterol subfraction has shown inconsistent results.[3,5,6] In our study population we observed association of cholesterol in medium HDL with decreased risk of stroke and triglycerides in large and medium LDL particles with increased risk of stroke. None of these lipoproteins measurements was found to be associated with stroke in the China Kadoorie Biobank.[12] Interestingly, the China Kadoorie Biobank reported association of low-, intermediate-, and low-density lipoproteins with ischemic stroke.[12] However, we were not able to confirm these results in our study population. Again, lack of replication could be explained by environmental and ethnic differences of studied populations or the confounders adjusted for.

Interestingly, using metabolites discovered in our study, we were able to discriminate future stroke with the AUC of 0.54. The metabolites discovered in the China Kadoorie Biobank also showed to be relevant for discriminating future stroke in our study population. Finally, when using metabolites discovered in our study and the China Kadoorie Biobank together with the Framingham Stroke Risk Score, we found further increase in the AUC. This suggests that the metabolites may have better utility for prediction of stroke and asks for use of other metabolomics platforms in order to discover additional metabolite measurements which could improve the risk prediction.

Strengths of our study are large sample size, prospective study design with detailed data collection over a long period of follow-up and the similar quantification method of circulating metabolites across the studies. Our study also has several limitations. With new improved methods available many other metabolites can be measured, which can be of importance to stroke.[46] Another limitation is differences in methods used across the cohorts to identify cases of incident stroke. As most of the cohorts used electronic health registries, this may have limited sensitivity which subsequently influenced power to identify novel significant associations. Furthermore, statistical power was also reduced in analyses of stroke subtypes as some of the cohorts were unable to distinguish between these. Another limitation is limited sample size for the analysis of hemorrhagic stroke which influenced our ability to detect novel associations for this stroke type.

To conclude, we found association of ten metabolites associated with risk of stroke in 1,791 incident stroke events observed among 38,797 individuals from seven population-based studies. The biological mechanisms underlying these associations should be subject of further studies.

REFERENCES

1. Mozaffarian. Heart Disease and Stroke Statistics-2015 Update: A Report From the American Heart Association (vol 131, pg e29, 2015). *Circulation* **131**, E535-E535 (2015).
2. Andersen, K.K., Olsen, T.S., Dehlendorff, C. & Kammersgaard, L.P. Hemorrhagic and ischemic strokes compared: stroke severity, mortality, and risk factors. *Stroke* **40**, 2068-72 (2009).
3. Goldstein, L.B. *et al.* Guidelines for the primary prevention of stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* **42**, 517-84 (2011).
4. Bos, M.J., Koudstaal, P.J., Hofman, A. & Ikram, M.A. Modifiable etiological factors and the burden of stroke from the Rotterdam study: a population-based cohort study. *PLoS Med* **11**, e1001634 (2014).
5. Zhang, Y. *et al.* Total and high-density lipoprotein cholesterol and stroke risk. *Stroke* **43**, 1768-74 (2012).
6. Amarenco, P., Labreuche, J. & Touboul, P.J. High-density lipoprotein-cholesterol and risk of stroke and carotid atherosclerosis: a systematic review. *Atherosclerosis* **196**, 489-96 (2008).
7. Saenger, A.K. & Christenson, R.H. Stroke biomarkers: progress and challenges for diagnosis, prognosis, differentiation, and treatment. *Clin Chem* **56**, 21-33 (2010).
8. Makin, S.D., Doubal, F.N., Dennis, M.S. & Wardlaw, J.M. Clinically Confirmed Stroke With Negative Diffusion-Weighted Imaging Magnetic Resonance Imaging: Longitudinal Study of Clinical Outcomes, Stroke Recurrence, and Systematic Review. *Stroke* **46**, 3142-8 (2015).
9. Jung, J.Y. *et al.* 1H-NMR-based metabolomics study of cerebral infarction. *Stroke* **42**, 1282-8 (2011).
10. Wang, D., Kong, J., Wu, J., Wang, X. & Lai, M. GC-MS-based metabolomics identifies an amino acid signature of acute ischemic stroke. *Neurosci Lett* **642**, 7-13 (2017).
11. Lee, Y., Khan, A., Hong, S., Jee, S.H. & Park, Y.H. A metabolomic study on high-risk stroke patients determines low levels of serum lysine metabolites: a retrospective cohort study. *Mol Biosyst* **13**, 1109-1120 (2017).
12. Holmes, M.V. *et al.* Lipids, Lipoproteins, and Metabolites and Risk of Myocardial Infarction and Stroke. *J Am Coll Cardiol* **71**, 620-632 (2018).
13. Wieberdink, R.G., Ikram, M.A., Hofman, A., Koudstaal, P.J. & Breteler, M.M.B. Trends in stroke incidence rates and stroke risk factors in Rotterdam, the Netherlands from 1990 to 2008. *European Journal of Epidemiology* **27**, 287-295 (2012).
14. Kivimaki, M. *et al.* Validity of Cardiovascular Disease Event Ascertainment Using Linkage to UK Hospital Records. *Epidemiology* **28**, 735-739 (2017).
15. Britton, A. *et al.* Validating self-reported strokes in a longitudinal UK cohort study (Whitehall II): Extracting information from hospital medical records versus the Hospital Episode Statistics database. *Bmc Medical Research Methodology* **12** (2012).
16. Shepherd, J. *et al.* The design of a Prospective Study of Pravastatin in the Elderly at Risk (PROSPER). *American Journal of Cardiology* **84**, 1192-1197 (1999).
17. Carandang, R. *et al.* Trends in incidence, lifetime risk, severity, and 30-day mortality of stroke over the past 50 years. *JAMA* **296**, 2939-46 (2006).
18. Seshadri, S. *et al.* The lifetime risk of stroke: estimates from the Framingham Study. *Stroke* **37**, 345-50 (2006).
19. Soininen, P., Kangas, A.J., Wurtz, P., Suna, T. & Ala-Korpela, M. Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Cardiovascular Epidemiology and Genetics. *Circulation-Cardiovascular Genetics* **8**, 192-206 (2015).

20. Soininen, P. *et al.* High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. *Analyst* **134**, 1781-1785 (2009).
21. Freedman, D.S. *et al.* Sex and age differences in lipoprotein subclasses measured by nuclear magnetic resonance spectroscopy: the Framingham Study. *Clin Chem* **50**, 1189-200 (2004).
22. Kathiresan, S. *et al.* A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. *BMC Med Genet* **8 Suppl 1**, S17 (2007).
23. Delles, C. *et al.* Nuclear magnetic resonance-based metabolomics identifies phenylalanine as a novel predictor of incident heart failure hospitalisation: results from PROSPER and FINRISK 1997. *European Journal of Heart Failure* **20**, 663-673 (2018).
24. Akbaraly, T. *et al.* Association of circulating metabolites with healthy diet and risk of cardiovascular disease: analysis of two cohort studies. *Scientific Reports* **8**(2018).
25. Kettunen, J. *et al.* Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat Commun* **7**, 11122 (2016).
26. van der Lee, S.J. *et al.* Circulating metabolites and general cognitive ability and dementia: Evidence from 11 cohort studies. *Alzheimers Dement* (2018).
27. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* **95**, 221-227 (2005).
28. Ignesti, G. *et al.* Increased desensitization by picomolar phorbol ester of the endothelium-mediated effect of histamine in the perfused rat mesenteric bed. *Inflamm Res* **45**, 171-5 (1996).
29. Wang, L. *et al.* Plasma Amino Acid Profile in Patients with Aortic Dissection. *Sci Rep* **7**, 40146 (2017).
30. Toba, H. *et al.* Oral L-histidine exerts antihypertensive effects via central histamine H3 receptors and decreases nitric oxide content in the rostral ventrolateral medulla in spontaneously hypertensive rats. *Clin Exp Pharmacol Physiol* **37**, 62-8 (2010).
31. Tuttle, K.R., Milton, J.E., Packard, D.P., Shuler, L.A. & Short, R.A. Dietary amino acids and blood pressure: a cohort study of patients with cardiovascular disease. *Am J Kidney Dis* **59**, 803-9 (2012).
32. Jennings, A. *et al.* Amino Acid Intakes Are Inversely Associated with Arterial Stiffness and Central Blood Pressure in Women. *J Nutr* **145**, 2130-8 (2015).
33. Watanabe, M. *et al.* Consequences of low plasma histidine in chronic kidney disease patients: associations with inflammation, oxidative stress, and mortality. *Am J Clin Nutr* **87**, 1860-6 (2008).
34. Mihalik, S.J. *et al.* Metabolomic profiling of fatty acid and amino acid metabolism in youth with obesity and type 2 diabetes: evidence for enhanced mitochondrial oxidation. *Diabetes Care* **35**, 605-11 (2012).
35. Feng, R.N. *et al.* Histidine supplementation improves insulin resistance through suppressed inflammation in obese women with the metabolic syndrome: a randomised controlled trial. *Diabetologia* **56**, 985-994 (2013).
36. Adachi, N., Liu, K. & Arai, T. Prevention of brain infarction by postischemic administration of histidine in rats. *Brain Res* **1039**, 220-3 (2005).
37. Liao, R.J. *et al.* Histidine provides long-term neuroprotection after cerebral ischemia through promoting astrocyte migration. *Scientific Reports* **5**(2015).
38. Gray, L.R., Tompkins, S.C. & Taylor, E.B. Regulation of pyruvate metabolism and human disease. *Cell Mol Life Sci* **71**, 2577-604 (2014).
39. Wang, Q. *et al.* Pyruvate protects against experimental stroke via an anti-inflammatory mechanism. *Neurobiol Dis* **36**, 223-31 (2009).
40. Ryou, M.G. *et al.* Pyruvate protects the brain against ischemia-reperfusion injury by activating the erythropoietin signaling pathway. *Stroke* **43**, 1101-7 (2012).

41. Wurtz, P. *et al.* Metabolite profiling and cardiovascular event risk: a prospective study of 3 population-based cohorts. *Circulation* **131**, 774-85 (2015).
42. Connelly, M.A., Gruppen, E.G., Otvos, J.D. & Dullaart, R.P.F. Inflammatory glycoproteins in cardio-metabolic disorders, autoimmune diseases and cancer. *Clin Chim Acta* **459**, 177-186 (2016).
43. Lawler, P.R. *et al.* Circulating N-Linked Glycoprotein Acetyls and Longitudinal Mortality Risk. *Circ Res* **118**, 1106-15 (2016).
44. Wang, T.J. *et al.* Metabolite profiles and the risk of developing diabetes. *Nat Med* **17**, 448-53 (2011).
45. Floegel, A. *et al.* Identification of Serum Metabolites Associated With Risk of Type 2 Diabetes Using a Targeted Metabolomic Approach. *Diabetes* **62**, 639-648 (2013).
46. Wishart, D.S. *et al.* HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic Acids Res* **41**, D801-7 (2013).

Chapter 3.4

Relationship between gut microbiota and circulating metabolites in population-based cohorts

Dina Vojinovic*, Djawad Radjabzadeh*, Alexander Kurilshikov*, Najaf Amin, Cisca Wijmenga, Lude Franke, Andre G. Uitterlinden, Alexandra Zhernakova, Jingyaun Fu**,Robert Kraaij**, Cornelia M. van Dujin**

*These authors contributed equally to this work
**These senior authors contributed equally to this work

This chapter is in preparation.

The supplemental information for this paper is available at https://drive.google.com/file/d/1MJW0tieyfMGBWLnqS7xTAIxq_cEzNuXb/view?usp=sharing

ABSTRACT

Gut microbiota has been implicated in the major diseases affecting the human population and has also been linked to triglycerides and high-density lipoprotein levels in the circulation. As recent development in metabolomics allows to classify the lipoprotein particles into more details, we aim to examine the impact of gut microbiota on circulating metabolites measured by Nuclear Magnetic Resonance (^{1}H-NMR) technology in 2,309 individuals from the Rotterdam Study and LifeLines DEEP cohort in whom gut microbiota was profiled using 16S rRNA gene sequencing. The relationship between gut microbiota and metabolites was assessed by linear regression analysis while adjusting for age, gender, body-mass index, technical covariates and medication use. Our analysis revealed association of 32 microbial families and genera with very-low-density and high-density lipoprotein subfractions, serum lipid measures, glycolysis-related metabolites, amino acids, and acute phase reaction markers. These observations provide novel insights into the role of microbiota in host metabolism and support the potential of gut microbiota as a target for therapeutic and preventive interventions.

INTRODUCTION

There is increasing interest in the role of the gut microbiota in the major diseases affecting the human population. For a large part, these association can be attributed to metabolic and immune signals of the microbiota that enter the circulation.[1] The gut microbiota has been implicated in obesity and diabetes,[2] but recently Zhernakova *et al.* have shown that the microbiota is also a major driver of circulating lipid levels, including triglycerides and high-density lipoproteins (HDL).[3] The association with low-density lipoprotein (LDL) cholesterol levels, the major target for treatment of dyslipidemia, or total cholesterol was weaker.[3,4] Recent development in metabolomics allows subclassifying the lipoprotein classes into more detail based on their particle size, composition, and concentration. Various studies further linked the gut microbiota to various amino acids, which have been implicated in diabetes and cardiovascular disease.[5-9]

To provide novel insights into the relation of gut microbiota and circulating metabolites, we have performed an in-depth study of the metabolome characterized by nuclear magnetic resonance (^{1}H-NMR) technology and the microbiota. To obtain sufficient power, we combined the data of two large population-based prospective studies, which have a rich amount of data on risk factors and treatment of disease.

METHODS

Study population

Our study population included participants from Rotterdam Study and LifeLines-DEEP cohort.

The Rotterdam Study is a prospective population-based cohort study that includes participants from the well-defined district of Rotterdam.[10] The initial cohort was defined among 7,983 persons, aged 55 years or older in 1990 (RS-I).[10] The cohort was further extended in 2000/2001 by additional 3,011 individuals, aged 55 years and older (RS-II), and in 2006/2008 by adding 3,932 individuals, aged 45 years and older (RS-III).[10] All participants provided written informed consent. The Medical Ethics Committee of the Erasmus Medical Center, Rotterdam, approved the study.

The LifeLines-DEEP cohort is a sub-cohort of LifeLines study, a prospective population-based cohort study in the north of the Netherlands.[11] The LifeLines cohort was established in 2006 among participants aged 20-50 years.[12] After completion of inclusion in 2013, the cohort includes 165,000 participants.[11] A subset of approximately 1,500 Life-

Lines participants participated in Lifelines-DEEP.[12] The LifeLines-DEEP study is approved by the Ethical Committee of the University Medical Center Groningen.[12] All participants provided written informed consent.

Metabolite profiling

Quantification of small compounds in fasting plasma samples was performed using ^{1}H-NMR technology in both participating studies.[13-15] Simultaneous quantification of a wide range of metabolites, including amino acids, glycolysis-related metabolites, ketone bodies, fatty acids, routine lipids and lipoprotein subclasses was done using the Nightingale Health metabolomics platform (Helsinki, Finland). The detailed description of the method can be found elsewhere.[13,16] In total there were 145 non-derived metabolite measures quantified in absolute concentration units across the participating studies (**Supplementary Table 1**).

Gut microbiota profiling

In order to study gut microbiota, fecal samples were collected from participants of Rotterdam Study and LifeLines-DEEP study. 16S rRNA gene sequencing of the V4 variable region was performed using the Illumina MiSeq platform.[12] A closed reference Operational Taxonomic Unit (OTU) mapped to a Silva (128) database as implemented by RDP classifier (2.12) was used to infer taxonomy.[12] Detail information regarding the gut microbiota profiling is described elsewhere.[12] Absolute values of taxonomy abundance were used. Furthermore, the microbial Shannon diversity index was calculated. Gut microbiota composition dataset included 1,427 participants from the RS-III cohort that participated in the second examination round at the study center. Metabolite measurements were available for 1,390 RS-III participants. In the LifeLines-DEEP study, gut microbiota composition dataset included 1,248 participants and metabolite measurements were available for 915 participants.[12]

Statistical analysis

Prior to the analysis, metabolites were natural logarithmic transformed to obtain approximately normal distribution. To deal with metabolite concentration of zero, half of the minimum detectable value of the metabolite was added to metabolites before the transformation. The metabolite measures were scaled to standard deviation units (SD). Similarly, to obtain approximately normal distribution of microbial taxa, we first added 1 to the abundance values and subsequently performed natural logarithmic transformation.

The relationship between metabolites and microbial taxa was assessed by linear regression analysis while adjusting for age, gender, body-mass index (BMI), technical covariates

including time in mail and storage time, and medication use including lipid-lowering medication, protein-pump inhibitors, and metformin. Furthermore, the analyses were adjusted for smoking and alcohol consumption. Participants using antibiotics were excluded from the analysis. The summary statistics of participating studies were combined using inverse variance-weighted fixed-effect meta-analysis in R (https://www.r-project.org/). In total, 145 overlapping metabolite measures and 455 overlapping microbial taxa were tested for association. As measurements in both metabolomics and gut microbiota datasets are highly correlated, we used a method of Li and Ji to calculate a number of independent tests.[17] There were 37 independent tests among the metabolite measures and 152 independent tests among microbial taxa. The significance threshold was set at $0.05/(37 \times 152) = 8.89 \times 10^{-6}$.

The relationship between metabolites and microbial diversity was also assessed by linear regression analysis while adjusting for age, gender, BMI, technical covariates and medication use (lipid-lowering medication, protein-pump inhibitors, and metformin) in each of the participating studies and summary statistics results were combined using inverse variance-weighted fixed-effect meta-analysis.

RESULTS

Participants from Rotterdam Study (n = 1,390, mean age 56.9±5.9, 57.5% women) were older compared to the participants from LifeLines-DEEP study (n = 915, mean age 44±13.9, 58.7% women), while gender distribution in the two cohorts was comparable.

The results of association analysis between circulating metabolites and composition of gut microbiota are shown in **Supplementary Table 2.** Multiple significant associations were detected for very low-density lipoprotein (VLDL) particles of various sizes (extra small, small, medium, large, very large, extremely large) and HDL particles (small, medium, large, very large) when adjusting for age, gender, BMI, medication use, technical covariates, and multiple testing (**Figure 1A**). When adjusting for smoking and alcohol intake in addition, similar association pattern was observed (**Figure 1B**). Family *Christensenellaceae* and genera *Christensenellaceae R7 group*, *Ruminococcaceae UCG-005,* and *Eubacterium xylanophilum group* were found to be generically associated with VLDL particles of various sizes, small HDL particles and triglycerides in medium HDL. Of note is that the association pattern of very large and large HDL particles including concentration of particles and its total lipids, cholesterol, free cholesterol, cholesterol esters was opposite compared to the association pattern of small and medium HDL (**Figure 1**). Monounsaturated fatty acids (MUFA), serum triglycerides (TG), saturated fatty acids

(SFA), and total fatty acids (TotFA) followed the same direction of association of VLDL (**Figure 1**). Focusing further on the lipoprotein particles, we found that genus *Clostridium sensu stricto 1*, family *Clostridiaceae1*, and one unknown family and genus followed a similar pattern as described above, i.e., being inversely associated to VLDL particles or various size, small HDL subfractions, and triglycerides in medium HDL and positively associated to very large and large HDL particles (**Figure 1B)**. Genera *Ruminococcaceae UCG-003*, *Ruminococcaceae UCG-002,* and *Ruminococcaceae UCG-010, Marvinbryantia* and *Lachnospiraceae FCS020 group* were again inversely associated with VLDL particles of various size and small HDL but the positive association to very large and large HDL was not significant when adjusting for multiple testing (**Figure 1B**). In addition to these generic effect, there are more targeted associations, for instance of family *Lachnospiraceae* and genus *Blautia* with small HDL particles and genus *Ruminococcus gnavus group*

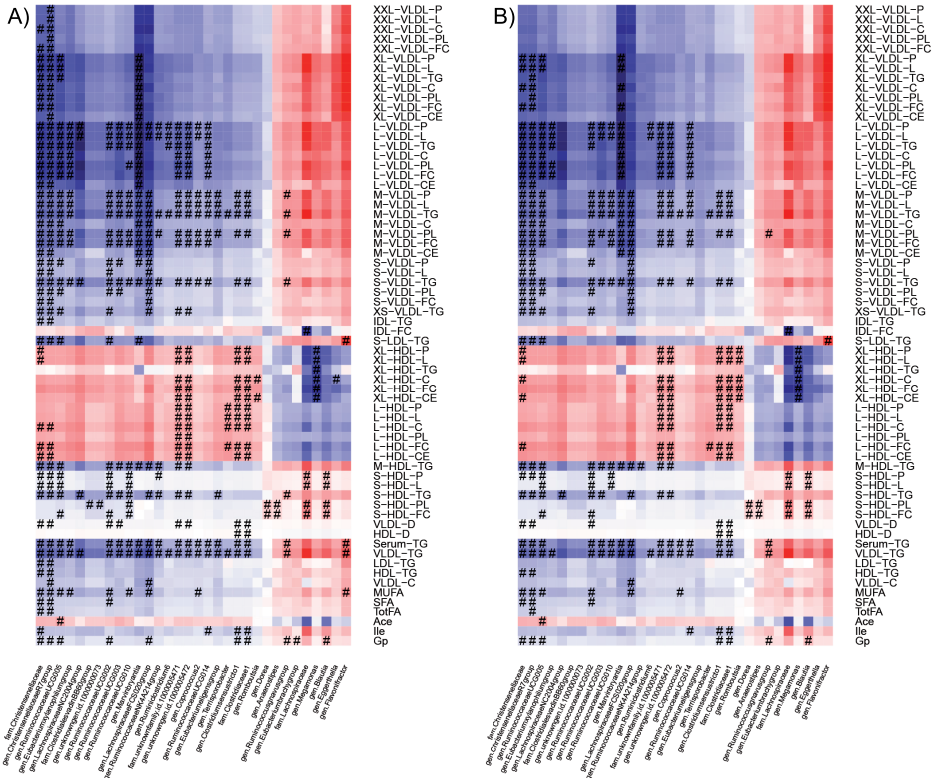


Figure 1. A) Results of association analysis between metabolites and microbial genera and families. The colors represent effect estimates of the metabolites and microbial taxa after adjustment for age, gender, body-mass index, technical covariates and medication use. Blue color stands for inverse association. Red color denotes positive associations. Symbols on the plot represent the level of significance with hash denoting Bonferroni significant associations. B) Association between metabolites and microbial genera and families after additional adjustment for smoking and alcohol consumption.

to phospholipids in medium VLDL, triglycerides in VLDL and serum triglycerides (**Figure 1B**). Family *Clostridiaceae1* and genus *Clostridium sensu stricto 1* were associated with both the HDL diameter and VLDL diameter (**Figure 1B**). The VLDL diameter was further associated with family *Christensenellaceae* and genera *Christensenellaceae R7 group* and *Ruminococcaceae UCG-002*. There is further a specific association of free cholesterol in IDL with family *Lachnospiraceae* and triglycerides in small LDL with genus *Flavonifractor*.

Beyond the lipoprotein fractions, we found three other metabolites, including the ketone body acetate, amino acid isoleucine, and acute phase reaction marker glycoprotein acetyl (mainly alpha 1), to be significantly associated with the microbiota when adjusting for multiple testing and age, sex, BMI, technical covariates, medication, smoking, and alcohol consumption. Genus *Ruminococcaceae UCG-005* was associated to acetate levels, family *Clostridiaceae1* and genera *Clostridium sensu stricto 1* and *Ruminococcaceae UCG-014* with isoleucine and genera *Clostridium sensu stricto1*, *Christensenellaceae R7 group, Ruminococcaceae UCG-005, Ruminococcus gnavus group*, *Blautia* and families *Clostridiaceae1* and *Christensenellaceae*, all associated to glycoprotein levels.

We next determined whether microbial diversity of gut microbiota was associated with lipoprotein particles or other metabolites (**Figure 2**). When adjusting for multiple testing and age, sex, BMI, technical covariates, and medication use, the pattern emerging is that higher microbiome diversity is significantly associated with lower levels of VLDL particles (small, large, medium, very large, extra-large), TotFA, MUFA, and SFA and increased levels of large and extra-large HDL particles and an increased diameter of HDL (**Figure 2**). As to the other metabolites, higher microbiome diversity is significantly associated with lower levels of glycoprotein acetyl, alanine, isoleucine, and lactate (**Figure 2)**.

DISCUSSION

We have examined the impact of gut microbiota on host circulating metabolites in 2,300 individuals from Rotterdam Study and LifeLines-DEEP cohort using ^{1}H-NMR technology. We identified associations between the gut microbiota composition and various metabolites including specific VLDL and HDL lipoprotein subfractions, serum lipid measures including triglycerides and fatty acids, glycolysis-related metabolite lactate, ketone body acetate, amino acids including alanine and isoleucine, and acute phase reaction marker including the glycoprotein acetyls independent on age, gender, BMI, and medication use. No associations were found to LDL subfractions and glucose levels measured by ^{1}H-NMR.

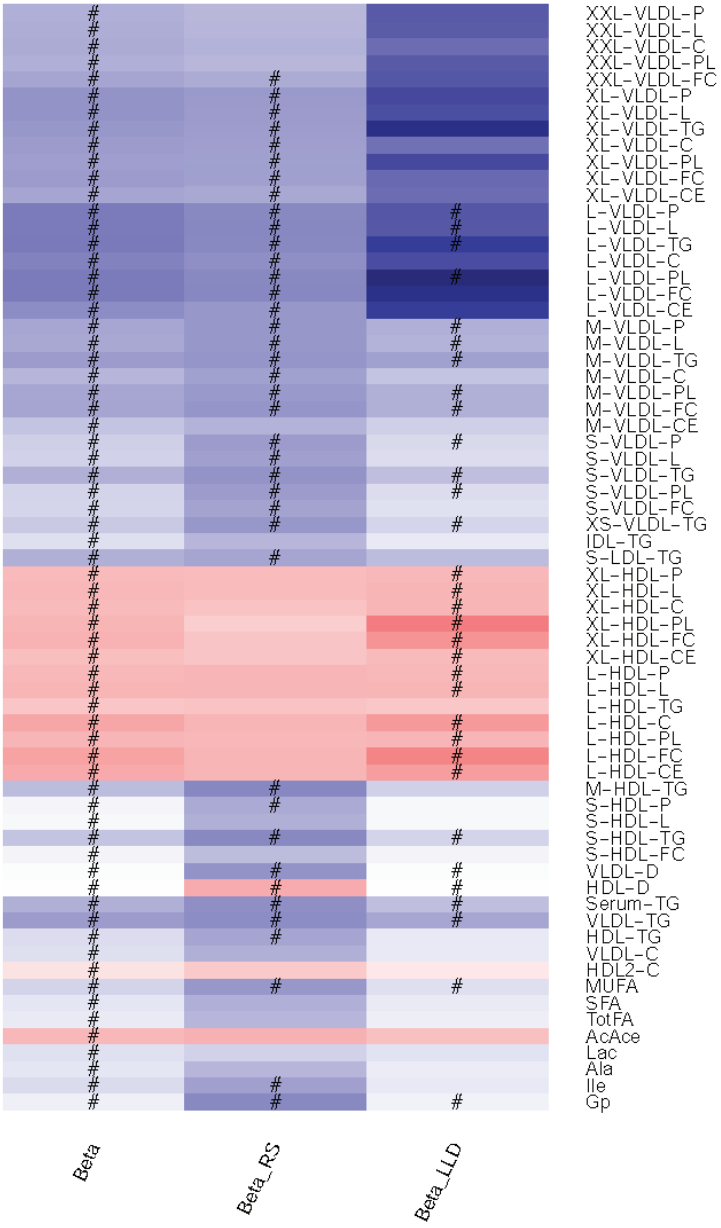


Figure 2. Results of association analysis between metabolites and alpha diversity. The colors represent effect estimates of the metabolites with alpha diversity. Effect estimates from meta-analysis (Beta), and in each of the participating studies are shown (effect estimate in Rotterdam Study - Beta_RS, effect estimate in LifeLines DEEP - Beta_LLD). Blue color stands for inverse association. Red color denotes positive associations. Symbols on the plot represent level of significance with hash denoting Bonferroni significant associations.

Our results based on two large population-based studies identified novel associations between the gut microbiota composition and various lipoprotein particles. We observed inverse association of family *Christensenellaceae* with VLDL particles of various sizes, small HDL particles, and triglycerides in medium HDL (**Figure 1B**). The family *Christensenellaceae* was previously linked to BMI and was associated with the reduced weight gain as reported in the mice study in which germfree mice were inoculated with lean and obese human fecal samples.[18] Furthermore, the family *Christensenellaceae* was reported to be the most heritable microbial taxon in the study by Goodrich *et al.* independently of the effect of BMI.[18]

Interestingly, the gut microbiota composition showed association with VLDL and HDL particles of various sizes, however weak association has been found for LDL and IDL particles suggesting that gut microbiota affects distinct classes of lipoproteins.[19] While VLDL particles of various sizes showed the same pattern of association, differences were noticed between large, medium, and small HDL particles suggesting that they are heterogeneous structures.[20] Small HDL particles are dense, protein-rich, and lipid-poor, whereas large HDL particles are large, lipid-rich particles.[21,22] Despite the fact that HDL is consistently associated with a reduced risk of cardiovascular disease, the past decade has seen major controversies on the clinical relevance of HDL interventions. Most trials aiming to increase HDL levels in the aggregate have been unsuccessful and were even stopped because of adverse effects.[23-25] The heterogeneity of HDL classes has been long recognized but can now be assessed on a large scale. This compositional heterogeneity of HDL results in functional heterogeneity such that small and large HDL particles are negatively correlated and display inverse relationship with various diseases including cardiovascular disease, as reported previously.[20,21] As observed in our study the small HDL particles were driven by genus *Blautia* and family *Lachnospiraceae* and were associated with lower diversity. Indeed the higher levels of small lipoprotein particle concentration have previously been associated with increased risk of stroke as reported in a recently published study of Holmes *et al.*, while the large and extra-large HDL particles that were driven by family *Clostridiacae1,* genus *Clostridium sensu stricto 1* and unknown family and genus and were associated with decreased risk of cardiovascular disease and stroke.[6] Interestingly, family *Clostridiacae1* was previously inversely correlated with BMI, serum triglycerides and is known to be involved in bile acid metabolism.[4,26]

Furthermore, we confirmed association of genus *Ruminococcus gnavus group* and serum triglycerides level,[27] and additionally reported association with triglycerides in VLDL particles and phospholipids in medium VLDL. *Ruminococcus gnavus group* was previously associated with low gut microbial richness[28] and its abundance was higher in patients with atherosclerotic cardiovascular disease.[29]

In addition to circulating lipids and lipoprotein particles, an association was found between gut microbiota and ketone bodies including acetate, amino acids including isoleucine, and acute phase reaction marker including glycoprotein acetyls mainly alpha 1. Circulating levels of acetate were specifically associated with genus *Ruminococcaceae UCG-005*. Acetate is the most common short-chain fatty acid (SCFA) formed by bacterial species in the colon.[30] SCFA can serve as an energy source, predominately via metabolism in liver.[31,32] Previous studies suggested that acetate mediates a microbiota-brain axis and promotes metabolic syndrome.[33] Circulating levels of isoleucine, an essential branched-chain amino acid, were inversely associated with family *Clostridiaceae1* and genera *Clostridium sensu stricto 1* and *Ruminococcaceae UCG-014* in our sample. Recent studies reported association of circulating levels of isoleucine with diabetes and cardiovascular disease.[7,34] Furthermore, isoleucine was reported to be negatively associated with *Christensenellaceae* and positively with *Blautia*.[35] Even though we observed the same pattern of association between isoleucine and these taxa, the associations did not reach the significance threshold. Also recently, a study focusing on relation of fecal metabolites using mass spectroscopy (Metabolon) and the gut microbiota was published.[5] Even though the overlap of measured metabolites is limited, amino acids are measured on both platforms. Other amino acids showed a strong association with the gut microbiota but not isoleucine.[5] However, the concentration of metabolite levels in feces and blood may differ. This an important field of future research. Lastly, glycoprotein acetyls, a composite marker that integrates protein levels and glycosylation states of the most abundant acute phase proteins in circulation,[36,37] was positively associated with genus *Blautia* and *Ruminococcus gnavus group*. Genus *Blautia* is one of the microbial taxa with substantial heritability in twin study,[18] and showed strong association with the host genetic determinants which has been associated with BMI and obesity.[38] Glycoprotein acetyls are associated with other common markers of inflammation.[36,37] Circulating level of glycoprotein acetyls have been implicated in inflammatory diseases and cancer, and have been associated with mortality and cardiovascular disease.[6,7,39,40]

The strengths of our study are large sample size, population-based study design, extensive phenotyping of study participants, and harmonized analysis in participating studies while correcting for factors such as use of medication and BMI. Merging the data of two large population-based studies allowed us to internally validate the findings. However, our study has also limitations. When exploring circulating molecules, we focused on metabolites measured by Nightingale platform which covers a wide range of circulating compounds.[14] However, these compounds represent a limited proportion of circulating metabolites, therefore, future studies should focus on metabolites detected by other more detailed techniques.[41] Further, the gut microbial composition was determined from fecal samples. As gut microbial composition varies throughout the gut with

respect to the anatomic location along the gut and at the given site, more complete picture of the gut microbiota could be obtained by getting samples from different locations along the intestines in the future.[19,42] Furthermore, when exploring gut microbiota, we focused on 16S rRNA sequencing. Even though broad shifts in community diversity could be captured by 16S rRNA, metagenomics approaches provide better resolution and sensitivity.[43] With the decreasing costs of metagenome sequencing, our knowledge can be extended in the future. Finally, although our analyses were adjusted for various known confounders, residual confounding remains possible.

To conclude, we found association between gut microbiota composition and various circulating metabolites including lipoprotein subfractions, serum lipid measures, glycolysis-related metabolites, ketone bodies, amino acids, and acute phase reaction markers. Association between gut microbiota and specific lipoprotein subfractions of VLDL and HDL particles provides novel insights into the role of microbiota in influencing host lipid levels. These observations support the potential of gut microbiota as a target for therapeutic and preventive interventions.

REFERENCES

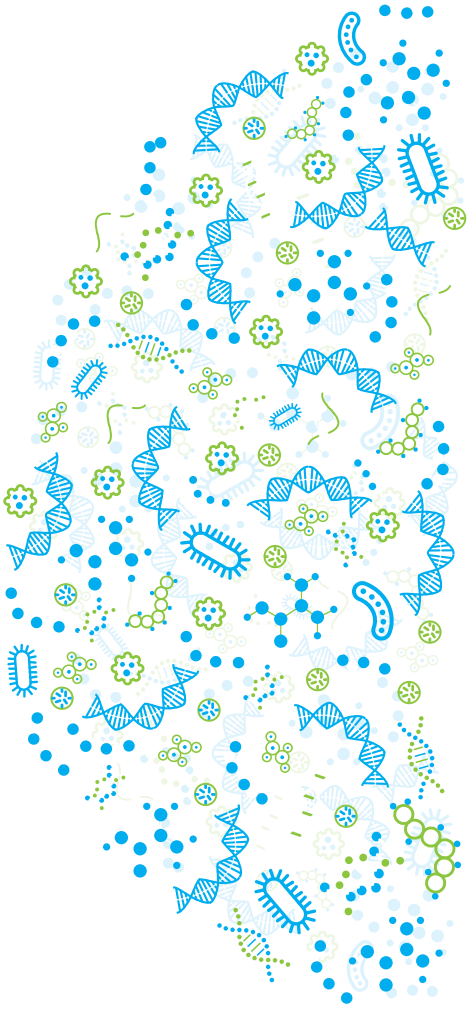
1. Holmes, E., Li, J.V., Marchesi, J.R. & Nicholson, J.K. Gut microbiota composition and activity in relation to host metabolic phenotype and disease risk. *Cell Metab* **16**, 559-64 (2012).
2. Komaroff, A.L. The Microbiome and Risk for Obesity and Diabetes. *JAMA* **317**, 355-356 (2017).
3. Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565-9 (2016).
4. Fu, J. *et al.* The Gut Microbiome Contributes to a Substantial Proportion of the Variation in Blood Lipids. *Circ Res* **117**, 817-24 (2015).
5. Zierer, J. *et al.* The fecal metabolome as a functional readout of the gut microbiome. *Nature Genetics* **50**, 790-+ (2018).
6. Holmes, M.V. *et al.* Lipids, Lipoproteins, and Metabolites and Risk of Myocardial Infarction and Stroke. *J Am Coll Cardiol* **71**, 620-632 (2018).
7. Wurtz, P. *et al.* Metabolite profiling and cardiovascular event risk: a prospective study of 3 population-based cohorts. *Circulation* **131**, 774-85 (2015).
8. Nakamura, H. *et al.* Plasma amino acid profiles are associated with insulin, C-peptide and adiponectin levels in type 2 diabetic patients. *Nutrition & Diabetes* **4** (2014).
9. Magnusson, M. *et al.* A diabetes-predictive amino acid score and future cardiovascular disease. *European Heart Journal* **34**, 1982-1989 (2013).
10. Ikram, M.A. *et al.* The Rotterdam Study: 2018 update on objectives, design and main results. *Eur J Epidemiol* **32**, 807-850 (2017).
11. Scholtens, S. *et al.* Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int J Epidemiol* **44**, 1172-80 (2015).
12. Tigchelaar, E.F. *et al.* Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open* **5**, e006772 (2015).
13. Soininen, P. *et al.* High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. *Analyst* **134**, 1781-5 (2009).
14. van der Lee, S.J. *et al.* Circulating metabolites and general cognitive ability and dementia: Evidence from 11 cohort studies. *Alzheimers Dement* (2018).
15. Vojinovic, D. *et al.* Metabolic profiling of intra- and extracranial carotid artery atherosclerosis. *Atherosclerosis* **272**, 60-65 (2018).
16. Soininen, P., Kangas, A.J., Wurtz, P., Suna, T. & Ala-Korpela, M. Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circ Cardiovasc Genet* **8**, 192-206 (2015).
17. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity (Edinb)* **95**, 221-7 (2005).
18. Goodrich, J.K. *et al.* Human genetics shape the gut microbiome. *Cell* **159**, 789-99 (2014).
19. Ghazalpour, A., Cespedes, I., Bennett, B.J. & Allayee, H. Expanding role of gut microbiota in lipid metabolism. *Curr Opin Lipidol* **27**, 141-7 (2016).
20. Kontush, A. HDL particle number and size as predictors of cardiovascular disease. *Front Pharmacol* **6**, 218 (2015).
21. Camont, L., Chapman, M.J. & Kontush, A. Biological activities of HDL subpopulations and their relevance to cardiovascular disease. *Trends in Molecular Medicine* **17**, 594-603 (2011).
22. Camont, L. *et al.* Small, Dense High-Density Lipoprotein-3 Particles Are Enriched in Negatively Charged Phospholipids Relevance to Cellular Cholesterol Efflux, Antioxidative, Antithrombotic,

Anti-Inflammatory, and Antiapoptotic Functionalities. *Arteriosclerosis Thrombosis and Vascular Biology* **33**, 2715-2723 (2013).
23. Clark, R.W. *et al.* Raising high-density lipoprotein in humans through inhibition of cholesteryl ester transfer protein: An initial multidose study of torcetrapib. *Arteriosclerosis Thrombosis and Vascular Biology* **24**, 490-497 (2004).
24. Keene, D., Price, C., Shun-Shin, M.J. & Francis, D.P. Effect on cardiovascular risk of high density lipoprotein targeted drug treatments niacin, fibrates, and CETP inhibitors: meta-analysis of randomised controlled trials including 117 411 patients. *Bmj-British Medical Journal* **349** (2014).
25. Berenson, A. Pfizer Ends Studies on Drug for Heart Disease. in *The New York Times* (2006).
26. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research* **40**, D742-D753 (2012).
27. Lahti, L. *et al.* Associations between the human intestinal microbiota, Lactobacillus rhamnosus GG and serum lipids indicated by integrated analysis of high-throughput profiling data. *Peerj* **1** (2013).
28. Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541-6 (2013).
29. Jie, Z. *et al.* The gut microbiome in atherosclerotic cardiovascular disease. *Nat Commun* **8**, 845 (2017).
30. Louis, P., Scott, K.P., Duncan, S.H. & Flint, H.J. Understanding the effects of diet on bacterial metabolism in the large intestine. *J Appl Microbiol* **102**, 1197-208 (2007).
31. Rios-Covian, D. *et al.* Intestinal Short Chain Fatty Acids and their Link with Diet and Human Health. *Frontiers in Microbiology* **7** (2016).
32. Turnbaugh, P.J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**, 1027-1031 (2006).
33. Perry, R.J. *et al.* Acetate mediates a microbiome-brain-beta-cell axis to promote metabolic syndrome. *Nature* **534**, 213-7 (2016).
34. Wang, T.J. *et al.* Metabolite profiles and the risk of developing diabetes. *Nat Med* **17**, 448-53 (2011).
35. Org, E. *et al.* Relationships between gut microbiota, plasma metabolites, and metabolic syndrome traits in the METSIM cohort. *Genome Biology* **18** (2017).
36. Otvos, J.D. *et al.* GlycA: A Composite Nuclear Magnetic Resonance Biomarker of Systemic Inflammation. *Clin Chem* **61**, 714-23 (2015).
37. Connelly, M.A. *et al.* GlycA, a marker of acute phase glycoproteins, and the risk of incident type 2 diabetes mellitus: PREVEND study. *Clin Chim Acta* **452**, 10-7 (2016).
38. Bonder, M.J. *et al.* The effect of host genetics on the gut microbiome. *Nat Genet* **48**, 1407-1412 (2016).
39. Connelly, M.A., Gruppen, E.G., Otvos, J.D. & Dullaart, R.P.F. Inflammatory glycoproteins in cardiometabolic disorders, autoimmune diseases and cancer. *Clinica Chimica Acta* **459**, 177-186 (2016).
40. Lawler, P.R. *et al.* Circulating N-Linked Glycoprotein Acetyls and Longitudinal Mortality Risk. *Circulation Research* **118**, 1106-1115 (2016).
41. Wishart, D.S. *et al.* HMDB 3.0-The Human Metabolome Database in 2013. *Nucleic Acids Research* **41**, D801-D807 (2013).
42. Allayee, H. & Hazen, S.L. Contribution of Gut Bacteria to Lipid Levels Another Metabolic Role for Microbes? *Circulation Research* **117**, 750-754 (2015).
43. Poretsky, R., Rodriguez, R.L., Luo, C., Tsementzi, D. & Konstantinidis, K.T. Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. *PLoS One* **9**, e93827 (2014).

SUPPLEMETARY TABLES

Supplementary Table 1. List of all circulating metabolites tested for association with gut microbiota.

Supplementary Table 2. Results of association analysis between gut microbiota and circulating metabolites.



Chapter 4

Genomic studies of psychiatric diseases

Chapter 4.1

Variants in *TTC25* affect autistic trait in patients with autism spectrum disorder and general population

Dina Vojinovic, Nathalie Brison, Shahzad Ahmad, Ilse Noens, Irene Pappa, Lennart C Karssen, Henning Tiemeier, Cornelia M. van Duijn, Hilde Peeters*, Najaf Amin*

* These authors contributed equally to this work.

This chapter was published in Eur J Hum Genet. 2017 Aug;25(8):982-987.

The supplemental information for this paper is available online on European Journal of Human Genetics website (http://www.nature.com/ejhg)

ABSTRACT

Autism spectrum disorder (ASD) is a highly heritable neurodevelopmental disorder with a complex genetic architecture. To identify genetic variants underlying ASD, we performed single-variant and gene-based genome-wide association studies using a dense genotyping array containing over 2.3 million single-nucleotide variants in a discovery sample of 160 families with at least one child affected with non-syndromic ASD using a binary (ASD yes/no) phenotype and a quantitative autistic trait. Replication of the top findings was performed in Psychiatric Genomics Consortium and Erasmus Rucphen Family (ERF) cohort study. Significant association of quantitative autistic trait was observed with the *TTC25* gene at 17q21.2 (effect size = 10.2, *p*-value = 3.4×10^{-7}) in the gene-based analysis. The gene also showed nominally significant association in the cohort-based ERF study (effect = 1.75, *p*-value = 0.05). Meta-analysis of discovery and replication improved the association signal ($p\text{-value}_{meta} = 1.5\times10^{-8}$). No genome-wide significant signal was observed in the single-variant analysis of either the binary ASD phenotype or the quantitative autistic trait. Our study has identified a novel gene *TTC25* to be associated with quantitative autistic trait in patients with ASD. The replication of association in a cohort-based study and the effect estimate suggest that variants in *TTC25* may also be relevant for broader ASD phenotype in the general population. *TTC25* is overexpressed in frontal cortex and testis and is known to be involved in cilium movement and thus an interesting candidate gene for autistic trait.

INTRODUCTION

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized by deficits in social communication and social interaction and restricted and repetitive patterns of activities and behavior with an onset in early development.[1] However ASD is a psychiatric diagnosis based on clinical criteria, and the severity of these characteristics can be measured as quantitative traits that represent a continuum that extends into the general population, with ASD at the extreme end of the distribution.[2] Other associated, but not core features are intellectual disability, attention-deficit disorder and medical comorbidities.[3] The prevalence of ASD is estimated to be 62/10,000[4] with boys-to-girls ratio of ~ 4:1.[5] The importance of a genetic aetiology is established with heritability estimates ranging from 37 to 90%.[6–9] Despite genetic heterogeneity, considerable progress in understanding the genetic architecture of ASD has been made by identifying monogenetic causes through genetic syndromes,[10] rare chromosomal abnormalities,[11,12] rare copy-number variants[13–16] and rare penetrant gene mutations.[3] Several genomic regions, including 2q, 3q25–27, 3p25, 6q14–21, 7q31–36, and 17q11–21[17] have been linked to ASD. The role of rare genetic variants in the aetiology of ASD has been established by high-throughput technologies.[18,19] More recently, the theory of excess of de novo loss-of-function variants in ASD patients has gained popularity after some initial successes.[18–20] Around 1,000 genes have been identified to be enriched with de novo loss-of-function mutations in ASD patients.[21] However, de novo genetic variants do not contribute to the estimated heritability as these are not inherited. On the other hand, most genetic variance in ASD is attributed to common genetic variants.[9,22] Their role has been demonstrated by several genome-wide association studies (GWAS)[23–29] (**Supplementary Table S1**). Even though not many common susceptibility variants have been identified, significant association has been reported at 5p14.1,[23] at 5p15.31 between *SEMA5A* and *TAS2R1* genes,[28] within *MACROD2* at 20p12.1,[26] and at 1p13.2.[29] However, there is a significant overlap of the discovery samples used and little replication of specific loci between studies.[30]

Although the individual effect of common variants is modest, their joint effect may be substantial.[25] In this study besides assessing the effect of single variants on ASD, we evaluated the joint effect of multiple single variants in a gene in a genome-wide gene-based association analysis in patients with ASD from a Belgian Flemish cohort who were genotyped on a dense genotyping array.

MATERIALS AND METHODS

General overview of the study design and workflow are illustrated in **Supplementary Figure S1**.

Discovery sample

The discovery sample consisted of 160 nuclear Belgian Flemish families (657 individuals; **Supplementary Table S2**). The families were recruited to participate in the prospective study through the Expert Center for Autism (ECA) Leuven. All probands had been seen multiple times as part of their clinical care program in the ECA before recruitment. The families were asked to participate if there was at least one child with the diagnosis of non-syndromic ASD of unknown origin after a clinical genetics workup. Out of the 160 families, 55 were multiplex (two or more siblings with ASD) and 105 simplex. In six families the father had also been diagnosed with ASD such that there were 77.7% affected males and 22.3% affected females with male-to-female ratio of 3.5:1. Among them, 88.4% had normal and high intelligence, whereas 11.6% had mild, moderate or severe intellectual disability.

Diagnoses of ASD were made by a multidisciplinary team in the ECA Leuven according to DSM-IV-TR (American Psychiatric Association, 2000) criteria. Additionally, participants were assessed for quantitative autistic trait using the Dutch version of the Social Responsiveness Scale (SRS) and the Social Responsiveness Scale for Adults (SRS-A) designed to measure social impairment associated with ASD across a wide range of severity.[31,32] Completed questionnaires were obtained for 490 probands, parents and siblings. Among the affected patients that had the SRS score available, the majority had normal and high intelligence (86%). For all participants, we received written informed consent. This study was approved by the Medical Ethical Committee of the University Hospitals Leuven.

Genotyping

Genotyping of 657 individuals from the discovery cohort was performed at the Center for Human Genetics at the KU Leuven, Belgium using the HumanOmni2.5-8 BeadChip, which contains more than 2.3 million common and less-frequent single-nucleotide polymorphisms (SNPs) from the 1000 Genome Project (minor allele frequency 42.5%). SNP calling was performed in Genome Studio 2011.1 using the genotyping module v1.9. Markers with call rate < 95%, or which were monomorphic or which failed an exact test of Hardy–Weinberg equilibrium (HWE) (p-value < 1×10^{-7}) were removed from the analysis. Samples with low call rate < 95% or high identity-by-state (≥ 95%) were also removed from the analysis. Ethnic outliers were determined using multidimensional

scaling analysis with 1000 Genomes dataset (**Supplementary Figure S2**). All samples clustered tightly with the Europeans and no ethnic outlier was identified. In total 1,646,898 markers and 654 genotyped individuals were retained for further statistical analysis. Lastly, Mega2 tool v4.4[33] was used to identify Mendelian inconsistencies, which were later set to missing.

Statistical analysis

Baseline descriptive analysis was performed with SPSS v21 (IBM Corporation, Armonk, NY, USA) and PEDSTATS v0.6.12.[34] Genome-wide association analyses of the binary ASD phenotype were performed through joint modeling of linkage and association, using the LAMP software v0.0.9 (School of Public Health, Ann Arbor, MI, USA). LAMP uses a maximum likelihood model to extract information on genetic linkage and association from samples of unrelated individuals, sib pairs, trios and larger pedigrees in settings where population stratification is not a concern (**Supplementary Figure S2**).[35] Odds ratios and 95% confidence intervals were estimated using PLINK v1.07.[36] The association tests for markers on sex chromosomes were performed by transmission disequilibrium test for chromosome X and by logistic regression for chromosome Y. Single-variant and gene-based genome-wide association analyses of the quantitative autistic trait adjusted for age, gender and familial relationships were performed using the RVtests software tool version 20150630 (http://zhanxw.github.io/rvtests/). The gene-based analysis included Combined Multivariate and Collapsing method which is robust and powerful in the presence of wide spectrum of variant allele frequencies.[37] The genes were defined according to human reference genome hg19. All association analyses were performed for entire discovery sample, and simplex and multiplex families separately. The standard genome-wide significance threshold of 5×10^{-8} was used to declare significance in the single-variant analyses, while the genome-wide significance threshold for the gene-based analysis was set at 2.5×10^{-6} based on 19,650 genes tested. PLINK/SEQ v0.10 (https://atgu.mgh.harvard.edu/plinkseq/) was used to convert PLINK files into variant call format files. All genome maps were updated to human genome build 19 (hg19). Gene pathway enrichment analysis of all nominally significant genes (p-value < 0.01) in the gene-based analysis was performed using the web-based gene network pathway enrichment tool (http://129.125.135.180:8080/ GeneNetwork/pathway.html).

The data were deposited in the GWAS Central database (http://www.gwascentral.org/study/HGVST1847).

Bioinformatic analysis

To annotate SNPs with regulatory information, we used RegulomeDB v1.1 database (http://www.regulomedb.org/index) that combines information from ENCODE and

other sources, as well as computational predictions and manual annotations into a tool that classifies SNPs into six categories, where Category 1 variants are 'likely to affect binding and linked to expression of a gene target', whereas category 6 variants have 'minimal binding evidence'.[38] Furthermore, regulatory information on SNPs in haplotype blocks was explored using a HaploReg v4.1 tool.[39] For these analysis r^2 was set to 1 and the population of European descent was chosen.

Replication samples

Psychiatric Genomics Consortium (PGC). A lookup of top findings from the single-variant analyses of the binary ASD phenotype and quantitative autistic trait was performed in the latest PGC GWAS. This dataset consists of a total of 6,495 parent-child trios who met diagnostic criteria for ASD and had genome-wide SNP data available (https://www.med.unc.edu/pgc).

Erasmus Rucphen Family (ERF) study. Replication of the gene-based analysis of quantitative autistic trait was performed in the ERF study as 1,250 participants from this cohort have been assessed for quantitative autistic trait using Baron-Cohen's Autism-Spectrum Quotient (AQ) test[40] and exomes of half of these participants (n = 615) have been sequenced, thus providing a greater resolution at the gene level. Individuals whose exome were sequenced were selected based on having good quality phenotype information on a wide range of topics, and therefore random with regards to AQ scores (**Supplementary Table S2**). ERF is a family-based cohort originating from 22 couples and spread over 23 generations.[41,42] The ERF study was approved by the Medical Ethics Committee of the Erasmus MC which is constituted according to the WMO (Wet Medisch-wetenschappelijk Onderzoek met mensen). A written informed consent was obtained from all study participants.

Sequencing was done at a mean depth of 74× using the Nimblegen SeqCap EZ V2 capture kit on an Illumina Hiseq2000 sequencer (Illumina, San Diego, CA, USA) using the TruSeq Version 3 protocol at the Human Genotyping facility of the Internal Medicine department, at the Erasmus MC, The Netherlands.[43,44] The sequence reads were aligned to the human genome build 19 (hg19), using Burrows-Wheeler Aligner and the NARWHAL pipeline.[45,46] After processing, genetic variants were called, using the Unified Genotyper tool from the GATK.[47] Variants with a low quality (QUAL < 150), which were out of HWE (p-value < 10^{-6}) or with low call rate (< 90%), as well as samples with a low call rate (< 90%), and duplicates, were removed.[44] Functional annotations were also performed using the SeattleSeq annotation 138 database (http://snp.gs.washington. edu/SeattleSeqAnnotation138/).

Association analyses of the quantitative autistic trait adjusted for age, gender and familial relationship were performed using the RVtests software. Meta-analysis of gene-based results of discovery sample and ERF study was performed using Fisher's combined probability test.

RESULTS

Results of genome-wide association analysis for the binary ASD phenotype and quantitative autistic trait are illustrated in **Supplementary Figures S3** and **S4**. No single-variant surpassed the genome-wide significance threshold. Top findings from the association analyses are shown in **Tables 1** and **2** for the discovery sample, and in **Supplementary Tables S3** and **S4** for simplex and multiplex families separately. Suggestive association of binary ASD phenotype was observed with two common variants (rs6452310; *p*-value = 7.8×10^{-8} and rs7700465; *p*-value = 8.70×10^{-6}; **Table 1**) at chromosome 5p14.1 - a region previously known to be associated with ASD (**Figure 1**). The two variants were in strong linkage disequilibrium (LD) ($r^2 = 0.85$) with each other but not in LD with any of the previously identified variants in this region associated with ASD[23,24] (r^2 ranged from 0.002 to 0.009, D' ranged from 0.02 to 0.33). None of the top variants from this analysis showed evidence of association in the replication sample (**Table 1**).

Results of gene-based genome-wide association analysis are illustrated in **Supplementary Figure S5** for the discovery sample, and in **Supplementary Table S5** for simplex and multiplex families separately. The gene-based association analysis revealed significant association of quantitative autistic trait and *TTC25* gene (*p*-value = 3.4×10^{-7}) on chromosome 17 (**Table 3**). This association was not driven by any single variant but by nine variants, four of which showed nominally significant association with quantitative autistic trait (*p*-value < 0.05) (**Supplementary Table S6**). The combined effect of these variants on the SRS score was large (effect size = 10.2). The functional annotation of nine variants revealed that five variants are likely to have regulatory functions (Category 1 RegulomeDB score; **Supplementary Table S6**). Furthermore, the surrounding variants in strong LD ($r^2 = 1$) with nine variants lie in enhancer histone marks, and protein-binding regions and change regulatory motifs based on the variant allele changes (**Supplementary Table S7**). The gene was also nominally associated with autistic trait in the ERF cohort (*p*-value = 0.045). The combined effect of 15 variants in the *TTC25* gene (**Supplementary Table S8**) in the replication sample was much smaller (effect size = 1.75). Meta-analysis of discovery and replication samples resulted in an improved association signal (*p*-value = 1.5×10^{-8}).

Table 1. Top findings (p-value < 1×10^{-5}) from the binary ASD phenotype association analysis

SNP	Chr	Position[a]	Function	CADD	Gene	Reference Allele	Variant Allele	Discovery cohort						PGC	
								N	Call Rate	MAF	HWE	LOD	P	OR	P
rs6452310	5	25996104	intergenic	0.58	*MSNP1AS*	G	A	654	1	0.32	0.9	6.27	7.80×10^{-8}	0.996	0.9
rs72996425	18	59451206	intergenic	5.32	*RFN125/CHD20*	A	C	654	1	0.03	1	4.37	7.20×10^{-6}	0.96	0.7
rs7147213	14	52932731	intron	2.13	*TXNDC16*	A	G	653	0.998	0.12	1	4.32	8.20×10^{-6}	1.03	0.34
rs7700465	5	26003284	intergenic	3.03	*MSNP1AS*	G	A	654	1	0.28	0.89	4.29	8.70×10^{-6}	0.99	0.69
rs7142973	14	52926750	intron	1.02	*TXNDC16*	A	G	654	1	0.12	1	4.26	9.50×10^{-6}	1.04	0.36
rs4901281	14	52980190	intron	7.66	*TXNDC16*	C	A	654	1	0.12	1	4.26	9.50×10^{-6}	0.97	0.39
rs12435289	14	52981240	intron	3.54	*TXNDC16*	C	A	654	1	0.12	1	4.26	9.50×10^{-6}	0.97	0.37

Abbreviations: Chr - Chromosome, CADD - phred-like Combined Annotation Dependent Depletion scores from Kircher *et al*.,[70] University of Washington, N - number of individuals, MAF - minor allele frequency, HWE - Hardy-Weinberg equilibrium in founders, P - p-value; PGC - Psychiatric Genomics Consortium.

[a]Genomic position is according the hg19 assembly.

Table 2. Top findings (p-value < 1×10^{-5}) from the quantitative autistic trait association analysis

								Discovery cohort								PGC	
SNP	Chr	Position[a]	Function	CADD	Gene	Reference allele	Variant Allele	N	AF	Call Rate	HWE	U_ Stat	Sqrt_V_ Stat	Effect	P	OR	P
rs77782810	14	93169444	downstream	0.81	none	T	G	486	0.09	0.99	0.35	2.11	0.42	12.13	4.28×10^{-7}	1.03	0.54
rs75822891	12	551394	3-prime-UTR	7.37	*CCDC77*	A	C	486	0.02	1	1	1.11	0.23	21.81	8.78×10^{-7}	0.89	0.12
rs61867303	10	106594496	intron	3.92	*SORCS3*	G	A	486	0.04	1	0.03	1.44	0.30	16.34	1.19×10^{-6}	1.03	0.70
rs202183827	14	93162861	intergenic	1.14	none	G	A	486	0.10	0.99	0.12	2.11	0.45	10.25	3.30×10^{-6}	-	-
rs75820392	12	526569	intron	1.64	*CCDC77*	C	T	486	0.02	0.99	1	0.98	0.21	21.86	3.61×10^{-6}	1.08	0.44
rs76459666	14	93161327	intergenic	4.74	none	C	T	486	0.1	0.99	0.13	2.11	0.46	10.11	3.88×10^{-6}	1.01	0.85
rs12433198	14	78042307	intron	2.51	*SPTLC2*	G	A	486	0.08	1	0.52	1.81	0.40	11.22	6.49×10^{-6}	1.13	0.01
rs115601680	6	46378057	intron	0.12	*RCAN2*	C	T	486	0.01	1	1	0.60	0.13	32.90	8.99×10^{-6}	NA	NA
rs77988177	20	21972232	intergenic	2.47	none	G	A	486	0.03	1	0.27	1.06	0.24	18.52	9.18×10^{-6}	0.94	0.52

Abbreviations: Chr - Chromosome, CADD - phred-like Combined Annotation Dependent Depletion scores from Kircher et al., University of Washington, N - number of individuals, AF- allele frequency, HWE - Hardy-Weinberg equilibrium in founders, Effect - Effect of variant allele, P - p-value, PGC - Psychiatric Genomics Consortium. U and V statistics are score statistics and their covariance matrix.

[a]Genomic position is according the hg19 assembly.

Pathway analysis

Pathway enrichment analysis based on nominally significant genes in the gene-based association analysis (**Supplementary Table S9**) showed significantly enriched SMAD

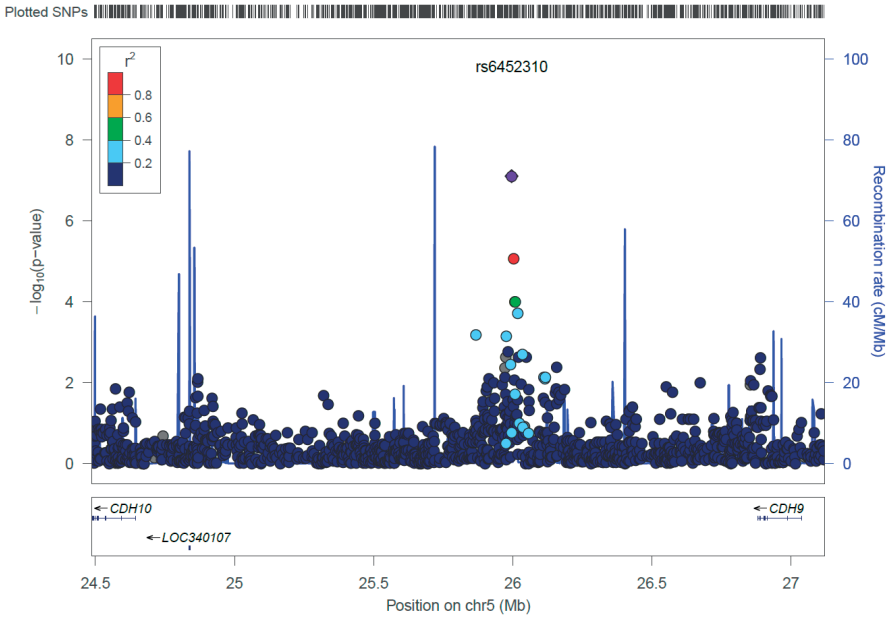


Figure 1. Regional association and recombination rate plot of the 5p14.1 region of the binary ASD phenotype association analysis in the discovery cohort. The left y-axis represents –log10 p-values for association with binary ASD phenotype in the discovery cohort. The right y-axis represents the recombination rate, and the x-axis represents chromosomal position (genomic position is according to the hg19 assembly). The most significantly associated single SNP in this region (rs6452310) is denoted with a purple diamond. Surrounding SNPs are shaded according to their pairwise correlation (r^2) with rs6452310. The gene annotations are shown below the figure.

Table 3. Top results (*p*-value < 1×10^{-3}) from the gene-based association analysis of quantitative autistic trait in the discovery cohort

Gene	Chr	N	NumVar	NumSite	AF*	Effect	P
TTC25	17	486	13	9	0.338	10.23	3.44×10^{-7}
SH2D5	1	486	9	5	0.440	-10.88	1.49×10^{-4}
PSMC5	17	486	4	3	0.458	-11.13	6.81×10^{-4}
LOC729177	6	486	22	14	0.434	-9.15	9.37×10^{-4}

Abbreviations: Chr - Chromosome, N - sample size, NumVar - Number of variants, NumSite - Number of sites, AF - allele frequency, Effect - effect size, P - *p*-value.
*Allele frequency is calculated as an average of individual allele frequency adjusted for relatedness (http://zhanxw.github.io/rvtests).

protein signal transduction (p-value = 2×10^{-5}) pathway and series of digestive system development enrichment categories (**Supplementary Table S10**).

DISCUSSION

In this study, we have identified a novel gene *TTC25* associated with quantitative autistic trait. The association of the gene *TTC25* in a cohort-based study suggests that the gene may be relevant for broader ASD phenotype in the general population. Further, we identified SMAD protein signal transduction pathway and series of digestive system development categories as being significantly enriched with genes nominally associated with quantitative autistic trait. Moreover, our study provides additional evidence for the previously identified association of the intergenic loci at 5p14.1 with the binary ASD phenotype.

TTC25 gene is located on chromosome 17q21.2 and encodes Tetratricopeptide Repeat Domain 25. 17q21.2 locus has previously been linked to ASD in a genome-wide linkage scan,[48,49] although a gene was never implicated. *TTC25* is overexpressed in testis, frontal cortex, and rectum (http://www.genecards.org/cgi-bin/carddisp.pl?gene=TTC25). *TTC25* is involved in cilium movement, organization and morphogenesis.[50] Cilia are specialized organelles protruding from the cell surface of almost all mammalian cells.[51] Mutations in ciliary proteins cause ciliopathies which can affect many organs at different levels of severity and are characterized by a wide spectrum of phenotypes.[51] In the vertebrate nervous system, the primary cilium is increasingly viewed as hub for certain neural developmental signalling pathways, and growing data suggest this is also true for several types of adult neuronal signalling.[52] The capacity of the brain to interpret the sensory input is often affected in ciliopathies, resulting in neurological disorders; cognitive impairment, anosmia, intellectual disability, ASD, and obesity are apparent in various degrees in many of the ciliopathies.[51,53] Further Joubert syndrome (JS) is a well-known ciliopathy of the central nervous system.[52,54] Features of ASD, such as problems in social behaviour, communication problems, and repetitive behaviours, have been described in up to 40% of JS patients[55–57] and about 25% of JS patients meet criteria for the DSM-IV diagnosis of ASD.[55,58] Multiple variants mapped to this gene in our sample appear to have a regulatory function.

Our identification of SMAD protein signal transduction pathway as being significantly enriched with genes nominally associated with quantitative autistic trait reinforces the role of the transforming growth factor-β (TGFβ) in ASD. The Smad pathways are the

major mediators of transcriptional responses induced by the TGFβ family, which control cell-fate determination, cell cycle arrest, apoptosis and actin rearrangements.[59] While decreased levels of TGFβ have been reported in blood samples from individuals with ASD[60] and associated with more severe behavioural scores in ASD children,[61] higher levels of TGFβ have been reported in postmortem brain and cerebrospinal fluid samples of ASD patients.[62] In addition, series of digestive system development categories were enriched with genes nominally associated with ASD. Gastrointestinal (GI) disturbances are 4-fold more common in ASD[63] and available scientific evidence supports combination of changes in the areas of immune function, gut microbiome and gut and brain signalling pathways.[64] Recent studies in animal models suggests that GI difficulties may originate from the same genetic changes that lead to the behavioural characteristics of ASD.[65]

In addition, we found the association of ASD with the known region on 5p14.1. This is one of the few replicated GWA regions that implicates long noncoding RNA gene, *MSNP1AS* (moesin pseudogene 1, antisense) in ASD risk.[66] This region has also been associated with social communication spectrum phenotypes in the general population supporting the role of 5p14.1 as a quantitative trait locus for ASD.[49] *MSNP1AS* shows a very high sequence homology to the chromosome X transcript of *MSN* that is involved in brain development.[66] *MSNP1AS* is highly overexpressed (12.7-fold) in the postmortem cerebral cortex of individuals with ASD.[66] Interestingly, our top hit did not replicate in the PGC sample and vice versa. The multiple different variants discovered in the 5p14.1 region[23,24] may suggest that multiple alleles in the same region are implicated in ASD.

Although our study sample was small, the strength of our study is that we used a data set with high-quality phenotypes, in which participants were assessed for both binary ASD phenotype and a quantitative autistic trait. The use of quantitative endophenotype provides additional power to find genetic signals by focusing on less complex aspects of complex phenotypes such as ASD.[67] The identification of new loci associated with quantitative autistic trait in our study validates this approach. Another strength is phenotypic homogeneity in the sense that the majority of patients in the current study had a normal intelligence, unlike most ASD cohorts with typical rates of intellectual disability ranging from 30 to 50%.[68] Further, the genomes of our study participants were genotyped on a very dense SNP array that contains not only common but also less-frequent SNPs. This gave us the opportunity to make a comprehensive overview of how common and less-frequent variants, both individually and taken together, affect ASD. Our study shows the advantage of a gene-based test as the more powerful approach compared to single-variant analysis and demonstrated the use of gene-based pathway and enrichment analysis in understanding the molecular mechanisms of the disorder. One of the

possible limitations of our study is that the assessment of quantitative autistic trait in the discovery and replication cohort used two different questionnaires, the SRS and the AQ. As the questionnaires have both been designed to measure the severity of social responsiveness problems across clinical cases and the general population and as their ratings are significantly correlated,[69] we were able to compare results from the two cohorts.

To conclude, our study has identified a novel gene *TTC25* to be associated with autistic trait in the ASD population where majority of patients have a normal intelligence. The replication of *TTC25* association in a cohort-based study suggests that this gene may also be relevant for broader ASD phenotype in the general population. However, whether these findings hold true also for ASD patients with intellectual disability remains to be evaluated. *TTC25* is overexpressed in frontal cortex and testis and is known to be involved in cilium movement and thus an interesting candidate gene for autistic trait. Furthermore, we discovered significantly enriched SMAD protein signal transduction pathway and series of digestive system development categories in the pathway analysis of quantitative autistic trait. Our finding provides new insights into the genetic background of quantitative autistic trait.

REFERENCES

1. Association, A.P. Diagnostic and statistical manual of mental disorders (5th ed.), (Arlington, VA: American Psychiatric Publishing, 2013).
2. Constantino, J.N. & Todd, R.D. Autistic traits in the general population: a twin study. *Arch Gen Psychiatry* **60**, 524-30 (2003).
3. Devlin, B. & Scherer, S.W. Genetic architecture in autism spectrum disorder. *Curr Opin Genet Dev* **22**, 229-37 (2012).
4. Elsabbagh, M. *et al*. Global prevalence of autism and other pervasive developmental disorders. *Autism Res* **5**, 160-79 (2012).
5. Werling, D.M. & Geschwind, D.H. Sex differences in autism spectrum disorders. *Curr Opin Neurol* **26**, 146-53 (2013).
6. Hallmayer, J. *et al*. Genetic heritability and shared environmental factors among twin pairs with autism. *Arch Gen Psychiatry* **68**, 1095-102 (2011).
7. Bailey, A. *et al*. Autism as a strongly genetic disorder: evidence from a British twin study. *Psychol Med* **25**, 63-77 (1995).
8. Sandin, S. *et al.* The familial risk of autism. *JAMA* **311**, 1770-7 (2014).
9. Gauglerl, T. et al. Most genetic risk for autism resides with common variation. *Nature Genetics* **46**, 881-885 (2014).
10. Betancur, C. Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. *Brain Res* **1380**, 42-77 (2011).
11. Marshall, C.R. *et al*. Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet* **82**, 477-88 (2008).
12. Shen, Y. *et al*. Clinical genetic testing for patients with autism spectrum disorders. *Pediatrics* **125**, e727-35 (2010).
13. Pinto, D. *et al*. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368-72 (2010).
14. Gai, X. *et al*. Rare structural variation of synapse and neurotransmission genes in autism. *Mol Psychiatry* **17**, 402-11 (2012).
15. Sebat, J. *et al*. Strong association of de novo copy number mutations with autism. *Science* **316**, 445-9 (2007).
16. Girirajan, S. *et al*. Relative burden of large CNVs on a range of neurodevelopmental phenotypes. *PLoS Genet* **7**, e1002334 (2011).
17. Freitag, C.M. The genetics of autistic disorders and its clinical relevance: a review of the literature. *Mol Psychiatry* **12**, 2-22 (2007).
18. Sanders, S.J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-41 (2012).
19. O'Roak, B.J. *et al*. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-50 (2012).
20. O'Roak, B.J. *et al*. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet* **43**, 585-9 (2011).
21. Samocha, K.E. *et al*. A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-50 (2014).
22. Cross-Disorder Group of the Psychiatric Genomics, C. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* **381**, 1371-9 (2013).

23. Wang, K. *et al*. Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* **459**, 528-33 (2009).
24. Ma, D. *et al*. A genome-wide association study of autism reveals a common novel risk locus at 5p14.1. *Ann Hum Genet* **73**, 263-73 (2009).
25. Anney, R. *et al*. Individual common variants exert weak effects on the risk for autism spectrum disorderspi. *Hum Mol Genet* **21**, 4781-92 (2012).
26. Anney, R. *et al*. A genome-wide scan for common alleles affecting risk for autism. *Hum Mol Genet* **19**, 4072-82 (2010).
27. Connolly, J.J., Glessner, J.T. & Hakonarson, H. A genome-wide association study of autism incorporating autism diagnostic interview-revised, autism diagnostic observation schedule, and social responsiveness scale. *Child Dev* **84**, 17-33 (2013).
28. Weiss, L.A. *et al.* A genome-wide linkage and association scan reveals novel loci for autism. *Nature* **461**, 802-8 (2009).
29. Xia, K. *et al*. Common genetic variants on 1p13.2 associate with risk of autism. *Mol Psychiatry* **19**, 1212-9 (2014).
30. Holt, R. & Monaco, A.P. Links between genetics and pathophysiology in the autism spectrum disorders. *EMBO Mol Med* **3**, 438-50 (2011).
31. Roeyers, H., Thys, M., Druart, C., De Schryver, M. & Schittekatte, M. SRS Screeningslijst voor autismespectrumstoornissen. (Hogrefe Uitgevers B.V., Amsterdam, 2011).
32. Noens, I., De la Marche, W. & Scholte, E. SRS Screeningslijst voor autismespectrumstoornissen bij volwassenen. Amsterdam: Hogrefe Uitgevers B.V. (2012).
33. Mukhopadhyay, N., Almasy, L., Schroeder, M., Mulvihill, W.P. & Weeks, D.E. Mega2: data-handling for facilitating genetic linkage and association analyses. *Bioinformatics* **21**, 2556-7 (2005).
34. Wigginton, J.E. & Abecasis, G.R. PEDSTATS: descriptive statistics, graphics and quality assessment for gene mapping data. *Bioinformatics* **21**, 3445-7 (2005).
35. Li, M., Boehnke, M. & Abecasis, G.R. Joint modeling of linkage and association: identifying SNPs responsible for a linkage signal. *Am J Hum Genet* **76**, 934-49 (2005).
36. Purcell, S. *et al*. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J of Hum Genet* **81**, 559-575 (2007).
37. Li, B. & Leal, S.M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* **83**, 311-21 (2008).
38. Boyle, A.P. *et al*. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research* **22**, 1790-1797 (2012).
39. Ward, L.D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Research* **40**, D930-D934 (2012).
40. Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J. & Clubley, E. The autism-spectrum quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *J Autism Dev Disord* **31**, 5-17 (2001).
41. Aulchenko, Y.S. *et al*. Linkage disequilibrium in young genetically isolated Dutch population. *Eur J Hum Genet* **12**, 527-34 (2004).
42. Pardo, L.M., MacKay, I., Oostra, B., van Duijn, C.M. & Aulchenko, Y.S. The effect of genetic drift in a young genetically isolated population. *Ann Hum Genet* **69**, 288-95 (2005).
43. Vojinovic, D. *et al*. The dystrophin gene and cognitive function in the general population. *Eur J Hum Genet* **23**, 837-43 (2015).

44. Amin, N. *et al*. Exome-sequencing in a large population-based study reveals a rare Asn396Ser variant in the LIPG gene associated with depressive symptoms. *Mol Psychiatry* (2016).
45. Brouwer, R.W., van den Hout, M.C., Grosveld, F.G. & van Ijcken, W.F. NARWHAL, a primary analysis pipeline for NGS data. *Bioinformatics* **28**, 284-5 (2012).
46. Bucan, M. *et al*. Genome-wide analyses of exonic copy number variants in a family-based study point to novel autism susceptibility genes. *PLoS Genet* **5**, e1000536 (2009).
47. McKenna, A. *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-303 (2010).
48. McCauley, J.L. *et al*. Genome-wide and Ordered-Subset linkage analyses provide support for autism loci on 17q and 19p with evidence of phenotypic and interlocus genetic correlates. *BMC Med Genet* **6**, 1 (2005).
49. Lowe, J.K., Werling, D.M., Constantino, J.N., Cantor, R.M. & Geschwind, D.H. Social responsiveness, an autism endophenotype: genomewide significant linkage to two regions on chromosome 8. *Am J Psychiatry* **172**, 266-75 (2015).
50. Xu, Y. *et al*. Characterization of tetratricopeptide repeat-containing proteins critical for cilia formation and function. *PLoS One* **10**, e0124378 (2015).
51. D'Angelo, A. & Franco, B. The dynamic cilium in human diseases. *Pathogenetics* **2**, 3 (2009).
52. Louvi, A. & Grove, E.A. Cilia in the CNS: the quiet organelle claims center stage. *Neuron* **69**, 1046-60 (2011).
53. Basten, S.G. & Giles, R.H. Functional aspects of primary cilia in signaling, cell cycle and tumorigenesis. *Cilia* **2**, 6 (2013).
54. Doherty, D. Joubert syndrome: insights into brain development, cilium biology, and complex disease. *Semin Pediatr Neurol* **16**, 143-54 (2009).
55. Ozonoff, S., Williams, B.J., Gale, S. & Miller, J.N. Autism and autistic behavior in Joubert syndrome. *J Child Neurol* **14**, 636-41 (1999).
56. Artigas-Pallares, J., Gabau-Vila, E. & Guitart-Feliubadalo, M. [Syndromic autism: II. Genetic syndromes associated with autism]El autismo sindromico: II. Sindromes de base genetica asociados a autismo. *Rev Neurol* **40** Suppl 1, S151-62 (2005).
57. Holroyd, S., Reiss, A.L. & Bryan, R.N. Autistic features in Joubert syndrome: a genetic disorder with agenesis of the cerebellar vermis. *Biol Psychiatry* **29**, 287-94 (1991).
58. Alvarez Retuerto, A.I. *et al*. Association of common variants in the Joubert syndrome gene (AHI1) with autism. *Hum Mol Genet* **17**, 3887-96 (2008).
59. Wrana, J.L. Signaling by the TGFbeta superfamily. *Cold Spring Harb Perspect Biol* **5**, a011197 (2013).
60. Okada, K. *et al*. Decreased serum levels of transforming growth factor-beta1 in patients with autism. *Prog Neuropsychopharmacol Biol Psychiatry* **31**, 187-90 (2007).
61. Ashwood, P. *et al*. Decreased transforming growth factor beta1 in autism: a potential link between immune dysregulation and impairment in clinical behavioral outcomes. *J Neuroimmunol* **204**, 149-53 (2008).
62. Vargas, D.L., Nascimbene, C., Krishnan, C., Zimmerman, A.W. & Pardo, C.A. Neuroglial activation and neuroinflammation in the brain of patients with autism. *Ann Neurol* **57**, 67-81 (2005).
63. McElhanon, B.O., McCracken, C., Karpen, S. & Sharp, W.G. Gastrointestinal symptoms in autism spectrum disorder: a meta-analysis. *Pediatrics* **133**, 872-83 (2014).
64. Coury, D.L. *et al*. Gastrointestinal conditions in children with autism spectrum disorder: developing a research agenda. *Pediatrics* **130** Suppl 2, S160-8 (2012).
65. Margolis, K.G. *et al*. Serotonin transporter variant drives preventable gastrointestinal abnormalities in development and function. *J Clin Invest* (2016).

66. Kerin, T. *et al*. A noncoding RNA antisense to moesin at 5p14.1 in autism. *Sci Transl Med* **4**, 128ra40 (2012).
67. Almasy, L. The role of phenotype in gene discovery in the whole genome sequencing era. *Hum Genet* **131**, 1533-40 (2012).
68. Geschwind, D.H. Advances in autism. *Annu Rev Med* 60, 367-80 (2009).
69. Armstrong, K. & Iarocci, G. Brief report: the autism spectrum quotient has convergent validity with the social responsiveness scale in a high-functioning sample. *J Autism Dev Disord* **43**, 2228-32 (2013).
70. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J: A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310–315 (2014).

Chapter 4.2

STXBP5 Antisense RNA 1 gene and adult ADHD symptoms

Alejandro Arias-Vásquez*, Alexander J. Groffen*, Sabine Spijker*, Klaasjan G. Ouwens*, Marieke Klein*, Dina Vojinovic*, Tessel E. Galesloot, Janita Bralten, Jouke-Jan Hottenga, Peter J. van der Most, V. Mathijs Kattenberg, Rene Pool, Ilja M. Nolte, Brenda W.J.H. Penninx, Iryna O. Fedko, Conor V. Dolan, Michel G. Nivard, Anouk den Braber, Cornelia M. van Duijn, Pieter J. Hoekstra, Jan K. Buitelaar, Bart Kiemeney, Martine Hoogman, Christel M. Middeldorp, Harmen H.M. Draisma, Sit H. Vermeulen, Cristina Sánchez-Mora, J. Antoni Ramos-Quiroga, Marta Ribasés, The EAGLE-ADHD Consortium, Catharina A. Hartman, J.J. Sandra Kooij, Najaf Amin, August B. Smit**, Barbara Franke**, Dorret I. Boomsma**

* These authors contributed equally to this work
** These authors shared final responsibility

This chapter is submitted.

The supplemental information for this paper is available at https://drive.google.com/drive/folders/1POG-CJ_RV52bSRwAtbc4dOJFfM7RAYzs?usp=sharing

ABSTRACT

Attention-deficit/hyperactivity disorder (ADHD) is characterized by age-inappropriate levels of inattention and/or hyperactivity-impulsivity and persists into adulthood in a substantial proportion of cases. ADHD is heritable and is thought to represent the clinical extreme of a continuous distribution of ADHD symptoms in the general population. We aimed to detect ADHD risk conferring genes leveraging the power of population studies of ADHD symptoms in adults. Within the SAGA (Study of ADHD trait Genetics in Adults) consortium, we estimated the SNP-based heritability of self-reported ADHD symptoms and carried out a genome-wide association meta-analysis in nine adult population-based and case-only cohorts of unrelated adults. A total of n = 14,689 individuals were included. We found a significant SNP-based heritability for self-rated ADHD symptom scores of respectively 15% (n = 3,656) and 30% (n = 1,841) in the two cohorts. The top-hit of the genome-wide meta-analysis (SNP rs12661753) was present in the hitherto uncharacterized long non-coding RNA *STXBP5-AS1* gene. This association was also observed in a meta-analysis of childhood ADHD symptom scores in eight population-based pediatric cohorts from the EAGLE ADHD consortium (n = 14,776). Genome-wide meta-analysis of the SAGA and EAGLE data (n = 29,465) increased the strength of the association on the *STXBP5-AS1* gene. In human HEK293 cells, expression of *STXBP5-AS1* enhanced the expression of a reporter-construct of *STXBP5*, a gene known to be involved in SNARE complex formation. In mouse strains featuring different levels of impulsivity, *Stxbp5-AS1* transcript levels in the prefrontal cortex strongly correlated with motor impulsivity as measured in the 5-choice serial reaction time task (r^2 = 0.55). Our results implicate the *STXBP5-AS1* gene in ADHD symptom scores and point to vesicle transport as a biological mechanism involved in ADHD-related impulsivity levels.

INTRODUCTION

Attention-deficit/hyperactivity disorder (ADHD) is a common neurodevelopmental disorder affecting 2–5% of children[1,2] and adults.[3,4] ADHD is characterized by age-inappropriate, sustained symptoms of inattention and/or hyperactivity-impulsivity. In children[5] and adults,[6] ADHD shows substantial heritability. Heritability estimates are largely independent of the phenotypic measurement scale (i.e., categorical or continuous) in children; in adults, estimates are lower when using self-report rating scales.[3] Twin studies[5] suggest that etiological influences on ADHD symptoms are distributed throughout the population, consistent with a liability model.[7] Inattention and hyperactivity-impulsivity symptoms can be reliably assessed in population-based cohorts based on rating scales,[8] creating the possibility to collect large samples for gene-finding studies. The genetic contributions to ADHD in children and adults are complex, with multiple different genetic variants contributing to the disorder,[4] both common and rare.[3] Recently, 16 genome-wide associations have been established in an ADHD Genome-Wide Association Studies meta-analysis (GWASMA) of childhood case-control studies from the Psychiatric Genomics Consortium (PGC) and The Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH) and population-based samples from the EArly Genetics and Lifecourse Epidemiology (EAGLE) consortium.[9,10]

Here, we sought to leverage the power of population studies of ADHD symptoms in adults to detect disease-relevant genes. Within the SAGA (Study of ADHD trait Genetics in Adults) consortium, we estimated the SNP-based heritability of self-reported adult ADHD symptoms and subsequently carried out a GWASMA in nine cohorts of European Caucasian origin (n = 14,689 individuals, age 18 years or older), in whom adult self-reported ADHD symptom scores were available. These samples included six population-based cohorts, two clinical ADHD samples and one clinical cohort ascertained for depressive and anxiety disorders (to enrich the clinical extreme of the ADHD symptom continuum). The locus with the strongest statistical association was followed-up in a replication analysis of quantitative childhood ADHD symptom scores (n = 14,776) from the EAGLE consortium.[9] Genetic correlations were obtained between the PGC and the iPSYCH sample of children[10] and the SAGA sample of adults. Finally, we conducted gene-based tests for genes with SNPs showing a p-value < 1×10^{-6} in the meta-analysis, making use of the common SNPs from SAGA and rare variant data from the Erasmus Rucphen Family (ERF) study (see **Table 1**), one of the adult cohorts.

Functional follow-up studies downstream of gene-finding in ADHD, e.g. in model systems, to determine the biological relevance of a genetic finding, are scarce.[11] Core features of ADHD, inattention, hyperactivity, and impulsivity are well defined e.g. in mouse

models.[12] Here we carried out functional follow-up studies for the hitherto uncharacterized top-gene of the GWASMA in three mouse inbred strains with large differences in motor impulsivity derived from reaction time tasks, and in a human cell assay.

METHODS

ADHD symptom scores and study populations

ADHD symptom scores were assessed by three instruments (see **Table 1**) in nine cohorts (for a complete description of each sample please follow the references in **Table 1**): the ADHD-index of the Conners Adult ADHD Rating Scale[13] (CAARS ADHD-index; 12 items), the total scores of the DSM-IV ADHD Rating Scale (ADHD-RS),[14] and the Attentional Deficit/Hyperactivity Problems subscale from the Adult Self Report (ASR-ADHD; 13 items).[15] The CAARS (used in NESDA, NTR, ERF) is an extensively tested psychometric instrument with high internal consistency and reliability. Five cohorts (NeuroIMAGE, BIG, IMpACT-NL, VHIR, NBS) collected information using the ADHD-RS,[14] which has high validity in population-based and case samples. For IMpACT-NL and VHIR, only affected individuals were included. One cohort (TRAILS) assessed ADHD problems through the ASR-ADHD (http://www.aseba.org/).[15-17]

Genetic Variant Calling and Quality Control

An overview of genome-wide single nucleotide polymorphism (SNP; for common variants) genotyping, quality control, and imputation is given in **Supplementary Table 1**. Exomes of 1,336 individuals from the ERF population, which is a genetically isolated population in the Netherlands,[18] were sequenced (see **Supplementary Methods**), and ADHD index data were available for 587 of these individuals. Detection of rare variants in the ERF study was done for those genes with SNPs with p-value $< 1\times10^{-5}$ in the GWASMA and variants identified in these exomes were used to estimate the contribution of rare variants in the genes of interest to ADHD behavior (see **Supplementary Methods**).

GCTA

Genome-wide Complex Trait Analysis (GCTA)[19] was used to compute the variance in the ADHD symptom score explained by common SNPs in the two largest cohorts included in the meta-analysis, the NTR and NESDA (n > 1,500 unrelated subjects). A genetic relationship matrix (GRM) for all individuals in the dataset was estimated based on SNPs with high imputation quality (see **Supplementary Methods**). Bivariate GCTA[19] was additionally run on the ADHD-index of the CAARS and ASR-ADHD data also available in the NTR cohort, to assess the genetic correlation (rg) between the two diagnostic instruments.

Genome-wide association and meta-analysis

GWAS was conducted in each cohort by linear regression under an additive model. Age was included as a covariate, but not gender, which was not significantly associated with the ADHD scores in any study. Four principal components were added to account for possible population stratification effects. Information on software packages is provided in **Supplementary Table 1**. In all analyses, the uncertainty of the imputed genotypes was taken into account. Location of SNPs reported is from the build 37 (hg19) 1000G data. Meta-analysis was conducted in METAL (www.sph.umich.edu/csg/abecasis/ metal/ index.html) by the *p*-value-based method, given the intrinsic variability of the quantitative traits used (see **Supplementary Methods**). The meta-analysis was performed in the full sample (nine cohorts) and restricted to the population-based samples (seven cohorts; "restricted sample").

Replication in the EAGLE consortium

Within EAGLE, association of ADHD-related measures was assessed in nine population-based childhood cohorts with genotype data imputed against the 1000 Genomes reference panel.[9] Linear regression of the phenotype on sex, age, genotype dose, and principal components was performed in all cohorts, followed by meta-analysis based on *p*-values in METAL. The TRAILS cohort is part of both consortia and was excluded from the EAGLE consortium for replication analysis, leaving a total of 14,776 children from eight cohorts.

Look-up of significant GWAS loci

Evidence for an effect of the 12 independent ADHD-associated SNPs from the PGC+iPSYCH GWASMA on adult ADHD symptoms was studied through a look-up of results. LD-independent loci with corresponding index-SNPs were obtained from **Table 1** of Demontis *et al*.[10] If the index variant was not present in the SAGA data set, a proxy variant was selected using LDlink (https://analysistools.nci.nih.gov/LDlink/). The Bonferroni-corrected significance level was set at p-value = 0.05/12 = 0.00417.

Linkage disequilibrium score regression (LDSR) analysis

LDSR was used to estimate the genetic correlation between the PGC+iPSYCH sample of children[10] and the SAGA sample of adults. Each dataset underwent additional filtering for markers overlapping with HapMap Project Phase 3 SNPs, INFO score ≥ 0.9 (where available), and MAF ≥ 1%. Indels and strand-ambiguous SNPs were removed. LDSR analysis was performed using the LDSR package (https://github.com/bulik/ldsc[20], see **Supplementary Methods**).

Gene-wide analysis of common and rare variants

Genes containing SNPs with *p*-values < 1×10^{-6} in the meta-analysis of the nine cohorts were selected for gene-wide tests using common and rare variants. The common variant analysis was performed in MAGMA.[21] Flanking regions of 25kb for each gene were included in the analyses. The rare variant analysis was performed with the Sequence Kernel Association Test (SKAT; only in the ERF study) library of the R-software.[22]

Functional analyses

Follow-up functional analyses were performed on the locus containing the best association *p*-value. This locus contains *STXBP5-AS1*, representing a putative long noncoding RNA, predicted to be expressed in several species (**Supplementary Methods**; **Supplementary Table 5** & **Supplementary Fig. 2**). Human *STXBP5-AS1* encodes multiple splice variants, many of which lack a region that overlaps the *STXBP5* gene. To test for regulatory effects of *STXBP5-AS1* on the expression of *STXBP5*, a fluorescent reporter construct was designed to contain the region of antisense overlap (see **Supplementary Methods**).

Mouse models

RNA was derived from prefrontal cortex of adult male mice from the inbred strain C57/Bl6J (n = 7) and recombinant inbred strains BXD29 (n = 8) and BXD68 (n = 7), and gene expression was quantitated (see **Supplementary Methods**). Strains were bred in the facility of the Neuro-Bsik consortium of the VU University (Amsterdam, The Netherlands) and used for behavioral analysis.[12,23]

RESULTS

Quantitative assessment instruments are listed in **Table 1**. The quantitative phenotypes showed a weak, negative correlation with age and no association with gender in any cohort. Phenotypic and genetic correlations between symptom scores assessed with the different instruments were substantial: in a clinical sample of 120 adults with ADHD the phenotypic correlation between the CAARS[13] (ADHD-index) and the ADHD-RS[24] (total score) was high (r = 0.73; *p*-value < 0.01).[24] In 380 parents of children with ADHD, the correlation was of similar magnitude (r = 0.69; *p*-value < 0.001).[25] We estimated the phenotypic correlation between the CAARS ADHD-index and the ASR-ADHD[15,16] in the NTR (n = 15,226; average age 40 years, SD=16.1) to be 0.67 (*p*-value < 0.0001). In younger participants in the age range of the TRAILS cohort (18–22 years, n = 2,687), the correlation was similar (0.68, *p*-value < 0.0001).

Table 1. Descriptive information for all cohorts and for phenotype assessment in the SAGA consortium.

Cohort Name	N (% F)	Age (SD)	Symptom list (N items)	Score range*	Mean Score (SD)*	Ref
NTR	5935 (63%)	43.7 (15.2)	CAARS ADHD-index (12)	0–30	7.9 (3.7)	(Willemsen et al. 2010)
NESDA	1977 (66%)	46.5 (13.0)	CAARS ADHD-index (12)	0–32	8.7 (5.4)	(Boomsma et al. 2008)
ERF	1043 (53%)	45.6 (13.3)	CAARS ADHD-index (12)	0–25	7.8 (4.4)	(Aulchenko et al. 2004)
NeuroIMAGE	470 (51%)	42.3 (5.3)	ADHD-RS (23)	0–43	14.1 (8.9)	(von Rhein et al. 2014)
BIG	448 (63%)	22.3 (3.2)	ADHD-RS (23)	0–40	14.0 (6.4)	(Hoogman et al. 2012)
NBS	2925 (53%)	57.4 (16.3)	ADHD-RS (23)	0–15	1.4 (2.2)	(Galesloot et al. 2017)
IMpACT-NL±	113 (62.8%)	37.7 (11.5)	ADHD-RS (23)	1–18	12.04 (3.3)	(Franke and Reif 2013)
VHIR±	559 (32%)	33.3 (10.6)	ADHD-RS (18)	4–54	31.0 (9.7)	(Bosch R 2019)
TRAILS	1215 (48%)	19.0 (0.6)	ASR ADHD (13)	0–22	5.9 (4.4)	(Ormel et al. 2014)

Conners' Adult ADHD Rating Scale (CAARS ADHD-index), DSM-IV ADHD Rating Scale (ADHDRS), and Attentional Deficit/Hyperactivity Problems subscale from the ASR (ASR ADHD);
*Untransformed values observed per cohort;
± only affected individuals included.

A significant SNP-based heritability was estimated for the CAARS ADHD-index in a subsample of each of the two largest cohorts: 30% (SE = 16.7%, *p*-value = 0.035) in NESDA (n = 1,841 unrelated subjects) and 15% (SE = 7.8%, *p*-value = 0.020) in NTR (n = 3,881 unrelated participants). We also estimated the genetic correlation for the CAARS ADHD-index and the ASR-ADHD using bivariate GCTA. In all individuals from the NTR with genotype and phenotype data (n = 6,036 related and unrelated subjects), the genetic correlation was 0.818 (SE = 0.256). When analyzing the bivariate data in 2,921 unrelated subjects, the point estimate of the genetic correlation was 0.813 (SE = 0.364). The significant SNP-based heritability and the considerable phenotypic and genetic correlations between assessment instruments support the validity of our meta-analysis approach of GWA results obtained across contributing data sets.

For the nine separate GWAS, the genomic control inflation factors (lambda) ranged between 0.996 and 1.026 (mean lambda 1.009, **Supplementary Table 2**). Meta-analysis (**Supplementary Figure 1A**) of the full sample revealed the lowest *p*-value (3.03×10^{-7}) for the intronic SNP rs12661753 in *STXBP5-AS1* (**Supplementary Figure 3**); for the meta-analysis of the restricted sample, *p*-value for this SNP was 1.48×10^{-6} (**Supplementary Figure 1B**). Replication was observed for rs12661753 (*p*-value = 3.07×10^{-2}) for childhood

ADHD symptoms in the EAGLE-ADHD consortium.[9] The subsequent GWASMA between SAGA and EAGLE revealed the best association *p*-value = 2.05×10^{-7} for SNP rs12664716 (n = 29,465; **Supplementary Figure 3F**) located in the *STXBP5-AS1* gene, and in high LD (D' = 1.0, r^2 = 0.98) with rs12661753 (p-value $_{SAGA-EAGLE}$ = 3.55×10^{-7}; **Supplementary Figure 1C**).

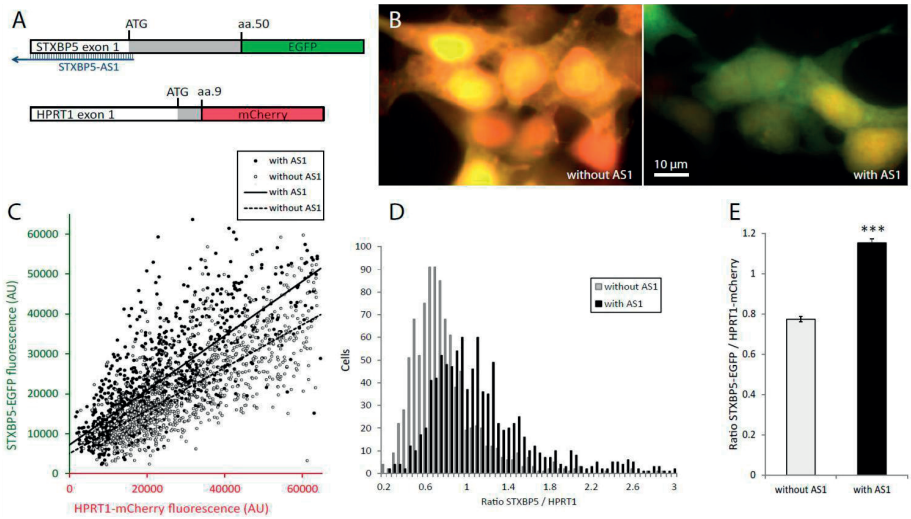


Figure 1. STXBP5-AS1 positively regulates the expression of its cognate mRNA. (A) Design of two reporter constructs. Top: Exon 1 of human *STXBP5*, containing the natural 5'UTR and encoding the first 50 amino acids, was fused in-frame to EGFP. The *Stxbp5-AS1* transcript including the region showing perfect (100%) sequence overlap with the encoded *Stxbp5* transcript is depicted schematically in blue. Bottom: To control for transfection efficiency and differences in cellular metabolic rates, we co-expressed a non-target mRNA comprised of human *HPRT1* exon 1 fused to mCherry. (B) Typical examples of HEK293 cells expressing both constructs with or without STXBP5-AS1. (C) Quantitation of EGFP and mCherry fluorescence in presence or absence of AS1 (947 and 974 cells respectively). (D-E) The ratio of STXBP5-EGFP and HPRT1-mCherry was calculated for each cell. Data are presented as a histogram (D) or as mean±SEM (E). ***, *p*-value = 6×10^{-51}; t_{946} = 4.4412, Student's t-test.

The index variant rs12661753 was not associated with ADHD risk in the recent case-control PGC+iPSYCH GWASMA of ADHD in a sample mainly consisting of children (*p*-value = 0.6316, n = 55,374). A look-up of genome-wide significant ADHD index SNPs from this PGC+iPSYCH GWASMA for association in the SAGA consortium also revealed no significant associations with adult ADHD symptoms (**Supplementary Table 6**).

We estimated the genetic correlation between PGC+iPSYCH and the complete SAGA sample to be 0.541 (SE = 0.447, *p*-value = 2.26×10^{-1}; the VHIR cohort present in both studies). We tested if the two rg values differed significantly from each other, which was not the case (X^2-based test *p*-value > 0.05) (**Supplementary Methods**).

In NTR and NESDA, a subset of participants (n = 6,678) had additional phenotype data on hyperactivity/impulsivity and inattention symptom subscales of the CAARS available. These scales of each 9 items are non-overlapping with the 12 ADHD-index items. For hyperactivity/impulsivity symptoms, the *p*-value for association with rs12661753 was 1.51×10^{-5}, whereas for inattention it was 3.53×10^{-2}, suggesting a stronger effect of the variant on hyperactivity/impulsivity.

As shown in **Table 2** and **Supplementary Table 3**, 50 common variants from 8 independent (clumped) loci showed *p*-value < 1×10^{-6}. Of these, four were also amongst the top-associated loci from the restricted SAGA GWASMA (no patients; **Supplementary Table 4**). The genes closest to these SNPs were selected for gene-wide analysis (**Table 2**). Analysis of common variants in seven genes (plus 25kb flanking regions) in the SAGA GWASMA showed significant association with ADHD symptoms. Two significant findings (*p*-value < 0.003) were for long intergenic non-protein coding RNA genes (*LINC01247, LINC00534*), and nominal significant associations (*p*-value < 0.05 gene-wide) were seen for *STXBP5-AS1*, *CALB1*, *GNG12-AS1*, *STXBP5* (**Supplementary Table 5**). It is important to note that *STXBP5* and *STXBP5-AS1* have no physical separation, thus their 25kb flanking regions overlap. The rare variant analysis also showed nominal association for *STXBP5*. For four genes (*GNG12-AS1*, *LINC01247, STXBP5-AS1*, *LINC00534*), rare variants were not observed/detected (**Supplementary Table 5**).

Table 2. Most strongly associated (clumped) SNPs (*p*-value < 1×10^{-6}) coming from the meta-analysis of nine cohorts from the SAGA consortium in physical position order (hg19).

SNP name	Chr	Locus	Pos*	*p*-value	Tested/ Non-Tested Allele	Frequency Tested Allele#	Gene(s) in locus
rs11209188	1	1p31.3	68455306	7.88×10^{-6}	A/G	0.534	*GNG12-AS1*
rs1930272	1	1p31.1	83491910	4.75×10^{-6}	T/C	0.544	*LOC107985037*
rs1564034	2	2p25.2	6510305	2.15×10^{-6}	T/G	0.670	*LINC01247*
rs28734069	4	4q26	120042409	5.77×10^{-6}	T/C	0.016	*LOC102723967; LOC105377395*
rs12661753	6	6q24.3	147409235	3.02×10^{-7}	A/G	0.962	*STXBP5-AS1*
rs13274695	8	8p23.2	3723378	6.00×10^{-6}	A/G	0.077	*CSMD1; LOC105377790*
rs2189255	8	8q21.3	91190297	9.61×10^{-6}	T/C	0.703	*CALB1; LINC00534*
rs73204517	13	13q21.33	69920315	7.19×10^{-6}	C/G	0.126	Downstream *LINC00383*

*bp position based on the GRCh37.p13 build;

#Allele frequency of tested allele based on N = 14,689.

Given that the *STXBP5-AS1* gene, which contains the top-hits, is hitherto uncharacterized, we investigated its function. *STXBP5-AS1* encodes a long noncoding RNA (lncRNA). Although human *STXBP5-AS1* does not have any orthologues listed in the UniGene

database, it is conserved in primates and shows a modest conservation in rodents (**Supplementary Table 7** and alignment in **Supplemental Figure 2**). In the hg19 genome release annotation *STXBP5-AS1* is located next to *STXBP5* in the opposite orientation, with antisense sequence overlap in exon 1 of *STXBP5* (**Figure 1A**). It may be hypothesized that *STXBP5-AS1* affects *STXBP5* expression. For such natural antisense RNAs, both repression and positive effects on the expression of cognate genes have been described.[26,27] We tested this hypothesis by designing a reporter gene fusing exon 1 of *STXBP5* to *EGFP*, and quantifying its expression in human HEK293 cells. Expression of the antisense lncRNA variant *STXBP5-AS1-003* (containing the overlap with *Stxbp5*) caused an increase in the fluorescence ratio between *STXBP5-EGFP* and the control (**Figure 1B-E**).

Given the *in vitro* effects on STXBP5-EGFP protein expression, we tested the relationship between gene expression of mouse *Stxbp5* and/or *Stxbp5-AS1* and measures of behavioral impulsivity. We analyzed gene expression in medial prefrontal cortex of three mouse inbred strains previously described to have large differences in motor impulsivity.[12] Here, we confirmed the strain difference in motor impulsivity between the BXD68, BXD29, and C57BL/6J strains ($F_{2,20}$ = 6.91, *p*-value = 0.005), measured as premature responses in the 5-choice serial reaction time task. In addition, these strains showed differences in errors of omission ($_{F2,20}$ = 5.18, *p*-value = 0.015), but not attention ($F_{2,20}$ = 0.35, *p*-value = 0.771) **(Figure 2A**). In these mice, we detected expression of a mouse *Stxbp5-AS1* transcript in the prefrontal cortex by real-time quantitative PCR, which differed across strains ($F_{2,19}$ = 11.73; *p*-value < 0.001). This transcript showed lowest expression in the most highly impulsive strain, BXD68 (BXD68: 4.58±0.11, C57BL/6J: 5.25±0.14, BXD29: 5.19±0.07, $p\text{-value}_{\text{BXD68 vs C57BL/6J}}$ = 0.003, t_{13} = 3.73; $p\text{-value}_{\text{BXD68 vs BXD29}}$ < 0.001, t_{14} = 4.63) (**Figure 2B**). Expression of *Stxbp5* mRNA was not different between the three strains (BXD68:9.89±0.24; C57BL/6J: 9.83±0.10; BXD29: 9.99±0.23). These results suggest that the role of *STXBP5-AS1* plays in impulsivity is not due to influencing the level of the *STXBP5* transcript. Examining correlations between *Stxbp5-AS1* transcript level and impulsivity/inattention measures, we found a significant correlation with motor impulsivity (r^2 = 0.55; *p*-value = 8.26×10^{-5}, Bonferroni-corrected *p*-value < 0.0083) and a nominally significant association with attention, when measured as errors of omission[28] (r^2 = 0.1765; *p*-value = 5.16×10^{-2}), but not when measured as percentage correct responses (r^2 = 0.0862; *p*-value = 1.85×10^{-1}). Expression of *Stxbp5* did not correlate with these parameters (**Figure 2C**).

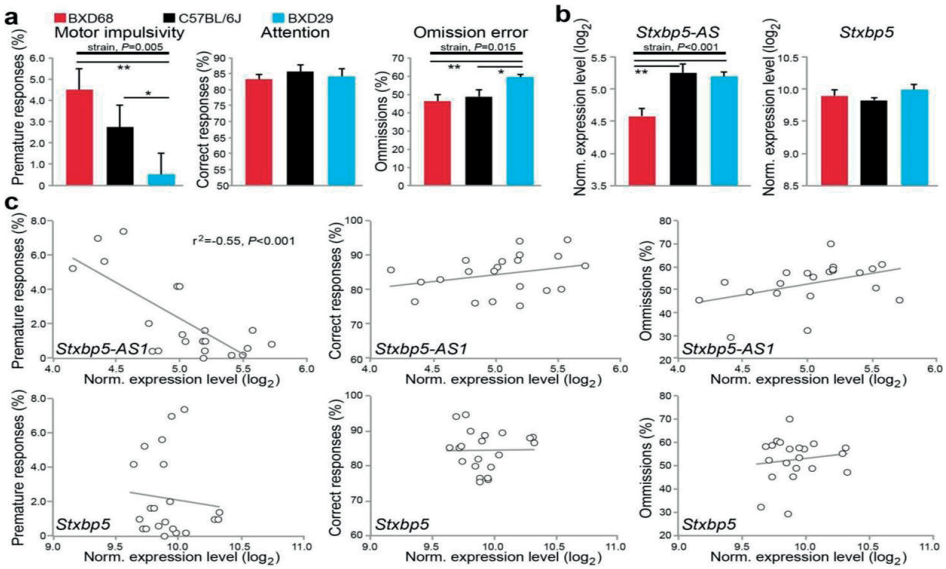


Figure 2. Prefrontal cortex gene expression of putative mouse *Stxbp5-AS1* is correlated with impulsivity. a. Mouse strains BXD68 (red, n = 7), C57BL/6J (black, n = 8) and BXD29 (blue, n = 8) were selected based on a difference in premature responses (motor impulsivity; BXD68 vs. BXD29, t_{14} = 3.71; C57BL/6J vs. BXD29, t_{15} = 2.78) and error of omissions (BXD68 vs. BXD29, t_{14} = 3.54; C57BL/6J vs. BXD29, t_{15} = 2.52), without being different on accuracy (Loos et al. 2014). Shown are data (mean±SEM) of the animals used for gene expression analysis (see b). b. Strain mean±SEM of prefrontal cortex gene expression in BXD68 (red, n=7), C57 (black, n=7), and BXD29 (blue, n=8) for *Stxbp5-AS1* (left) and *Stxbp5* (right). *Stxbp5-AS1* is differentially expressed between strains, with lower expression in BXD68. Yet, Stxbp5 shows no differential expression. No difference in variation was observed. c. Gene expression of *Stxbp5-AS1* (upper panels) in individual mice for which behavioral data was available (BXD68, n = 6; C57BL/6J, n = 7; BXD29, n = 8) correlated well with premature responses (motor impulsivity; left), not with accuracy (attention; middle), and showed a trend towards correlation with errors of omissions (right; *p*-value = 0.0516). *Stxbp5* expression (lower panels) did not correlate with any of these parameters. Trend lines are given in gray. **p*-value < 0.05; ***p*-value < 0.01; ****p*-value < 0.001.

DISCUSSION

We report a genetic variant associated with three different but correlated adult ADHD symptom lists in a meta-analysis of nine European adult population-based and case-only cohorts (n=14,689 individuals). The *STXBP5-AS1* gene (best SNP *p*-value = 3.02×10^{-7}) was the most strongly associated locus in a meta-analysis. This association was confirmed in the EAGLE meta-analysis ($p\text{-value}_{EAGLE} = 2.89\times10^{-2}$), and the top-hit from the full SAGA-EAGLE GWASMA was also located in the *STXBP5-AS1* gene and in almost perfect LD with the original finding (SNP rs12664716, $p\text{-value}_{SAGA\text{-}EAGLE} = 2.05\times10^{-7}$; n = 29,465).

For the adult ADHD-index, an earlier large twin family study estimated total heritability at 30%, and common SNPs thus contain substantial information concerning its genetic

variance. SNP-based heritability analyses, which were ran prior to GWASMA, provided estimates of 15-30% explained variance of adult ADHD symptom scores in the general population. Such estimates are comparable with the estimates obtained for ADHD and four additional categorically defined psychiatric disorders,[29] providing rationale for a gene-finding enterprise for adult ADHD symptoms in the general population.

The function of the *STXBP5-AS1*-encoded lncRNA is currently unknown. *STXBP5-AS1* has been proposed as a prognostic biomarker for survival of cancer patients,[30] but no information is available for its role in ADHD, related traits, or other psychiatric diseases. It overlaps in anti-sense with *STXBP5* encoding a protein involved in synaptic function by regulating neurotransmitter release through stimulating SNARE complex formation.[31,32] This complex plays a major role in intracellular vesicular trafficking in eukaryotic cells and is involved in the exocytotic release of neurotransmitters during synaptic transmission.[33] Genes related to the SNARE complex and its regulators have been investigated in ADHD,[34] and current results suggest that this complex may exert distinct roles throughout development, with age-specific effects of its genetic variants on ADHD behavior.[35] Specifically, deletions and mutations of *STXBP5* occur in autism[36] and epilepsy.[37] *STXBP5* has a presynaptic role that negatively regulates neurotransmitter release by forming syntaxin-SNAP25-tomosyn complex.[38] However, the postsynaptic role of STXBP5 has not been well elucidated.

Post-hoc analysis suggested that *STXBP5-AS1* affects hyperactivity-impulsivity more strongly than inattention. The stronger link with impulsivity was corroborated in behavioral studies in mice. Our experiments in HEK293 cells showed that the lncRNA does not cause antisense inhibition of *Stxbp5*. The increased fluorescence of a reporter protein containing mouse *Stxbp5* exon 1, together with unchanged *Stxbp5* mRNA levels in mouse strains expressing different *Stxbp5-AS1* levels, suggest that the lncRNA might enhance Stxbp5 protein translation or stability. Alternatively, *Stxbp5-AS1* might contribute to impulsivity by a Stxbp5-unrelated mechanism. In line with this idea, *Stxbp5-AS1*, expression (but not that of *Stxbp5*) correlated negatively with motor impulsivity in mice.

Our study should be viewed in the light of some strengths and limitations. A pro was the sample size that could be achieved for quantitative data available through a population-based approach. Moreover, the functional analyses provided a very strong candidate associated with adult and childhood ADHD symptoms. A limitation of our study was the combination of three different phenotyping instruments, but given the strong phenotypic and genetic correlations between the instruments, this might not have reduced power substantially.

The genetic correlation of PCG+iPSYCH with SAGA should be interpreted carefully because the standard error is high. The fact that the PCG+iPSYCH/SAGA rg (0.54), did not differ from the published rg estimate between the PCG+iPSYCH GWASMA and a GWAS of the 23andMe sample (0.65, SE = 0.114)[10] is encouraging but not unexpected given the low power to detect a difference. The estimated genetic correlation between the 23andMe and PGC+iPSYCH analyses was significant but lower than the genetic correlation of the EAGLE and PCG+iPSYCH childhood cohorts (rg = 0.943, SE = 0.204, *p*-value = 3.65×10^{-6}).[10] The ADHD diagnosis (yes/no) in 23andMe is based on the self-reported answer to a single question about presence of a lifetime diagnosis of ADHD[10] and we do not know if the 23andMe participants were diagnosed in childhood or as adults. With a further increase in GWAS sample size update rg results could suggest that there are different genetic correlation patterns between the association results estimated from the GWAS of adult (population-based) ADHD behavior and the GWAS from children, at this point the lack of power makes these analyses inconclusive.

Our study shows that self-reported adult ADHD symptoms measured in the general population have a genetic component and that performing population-based GWASMA of adult ADHD symptoms provides novel insights into the genetic underpinnings of hyperactivity/impulsivity symptoms that are a hallmark of ADHD. Our findings implicate synaptic function regulation through *STXBP5-AS1* and potentially *STXBP5* in ADHD symptom etiology.

REFERENCES

1. Polanczyk, G. & Rohde, L.A. Epidemiology of attention-deficit/hyperactivity disorder across the lifespan. *Curr Opin Psychiatry* **20**, 386-92 (2007).
2. Association, A.P. *Diagnostic and statistical manual of mental disorders*, (Washington, DC, 1994).
3. Franke, B. *et al.* The genetics of attention deficit/hyperactivity disorder in adults, a review. *Mol Psychiatry* **17**, 960-87 (2012).
4. Faraone, S.V. *et al.* Attention-deficit/hyperactivity disorder. *Nat Rev Dis Primers* **1**, 15020 (2015).
5. Faraone, S.V. *et al.* Molecular genetics of attention-deficit/hyperactivity disorder. *Biol Psychiatry* **57**, 1313-23 (2005).
6. Saviouk, V. *et al.* ADHD in Dutch adults: heritability and linkage study. *Am J Med Genet B Neuropsychiatr Genet* **156B**, 352-62 (2011).
7. Caspi, A. *et al.* A replicated molecular genetic basis for subtyping antisocial behavior in children with attention-deficit/hyperactivity disorder. *Arch Gen Psychiatry* **65**, 203-10 (2008).
8. Larsson, H. *et al.* Genetic and environmental influences on adult attention deficit hyperactivity disorder symptoms: a large Swedish population-based study of twins. *Psychol Med* **43**, 197-207 (2013).
9. Middeldorp, C.M. *et al.* A Genome-Wide Association Meta-Analysis of Attention-Deficit/Hyperactivity Disorder Symptoms in Population-Based Pediatric Cohorts. *J Am Acad Child Adolesc Psychiatry* **55**, 896-905 e6 (2016).
10. Demontis, D. *et al.* Discovery of the first genome-wide significant risk loci for ADHD. *bioRxiv* (2017).
11. Klein, M. *et al.* Brain imaging genetics in ADHD and beyond - Mapping pathways from gene to disorder at different levels of complexity. *Neuroscience and Biobehavioral Reviews* **80**, 115-155 (2017).
12. Loos, M. *et al.* Neuregulin-3 in the Mouse Medial Prefrontal Cortex Regulates Impulsive Action. *Biological Psychiatry* **76**, 648-655 (2014).
13. Conners, C.K. Clinical use of rating scales in diagnosis and treatment of attention-deficit/hyperactivity disorder. *Pediatric Clinics of North America* **46**, 857 (1999).
14. Kooij, J.J.S. *et al.* Reliability, Validity, and Utility of Instruments for Self-Report and Informant Report Concerning Symptoms of ADHD in Adult Patients. *Journal of Attention Disorders* **11**, 445-458 (2008).
15. Kessler, R.C. *et al.* Validity of the World Health Organization Adult ADHD Self-Report Scale (ASRS) Screener in a representative sample of health plan members. *International Journal of Methods in Psychiatric Research* **16**, 52-65 (2007).
16. Kessler, R.C. *et al.* The World Health Organization adult ADHD self-report scale (ASRS): a short screening scale for use in the general population. *Psychological Medicine* **35**, 245-256 (2005).
17. Achenbach, T.M. & Rescorla, L.A. *Manual for the ASEBA Adult Forms & Profiles*, (Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families, 2003).
18. Aulchenko, Y.S. *et al.* Linkage disequilibrium in young genetically isolated Dutch population. *Eur J Hum Genet* **12**, 527-34 (2004).
19. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
20. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).

21. de Leeuw, C.A., Mooij, J.M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *Plos Computational Biology* **11** (2015).
22. Wu, M.C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**, 82-93 (2011).
23. Spijker, S. *et al.* Morphine exposure and abstinence define specific stages of gene expression in the rat nucleus accumbens. *FASEB J* **18**, 848-50 (2004).
24. Sandra Kooij, J.J. *et al.* Reliability, validity, and utility of instruments for self-report and informant report concerning symptoms of ADHD in adult patients. *J Atten Disord* **11**, 445-58 (2008).
25. Thissen, A.J., Rommelse, N.N., Altink, M.E., Oosterlaan, J. & Buitelaar, J.K. Parent-of-origin effects in ADHD: distinct influences of paternal and maternal ADHD on neuropsychological functioning in offspring. *J Atten Disord* **18**, 521-31 (2014).
26. Kimura, T. *et al.* Stabilization of human interferon-alpha1 mRNA by its antisense RNA. *Cell Mol Life Sci* **70**, 1451-67 (2013).
27. Matsui, K. *et al.* Natural antisense transcript stabilizes inducible nitric oxide synthase messenger RNA in rat hepatocytes. *Hepatology* **47**, 686-97 (2008).
28. Guillem, K. *et al.* Nicotinic acetylcholine receptor beta2 subunits in the medial prefrontal cortex control attention. *Science* **333**, 888-91 (2011).
29. Lee, S.H. *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics* **45**, 984 (2013).
30. Guo, W. *et al.* Transcriptome sequencing uncovers a three-long noncoding RNA signature in predicting breast cancer survival. *Sci Rep* **6**, 27931 (2016).
31. Sakisaka, T. *et al.* Dual inhibition of SNARE complex formation by tomosyn ensures controlled neurotransmitter release. *J Cell Biol* **183**, 323-37 (2008).
32. Yizhar, O. *et al.* Tomosyn inhibits priming of large dense-core vesicles in a calcium-dependent manner. *Proc Natl Acad Sci U S A* **101**, 2578-83 (2004).
33. Antonucci, F. *et al.* SNAP-25, a Known Presynaptic Protein with Emerging Postsynaptic Functions. *Front Synaptic Neurosci* **8**, 7 (2016).
34. Bonvicini, C., Faraone, S.V. & Scassellati, C. Attention-deficit hyperactivity disorder in adults: A systematic review and meta-analysis of genetic, pharmacogenetic and biochemical studies. *Mol Psychiatry* **21**, 872-84 (2016).
35. Cupertino, R.B. *et al.* SNARE complex in developmental psychiatry: neurotransmitter exocytosis and beyond. *Journal of Neural Transmission* **123**, 867-883 (2016).
36. Davis, L.K. *et al.* Novel copy number variants in children with autism and additional developmental anomalies. *Journal of Neurodevelopmental Disorders* **1**, 292-301 (2009).
37. Dhillon, S., Hellings, J.A. & Butler, M.G. Genetics and mitochondrial abnormalities in autism spectrum disorders: a review. *Curr Genomics* **12**, 322-32 (2011).
38. Sakisaka, T. *et al.* Regulation of SNAREs by tomosyn and ROCK: implication in extension and retraction of neurites. *J Cell Biol* **166**, 17-25 (2004).

SUPPLEMETARY METHODS AND TABLES

Supplementary Methods

Supplementary Table 1. Quality control, filtering, pre-imputation, and imputation algorithms used.

Supplementary Table 2. Number of SNPs tested after imputation and the lambda values per sample.

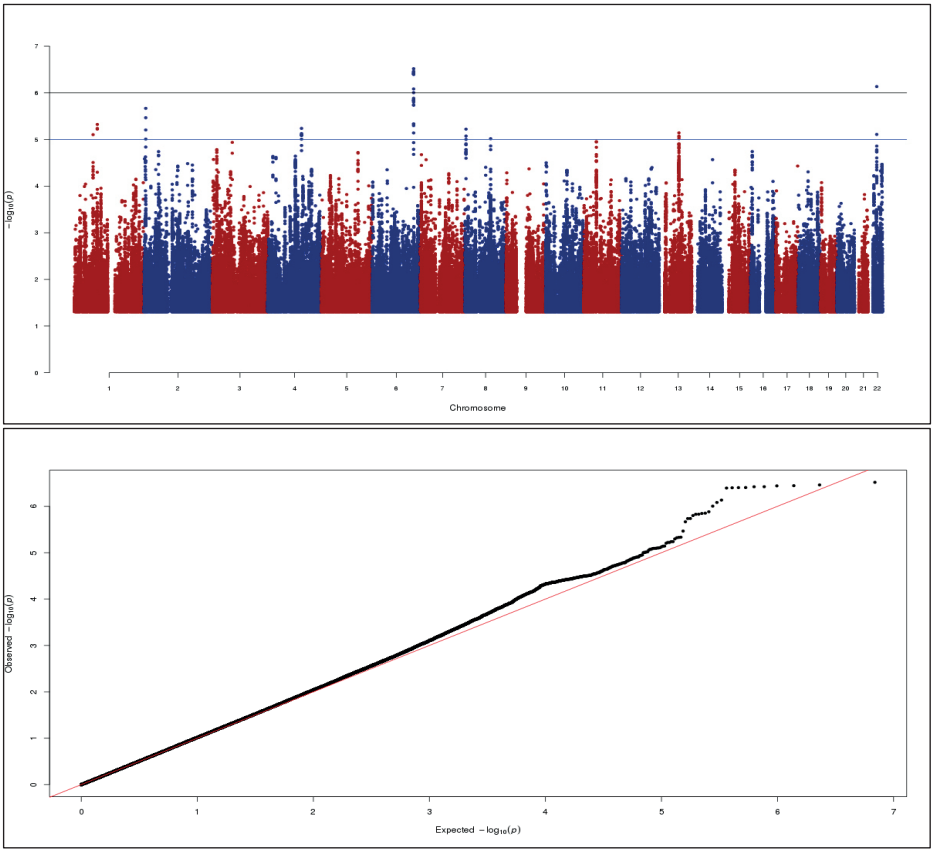
Supplementary Table 3. Most strongly associated SNPs (p-value < 1×10^{-5}) coming from the meta-analysis of nine cohorts from the SAGA consortium in physical position order (hg19).

Supplementary Table 4. Association results for most strongly associated (clumped) SNPs (p-value < 1×10^{-6}) coming from the meta-analysis of seven cohorts from the SAGA consortium without including patients in physical position order (hg19).

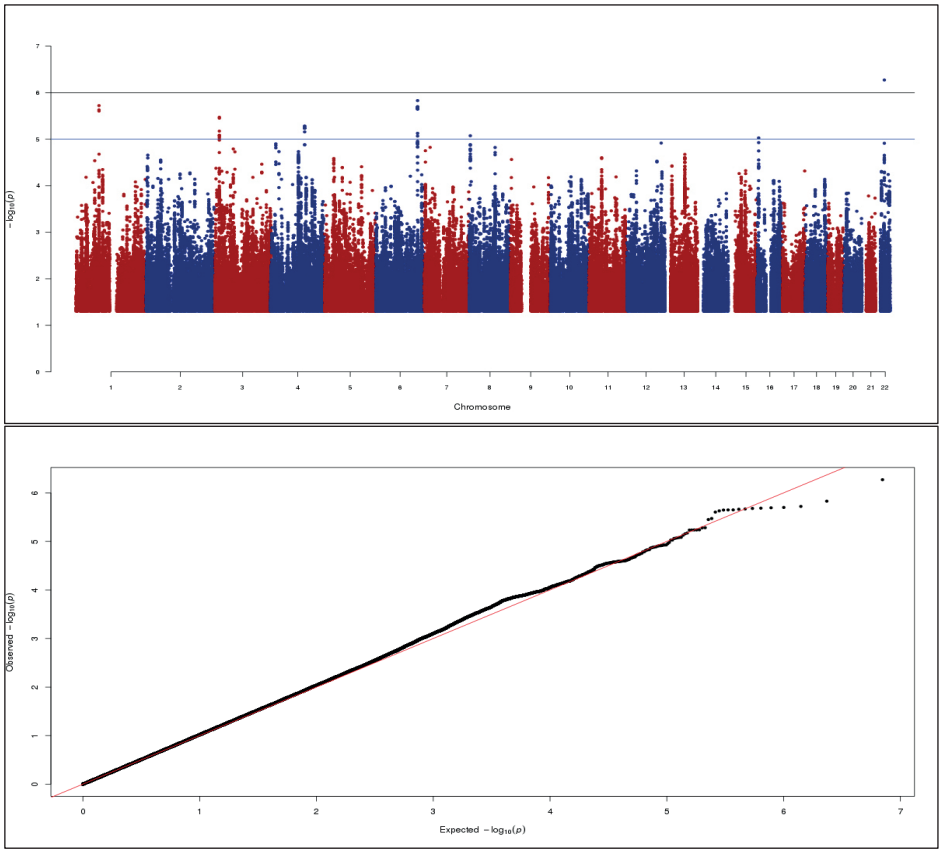
Supplementary Table 5. Gene-wide association p-values for common and rare variants present in the top loci from the SAGA meta-analysis of ADHD Symptom Total Score.

Supplementary Table 6. Best associated loci from the PGC+iPSYCH ADHD GWASMA and their association in the SAGA GWASMA.

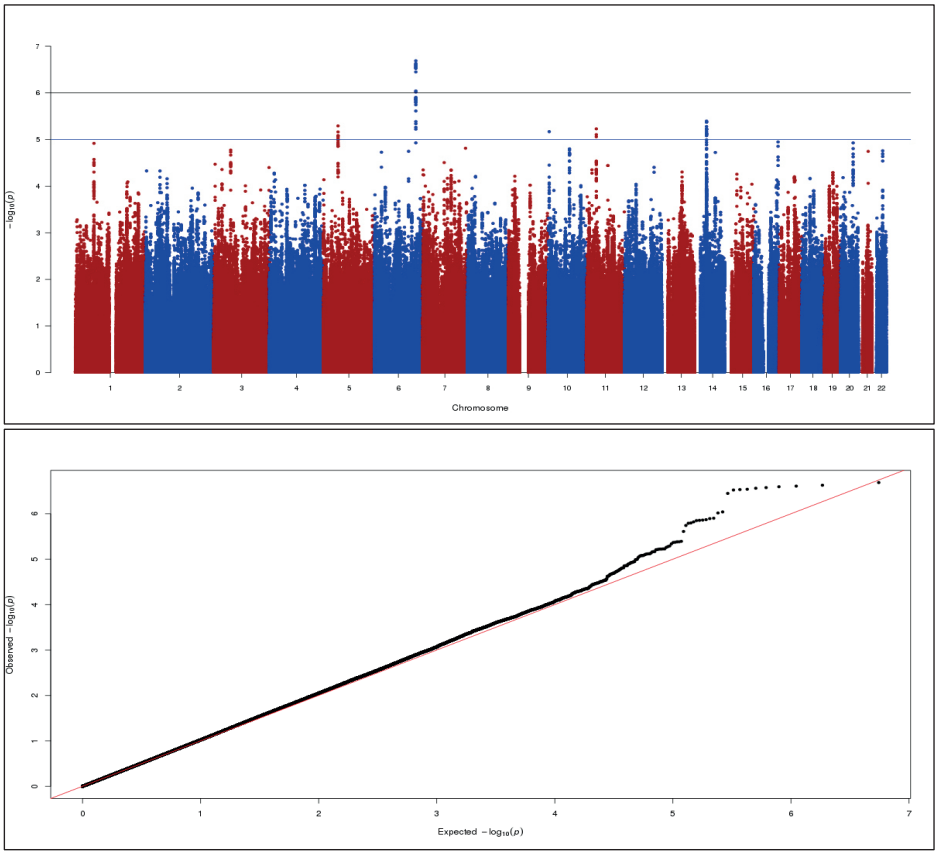
Supplementary Table 7. Conservation and mapping of genomic sequences in primates with similarity to exons encoding human *STXBP5-AS1-003* (Ensembl accession# ENST00000427394).



Supplementary Figure 1A. Manhattan & QQ plot of the ADHD Symptom Total Score meta-analysis from the complete SAGA consortium.



Supplementary Figure 1B. Manhattan & QQ plot of the ADHD Symptom Total Score meta-analysis from the SAGA consortium without patient cohorts.



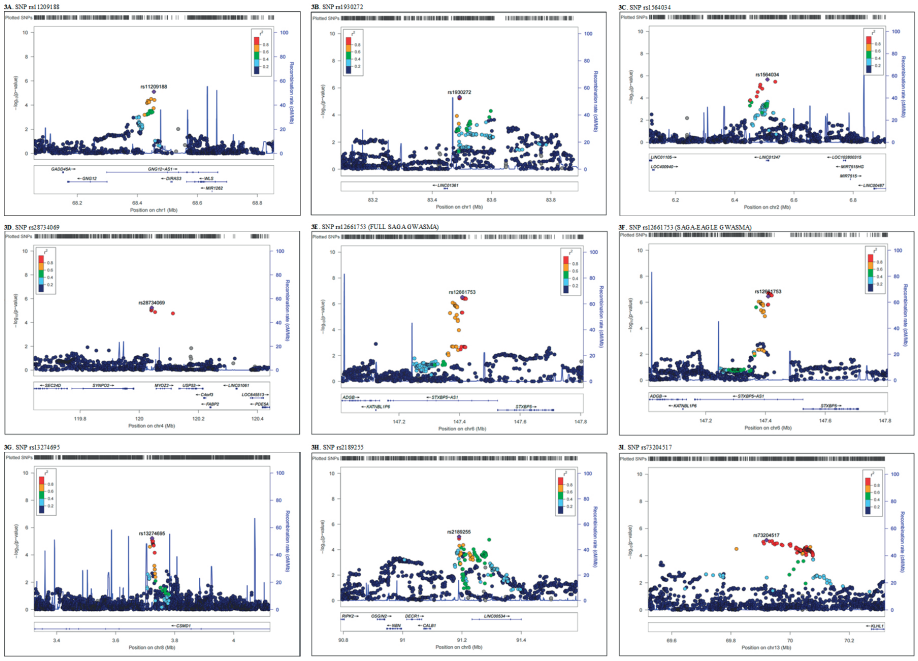
Supplementary Figure 1C. Manhattan & QQ plot of the ADHD Symptom Total Score meta-analysis from the SAGA & EAGLE consortia.

STXBP5-AS1: Exon 6

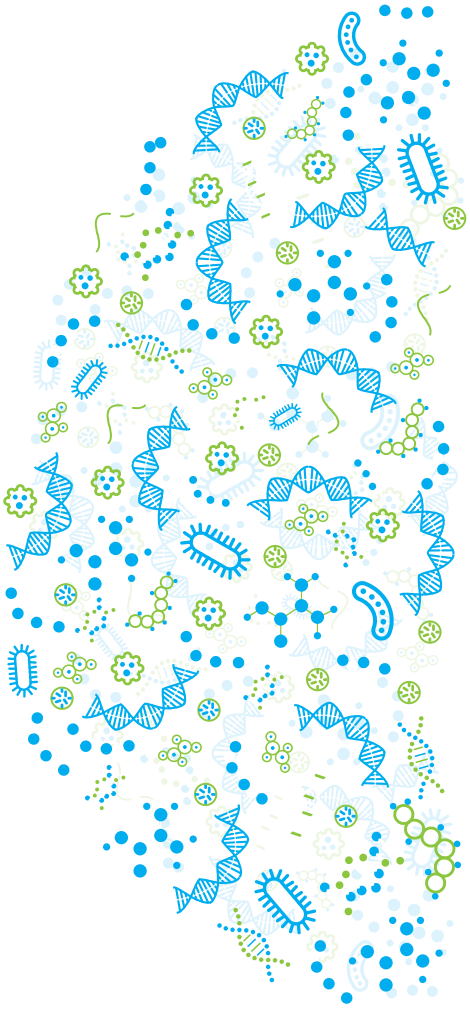
```
man          agATGCCACTGTTGCTGGAGAAATGTCACTGCTAGAGAACAGAAGACTATTATT
chimp        agATGCCACTGTTGCTGGAGAAATGTCACTGCTAGAGAATAGAAGACTATTATT
gorilla      agATGCCACTGTTGCTGGAGAAATGTCACTGCTAGAGAATAGAAGACTATTATT
orangutan    agATGCCACTGTTGCTGCAGAAATGTCACTGCTAGAGAATAGAAGACTATTATT
rhesus       agATGCCACTGTTGCTGGAGAAATGTCACTGCTAGAGAATAGAAGACTATTATT
gibbon       agATGCCACTGTTGCTGGAGAAATGTCACTGCTAGAGAATAGAAGACTATTATT
baboon       agATGCCACTGTTGCTGGAGAAATGTCACTGCTAGAGAATAGAAGACTATTATT
gr.monkey    agATGCCACTGTTGCTGGAGAAATGTCACTGCTAGAGAATAGAAGACTATTATT
mus          agATGTTGCTGTTGCTGGAGAAATGTCACTGGGAAAGATAA-AGGACCATTATT
rat          agATGTTACTGTTGCTGGAGAAATGTCACTATGAAAGATAA-AGGACCATTATT

man          CCAAAGGGTATTTGAGACTGACTGAATCAGGTCTGGAACATTATTGAAATGgt
chimp        CCAAAGGGTATTTGAGACTGACTGAATCAGGTCTGGAACATTATTGAAATGgt
gorilla      CCAAAGGGTATTTGAGACTGACTGAATCAGGTTTGGAACATTATTGAAATGgt
orangutan    CCAAAAGGTATTTGAGACTGACTGAATCAGGTCTGGAACATTATTGAAATGgt
rhesus       CCAAAGGGTATTTGAGACTGACTGAATCAGGTCTGGAATATTATTGAAATGgt
gibbon       CCAAAGGGCATTTGAGACTGACTGAATCAGGTCTGGAACATTATTGAAATGgt
baboon       CCAAAGGGTATTTGAGACTGACTGAATCAGGTCTGGAATATTATTGAAATGgt
gr.monkey    CCAAAGGGTATTTGAGACTGACTGAATCAGGTCTGGAATATTATTGAAATGgt
mus          CTAAAGAGTGTTTCAGGCTGACTGAATCAGGTCTGAAACGTTATTGAAATGgg
rat          CTAAAGAGTGTTTCAGGCTGACTGAATCAGGTCTGAAACCTTATTGAAATGgt
```

Supplementary Figure 2. ClustalW2 alignment of primate and rodent genomic sequences sharing similarity with human *STXBP5-AS1*. Splice acceptor and (possible) donor sites are indicated in lowercase font. See **Supplementary Table 5** for exon start/end positions and sequence accession numbers. The locations of mouse real-time PCR primers, as used in **Fig. 3**, are indicated in bold.



Supplementary Figure 3. Locus Zoom plots for SNPs from the SAGA Meta-analysis associated with Total ADHD Symptom Score



Chapter 5

General discussion

Chapter 5.1

Findings Of This Thesis

This thesis aimed to identify genomic and metabolomic determinants of neurological and psychiatric disorders and their related endophenotypes by making use of various omics approaches. This chapter summarizes the main findings of this thesis, discusses the implication towards the understanding of molecular processes and pathways underlying these disorders and comments on future research.

OMICS OF NEURODEGENERATION

The five projects described in **Chapter 2** focus on endophenotypes of neurological and psychiatric disorders including brain volumetric measures obtained by magnetic resonance imaging (MRI) and cognitive ability.

Brain MRI provides an opportunity to study the complex architecture of brain structures and related disorders. In **Chapter 2.1**, I performed the genome-wide association study of lateral ventricular volume in 23,533 middle-aged to elderly individuals from 26 population-based cohorts participating in the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium. I identified, for the first time, association of lateral ventricular volume and genetic variants at 7 loci (3q28, 16q24.2, 7p22.3, 12q23.3, 22q13.1, 10p12.31 and 11q23.1) (**Figure 1**). Some of the identified loci have previously been linked to various pathologies including cerebrospinal tau/ptau levels, Alzheimer's disease risk, and cognitive decline (3q28),[1] tau pathology (12q23.3),[2] or small vessel disease and white matter lesions (16q24.2).[3] Additionally, several biological pathways emerged including regulation of cytoskeleton organization and S1P signaling.[4-6] The findings described in this chapter provide new insights into understanding complex genetic architecture underlying brain structures. However, identified associations even cumulatively do not account for a substantial fraction of heritability of lateral ventricular volume. In addition to increasing sample size, studying low-frequency and rare variants, incorporating interactions or investigating aggregate or multivariate effects holds promise to expand our knowledge about the genetic architecture of brain structures and related disorders.[7,8]

I next focused on cognitive function, an important predictor of health outcomes, including mortality and morbidity.[9-12] Even though some genes were discovered in specific domains of cognition,[13,14] capturing all cognitive domains into general cognitive function has been more successful in detecting genetic variants.[15] In **Chapter 2.2,** I performed the genome-wide association study of general cognitive function in 243,000 participants of European ancestry (EA) from CHARGE consortium and UK Biobank. I reported 32 novel genetic loci, bringing the total number of independent loci implicated in general cogni-

tive function up to 180.[15] I also showed that genetic risk score based on 180 loci was significantly associated with general cognitive function in 2,117 participants of African-American ancestry (AA) from CHARGE consortium, suggesting that findings in EA can be generalized to AA. Furthermore, I linked genes implicated in general cognitive function to circulating levels of metabolites and found association with tyrosine, an amino acid that plays an important role in synthesis of dopamine,[16,17] glycoprotein acetyl, a marker of acute phase reaction associated with future mortality and cognitive ability,[18,19] and 22:6 docosahexaenoic acid (DHA), a long-chain omega-3 polyunsaturated fatty acid, that has been associated with cognitive function and risk of Alzheimer's disease and dementia.[18] Using Mendelian randomization, I also showed that that genes determining circulating levels of tyrosine and glycoprotein acetyl also determine general cognitive function while DHA is rather a consequence of the physiological processes determining cognitive function. The findings described in this chapter provide new insights into general cognitive ability and demonstrate that further integration of genetic and molecular data with nongenetic data holds great potential to provide additional information about variation in cognitive ability. Discovery of casually associated metabolites provides insights into the pathways underlying general cognitive function and provides starting point for new preventive studies. However, the fact that studied metabolites represent only small proportion of circulating metabolites asks for future studies focusing on a wider spectrum of metabolites. Future efforts should also focus on improving the power of genome-wide association studies of metabolites as novel genetic instruments for running Mendelian randomization for these metabolites may be revealed. Last but not least the causal association between circulating metabolites and general cognitive function should be replicated in other ethnic groups. Future epidemiological research efforts should focus on longitudinal data to validate the cross –omics findings.

There is increasing interest in epigenetics studies that reflect the effect of the genome and exposome (e.g. diet, life style, medication). As cognitive function is also determined by environmental factors, and the complex balance between genes and environment is poorly understood, studying epigenetic signatures may provide insights into cognitive function. [20] In **Chapter 2.3**, we studied the association of blood-based DNA methylation and cognitive test scores in up to 6,809 healthy adults from 11 cohorts. We identified a significant association of two CpG sites and cognitive tests. Cg21450381, located in an intergenic region on chromosome 12 was associated with global cognitive function, whereas cg12507869, located in the *INPP5A* gene on chromosome 10, was associated with verbal fluency. The findings described in this chapter provided evidence for blood-based epigenetic signatures of cognitive function. However, methylation signatures for cognitive function are modest compared to other traits such as body mass index.[21] One of the reasons that may explain this is that the epigenetics in the blood is a poor

surrogate for the post-translational processes in the brain. Improving statistical power by adding additional samples and using newly developed methylation arrays with increased genome coverage may lead to novel discoveries in the future.[22] Furthermore, disentangling correlation from causation in epigenetics is also important from a genetic epidemiological perspective. Careful epidemiological follow-up studies could be used. Alternative approach is Mendelian randomization that makes use of the fact that the genetic drivers of methylation are being rapidly uncovered.[23] Mendelian randomization approach uses a genetic proxy for DNA methylation to evaluate causal relationship between the disease outcome or trait and epigenetic variation and has the potential to help distinguish between truly causal intervention targets and non-causal, which may be informative biomarkers.[24,25]

Even though genome-wide association studies, have been successful in identifying the genetic variants underlying cognitive ability, hypothesis-driven candidate gene design in which biologically relevant regions of the genome are studied in relation to cognitive ability could also shed light on pathways involved in cognitive ability.[26] With recent advances in high-throughput technologies, candidate gene approach is making its re-appearance in genetic epidemiology.[27] In **Chapter 2.4,** I used exome-sequencing data in order to study impact of rare genetic variants in the dystrophin gene (*DMD)* on cognitive ability in about 2,700 participants from two studied populations including family-based Erasmus Rucphen Family (ERF) study and population-based Rotterdam Study. I found a suggestive association of rs147546024:A>G and visuospatial ability in ERF study. However, I was not able to replicate this finding in the Rotterdam Study. I also found a missense variant rs1800273:G>A to be nominally associated with cognitive tests in ERF and Rotterdam Study. The variant, predicted to have a damaging effect on the protein, is present in the different isoforms which are expressed in the brain and which have a stabilizing effect on the GABA receptors recognized for regulation of cognition, emotions, and memory.[28-31] This chapter highlights the challenges of search and replication of rare variant associations. The replication of rare variants is even more challenging if the variants are identified in family-based studies and validation of findings in general population requires extremely large studies. Family-based studies have unique advantages such as enrichment of rare variants and control of population stratification.[32] This design holds great promise for success in searching for rare variants and the chances of success are even higher in genetic isolates since due to genetic drift and inbreeding over several generations, rare variants become more frequent over generations.[33-36]

As *DMD* gene has several different isoforms, **Chapter 2.5** focused on studying association between intelligence and structural mutation location and affected dystrophin isoforms including full-length dystrophin isoform (Dp427) and shorter dystrophin isoforms

(Dp260, Dp140, Dp116, and Dp71/Dp40).[37-39] The study population included patients with Duchenne muscular dystrophy (DMD), a fatal muscular dystrophy during childhood that leads to progressive muscular weakness and less well described nonprogressive central nervous system manifestations.[40] We found that mutations affecting expression of several isoforms including Dp427, Dp140 and Dp71/Dp40 were associated with higher frequency and severe cognitive impairment confirming the findings of previous studies that cumulative loss of dystrophin isoforms has an impact on intellectual ability.[41-45] Furthermore, we observed that expression of Dp140 isoform is not mainly affected by the mutations located in 5'UTR.[41] The findings of this chapter are relevant for personalized medicine initiatives as they allow recognition of the subgroup of patients with great risk for cognitive problems in whom early intervention and support in cognitive, emotional and behavioral development could be very useful.

OMICS OF NEUROVASCULAR PATHOLOGY

Four projects described in **Chapter 3** focus on omics of neurovascular pathology. In **Chapter 3.1** and **3.2**, I focused on endophenotypes of neurodegenerative disorders and stroke characterized by imaging, including carotid intima-media thickness (cIMT) and carotid artery calcification.

Carotid-intima media thickness is an established heritable marker for subclinical atherosclerosis that has been shown to predict future cardiovascular events.[46-48] As previous genome-wide and exome-wide studies identified only a few genetic regions that explain a small proportion of trait variance, and sequencing study of candidate regions yielded inconclusive results due to limited power, more powerful approaches for uncovering the role of rare variants are needed. In **Chapter 3.1**, I performed a genome-wide linkage analysis of individuals in the extremes of cIMT trait distribution (>90th percentile) followed by fine-mapping using exome-sequencing in a large family-based study from a genetically isolated population in the Netherlands. I observed significant evidence of linkage on chromosomes 2p16.3, 19q13.43, 20p13, and 21q22.12. Fine-mapping using exome-sequencing data identified a variant under the linkage peak at 2p16 mapped to *PNPT1* gene which has been characterized as a type I interferon-inducible early response gene.[49-51] Interestingly, several plausible candidate genes were noted under 19q13.43, 20p13, and 21q22.12 peaks, which are highly expressed in tissues relevant for atherosclerosis and linked to pathways implicated in the development of atherosclerosis or cardiovascular diseases. The results of this chapter provide novel insights into genetic architecture of cIMT by making use of extreme phenotype approach. This approach was reported to be better powered in rare variant studies as it reduces phenotypic

heterogeneity.[52] Future functional studies of identified candidate genes are needed to validate and explore the findings ultimately leading to biological pathways involved in the etiology of subclinical atherosclerosis.

Another proxy of carotid artery atherosclerosis studied in this thesis is carotid artery calcification. Carotid artery atherosclerosis is associated with stroke, dementia, and cognitive decline.[53,54] As the specific location of carotid atherosclerosis, i.e. extracranial versus intracranial, may develop under the influence of different metabolic risk factors, in **Chapter 3.2** I performed further in-depth investigation of metabolic determinants and extra- and intracranial carotid artery calcification (ICAC and ECAC) in 1,111 participants from the Rotterdam Study. The significant evidence for association was found between 3-hydroxybutyrate, a ketone body, and ICAC volume. Additionally, the metabolic association pattern of ICAC was found to be different compared to that of ECAC providing further evidence for location-specific differences in the etiology of atherosclerosis.[55,56] However, our study was not designed to resolve the question of reverse causation, emphasizing need to explore metabolomics in longitudinal studies. Also here, Mendelian randomization may contribute substantially to separate associations that are a cause or rather a consequence of disease. This was illustrated by Liu *et al.* who used genetic determinants robustly associated with plasma metabolite levels in order to investigate causal relationship between circulating metabolites and fasting glucose and type 2 diabetes.[57] This may help in translating findings from observational studies from association to causation.

I next focused on stroke, a neurological deficit of sudden onset. As risk determinants of stroke are various complex modifiable risk factors, detailed profiling of metabolic status facilitated by development of high-throughput technologies, could provide novel insights into metabolic changes and identify individuals with higher risk of stroke. In the most comprehensive study to date conducted within the China Kadoorie Biobank, several circulating compounds were associated with stroke, including lipids and lipoprotein particles of various sizes, glycoprotein acetyls, ketone bodies, glucose and docosahexaenoic acid.[58] As large metabolomics studies of stroke in individuals of European ancestry are lacking, in **Chapter 3.3**, I investigated association of circulating metabolites and risk of stroke in seven population-based cohorts including more than 1,790 incident stroke events among 38,797 participants. The significant associations were found between incident stroke and amino acid histidine, glycolysis-related metabolite pyruvate, acute phase reaction marker glycoprotein acetyls, cholesterol in high-density lipoprotein-2 and several other lipoprotein particles. Furthermore, amino-acid phenylalanine and total and free cholesterol in large high-density lipoprotein particles were associated with risk of ischemic stroke. Our results confirmed the association of glycoprotein acetyl and

ischemic stroke that was observed in individuals within the China Kadoorie Biobank, however, we also observed associations that are specific for Western societies.[58] Environmental and ethnic differences across populations or the confounders adjusted for could explain lack of the replication. The results of this chapter provide insights into understanding metabolic determinants of stroke and highlight potential of metabolomics approach in identifying potential targets for the prevention strategies. Future studies should focus on wider range of metabolites and collection of blood samples at multiple time points. As identified associations are starting point for relating metabolites to their biological role, new efforts should focus on integrating metabolomics and other –omics data in order to shed light on metabolic pathways underlying stroke.

Finally, the last project described in **Chapter 3** focused on relation between gut microbiota and human metabolome. As described in **Chapter 3.4,** we examined the association of gut microbiota on circulating metabolites in 2,309 individuals from two population-based cohort studies, Rotterdam Study and LifeLines DEEP. We found 32 microbial families and genera to be associated with a wide range of circulating metabolites including specific very-low density and high-density lipoprotein subfractions, serum lipid measures, glycolysis-related metabolites, amino acids, and acute phase reaction markers. These results provide insights into the role of microbiota in human metabolome, supporting the role of gut microbiota as a target for therapeutic and preventive interventions. However, the current challenges to study microbiota are large, involving reverse causality and bias due to confounding (e.g. the association of the gut microbiome with diabetes was found to be contributed largely to the effects of metformin).[59] Capturing the gut microbiota composition is easy for the distal part of the gut through feces, but microbiota of the proximal part of the gut is more difficult to characterize.[60,61] Perhaps the largest hurdle to overcome is that to date there are no studies that have stored feces samples, allowing prospective studies.[62] An alternative approach may be to use Mendelian randomization, as also the gut microbiome is determined by the human host genome.[63]

GENETIC STUDIES OF PSYCHIATRIC DISEASES

The two projects described in **Chapter 4** focus on genetic determinants of neurodevelopmental disorders including autism spectrum disorder (ASD) and attention deficit hyperactivity disorder (ADHD).

Even though ASD is a heritable disorder and most genetic variance is attributed to common genetic variants, not many loci have been identified.[64,65] The largest effort to

date including more than 16,000 individuals failed to identify new common genetic variants.[66] This suggests that common variants individually have low impact in the ASD, as seen in other psychiatric disorders. However, their joint effect may be substantial.[67] Therefore, in **Chapter 4.1**, I performed a single-variant and gene-based genome-wide association studies in a sample of 160 families with at least one child affected with non-syndromic ASD using both binary phenotype and a quantitative autistic trait. The majority of patients in the study had a normal intelligence, unlike the most ASD cohorts in which rates of intellectual disability are ranging from 30 to 50%.[68] I identified a novel gene *TTC25* associated with quantitative autistic trait in gene-based analysis, replicated this association in an independent sample of general population, and confirmed association of ASD with the known 5p14.1 locus.[69,70] This chapter demonstrates power of endophenotypes to identify a genetic signal, the strength of phenotypic homogeneity of the sample (normal intelligence of majority of the sample) and advantage of gene-based test compared to single-variant analysis.

ADHD is also heritable disorder and a substantial proportion of genetic variance is attributed to common genetic variants.[71-73] However, the first risk loci have been described recently.[74] As ADHD is the extreme end of a continuous ADHD symptoms scores, novel variants could be discovered by focusing on quantitative ADHD symptoms. In **Chapter 4.2**, we sought to leverage the power of population studies of ADHD symptoms in adults in order to discover disease-relevant genes in nine cohorts including about 15,000 individuals. The most strongly associated variant in a genome-wide meta-analysis was mapped to *STXBP5-AS1*. This association was confirmed in the replication analysis of childhood ADHD symptom scores (n~15,000).[75] Even though the function of *STXBP5-AS1* is currently unknown, this lncRNA overlaps in anti-sense with *STXBP5* encoding a protein involved in synaptic function by regulating neurotransmitter release.[76,77] The results of this chapter provide novel insights into the genetic underpinnings of ADHD symptoms implicating synaptic function regulation through *STXBP5-AS1* and potentially *STXBP5* in ADHD symptom etiology.

New insights into the genetics of neurological and psychiatric disorders and related endophenotypes described in this thesis are summarized in **Figure 1.** Previously reported genomic regions are shaded.

FROM OMICS TO TRANSLATION

With the development of high-throughput technologies and omics approaches, our understanding of disease pathophysiology is improving. The major expectation is that

Figure 1. Association of neurological and psychiatric disorders with SNPs accross the genome. Previously reported genomic regions (GWAS catalog as of April 2018) are shaded, whereas the new insights are depicted by arrows. Colors represent studied traits.

these approaches will provide valuable information for prevention programs, earlier disease diagnosis, and personalized treatments taking into account individual variability in context of precision or personalized medicine.[78,79] The research presented in this thesis provides novel insight into the genetic and metabolic determinants or neurological and psychiatric disorders. Even though genetic variants have small effect sizes, going beyond genetic loci in combination with other –omics markers may help classify individuals with higher risk in precise manner.[80]

Genome-wide sequencing studies showed that most individuals carry at least some potentially deleterious variants in their genome.[81] However, the effects of these mutations on individuals are not well understood and to identify their function will be a tall order as there are millions of variants to be mapped in the medium-sized population. As the time and cost would be huge, a major question is the value which refers to doing things at high quality, safely, and at reasonable cost. However, there are certain medical condi-

tions for which genetic testing can provide new opportunities for patients' management but precision or personalized medicine has not seen widespread adoption because of difficulties in achieving a balance between providing personalized care at a population level and delivering standardized care. In **Chapter 5.2** we present how outcome-based healthcare system design developed by the Value-based Healthcare Programme at the University of Oxford could deliver better patient and population-level outcomes and personalized care for cardiovascular disease in a standard way. This chapter does not address a genetic disorder involved in neurodegeneration or neurovascular pathology but focuses on inherited heart rhythm disorder. As a proof of principle, we applied this approach to long QT syndrome (LQTS). We designed two outcome-based systems to focus on patient outcomes, which means that they are service agnostic, context-independent and applicable in a variety of healthcare organizations irrespective of resource constraints.

FUTURE RESEARCH

The extensive research efforts in the past decades have made a progress in understanding the complex architecture of neurological and psychiatric disorders. Identification of numerous common and rare variants underlying these disorders was facilitated by larger sample sizes and development of relatively inexpensive SNP arrays.[82] Resources such as large biobanks that collect biological material and phenotype data made it possible to dramatically increase sample size for some of the traits.[83,84] Population-wide biobanks have been developed in several countries including UK (UK Biobank, n = 0.5 million),[85] Estonia (Estonian Biobank, n = 52,000),[86] USA (Million Veteran Program, n = 1 million),[87] China (Kadoorie Biobank, n = 0.5 million).[88,89] These biobanks will be a large resource for studying neurological and psychiatric disorders and related endophenotypes in the future. Additionally, the majority of studies to date have been conducted in participants of European ancestry. Therefore, future studies should also focus on generalization of the findings in other ancestries and multi-ethnic studies.[90] Increasing the sample size by adding additional samples may lead to novel genetic discoveries and expansion of our knowledge on novel pathways underlying these disorders. Furthermore, novel genetic variants together may improve classification of neurological and psychiatric disorders and facilitate identification of individuals at high risk. As demonstrated by van der Lee *et al.* cumulative effect of common genetic variants modified the risk of Alzheimer's disease and all-cause dementia beyond the *APOE* genotype and contributed to better risk prediction.[91]

With development of SNP arrays and statistical imputation of unobserved variants, both common and less frequency variants could be assessed in the population increasing power of association studies and facilitating discovery of new loci.[92-94] On the other hand, whole-exome and whole-genome sequencing are expected to identify rare-variants. Even though studies of rare variants also require large datasets, these datasets are still small compared to datasets used for discovery of common variants. Focusing on family-based designs or studying extreme cases would be more efficient approach. For example, I showed in **Chapter 3.1** that studying extreme cases in a family-based study using linkage analysis could identify novel loci in a smaller sample. However, this approach should be complemented with deep sequencing. The power could also be boosted by combining the alleles of similar impact in a gene or a region.[82] This approach may also be highly relevant for personalized medicine - as discussed above - evaluating the health threat of damaging mutation is a tall order for rare variants in the population. However, within a family such a variant is not rare and segregates with a probability of 50%. Thus 50% of first-degree relatives and 25% of second-degree relative are carriers and those in other generations can provide key clues.

Apart from identifying new loci associated with neurological and psychiatric diseases focus in coming years should also be on understanding how these loci contribute to the disease. As illustrated in **Chapter 2.1**, most of the variants are located in non-coding regions of the genome. Understanding regulatory components of genome became recently available by projects such as ENCODE,[95] Epigenome RoadMap,[96] and GTEx.[82,97] For neurological and psychiatric disorders appropriate tissue-specific resources are also important and essential. These methods are useful for prioritizing genes from GWAS loci for functional follow-up with the ultimate objective to enable more effective prevention and treatment strategies of disease.[82]

Furthermore, other single omics approaches including proteomics, metabolomics, and microbiomics could also provide information about the biological processes. These fields could also greatly benefit from large population-based samples as increasing the sample size may lead to novel discoveries. In this thesis, we showed that large studies identified metabolites associated with stroke (**Chapter 3.3**) and gut microbiota (**Chapter 3.4**). The identified metabolites could be further studied, for example in relation to genetic determinants. Future studies should integrate multiple levels of data in multi-omics studies. This would overcome limited information derived from single omics approaches and could provide additional biological insight useful in understanding complex disorders (**Figure 2**). Exploration of genome, transcriptome, metabolome and microbiome levels could yield important conclusions that will be basis for precision medicine.[20]

Integrative multi-omics approaches should also focus on appropriate target tissues. One of the alternative approaches to examine tissue specificity for cells and tissues that

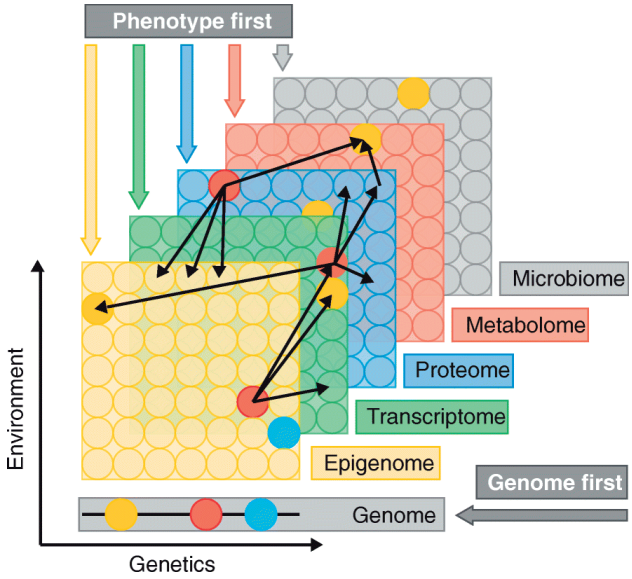


Figure 2. Multiple omics data. Multiple omics data types are depicted by layers. Pool of molecules collected within each of the layers is depicted by circles. All the layers except genome reflect genetic and environmental regulation. Arrows represent potential interactions between the molecules in different layers, whereas interaction within the layers are not shown. Source: Hasin *et al*.[20]

are difficult to obtain includes use of induced pluripotent stem (iPS) cell technology.[98] This technology may help to differentiate easily accessible cells into different brain cells. These cells with relevant genetic background or subjected to gene-editing through CRISP-Cas9 technology could be further used to reconstruct diseased brain models using organ-on-chip technology.[98-100]

Lastly, future translational research efforts may benefit from longitudinal measures. Collection of quantitative traits at different time points can reduce type I error and increase statistical power compared to a single measurement and could also identify determinants for age of onset.[101,102] Similarly, longitudinal profiling in a single individual could also be beneficial as it could provide consistent monitoring of dynamic changes in multi-omics components in relation to disease status and preventive interventions.[103]

CONCLUDING REMARKS

In this thesis, I have used various omics approaches in order to provide novel insights into the pathophysiology of complex neurological and psychiatric disorders and related endophenotypes. Discovery of novel genetic determinants underlying brain structures, cognitive ability, and neurodevelopmental disorders, as well as link between circulating metabolites and neurovascular pathology or gut microbiota are some of the highlights of this work. With the development of high-throughput technologies and integration of various approaches in the future, novel insights into molecular mechanisms underlying these disorders will be provided as well as valuable information for prevention programs, earlier disease diagnosis, and personalized treatments.

REFERENCES

1. Cruchaga, C. *et al.* GWAS of cerebrospinal fluid tau levels identifies risk variants for Alzheimer's disease. *Neuron* **78**, 256-68 (2013).
2. Lasagna-Reeves, C.A. *et al.* Reduction of Nuak1 Decreases Tau and Reverses Phenotypes in a Tauopathy Mouse Model. *Neuron* **92**, 407-418 (2016).
3. Traylor, M. *et al.* Genetic Variation at 16q24.2 Is Associated With Small Vessel Stroke. *Annals of Neurology* **81**, 383-394 (2017).
4. Mishra, A. & MacGregor, S. A Novel Approach for Pathway Analysis of GWAS Data Highlights Role of BMP Signaling and Muscle Cell Differentiation in Colorectal Cancer Susceptibility. *Twin Research and Human Genetics* **20**, 1-9 (2017).
5. Shen, H.Y. *et al.* Coupling between endocytosis and sphingosine kinase 1 recruitment. *Nature Cell Biology* **16**, 652 (2014).
6. Ma, S., Santhosh, D., Kumar, T.P. & Huang, Z. A Brain-Region-Specific Neural Pathway Regulating Germinal Matrix Angiogenesis. *Developmental Cell* **41**, 366 (2017).
7. Mufford, M.S. *et al.* Neuroimaging genomics in psychiatry-a translational approach. *Genome Med* **9**, 102 (2017).
8. Boyle, E.A., Li, Y.I. & Pritchard, J.K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177-1186 (2017).
9. Deary, I.J., Weiss, A. & Batty, G.D. Intelligence and personality as predictors of illness and death: How researchers in differential psychology and chronic disease epidemiology are collaborating to understand and address health inequalities. *Psychological Science in the Public Interest***.11**, pp (2010).
10. Batty, G.D., Deary, I.J. & Gottfredson, L.S. Premorbid (early life) IQ and later mortality risk: systematic review. *Ann Epidemiol* **17**, 278-88 (2007).
11. Wraw, C., Deary, I.J., Gale, C.R. & Der, G. Intelligence in youth and health at age 50. *Intelligence. Vol.53 2015, pp. 23-32. Nov-Dec Intelligence* (2015).
12. Hagenaars, S.P., Gale, C.R., Deary, I.J. & Harris, S.E. Cognitive ability and physical health: a Mendelian randomization study. *Scientific Reports* **7**(2017).
13. Ibrahim-Verbaas, C.A. *et al.* GWAS for executive function and processing speed suggests involvement of the CADM2 gene. *Mol Psychiatry* **21**, 189-197 (2016).
14. Debette, S. *et al.* Genome-wide studies of verbal declarative memory in nondemented older people: the Cohorts for Heart and Aging Research in Genomic Epidemiology consortium. *Biol Psychiatry* **77**, 749-63 (2015).
15. Davies, G. *et al.* Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nat Commun* **9**, 2098 (2018).
16. Jongkees, B.J., Hommel, B. & Colzato, L.S. People are different: tyrosine's modulating effect on cognitive control in healthy humans may depend on individual differences related to dopamine function. *Frontiers in Psychology* **5**(2014).
17. Jongkees, B.J., Hommel, B., Kuhn, S. & Colzato, L.S. Effect of tyrosine supplementation on clinical and healthy populations under stress or cognitive demands-A review. *Journal of Psychiatric Research* **70**, 50-57 (2015).
18. van der Lee, S.J. *et al.* Circulating metabolites and general cognitive ability and dementia: Evidence from 11 cohort studies. *Alzheimers Dement* (2018).
19. Lawler, P.R. *et al.* Circulating N-Linked Glycoprotein Acetyls and Longitudinal Mortality Risk. *Circ Res* **118**, 1106-15 (2016).

20. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biology* **18** (2017).
21. Dick, K.J. *et al.* DNA methylation and body-mass index: a genome-wide analysis. *Lancet* **383**, 1990-8 (2014).
22. Pidsley, R. *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol* **17**, 208 (2016).
23. Bonder, M.J. *et al.* Disease variants alter transcription factor levels and methylation of their binding sites. *Nature Genetics* **49**, 131-138 (2017).
24. Mill, J. & Heijmans, B.T. From promises to practical strategies in epigenetic epidemiology. *Nature Reviews Genetics* **14**, 585-594 (2013).
25. Relton, C.L. & Smith, G.D. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *International Journal of Epidemiology* **41**, 161-176 (2012).
26. Zondervan, K.T. & Cardon, L.R. Designing candidate gene and genome-wide case-control association studies. *Nat Protoc* **2**, 2492-501 (2007).
27. Patnala, R., Clements, J. & Batra, J. Candidate gene association studies: a comprehensive guide to useful in silico tools. *BMC Genet* **14**, 39 (2013).
28. Mohler, H. Role of GABAA receptors in cognition. *Biochem Soc Trans* **37**, 1328-33 (2009).
29. Millan, M.J. *et al.* Cognitive dysfunction in psychiatric disorders: characteristics, causes and the quest for improved therapy. *Nat Rev Drug Discov* **11**, 141-68 (2012).
30. Fritschy, J.M., Schweizer, C., Brunig, I. & Luscher, B. Pre- and post-synaptic mechanisms regulating the clustering of type A gamma-aminobutyric acid receptors (GABAA receptors). *Biochem Soc Trans* **31**, 889-92 (2003).
31. Muntoni, F., Torelli, S. & Ferlini, A. Dystrophin and mutations: one gene, several proteins, multiple phenotypes. *Lancet Neurology* **2**, 731-740 (2003).
32. Yan, Q. *et al.* The impact of genotype calling errors on family-based studies. *Sci Rep* **6**, 28323 (2016).
33. Gudmundsson, J. *et al.* A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nature Genetics* **44**, 1326-1329 (2012).
34. Stacey, S.N. *et al.* A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. *Nature Genetics* **43**, 1098-U85 (2011).
35. Jonsson, T. *et al.* A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature* **488**, 96-99 (2012).
36. Pardo, L.M., MacKay, I., Oostra, B., van Duijn, C.M. & Aulchenko, Y.S. The effect of genetic drift in a young genetically isolated population. *Annals of Human Genetics* **69**, 288-295 (2005).
37. Lidov, H.G., Byers, T.J., Watkins, S.C. & Kunkel, L.M. Localization of dystrophin to postsynaptic regions of central nervous system cortical neurons. *Nature* **348**, 725-8 (1990).
38. Record Owner, N.L.M. Dystrophins, utrophins, and associated scaffolding complexes: role in mammalian brain and implications for therapeutic strategies.
39. Anderson, J.L., Head, S.I., Rae, C. & Morley, J.W. Brain function in Duchenne muscular dystrophy. *Brain* **125**, 4-13 (2002).
40. Emery, A. & Muntoni, F. *Duchenne muscular dystrophy*, (Oxford University Press, New York, 2003).
41. Taylor, P.J. *et al.* Dystrophin gene mutation location and the risk of cognitive impairment in Duchenne muscular dystrophy. *PLoS One* **5**, e8803 (2010).
42. Daoud, F. *et al.* Analysis of Dp71 contribution in the severity of mental retardation through comparison of Duchenne and Becker patients differing by mutation consequences on Dp71 expression. *Human Molecular Genetics* **18**, 3779-3794 (2009).

43. Moizard, M.P. *et al.* Severe cognitive impairment in DMD: obvious clinical indication for Dp71 isoform point mutation screening. *Eur J Hum Genet* **8**, 552-6 (2000).
44. Moizard, M.P. *et al.* Are Dp71 and Dp140 brain dystrophin isoforms related to cognitive impairment in Duchenne muscular dystrophy? *Am J Med Genet* **80**, 32-41 (1998).
45. Chelly, J., Khelfaoui, M., Francis, F., Cherif, B. & Bienvenu, T. Genetics and pathophysiology of mental retardation. *Eur J Hum Genet* **14**, 701-13 (2006).
46. Polak, J.F. *et al.* Carotid-Wall Intima-Media Thickness and Cardiovascular Events. *New England Journal of Medicine* **365**, 213-221 (2011).
47. Lorenz, M.W., Markus, H.S., Bots, M.L., Rosvall, M. & Sitzer, M. Prediction of clinical cardiovascular events with carotid intima-media thickness - A systematic review and meta-analysis. *Circulation* **115**, 459-467 (2007).
48. Den Ruijter, H.M. *et al.* Common Carotid Intima-Media Thickness Measurements in Cardiovascular Risk Prediction A Meta-analysis. *Jama-Journal of the American Medical Association* **308**, 796-803 (2012).
49. Leszczyniecka, M. *et al.* Identification and cloning of human polynucleotide phosphorylase, hPNPase old-35, in the context of terminal differentiation and cellular senescence. *Proc Natl Acad Sci U S A* **99**, 16636-41 (2002).
50. Leszczyniecka, M., Su, Z.Z., Kang, D.C., Sarkar, D. & Fisher, P.B. Expression regulation and genomic organization of human polynucleotide phosphorylase, hPNPase(old-35), a Type I interferon inducible early response gene. *Gene* **316**, 143-56 (2003).
51. Goossens, P. *et al.* Myeloid Type I Interferon Signaling Promotes Atherosclerosis by Stimulating Macrophage Recruitment to Lesions. *Cell Metabolism* **12**, 142-153 (2010).
52. Peloso, G.M. *et al.* Phenotypic extremes in rare variant study designs. *Eur J Hum Genet* **24**, 924-30 (2016).
53. Bos, D. *et al.* Intracranial Carotid Artery Atherosclerosis and the Risk of Stroke in Whites The Rotterdam Study. *Jama Neurology* **71**, 405-411 (2014).
54. Bos, D. *et al.* Atherosclerotic calcification is related to a higher risk of dementia and cognitive decline. *Alzheimers & Dementia* **11**, 639-647 (2015).
55. Bos, D. *et al.* Genetic Loci for Coronary Calcification and Serum Lipids Relate to Aortic and Carotid Calcification. *Circulation-Cardiovascular Genetics* **6**, 47-U72 (2013).
56. Bos, D. *et al.* Comparison of Atherosclerotic Calcification in Major Vessel Beds on the Risk of All-Cause and Cause-Specific Mortality The Rotterdam Study. *Circulation-Cardiovascular Imaging* **8** (2015).
57. Liu, J. *et al.* A Mendelian Randomization Study of Metabolite Profiles, Fasting Glucose, and Type 2 Diabetes. *Diabetes* **66**, 2915-2926 (2017).
58. Holmes, M.V. *et al.* Lipids, Lipoproteins, and Metabolites and Risk of Myocardial Infarction and Stroke. *J Am Coll Cardiol* **71**, 620-632 (2018).
59. Forslund, K. *et al.* Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* **528**, 262-266 (2015).
60. Donaldson, G.P., Lee, S.M. & Mazmanian, S.K. Gut biogeography of the bacterial microbiota. *Nat Rev Microbiol* **14**, 20-32 (2016).
61. Sartor, R.B. Gut microbiota: Optimal sampling of the intestinal microbiota for research. *Nat Rev Gastroenterol Hepatol* **12**, 253-4 (2015).
62. Mai, V. & Morris, J.G., Jr. Need for prospective cohort studies to establish human gut microbiome contributions to disease risk. *J Natl Cancer Inst* **105**, 1850-1 (2013).

63. Bonder, M.J. *et al.* The effect of host genetics on the gut microbiome. *Nat Genet* **48**, 1407-1412 (2016).
64. Gaugler, T. *et al.* Most genetic risk for autism resides with common variation. *Nature Genetics* **46**, 881-885 (2014).
65. Smoller, J.W. *et al.* Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* **381**, 1371-1379 (2013).
66. Anney, R.J.L. *et al.* Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. *Molecular Autism* **8** (2017).
67. Anney, R. *et al.* Individual common variants exert weak effects on the risk for autism spectrum disorderspi. *Human Molecular Genetics* **21**, 4781-4792 (2012).
68. Geschwind, D.H. Advances in Autism. *Annual Review of Medicine* **60**, 367-380 (2009).
69. Wang, K. *et al.* Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* **459**, 528-533 (2009).
70. Ma, D.Q. *et al.* A Genome-wide Association Study of Autism Reveals a Common Novel Risk Locus at 5p14.1. *Annals of Human Genetics* **73**, 263-273 (2009).
71. Brikell, I., Kuja-Halkola, R. & Larsson, H. Heritability of attention-deficit hyperactivity disorder in adults. *Am J Med Genet B Neuropsychiatr Genet* **168**, 406-413 (2015).
72. Lee, S.H. *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics* **45**, 984 (2013).
73. Anttila, V. *et al.* Analysis of shared heritability in common disorders of the brain. *bioRxiv* (2017).
74. Demontis, D. *et al.* Discovery of the first genome-wide significant risk loci for ADHD. *bioRxiv* (2017).
75. Middeldorp, C.M. *et al.* A Genome-Wide Association Meta-Analysis of Attention-Deficit/Hyperactivity Disorder Symptoms in Population-Based Pediatric Cohorts. *J Am Acad Child Adolesc Psychiatry* **55**, 896-905 e6 (2016).
76. Yizhar, O. *et al.* Tomosyn inhibits priming of large dense-core vesicles in a calcium-dependent manner. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 2578-2583 (2004).
77. Sakisaka, T. *et al.* Dual inhibition of SNARE complex formation by tomosyn ensures controlled neurotransmitter release. *Journal of Cell Biology* **183**, 323-337 (2008).
78. Chen, R. & Snyder, M. Promise of personalized omics to precision medicine. *Wiley Interdisciplinary Reviews-Systems Biology and Medicine* **5**, 73-82 (2013).
79. Collins, F.S. & Varmus, H. A new initiative on precision medicine. *N Engl J Med* **372**, 793-5 (2015).
80. Fall, T., Mendelson, M. & Speliotes, E.K. Recent Advances in Human Genetics and Epigenetics of Adiposity: Pathway to Precision Medicine? *Gastroenterology* **152**, 1695-1706 (2017).
81. Sohail, M. *et al.* Negative selection in humans and fruit flies involves synergistic epistasis. *Science* **356**, 539-542 (2017).
82. Visscher, P.M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics* **101**, 5-22 (2017).
83. Nielsen, J.B. *et al.* Genome-wide association study of 1 million people identifies 111 loci for atrial fibrillation. *bioRxiv* (2018).
84. Jansen, P.R. *et al.* Genome-wide Analysis of Insomnia (N=1,331,010) Identifies Novel Loci and Functional Pathways. *bioRxiv* (2018).

85. Elliott, P., Peakman, T.C. & Biobank, U.K. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int J Epidemiol* **37**, 234-44 (2008).
86. Leitsalu, L. *et al.* Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *International Journal of Epidemiology* **44**, 1137-1147 (2015).
87. Gaziano, J.M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol* **70**, 214-23 (2016).
88. Chen, Z. *et al.* China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int J Epidemiol* **40**, 1652-66 (2011).
89. De Souza, Y.G. & Greenspan, J.S. Biobanking past, present and future: responsibilities and benefits. *AIDS* **27**, 303-12 (2013).
90. Carlson, C.S. *et al.* Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. *PLoS Biol* **11**, e1001661 (2013).
91. van der Lee, S.J. *et al.* The effect of APOE and other common genetic variants on the onset of Alzheimer's disease and dementia: a community-based cohort study. *Lancet Neurol* **17**, 434-444 (2018).
92. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**, 1084-97 (2007).
93. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes. *Genetic Epidemiology* **34**, 816-834 (2010).
94. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics* **39**, 906-913 (2007).
95. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
96. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).
97. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-5 (2013).
98. Wijmenga, C. & Zhernakova, A. The importance of cohort studies in the post-GWAS era. *Nat Genet* **50**, 322-328 (2018).
99. Yi, Y., Park, J., Lim, J., Lee, C.J. & Lee, S.H. Central Nervous System and its Disease Models on a Chip. *Trends Biotechnol* **33**, 762-776 (2015).
100. Ran, F.A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nature Protocols* **8**, 2281-2308 (2013).
101. Wu, Z., Hu, Y. & Melton, P.E. Longitudinal data analysis for genetic studies in the whole-genome sequencing era. *Genet Epidemiol* **38 Suppl 1**, S74-80 (2014).
102. Rasmussen-Torvik, L.J. *et al.* Impact of repeated measures and sample selection on genome-wide association studies of fasting glucose. *Genet Epidemiol* **34**, 665-73 (2010).
103. Lau, E. & Wu, J.C. Omics, Big Data, and Precision Medicine in Cardiovascular Sciences. *Circ Res* **122**, 1165-1168 (2018).

Chapter 5.2

A model for mass personalization in cardiology: standard outcomes-based systems that can deliver personalized care

Dina Vojinovic*, Anna Puggina*, Christian van der Werf, Carla G. van El, Olga C. Damman, Najaf Amin, Ayse Demirkan, Bruno H. Stricker, Muir Gray, Stefania Boccia, Martina C. Cornel, Cornelia M. van Duijn, Anant Jani

* These authors contributed equally to this work

This chapter is submitted.

ABSTRACT

Background

Even though there is excitement in the current healthcare environment on the potential of personalized medicine to utilize individuals' genomic data to improve patient outcomes and improve resource utilization, personalized medicine has not seen widespread adoption.

Main body

We explore how we can use the well-established principles of outcome-based healthcare system designed by the Value-Based Healthcare Programme at the University of Oxford to deliver better patient and population-level outcomes, encourage shared decision making, optimize resource utilization and to also deliver personalized care. This approach has been used to improve service delivery, improve outcomes and, importantly, drive culture change in England for a variety of different conditions since 2011. The approach, as applied to long QT Syndrome (LQTS), yielded two outcomes-based system specifications: one which outlines how to improve outcomes for patients with known LQTS; and a second which leverages genomic testing to identify people with unknown LQTS.

Conclusion

The simple approach outlined in this manuscript along with the context-independent and service agnostic systems presented have the potential to help deliver personalized care for cardiovascular diseases in a standard way.

BACKGROUND

There is much attention and excitement in the current healthcare environment on the potential of personalized medicine to utilize individuals' genomic data to improve patient outcomes and resource utilization. Despite tremendous promise, personalized medicine has not seen widespread adoption because of difficulties in achieving a balance between providing personalized care at a population level and delivering standardized care.

In this manuscript, we explore how we can use the well-established principles of outcomes-based healthcare system design developed by the Value-based Healthcare Programme at the University of Oxford[1,2] to deliver better patient and population-level outcomes, encourage shared decision making, optimize resource utilization, and to also deliver personalized care - i.e. mass personalization.

As a proof of concept, we apply this approach to long QT Syndrome (LQTS) and present outcomes-based systems which can be used for the effective reduction of risk of cardiac events (syncope, aborted cardiac arrest, or sudden cardiac death) in people with LQTS and their first-degree relatives.

MAIN TEXT

Designing outcomes-based systems

We used the 10 step model created by the Value-Based Healthcare Programme at the University of Oxford, which has been validated for several clinical conditions in England.[1] The model aims to maximize value and equity by focusing on populations defined by a common condition or characteristic through using a system approach.[2] According to this model, designing an ideal outcome-driven population-based system of care that delivers value to patients and populations requires 10 steps as illustrated in **Figure 1**.

Because of the focus on outcomes, the system design is flexible and it can be changed and adapted when new guidelines and best practices are revealed and also when innovative diagnostics and treatments are introduced. It takes a dedicated core group to take the initiative and start elaborating the subsequent steps. The scope of the system of care might be a symptom, a subgroup of population or condition as in this case. It is also essential to define the population to be served precisely, not only by naming it but also by specifying the practice and/or local authority. Although the aim of our system as applied to LQTS is set to reduce the risk of cardiac events in LQTS, it is important

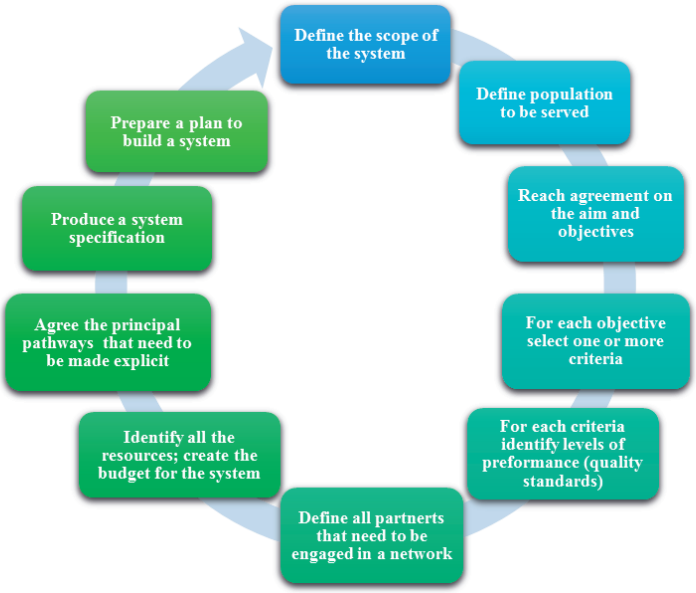


Figure 1. Steps required to design an outcome-driven population-based system of care according to the Value-Based Healthcare Programme at the University of Oxford.

to complement and supplement this aim with a set of objectives and one or more appropriate criteria to measure progress towards the objectives. The specific objectives and criteria for the LQTS system were defined based on expert consensus statement on the diagnosis and management of patients with LQTS and the Value-Based Healthcare Programme methodology.[2-4]

Long QT Syndrome

The value of screening for heritable cardiovascular diseases has been acknowledged by public health officials.[5] LQTS is an inherited heart rhythm disorder characterized by a prolonged ventricular repolarization (prolonged heart rate-corrected QT interval (QTc interval)) and T-wave abnormalities on the resting electrocardiogram (ECG), most commonly associated with specific ventricular tachyarrhythmia named torsade de pointes (TdP) which can cause syncope, aborted cardiac arrest and sudden cardiac death.[6-10] Occurring in approximately 1 individual in 2,500 worldwide,[11] LQTS is considered to be responsible for as many as 2,000-3,000 sudden deaths in children and young adults in the United States each year and 10-year mortality in untreated symptomatic cases is ~50%.[3,7,12]

The diagnosis of LQTS is either made when several ECGs with a clearly prolonged QTc interval are observed in the absence of acquired QTc interval prolonging factors, or

by use of a scoring system of clinical and ECG parameters.[3] Genetic testing is useful to make or exclude the diagnosis in borderline cases. In addition, genetic testing allows classification into LQTS subtypes by identifying the mutations in the genes coding for the ion channel subunits or the associated proteins.[4] At least 15 different genes are implicated in the development of 15 different LQTS subtypes (LQT1, LQT2, LQT3, and up to LQT15). The most common subtypes are due to mutations in three genes coding for pore-forming subunits of two potassium channels (*KCNQ1* and *KCNH2*) and a sodium channel (*SCN5a*) giving rise to LQT1, LQT2, and LQT3, respectively.[13]

The timely and accurate evaluation of the LQTS genotype has diagnostic, prognostic and therapeutic value, and thus an increased potential within clinical decision making.[14] In 2011, the Heart Rhythm Society (HRS) and the European Heart Rhythm Association (EHRA) developed an expert consensus statement on the state of genetic testing for the channelopathies and cardiomyopathies.[3] The document provides a detailed analysis of the diagnostic, prognostic, and therapeutic impact of genetic test results for LQTS. First, the consensus statement recognizes its diagnostic value and recommends genetic testing for any index case in which LQTS is suspected by a cardiologist based on a patient's clinical history, family history, QTc interval, T-wave morphology and/or response to either cycle/treadmill or catecholamine stress testing. In addition, when a putative causative mutation is identified in clinically affected index cases, mutation-specific genetic testing of all first-degree relatives is recommended, even in the absence of a clinical and electrocardiographic phenotype.[3] Second, since numerous genotype-phenotype relationships pertain to the most frequent (i.e., LQT1, LQT2, and LQT3) subtypes, the LQTS genetic tests join traditional risk factors (i.e., gender, age, QTc interval at rest, syncope) as independent prognostic risk factors.[3] Third, LQTS genetic tests can influence clinical treatment decisions and it is recommended to incorporate genotype and mutation data with all other non-genetic risk factors in assessing the patient's risk and personalizing the patient's treatment plan.[3]

Genetic testing for the three most common LQTS subtypes in symptomatic index cases appears to be a cost-effective option as compared with no testing,[15] but further economic evaluations are needed to evaluate the value for money of testing asymptomatic first-degree relatives of a patient with established LQTS.[16]

Despite the fact that timely and accurate testing for the LQTS genotype has high positive predictive value and seems to be cost-effective, in many countries it is not used regularly in practice because of a lack of knowledge and service-level barriers to implementation. Furthermore, due to different standards, opinions and possibilities, it is not certain which intervention is optimal for every LQTS subtype, for example, different opinions

exist among experts on the treatment of LQTS3 (beta-blockers, flecainide, mexiletine, ranolazine).[17] To begin to tackle these issues, and improve transparency of choices and outcomes within and across services, we designed two outcomes-based systems: one for management of patients with identified LQTS; and the other to identify patients with LQTS who have not yet been identified. Our hope is that these systems will give a context-independent and service agnostic template for healthcare services to improve and personalize care for patients with LQTS and identify unmet need in their population.

Outcomes-based system for patients with known LQTS

The first system focuses on people with known LQTS and aims to reduce the risk of cardiac events in these patients. The population to be served should be defined by all the practices in the region. The objectives of the service, as well as the criteria used to measure progress towards the objectives, are listed in **Table 1**.

Table 1. Criteria defined to measure progress of each objective in the system aiming to reduce risk of cardiac events in patients diagnosed with LQTS

Objective	Criteria
To treat people with LQTS safely and effectively	- % of asymptomatic patients stratified by LQTS-subtype with a QTc-interval ≥ 470 ms who are on beta-blocker; - % of symptomatic patients stratified by LQTS-subtype who are on beta-blocker therapy; - % of patients in whom avoidance of QT-prolonging drugs is recommended; - % of patients who stopped beta-blocker therapy; - % of patients stratified by LQTS-subtype who had a cardiac event; - % of patients stratified by LQTS-subtype who are survivors of an aborted cardiac arrest in whom an implantable cardioverter-defibrillator (ICD) is implanted; - % of patients stratified by LQTS-subtype with ICD who received at least one inappropriate (not needed) shock; - Number of inappropriate shock/ICD complications; - % of patients with left cardiac sympathetic denervation who had a cardiac event; - % of genotype-positive phenotype-negative LQTS patients who are advised against participating in competitive sports;
To accurately assess the risk of cardiac events in patients with LQTS.	- Number of people known to have LQTS; - % of people diagnosed with LTQS who had age-stratified risk assessment by year-end using constellation of electrocardiographic, clinical, and genetic factors; - % of patients with LQTS who never had a risk assessment; - % of people with LQTS who had a risk assessment in the first year of treatment and who are in the second or subsequent year who have a review during the course of the year using age-stratified risk assessment based upon constellation of electrocardiographic, clinical, and genetic factors;

Table 1. Criteria defined to measure progress of each objective in the system aiming to reduce risk of cardiac events in patients diagnosed with LQTS (*continued*)

Objective	Criteria
To ensure patients with LQTS make informed decisions that take their values into account.	- % of patients who were explicitly told that a choice for treatment is to be made and that the patient's opinion is important; - % of patients whom the options and pros and cons of each relevant option were discussed with using the available information aids (graphics, decision aids, decision grids); - % of patients whose patients preferences and underlying values were discussed; - % of patients whose decisional role preference was discussed as well as possible follow-up; - % of patients who feel they were adequately involved in decision making; - % of patients in whom beta-blockers are indicated who know the main pros and cons of beta-blocker therapy; - % of patients in whom ICD is indicated who know the main pros and cons of ICD implant;
To make the best use of resources.	- Mean cost of beta-blocker therapy; - Mean cost of ICD implantations; - Mean cost of molecular genetic testing; - Estimated cost of avoidable cardiac events; - Service cost/patient;
To promote and support research.	- Capture awareness of research undertaken; - Proportion of units with a defined person having a lead role to promote research and number of research publications; - % of staff undertaking research related course at university;
To train the professionals who support patients with LQTS.	- Structured education, consultation skills, and attitudes; - % of patients that are seen by an integrated, multidisciplinary team and expertise assessing them (cardiologist, nurses, mental health professionals, pharmacists); - % of staff trained in ECG;
To produce an annual report for the population served and to support quality improvement.	

Outcomes-based system for patients with LQTS who have not been identified

The second system focuses on family members of patients with LQTS in whom LQTS has not been recognized. The aim is to reduce the risk of cardiac events in these unidentified patients. The population to be served should be defined by all the practices in the region and the objectives and criteria of the service are listed in **Table 2**.

Table 2. Criteria defined to measure progress of each objective in the system aiming to reduce risk of cardiac events in family members of people with LQTS in whom LQTS has not been recognized yet

Objective	Criteria
To diagnose LQTS accurately in asymptomatic family members of LQTS patients	- Number of people with known LQTS; - Number of patients with known LQTS and confirmed genetic mutation; - Number of first degree relatives (parents, siblings and/or children) of LQTS patient that were informed and choosing to have or not to have molecular genetic testing (in a case mutation is known); - Number of first degree relatives (parents, siblings and/or children) of LQTS patient in whom molecular genetic testing confirmed genetic mutation and choosing to visit a cardiologist; - Number of first degree relatives (parents, siblings and/or children) of LQTS patient that were informed and choosing to have a cardiological examination (if no mutation is known); - % of first degree relatives (parents, siblings and/or children) of LQTS patient with prolonged QTc interval on ECG;
To treat an asymptomatic family member of LQTS patients safely and effectively.	- % of first degree relatives (parents, siblings and/or children) of patients with LQTS stratified on the basis of the LQTS-subtype with prolonged QTc-interval on ECG (≥ 470 ms) who are on beta-blocker therapy; - Number of first-degree relatives (parents, siblings and/or children) of patients with LQTS with failure to tolerate beta-blocker therapy; - % of first degree relatives of patients with LQTS stratified by age (children/adults) with normal QTc interval on ECG and positive genetic diagnosis who are on beta-blocker therapy (all and stratified by LQTS-subtype);
To accurately assess the risk of cardiac events in an asymptomatic family member of LQTS patients.	- % of first degree relatives of patients with LQTS who had age-stratified risk assessment by year-end using constellation of electrocardiographic, clinical, and genetic factors; - % of first degree relatives of patients with LQTS who had a risk assessment in the first year and who are in the second or subsequent year who have a review during the course of the year using age-stratified risk assessment based upon constellation of electrocardiographic, clinical, and genetic factors;
To ensure that asymptomatic family members of patients with LQTS make informed decisions that take their values into account.	- % of first degree relatives of patients with LQTS who were told their disease risk; - % of first degree relatives of patients with LQTS who participated in the decision to either undergo a particular form of screening and genetic testing or not; - % of first degree relatives of patients with LQTS who were explicitly told that a choice for treatment is to be made and that their opinion is important; - % of first degree relatives of patients with LQTS whom the options and pros and cons of each relevant treatment option were discussed with; - % of first degree relatives of patients with LQTS whose patients' preferences and underlying values were discussed; - % of first degree relatives of patients with LQTS whose decisional role preference was discussed as well as possible follow-up; - % of first degree relatives of patients with LQTS who feel they were adequately involved in decision making; - % of first degree relatives of patients with LQTS who know the main benefit and main risk of beta-blocker therapy;
To make the best use of resources.	- Mean cost of beta-adrenergic blockade therapy; - Mean cost of molecular genetic testing; - Service cost/patient;

Table 2. Criteria defined to measure progress of each objective in the system aiming to reduce risk of cardiac events in family members of people with LQTS in whom LQTS has not been recognized yet (*continued*)

Objective	Criteria
To promote and support research.	- Capture awareness of research undertaken; - Proportion of units with a defined person having a lead role to promote research; - % of staff undertaking research related courses at university;
To train the professionals	- Structured education, consultation skills, and attitudes; - Integrated, multidisciplinary team, and expertise (cardiologist, nurses, mental health professionals, pharmacists);
To produce an annual report for the population served and to support quality improvement.	

CONCLUSIONS

A major promise of the information deriving from 'omics' research is the transformation of healthcare and clinical decision-making through effective prevention programs, earlier diseases diagnosis and prognosis, and personalized treatments.[18] In this manuscript, we present an approach that can be used to deliver personalized care in a standardized way for LQTS, a condition for which genetic testing can provide new opportunities for patients' management, as stated by the Heart Rhythm Society (HRS) and the European Heart Rhythm Association (EHRA).[3]

Our work yielded two outcomes-based systems designed to reduce the risk of cardiac events in people with known LQTS and those who have LQTS but have not been identified. The systems are specifically designed to focus on the patient outcomes, which means that the systems are service agnostic, context-independent and applicable in a variety of healthcare organizations irrespective of resource constraints. Healthcare services can use these systems as a starting point to design their LQTS-focused healthcare services to focus more on patient outcomes and personalized care, while also tracking resource utilization for their services.

A key aspect of the systems is the requirement to produce an annual report that records data on outcomes delivered as well as resources used - thus giving an indication of the value (outcomes/resources used) of the service. We acknowledge that initially, the data will not be perfect - it may not be complete and the quality may not be great. Furthermore, even when there is agreement with the objectives and criteria, getting everyone in the system to work in a coordinated way and break down artificial silos may also be difficult. However, it is important to start shifting the culture and working practice of

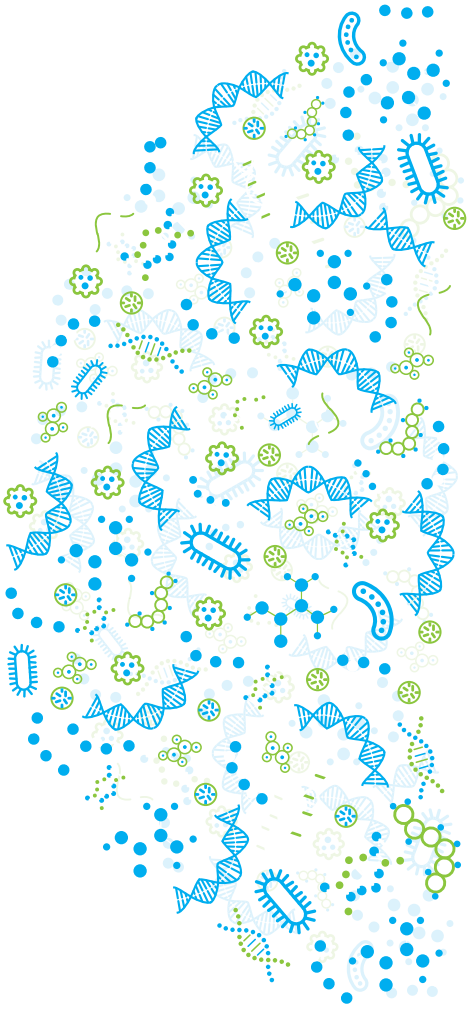
one's healthcare service to begin to think in a different way about how their service is designed and delivered and, most importantly, what it is accountable for. The data from annual reports can be used to:

- Determine how your service is evolving over time
- Identify gaps and/or areas where your service is not doing well (e.g. underuse/ underdiagnosis)
- Identify wasted resources in your service
- Determine how your service compares to other services serving similar demographics
- Improve transparency of choices and outcomes

The ultimate ambition in presenting this work is to create a learning and sharing network to identify new best practices as well as innovations, service level as well as technical, which can be used to deliver better outcomes, and optimize resource utilization, to patients and populations with LQTS globally.

REFERENCES

1. Jani, A. & Gray, M. Outcomes as a foundation for designing and building population healthcare systems in England. *BMJ Outcomes*, 16-19 (2015).
2. Gray, M. *How to practise population medicine*, (Offox Press, 2013).
3. Ackerman, M.J. *et al.* HRS/EHRA expert consensus statement on the state of genetic testing for the channelopathies and cardiomyopathies: this document was developed as a partnership between the Heart Rhythm Society (HRS) and the European Heart Rhythm Association (EHRA). *Europace* **13**, 1077-109 (2011).
4. Priori, S.G. *et al.* HRS/EHRA/APHRS expert consensus statement on the diagnosis and management of patients with inherited primary arrhythmia syndromes: document endorsed by HRS, EHRA, and APHRS in May 2013 and by ACCF, AHA, PACES, and AEPC in June 2013. *Heart Rhythm* **10**, 1932-63 (2013).
5. Sturm, A.C. Cardiovascular Cascade Genetic Testing: Exploring the Role of Direct Contact and Technology. *Front Cardiovasc Med* **3**, 11 (2016).
6. Alders, M. & Christiaans, I. Long QT Syndrome. *In: Pagon RA, Adam MP, Ardinger HH, et al., eds. GeneReviews® [Internet]. Seattle, WA: University of Washington* (2003: 1993-2017).
7. Modell, S.M., Bradley, D.J. & Lehmann, M.H. Genetic testing for long QT syndrome and the category of cardiac ion channelopathies. *PLoS Curr*, e4f9995f69e6c7 (2012).
8. Perez, M.V., Kumarasamy, N.A., Owens, D.K., Wang, P.J. & Hlatky, M.A. Cost-effectiveness of genetic testing in family members of patients with long-QT syndrome. *Circ Cardiovasc Qual Outcomes* **4**, 76-84 (2011).
9. Schwartz, P.J., Periti, M. & Malliani, A. The long Q-T syndrome. *Am Heart J* **89**, 378-90 (1975).
10. Schwartz, P.J. *et al.* Molecular diagnosis in a child with sudden infant death syndrome. *Lancet* **358**, 1342-1343 (2001).
11. Schwartz, P.J. *et al.* Prevalence of the congenital long-QT syndrome. *Circulation* **120**, 1761-7 (2009).
12. Schwartz, P.J. Idiopathic long QT syndrome: progress and questions. *Am Heart J* **109**, 399-411 (1985).
13. Nakano, Y. & Shimizu, W. Genetics of long-QT syndrome. *J Hum Genet* **61**, 51-5 (2016).
14. Wilde, A.A. & Pinto, Y.M. Cost-effectiveness of genotyping in inherited arrhythmia syndromes: are we getting value for the money? *Circ Arrhythm Electrophysiol* **2**, 1-3 (2009).
15. Phillips, K.A., Ackerman, M.J., Sakowski, J. & Berul, C.I. Cost-effectiveness analysis of genetic testing for familial long QT syndrome in symptomatic index cases. *Heart Rhythm* **2**, 1294-300 (2005).
16. Gonzalez, F.M., Veneziano, M.A., Puggina, A. & Boccia, S. A Systematic Review on the Cost-Effectiveness of Genetic and Electrocardiogram Testing for Long QT Syndrome in Infants and Young Adults. *Value Health* **18**, 700-8 (2015).
17. Wilde, A.A. *et al.* Clinical Aspects of Type 3 Long-QT Syndrome: An International Multicenter Study. *Circulation* **134**, 872-82 (2016).
18. Boccia, S. Why is personalized medicine relevant to public health? *European Journal of Public Health* **24**, 349-350 (2014).



Chapter 6

Summary/Samenvatting

SUMMARY

A substantial proportion of the world's disease burden arises from neurological and psychiatric disorders which are considered to be an important cause of death and disability worldwide. As these disorders are caused by a combination of genetic, environmental, and lifestyle factors, extensive research efforts have been invested to identify the molecular processes and pathways underlying these disorders and their related endophenotypes. As expanding our knowledge on the pathophysiology of these disorders and their endophenotypes may facilitate development of new prevention and treatment strategies, this thesis aimed to provide novel insights in the molecular mechanisms underlying most common neurological disorders including neurodegeneration and cerebrovascular pathology and the most common neurodevelopmental disorders by using several omics approaches, including genomics, epigenomics, metabolomics, and microbiomics.

Chapter 2 of the thesis is focused on neurodegeneration and its related endophenotypes including cognitive ability and brain volumetric measures obtained by brain imagining techniques. In **Chapter 2.1**, we studied common genetic determinants of lateral ventricular volume in participants included in the Cohorts for Heart and Aging Research in Genetic Epidemiology (CHARGE) consortium. We identified seven genetic loci that were significantly associated with lateral ventricular volume. Furthermore, several biological pathways including tau pathology, cytoskeleton organization, and S1P signaling emerged as relevant for lateral ventricular volume. In **Chapter 2.2**, we expanded the knowledge on genetic determinants of general cognitive function by identifying 32 novel genetic loci in the participants of European ancestry from the CHARGE consortium. These findings were generalized to participants of African American ancestry. Furthermore, we reported that genes underlying general cognitive function could be linked to circulating metabolites including tyrosine, creatinine, 22:6 docosahexaenoic acid (DHA), glycoprotein acetyls, acetate, and citrate. In **Chapter 2.3**, we explored blood-based DNA methylation in relation to cognitive ability in several cohorts. Significant associations were found for two CpG sites and different cognitive tests including global cognitive ability and executive function. In **Chapter 2.4**, I evaluated the effect of rare variants in dystrophin gene (*DMD*) on cognitive ability in general population. Suggestive associations were observed between cognitive ability and two rare variants among which one was predicted to have a damaging effect on the protein. Finally, in **Chapter 2.5**, we investigated the effect of structural variation in *DMD* gene on intellectual ability in patients with Duchenne muscular dystrophy. The results suggested that cumulative loss of dystrophin isoforms has an impact on intellectual ability.

Chapter 3 of the thesis reflected on neurovascular pathology. In **Chapter 3.1**, I performed a genome-wide linkage analysis of individuals in the extremes of intima-media thickness distribution (> 90th percentile) in a large family-based study from a genetically isolated population. Significant evidence of linkage was observed on four chromosomes and several plausible candidate genes were identified under the linkage peaks. In **Chapter 3.2**, I explored association of a broad range of metabolites with extra- and intracranial carotid artery calcification (ECAC and ICAC), as a proxy of carotid atherosclerosis, in a population-based setting. Significant evidence for association was found between 3-hydroxybutyrate and ICAC. Furthermore, the metabolic association pattern of ICAC was found to be different than that of ECAC providing evidence for location-specific differences in the etiology of atherosclerosis. In **Chapter 3.3,** I provided novel insights into association of circulating metabolites and risk of stroke in several population-based cohorts. The results suggested that several metabolites including amino acid histidine, glycolysis-related metabolite pyruvate, marker of acute phase reactions glycoprotein acetyls, two high-density lipoprotein (HDL) subfractions, and two low-density lipoprotein (LDL) subfractions were associated with risk of stroke. Three additional metabolites were identified when focusing on the risk of ischemic stroke including phenylalanine and two HDL lipoprotein subfractions. In **Chapter 3.4**, I explored relationship between gut microbiota and circulating metabolites. We found association between 41 microbial taxa and metabolite measures including specific lipoproteins subfractions such as very-low-density (VLDL) and HDL, serum lipid measures, glycolysis-related metabolites, amino acids, and acute phase reaction markers. These findings support the potential of gut microbiota as a target for therapeutic and preventive interventions.

In **Chapter 4**, I described genetic determinants of neurodevelopmental disorders including autism spectrum disorders (ASD) and attention deficit hyperactivity disorder (ADHD). In **Chapter 4.1**, I explored genetic determinants of ASD and quantitative autistic trait in several families with at least one child affected with ASD and reported association of a novel gene *TTC25* and quantitative autistic trait. Lastly, in **Chapter 4.2**, we described genetic determinants of ADHD symptoms in adults from several cohorts. Suggestive association was found between *STXBP5-AS1* and ADHD symptom scores in studied populations.

Finally, in **Chapter 5.1** of the thesis, I discussed main findings and commented on future research, while in **Chapter 5.2**, I proposed translational approach to help deliver personalized care for cardiovascular disorders at the population level. Using outcome-based healthcare system design, I created systems that could be used for effective reduction of risk of cardiac events in people with a condition for which genetic testing can provide new opportunities for patients' management and their first-degree relatives.

SAMENVATTING

Een substantieel deel van de ziektelast komt ten gevolge van neurologische en psychiatrische aandoeningen, die worden beschouwd als een belangrijke doodsoorzaak wereldwijd. Aangezien deze aandoeningen worden veroorzaakt door een combinatie van genetische, omgevings, en leefstijl factoren, is uitgebreid onderzoek gedaan naar de onderliggende moleculaire processen en pathways onderliggend aan deze aandoeningen en gerelateerde endophenotypen.

Omdat het uitbreiden van onze kennis over de pathofysiologie van deze aandoeningen en hun endofenotypen de ontwikkeling van nieuwe preventie- en behandelingsstrategieën kan bevorderen, heeft dit proefschrift als doel om nieuwe inzichten te verschaffen in de moleculaire mechanismen die ten grondslag liggen aan de meest voorkomende neurologische aandoeningen, waaronder neurodegeneratie en cerebrovasculaire pathologie en de meest voorkomende neurologische ontwikkelingsstoornissen. Dit wordt gedaan door verschillende benaderingen van omics te gebruiken, waaronder genomica, epigenomica, metabolomics en microbiomics.

Hoofdstuk 2 van dit proefschrift focust op neurodegeneratie en de daaraan gerelateerde endofenotypen, inclusief volumetrische metingen van de hersenen, verkregen door beeldvorming van de hersenen en cognitieve vaardigheidstesten. In **hoofdstuk 2.1** hebben wij gemeenschappelijke genetische determinanten van het laterale ventriculaire volume bestudeerd bij deelnemers van het Cohorts for Heart and Aging Research in Genetic Epidemiology (CHARGE) consortium. Wij ontdekten zeven genetische loci die significant geassocieerd waren met het laterale ventriculaire volume. Bovendien bleken verschillende biologische pathways waaronder tau-pathologie, cytoskeletorganisatie en S1P-signalering relevant voor het laterale ventriculaire volume.

In **hoofdstuk 2.2** hebben wij de kennis over genetische determinanten van algemene cognitieve functies uitgebreid door ontdekking van 32 nieuwe genetische loci in deelnemers van Europese afkomst binnen het CHARGE-consortium. Deze bevindingen waren te generaliseren naar deelnemers van Afro-Amerikaanse afkomst. Verder rapporteerden wij dat genen die ten grondslag liggen aan de algemene cognitieve functie kunnen worden gelinkt aan circulerende metabolieten, waaronder tyrosine, creatinine, 22:6 docosahexaeenzuur (DHA), glycoproteïne-acetyls, acetaat en citraat. In **hoofdstuk 2.3** hebben wij DNA-methylatie in het bloed onderzocht in relatie tot cognitieve vaardigheden in verschillende cohorten. Significante associaties werden gevonden voor twee CpG-sites en verschillende cognitieve testen waaronder globale cognitieve vaardigheid en executieve functie. In **Hoofdstuk 2.4** evalueerde ik het effect van zeldzame varianten

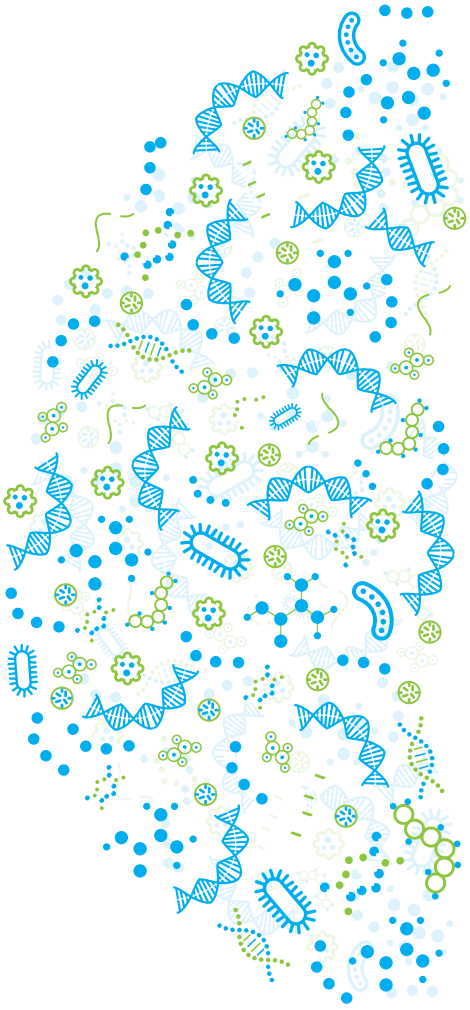
in het dystrofinegen (DMD) op cognitieve vaardigheden in de algemene populatie. Suggestieve associaties werden gezien tussen cognitieve vaardigheid en twee zeldzame varianten waarvan men voorspelde dat het een schadelijk effect op het eiwit had. Ten slotte onderzochten wij in **hoofdstuk 2.5** het effect van structurele variatie in het DMD gen op intellectuele capaciteiten bij patiënten met Duchenne spierdystrofie. De resultaten suggereerden dat cumulatief verlies van dystrofine isovormen van invloed is op het intellectuele vermogen.

Hoofdstuk 3 van dit proefschrift behandelt de neurovasculaire pathologie. In **Hoofdstuk 3.1** heb ik een genoom-wijde linkage analyse uitgevoerd op individuen in de extremen van de distributie van de intima media dikte (> 90e percentiel) in een groot familieonderzoek van een genetisch geïsoleerde populatie. Significant bewijs van linkage werd waargenomen op vier chromosomen en verschillende plausibele kandidaat genen werden gevonden. In **Hoofdstuk 3.2**, onderzocht ik de associatie van een breed scala van metabolieten met extra- en intracraniële carotis aderverkalking (ECAC en ICAC), als een proxy voor atherosclerose van de carotis, in een populatie-gebaseerde setting. Significant bewijs voor associatie werd gevonden tussen 3-hydroxybutyraat en ICAC. Bovendien bleek het metabole associatiepatroon van ICAC anders te zijn dan dat van ECAC, wat bewijs leverde voor locatie specifieke verschillen in de etiologie van atherosclerose. In **Hoofdstuk 3.3** verschafte ik nieuwe inzichten in de associatie van circulerende metabolieten en het risico op beroerte in verschillende populatie-gebaseerde cohorten. De resultaten suggereerden dat verschillende metabolieten, waaronder het aminozuur histidine, het aan glycolyse gerelateerde metaboliet pyruvaat, de marker voor acute fase-reacties glycoproteïne-acetyls, twee lipoproteïne subfracties met hoge dichtheid (HDL) en twee lipoproteïne subfracties met lage dichtheid (LDL), geassocieerd waren met het risico op beroerte. Drie andere metabolieten werden gevonden bij het focussen op het risico van ischemische beroerte waaronder fenylalanine en twee HDL lipoproteïne subfracties. In **hoofdstuk 3.4** heb ik de relatie tussen darmmicrobiota en circulerende metabolieten onderzocht. Wij vonden een associatie tussen 41 microbiële taxa en metabolietmetingen inclusief specifieke lipoproteïnen subfracties, zoals lipoproteïnen met zeer lage dichtheid (VLDL) en HDL, serumlipiden, glycolyse-gerelateerde metabolieten, aminozuren en markers voor acute fase. Deze bevindingen ondersteunen de potentie van darmmicrobiota als aangrijpingspunt voor therapeutische en preventieve interventies.

In **Hoofdstuk 4** beschreef ik genetische determinanten van neurologische ontwikkelingsstoornissen waaronder autismespectrumstoornissen (ASS) en aandachtstekortstoornis met hyperactiviteit (ADHD). In **Hoofdstuk 4.1** onderzocht ik genetische determinanten van ASS en de kwantitatieve autistische eigenschap in verschillende

families met ten minste één kind met ASS en een associatie van een nieuw gen *TTC25* en kwantitatieve autistische eigenschap. Ten slotte in **Hoofdstuk 4.2** beschreven wij genetische determinanten van symptomen van ADHD bij volwassenen uit verschillende cohorten. Een suggestieve associatie werd gevonden tussen *STXBP5-AS1* en ADHD-symptoomscores in onderzochte populaties.

Tot slot, in **hoofdstuk 5.1** van het proefschrift, besprak ik in de belangrijkste bevindingen en becommentarieerde ik toekomstig onderzoek, terwijl ik in **hoofdstuk 5.2** een translationele benadering voorstelde voor het leveren van gepersonaliseerde zorg voor cardiovasculaire aandoeningen op populatieniveau. Door gebruik te maken van een resultaatgebaseerd ontwerp van zorgsystemen, heb ik systemen ontwikkeld die kunnen worden gebruikt voor effectieve vermindering van het risico op cardiovasculaire events bij mensen met een aandoening waarvoor genetische testen nieuwe mogelijkheden kunnen bieden bij de management van patiënten en hun eerstegraads familieleden.



Chapter 7

Appendix

Chapter 7.1

Acknowledgements/Dankwoord

"It is good to have an end to journey toward; but it is the journey that matters, in the end."
Ursula K. Le Guin, The Left Hand of Darkness

Writing this thesis was possible with the help and great support of many, therefore, I would like to thank everyone who directly and indirectly contributed to the realization of this dissertation.

First of all, I would like to extend my deepest gratitude to my promotor Prof. Cornelia van Duijn and copromotor Dr. Najaf Amin. Dear Cornelia, thank you for your support, encouragement, and guidance. I am grateful for all the opportunities I was given to develop as a researcher and a person. Your energy, experience, broad knowledge, and ambition to make it to the top in science have been truly inspiring and motivating. Dear Najaf, thank you for all discussions, practical suggestions, and valuable advice. Your help and support with the projects have been of immense value. I am glad to have had a copromotor like you. Without you both, I would not be where I am now.

I would also like to express my deepest appreciation to the members of the reading committee Prof. Peter Koudstaal, Prof. Stéphanie Debette, and Prof. Arfan Ikram. Dear Prof. Koudstaal, thank you for not only critically reviewing my thesis but also accepting to be the secretary of the committee. Dear Prof. Debbate, thank you for agreeing to read this dissertation. It has been a real pleasure collaborating with you on several projects. I am thankful for all the discussions and efforts you have put into our project. Dear Prof. Ikram thank you for not only agreeing to be part of this committee but also for your feedback that helped me improve many of the papers included in this thesis.

I am extremely grateful to the plenary committee members Prof. Eline Slagboom, Prof. Aad van der Lugt, and Prof. Steven Kushner. Dear Prof. Slagboom, thank you the collaboration over recent years. Dear Prof. van der Lugt, thank you for the time and effort taken to read my manuscript. Dear Prof. Kushner, thank you for being part of the committee.

Posebnu zahvalnost dugujem Prof. Vedrani Milić Rašić. Draga Prof. Vedrana, hvala puno na nesebičnom deljenju znanja, podstreku da se bavim naučno-istrazivačkim radom i podrški tokom aplikovanja za ERAWEB projekat.

A special thank you to the team of ERAWEB project. Dear Prof. Hofman and Prof. Franco, thank you for the opportunity to study and research abroad. Dear Monique, Astrid, and Lennie, thank you for your help and support during my master and doctor of science studies.

This journey would not have been the same without the support of colleagues from GenEpi group. Dear Adriana, Ayse, Ashley, Andrea, Andy, Bernadette, Carla, Elisa, Dream, Hata, Ivana, Sven, Shahzad, Sara, Jeannette, Linda, Lennart, and Maaike, thank you for the wonderful time over recent years. Dear Ashley and Hata, first of all, thank you for being my paranymphs. Dear Ashely, I really appreciated our everyday conversations. Thank you for all discussions about the projects and for taking time to read my papers. Dear Hata, I enjoyed sharing the office with you. Thank you for listening to my concerns, taking time to read my papers, and for all the suggestions. Dear Adriana, I am grateful for all the moments and the time we shared together at the office and in private life. Thank you for being such a wonderful friend. Dear Sven, without you I would not have enjoyed this journey so much. Thank you for the constructive discussions, help, suggestions, and for being an amazing friend. Dear Ivana, it was a great pleasure to have a Serbian colleague in the group. Thank you for checking my presentations and papers and for making amazing cakes. Dear Shahzad, many thanks for all discussions about metabolomics and linkage analyses. I wish you all the best in completing your projects. Dear Dream, thank you for bringing a pleasant atmosphere in our group. I hope the big projects you are currently working on will get you the high impact factor you want. Dear Ayse, I am very grateful for your guidance on several metabolomics projects and for sharing your knowledge in metabolomics field with me. Dear Elisa, thank you for all nice talks and all the advice. Dear Maaike, I am grateful for your help and support. I am also grateful to Jeannette, Andrea, Andy, and Bernadette.

I would also like to extend my gratitude to colleagues from epidemiology and internal medicine departments. Dear Hieab, Gennady, Daniel, Maria, Irene, Olivera, Fadila, Bibi, Kate, Pooja, Naty, Carolina, Jeroen, Djawad, Robert, Maryam, Mohsen, Paula, Kim, Trudy, Natalie, Pieter, thank you for the good atmosphere at the department and positive interactions. I would also like to thank Frank van Rooij for taking care of the data, Nano for his help with my computer, and Solange and Andreas for arranging the paperwork for my residence permit.

It has also been a privilege to work with many international collaborators including the members of prestigious Cohorts for Heart Aging Research in Genomic Epidemiology (CHARGE) consortium. I am extremely grateful to all the senior researchers. Of these, I would especially like to thank Prof. Sudha Seshadri and Prof. Myriam Fornage. Dear Sudha, thank you for your valuable advice and encouragement throughout the duration of several projects included in this thesis. Your passion for science, enthusiasm, extensive knowledge and profound belief in the abilities of young researchers are admirable. Dear Myriam, I am thankful for all valuable advice, insightful suggestions, and guidance on several projects. It has been a real pleasure collaborating with you. I would also like

to extend my sincere thanks to members of PRECeDI project. Dear Anant, Muir, Carla, Martina, Olga, Stefania, and Anna, it has been truly amazing to work with you.

Special thanks to Strahinja, who together with Adriana, was part of my Gedempte Zalmhaven family. I am very lucky we got to share the apartment and many more things. Strahinja, hvala za mnoge lepe trenutke koje smo podelili predhodnih godina. Divim se tvojoj upornosti, istrajnosti, uspesima i radujem se našim novim okupljanjima. Ryan, thank you for your help with the projects but also movie nights and delicious dinners. Lieve Linda, heel erg bedankt voor de Nederlandse lessen en je hulp met mijn examen. Ik kijk uit naar meer van jou leerzame lessen in de toekomst.

Milorade, Ana, Tanja, Jeco hvala što ste uvek tu za mene. Beskrajno se radujem našim porukama i novim sustretima.

Neizmernu zahvalnost dugujem porodici. Mama i tata, hvala vam na razumevanju, podršci i što ste uvek uz mene. Sve što sam postigla do sada ne bi bilo moguće bez vas. Vi ste me naučili da se svaki rad i trud isplati. Veliku zahvalnost dugujem i seki i zetu, hvala na svemu. Radujem se našim novim druženjima. I konačno, Nina i Ena, nisam mogla da poželim slađe sestričine. ☺

Finally, I would like to thank Marco for his support, tremendous understanding, unconditional love, and encouragement during the last months of my PhD. I am grateful to have you in my life.

Chapter 7.2

PhD Portfolio

PhD PORTFOLIO

Name PhD student: Dina Vojinovic	PhD period: 2014-2018
Research School: Netherlands Institute for Health Sciences (NIHES)	Promotor: Prof. C.M. van Duijn
Erasmus MC department: Epidemiology	Copromotor: Dr. N. Amin

1. PhD training

	Year	Workload (ECTS)*
Courses		
- Research Integrity	2017	0.3
Presentations		
International conferences and meetings		
- Joint Congress of European Neurology, Istanbul, Turkey: "*The dystrophin gene and cognitive function in the general population*" (poster)	2014	1
- European Society of Human Genetics, Glasgow, UK: "*Exome-wide association analysis of attention hyperactivity disorder in a genetically isolated population*" (poster)	2015	1
- CVON In-Control meeting, Utrecht, the Netherlands: "*A promising biomarker from gut microbiota: TMAO*" (oral)	2015	1
- CVON In-Control meeting, Utrecht, the Netherlands: "*A key biomarker from gut microbiota: TMAO - resolving the measurement issues-*" (oral)	2015	1
- Genomics of Brain Disorders, Cambridge, UK: "*Genome-wide association scan identifies novel genes associated with autism spectrum disorder*" (oral)	2016	1
- International Stroke Genomic Consortium, Milan, Italy: "*Exome sequencing and chip analyses of small subcortical MRI-defined brain infarcts*" (oral)	2016	1
- Netherlands Metabolomics Centre Meeting, Leiden, the Netherlands: "*TMAO and other microbiome associated metabolites in cardiovascular disease*" (oral)	2017	1
- CHARGE Investigator Meeting, New York, USA: "*Exome sequencing and chip analyses of small subcortical MRI-defined brain infarcts*" (poster)	2017	1
- MGC Symposium, Rotterdam, the Netherlands: "Genome-wide association study of brain ventricular volume" (oral)	2017	1
- European Stroke Organization Conference, Prague, Czech Republic: "*Exome sequencing and chip analyses of small subcortical MRI-defined brain infarcts*" (oral)	2017	1
- Translational congress Better together, Utrecht, the Netherlands: "*Metabolomics as an intermediate between the gut microbiome and cardiovascular disease*" (poster)	2017	1
Oral presentations at lab meetings		
- Genetic Epidemiology Unit, Erasmus MC, Rotterdam, the Netherlands	2014-2018	4.5
- 2020 meeting at the Department of Epidemiology, Erasmus MC, Rotterdam, the Netherlands	2016	1

International Conferences and Consortium meetings		
- CVON In-Control, multiple consortium meetings, the Netherlands	2014-2017	1
- Exploring Human Host-Microbiome Interactions in Health and Disease, Cambridge, UK	2015	1
- CHARGE Investigator Meeting, Houston, USA	2016	1.1
- CoSTREAM, multiple meetings	2016-2017	1
- CHARGE Investigator Meeting, New York, USA	2017	1.1
- International Stroke Genomic Consortium, Utrecht, the Netherlands	2017	0.8
- CHARGE Investigator Meeting, Boston, USA	2017	1.1
- PRECeDI consortium meeting, Oxford, UK	2017	1.1
- BBMRI-NL consortium meetings, the Netherlands	2018	0.8
- PRECeDI consortium meeting, Amsterdam, the Netherlands	2018	0.8
- CHARGE Investigator Meeting, Rotterdam, the Netherlands	2018	1.1
Seminars, symposia, and workshops		
- Weekly seminars and 2020 meetings at the Department of Epidemiology	2014-2018	2
- Weekly scientific seminars of Genetic Epidemiology Unit, Department of Epidemiology	2014-2018	1
- Molecular Epidemiology research meetings	2016-2018	1
Other		
- Research fellow at the BVHC and Linkcare as part of PRECeDI, a Marie Sklodowska–Curie Research and Innovation Staff Exchange (MCSA RISE) program	2016	3 months
- Travel award CHARGE meeting	2017	
2. Teaching activities		
Teaching assistant		
- Teaching assistant for the NIHES Summer Programme course "Principles of Genetic Epidemiology"	2014-2016	1.5
- Teaching assistant for the NIHES course "Biostatistical Methods I: basic principles"	2014-2015	1.5
Lecturing		
- "Introduction to next-generation sequencing data analysis"	2017-2018	3
Supervising Master's thesis		
- Daniel Tibussek (Master student): "*Impact of deleterious mutations in genes associated with severe intellectual disability on cognitive functioning in the general population*"	2014-2016	3
Other		
- Peer review of articles	2017-2018	0.5

* 1 ECTS (European Credit Transfer System) equals a workload of 28 hours

Chapter 7.3

List of publications and manuscripts

Vojinovic D, Adams HHH, Jian X, Yang Q, Smith A, Bis J, Teumer A, Scholz M, Armstrong N, Hofer E, Saba Y, Luciano M, Bernard M, Trompet S, Yang J, Gillespie N, van der Lee SJ, Neumann A, Ahmad S, Andreassen O, Ames D, Amin N, Arfanakis K, Bastin M, Becker D, Beiser A, Beyer F, Brodaty H, Bryan RN, Bülow R, Dale A, De Jager P, Deary I, DeCarli C, Fleischman D, Gottesman R, van der Grond J, Gudnason V, Harris T, Homuth G, Knopman D, Kwok J, Lewis C, Li S, Loeffler M, Lopez O, Maillard P, El Marroun H, Mather K, Mosley T, Muetzel R, Nauck M, Nyquist P, Panizzon M, Pausova Z, Psaty B, Rice K, Rotter J, Royle N, Satizabal C, Schmidt R, Schofield P, Schreiner P, Sidney S, Stott D, Thalamuthu A, Uitterlinden A, Valdés Hernández M, Vernooij M, Wen W, White T, Witte V, Wittfeld K, Wright M, Yanek L, Tiemeier H, Kremen W, Bennett, Jukema JW, Paus T, Wardlaw J, Schmidt H, Sachdev P, Villringer A, Grabe H, Longstreth W, van Duijn CM, Launer L, Seshadri S, Ikram MA, Fornage M. Genome-wide association study of 23,500 individuals identifies 7 loci associated with brain ventricular volume. Nat Commun. 2018

Feitosa MF, Kraja AT, Chasman DI, Sung YJ, Winkler TW, Ntalla I, Guo X, Franceschini N, Cheng CY, Sim X, **Vojinovic D**, Marten J, Musani SK, Li C, Bentley AR, Brown MR, Schwander K, Richard MA, Noordam R, Aschard H, Bartz TM, Bielak LF, Dorajoo R, Fisher V, Hartwig FP, Horimoto ARVR, Lohman KK, Manning AK, Rankinen T, Smith AV, Tajuddin SM, Wojczynski MK, Alver M, Boissel M, Cai Q, Campbell A, Chai JF, Chen X, Divers J, Gao C, Goel A, Hagemeijer Y, Harris SE, He M, Hsu FC, Jackson AU, Kähönen M, Kasturiratne A, Komulainen P, Kühnel B, Laguzzi F, Luan J, Matoba N, Nolte IM, Padmanabhan S, Riaz M, Rueedi R, Robino A, Said MA, Scott RA, Sofer T, Stančáková A, Takeuchi F, Tayo BO, van der Most PJ, Varga TV, Vitart V, Wang Y, Ware EB, Warren HR, Weiss S, Wen W, Yanek LR, Zhang W, Zhao JH, Afaq S, Amin N, Amini M, Arking DE, Aung T, Boerwinkle E, Borecki I, Broeckel U, Brown M, Brumat M, Burke GL, Canouil M, Chakravarti A, Charumathi S, Ida Chen YD, Connell JM, Correa A, de Las Fuentes L, de Mutsert R, de Silva HJ, Deng X, Ding J, Duan Q, Eaton CB, Ehret G, Eppinga RN, Evangelou E, Faul JD, Felix SB, Forouhi NG, Forrester T, Franco OH, Friedlander Y, Gandin I, Gao H, Ghanbari M, Gigante B, Gu CC, Gu D, Hagenaars SP, Hallmans G, Harris TB, He J, Heikkinen S, Heng CK, Hirata M, Howard BV, Ikram MA; InterAct Consortium, John U, Katsuya T, Khor CC, Kilpeläinen TO, Koh WP, Krieger JE, Kritchevsky SB, Kubo M, Kuusisto J, Lakka TA, Langefeld CD, Langenberg C, Launer LJ, Lehne B, Lewis CE, Li Y, Lin S, Liu J, Liu J, Loh M, Louie T, Mägi R, McKenzie CA, Meitinger T, Metspalu A, Milaneschi Y, Milani L, Mohlke KL, Momozawa Y, Nalls MA, Nelson CP, Sotoodehnia N, Norris JM, O'Connell JR, Palmer ND, Perls T, Pedersen NL, Peters A, Peyser PA, Poulter N, Raffel LJ, Raitakari OT, Roll K, Rose LM, Rosendaal FR, Rotter JI, Schmidt CO, Schreiner PJ, Schupf N, Scott WR, Sever PS, Shi Y, Sidney S, Sims M, Sitlani CM, Smith JA, Snieder H, Starr JM, Strauch K, Stringham HM, Tan NYQ, Tang H, Taylor KD, Teo YY, Tham YC, Turner ST, Uitterlinden AG, Vollenweider P, Waldenberger M, Wang L, Wang YX, Wei WB, Williams C, Yao J, Yu C, Yuan JM, Zhao W, Zonderman AB, Becker DM,

Boehnke M, Bowden DW, Chambers JC, Deary IJ, Esko T, Farrall M, Franks PW, Freedman BI, Froguel P, Gasparini P, Gieger C, Jonas JB, Kamatani Y, Kato N, Kooner JS, Kutalik Z, Laakso M, Laurie CC, Leander K, Lehtimäki T, Study LC, Magnusson PKE, Oldehinkel AJ, Penninx BWJH, Polasek O, Porteous DJ, Rauramaa R, Samani NJ, Scott J, Shu XO, van der Harst P, Wagenknecht LE, Wareham NJ, Watkins H, Weir DR, Wickremasinghe AR, Wu T, Zheng W, Bouchard C, Christensen K, Evans MK, Gudnason V, Horta BL, Kardia SLR, Liu Y, Pereira AC, Psaty BM, Ridker PM, van Dam RM, Gauderman WJ, Zhu X, Mook-Kanamori DO, Fornage M, Rotimi CN, Cupples LA, Kelly TN, Fox ER, Hayward C, van Duijn CM, Tai ES, Wong TY, Kooperberg C, Palmas W, Rice K, Morrison AC, Elliott P, Caulfield MJ, Munroe PB, Rao DC, Province MA, Levy D. Novel genetic associations for blood pressure identified via gene-alcohol interaction in up to 570K individuals across multiple ancestries. PLoS One. 2018 Jun 18;13(6):e0198166.

Vojinovic D, van der Lee SJ, van Duijn CM, Vernooij MW, Kavousi M, Amin N, Demirkan A, Ikram MA, van der Lugt A, Bos D. Metabolic profiling of intra- and extracranial carotid artery atherosclerosis. Atherosclerosis. 2018 Mar 8;272:60-65.

Ahmad S, Bannister C, van der Lee SJ, **Vojinovic D**, Adams HHH, Ramirez A, Escott-Price V, Sims R, Baker E, Williams J, Holmans P, Vernooij MW, Ikram MA, Amin N, van Duijn CM. Disentangling the biological pathways involved in early features of Alzheimer's disease in the Rotterdam Study. Alzheimers Dement. 2018 Feb 26. pii: S1552-5260(18)30024-4.

Sung YJ, Winkler TW, de Las Fuentes L, Bentley AR, Brown MR, Kraja AT, Schwander K, Ntalla I, Guo X, Franceschini N, Lu Y, Cheng CY, Sim X, **Vojinovic D**, Marten J, Musani SK, Li C, Feitosa MF, Kilpeläinen TO, Richard MA, Noordam R, Aslibekyan S, Aschard H, Bartz TM, Dorajoo R, Liu Y, Manning AK, Rankinen T, Smith AV, Tajuddin SM, Tayo BO, Warren HR, Zhao W, Zhou Y, Matoba N, Sofer T, Alver M, Amini M, Boissel M, Chai JF, Chen X, Divers J, Gandin I, Gao C, Giulianini F, Goel A, Harris SE, Hartwig FP, Horimoto ARVR, Hsu FC, Jackson AU, Kähönen M, Kasturiratne A, Kühnel B, Leander K, Lee WJ, Lin KH, 'an Luan J, McKenzie CA, Meian H, Nelson CP, Rauramaa R, Schupf N, Scott RA, Sheu WHH, Stančáková A, Takeuchi F, van der Most PJ, Varga TV, Wang H, Wang Y, Ware EB, Weiss S, Wen W, Yanek LR, Zhang W, Zhao JH, Afaq S, Alfred T, Amin N, Arking D, Aung T, Barr RG, Bielak LF, Boerwinkle E, Bottinger EP, Braund PS, Brody JA, Broeckel U, Cabrera CP, Cade B, Caizheng Y, Campbell A, Canouil M, Chakravarti A; CHARGE Neurology Working Group, Chauhan G, Christensen K, Cocca M; COGENT-Kidney Consortium, Collins FS, Connell JM, de Mutsert R, de Silva HJ, Debette S, Dörr M, Duan Q, Eaton CB, Ehret G, Evangelou E, Faul JD, Fisher VA, Forouhi NG, Franco OH, Friedlander Y, Gao H; GIANT Consortium, Gigante B, Graff M, Gu CC, Gu D, Gupta P, Hagenaars SP, Harris TB, He J, Heikkinen S, Heng CK, Hirata M, Hofman A, Howard BV, Hunt S, Irvin MR, Jia Y, Joehanes R, Justice

AE, Katsuya T, Kaufman J, Kerrison ND, Khor CC, Koh WP, Koistinen HA, Komulainen P, Kooperberg C, Krieger JE, Kubo M, Kuusisto J, Langefeld CD, Langenberg C, Launer LJ, Lehne B, Lewis CE, Li Y; Lifelines Cohort Study, Lim SH, Lin S, Liu CT, Liu J, Liu J, Liu K, Liu Y, Loh M, Lohman KK, Long J, Louie T, Mägi R, Mahajan A, Meitinger T, Metspalu A, Milani L, Momozawa Y, Morris AP, Mosley TH Jr., Munson P, Murray AD, Nalls MA, Nasri U, Norris JM, North K, Ogunniyi A, Padmanabhan S, Palmas WR, Palmer ND, Pankow JS, Pedersen NL, Peters A, Peyser PA, Polasek O, Raitakari OT, Renström F, Rice TK, Ridker PM, Robino A, Robinson JG, Rose LM, Rudan I, Sabanayagam C, Salako BL, Sandow K, Schmidt CO, Schreiner PJ, Scott WR, Seshadri S, Sever P, Sitlani CM, Smith JA, Snieder H, Starr JM, Strauch K, Tang H, Taylor KD, Teo YY, Tham YC, Uitterlinden AG, Waldenberger M, Wang L, Wang YX, Wei WB, Williams C, Wilson G, Wojczynski MK, Yao J, Yuan JM, Zonderman AB, Becker DM, Boehnke M, Bowden DW, Chambers JC, Chen YI, de Faire U, Deary IJ, Esko T, Farrall M, Forrester T, Franks PW, Freedman BI, Froguel P, Gasparini P, Gieger C, Horta BL, Hung YJ, Jonas JB, Kato N, Kooner JS, Laakso M, Lehtimäki T, Liang KW, Magnusson PKE, Newman AB, Oldehinkel AJ, Pereira AC, Redline S, Rettig R, Samani NJ, Scott J, Shu XO, van der Harst P, Wagenknecht LE, Wareham NJ, Watkins H, Weir DR, Wickremasinghe AR, Wu T, Zheng W, Kamatani Y, Laurie CC, Bouchard C, Cooper RS, Evans MK, Gudnason V, Kardia SLR, Kritchevsky SB, Levy D, O'Connell JR, Psaty BM, van Dam RM, Sims M, Arnett DK, Mook-Kanamori DO, Kelly TN, Fox ER, Hayward C, Fornage M, Rotimi CN, Province MA, van Duijn CM, Tai ES, Wong TY, Loos RJF, Reiner AP, Rotter JI, Zhu X, Bierut LJ, Gauderman WJ, Caulfield MJ, Elliott P, Rice K, Munroe PB, Morrison AC, Cupples LA, Rao DC, Chasman DI. A Large-Scale Multi-ancestry Genome-wide Study Accounting for Smoking Behavior Identifies Multiple Significant Loci for Blood Pressure. Am J Hum Genet. 2018 Mar 1;102(3):375-400.

Marioni RE*, McRae AF*, Bressler J*, Colicino E*, Hannon E*, Li S*, Prada D*, Smith JA*, Trevisi L*, Tsai PC*, **Vojinovic D***, Simino J, Levy D, Liu C, Mendelson M, Satizabal CL, Yang Q, Jhun MA, Kardia SLR, Zhao W, Bandinelli S, Ferrucci L, Hernandez DG, Singleton AB, Harris SE, Starr JM, Kiel DP, McLean RR, Just AC, Schwartz J, Spiro A 3rd, Vokonas P, Amin N, Ikram MA, Uitterlinden AG, van Meurs JBJ, Spector TD, Steves C, Baccarelli AA, Bell JT, van Duijn CM, Fornage M, Hsu YH, Mill J, Mosley TH, Seshadri S, Deary IJ. Meta-analysis of epigenome-wide association studies of cognitive abilities. Mol Psychiatry. 2018 Jan 8.

Koeks Z, Bladen CL, Salgado D, van Zwet E, Pogoryelova O, McMacken G, Monges S, Foncuberta ME, Kekou K, Kosma K, Dawkins H, Lamont L, Bellgard MI, Roy AJ, Chamova T, Guergueltcheva V, Chan S, Korngut L, Campbell C, Dai Y, Wang J, Barišić N, Brabec P, Lähdetie J, Walter MC, Schreiber-Katz O, Karcagi V, Garami M, Herczegfalvi A, Viswanathan V, Bayat F, Buccella F, Ferlini A, Kimura E, van den Bergen JC, Rodrigues M, Roxburgh R, Lusakowska A, Kostera-Pruszczyk A, Santos R, Neagu E, Artemieva S, Rasic VM, **Vojinovic**

D, Posada M, Bloetzer C, Klein A, Díaz-Manera J, Gallardo E, Karaduman AA, Oznur T, Topaloğlu H, El Sherif R, Stringer A, Shatillo AV, Martin AS, Peay HL, Kirschner J, Flanigan KM, Straub V, Bushby K, Béroud C, Verschuuren JJ, Lochmüller H. Clinical Outcomes in Duchenne Muscular Dystrophy: A Study of 5345 Patients from the TREAT-NMD DMD Global Database. J Neuromuscul Dis. 2017;4(4):293-306.

Vojinovic D, Brison N, Ahmad S, Noens I, Pappa I, Karssen LC, Tiemeier H, van Duijn CM, Peeters H, Amin N. Variants in *TTC25* affect autistic trait in patients with autism spectrum disorder and general population. Eur J Hum Genet. 2017 Aug;25(8):982-987.

Natarajan P, Bis JC, Bielak LF, Cox AJ, Dörr M, Feitosa MF, Franceschini N, Guo X, Hwang SJ, Isaacs A, Jhun MA, Kavousi M, Li-Gao R, Lyytikäinen LP, Marioni RE, Schminke U, Stitziel NO, Tada H, van Setten J, Smith AV, **Vojinovic D**, Yanek LR, Yao J, Yerges-Armstrong LM, Amin N, Baber U, Borecki IB, Carr JJ, Chen YI, Cupples LA, de Jong PA, de Koning H, de Vos BD, Demirkan A, Fuster V, Franco OH, Goodarzi MO, Harris TB, Heckbert SR, Heiss G, Hoffmann U, Hofman A, Išgum I, Jukema JW, Kähönen M, Kardia SL, Kral BG, Launer LJ, Massaro J, Mehran R, Mitchell BD, Mosley TH Jr, de Mutsert R, Newman AB, Nguyen KD, North KE, O'Connell JR, Oudkerk M, Pankow JS, Peloso GM, Post W, Province MA, Raffield LM, Raitakari OT, Reilly DF, Rivadeneira F, Rosendaal F, Sartori S, Taylor KD, Teumer A, Trompet S, Turner ST, Uitterlinden AG, Vaidya D, van der Lugt A, Völker U, Wardlaw JM, Wassel CL, Weiss S, Wojczynski MK, Becker DM, Becker LC, Boerwinkle E, Bowden DW, Deary IJ, Dehghan A, Felix SB, Gudnason V, Lehtimäki T, Mathias R, Mook-Kanamori DO, Psaty BM, Rader DJ, Rotter JI, Wilson JG, van Duijn CM, Völzke H, Kathiresan S, Peyser PA, O'Donnell CJ; CHARGE Consortium. Multiethnic Exome-Wide Association Study of Subclinical Atherosclerosis. Circ Cardiovasc Genet. 2016 Dec;9(6):511-520.

Sung YJ, Winkler TW, Manning AK, Aschard H, Gudnason V, Harris TB, Smith AV, Boerwinkle E, Brown MR, Morrison AC, Fornage M, Lin LA, Richard M, Bartz TM, Psaty BM, Hayward C, Polasek O, Marten J, Rudan I, Feitosa MF, Kraja AT, Province MA, Deng X, Fisher VA, Zhou Y, Bielak LF, Smith J, Huffman JE, Padmanabhan S, Smith BH, Ding J, Liu Y, Lohman K, Bouchard C, Rankinen T, Rice TK, Arnett D, Schwander K, Guo X, Palmas W, Rotter JI, Alfred T, Bottinger EP, Loos RJ, Amin N, Franco OH, van Duijn CM, **Vojinovic D**, Chasman DI, Ridker PM, Rose LM, Kardia S, Zhu X, Rice K, Borecki IB, Rao DC, Gauderman WJ, Cupples LA. An Empirical Comparison of Joint and Stratified Frameworks for Studying G × E Interactions: Systolic Blood Pressure and Smoking in the CHARGE Gene-Lifestyle Interactions Working Group. Genet Epidemiol. 2016 Jul;40(5):404-15.

Olfson E, Saccone NL, Johnson EO, Chen LS, Culverhouse R, Doheny K, Foltz SM, Fox L, Gogarten SM, Hartz S, Hetrick K, Laurie CC, Marosy B, Amin N, Arnett D, Barr RG, Bartz

TM, Bertelsen S, Borecki IB, Brown MR, Chasman DI, van Duijn CM, Feitosa MF, Fox ER, Franceschini N, Franco OH, Grove ML, Guo X, Hofman A, Kardia SL, Morrison AC, Musani SK, Psaty BM, Rao DC, Reiner AP, Rice K, Ridker PM, Rose LM, Schick UM, Schwander K, Uitterlinden AG, **Vojinovic D**, Wang JC, Ware EB, Wilson G, Yao J, Zhao W, Breslau N, Hatsukami D, Stitzel JA, Rice J, Goate A, Bierut LJ. Rare, low frequency and common coding variants in CHRNA5 and their contribution to nicotine dependence in European and African Americans. Mol Psychiatry. 2016 May;21(5):601-7.

Milic Rasic V, **Vojinovic D**, Pesovic J, Mijalkovic G, Lukic V, Mladenovic J, Kosac A, Novakovic I, Maksimovic N, Romac S, Todorovic S, Savic Pavicevic D. Intellectual ability in the duchenne muscular dystrophy and dystrophin gene mutation location. Balkan J Med Genet. 2015 Apr 10;17(2):25-35.

Bladen CL, Salgado D, Monges S, Foncuberta ME, Kekou K, Kosma K, Dawkins H, Lamont L, Roy AJ, Chamova T, Guergueltcheva V, Chan S, Korngut L, Campbell C, Dai Y, Wang J, Barišić N, Brabec P, Lahdetie J, Walter MC, Schreiber-Katz O, Karcagi V, Garami M, Viswanathan V, Bayat F, Buccella F, Kimura E, Koeks Z, van den Bergen JC, Rodrigues M, Roxburgh R, Lusakowska A, Kostera-Pruszczyk A, Zimowski J, Santos R, Neagu E, Artemieva S, Rasic VM, **Vojinovic D**, Posada M, Bloetzer C, Jeannet PY, Joncourt F, Díaz-Manera J, Gallardo E, Karaduman AA, Topaloğlu H, El Sherif R, Stringer A, Shatillo AV, Martin AS, Peay HL, Bellgard MI, Kirschner J, Flanigan KM, Straub V, Bushby K, Verschuuren J, Aartsma-Rus A, Béroud C, Lochmüller H. The TREAT-NMD DMD Global Database: analysis of more than 7,000 Duchenne muscular dystrophy mutations. Hum Mutat. 2015 Apr;36(4):395-402.

Vojinovic D, Adams HH, van der Lee SJ, Ibrahim-Verbaas CA, Brouwer R, van den Hout MC, Oole E, van Rooij J, Uitterlinden A, Hofman A, van IJcken WF, Aartsma-Rus A, van Ommen GB, Ikram MA, van Duijn CM, Amin N. The dystrophin gene and cognitive function in the general population. Eur J Hum Genet. 2015 Jun;23(6):837-43.

Vojinovic D, Kavousi M, Ghanbari M, Brouwer RWW, Rooij JGJ, van den Hout MCGN, Kraaij R, van IJcken WFJ, Uitterlinden AG, van Duijn CM, Amin N. Whole-genome linkage scan combined with exome sequencing identifies novel candidate genes for carotid intima-media thickness. Front Genet. 2018

Vojinovic D, Puggina A, van der Werf C, van El CG, Damman OC, Amin N, Demirkan A, Stricker BH, Gray M, Boccia S, Cornel MC, van Duijn CM, Jani A. A model for mass personalization in cardiology: standard outcomes-based systems that can deliver personalized care. Submitted for publication.

Arias-Vásquez A*, Groffen AJ*, Spijker S*, Ouwens KG*, Klein M*, **Vojinovic D***, Galesloot TE, Bralten J, Hottenga JJ, van der Most PJ, Kattenberg VM, Pool R, Nolte IM, Penninx BWJH, Fedko IO, Dolan CV, Nivard MG, den Braber A, van Duijn CM, Hoekstra PJ, Buitelaar JK, Kiemeney B, Hoogman M, Middeldorp CM, Draisma HHM, Vermeulen SH, Sánchez-Mora C, Ramos-Quiroga JA, Ribasés M, The EAGLE-ADHD Consortium, Hartman CA, Kooij JJS, Amin N, Smit AB**, Franke B**, Boomsma DI**. STXBP5 Antisense RNA 1 gene and adult ADHD symptoms. Submitted for publication.

Sung YJ, de las Fuentes L, Winkler TW, Chasman DI, Bentley AR, Kraja AT, Ntalla I, Warren HR, Guo X, Schwander K, Manning AK, Brown MR, Aschard H, Feitosa MF, Franceschini N, Lu Y, Cheng CY, Sim X, **Vojinovic D**, Marten J, Musani SK, Kilpeläinen TO, Richard MA, Aslibekyan S, Bartz TM, Dorajoo R, Li C, Liu Y, Rankinen T, Smith AV , Tajuddin SM, Tayo BO, Zhao W, Zhou Y, Matoba N, Sofer T, Alver M, Amini M, Boissel M, Chai JF, Chen X, Divers J, Gandin I, Gao C, Giulianini F, Goel A, Harris SE, Hartwig FP, Horimoto AR, Hsu FC, Jackson AU, Kammerer CM, Kasturiratne A, Komulainen P, Kühnel B, Leander K, Lee WJ, Lin KH, Luan J, Lyytikainen LP, McKenzie CA, Meian H, Nelson CP, Noordam R, Scott RA, Sheu WHH, Stančáková A, Takeuchi F, van der Most PJ, Varga TV, Waken RJ, Wang H, Wang Y, Ware EB, Weiss S, Wen W, Yanek LR, Zhang W, Zhao JH, Afaq S, Alfred T, Amin N, Arking D, Aung T, Barr G, Bielak LF, Boerwinkle E, Bottinger EP, Braund PS, Brody JA, Broeckel U, Cade B, Caizheng Y, Campbell A, Canouil M, Chakravarti A, Cocca M, Collins FS, Connell JM, de Mutsert R, de Silva HJ, Dörr M, Duan Q, Eaton CB, Ehret G, Evangelou E, Faul JD, Forouhi NG, Franco OH, Friedlander Y, Gao H, Gigante B, Gu CC, Gupta P, Hagenaars SP, Harris TB, He J, Heikkinen S, Heng CK, Hofman A, Howard BV, Hunt S, Irving MR, Jia Y, Katsuya T, Kaufman J, Kerrison ND, Khor CC, Koh WP, Koistinen HA, Kooperberg CB, Krieger JE, Kubo M, Kutalik Z, Kuusisto J, Lakka TA, Langefeld CD, Langenberg C, Launer LJ, Lee JH, Lehne B, Levy D, Lewis CE, Li Y, Lifelines cohort study, Lim SH, Liu CT, Liu J, Liu J, Liu K, Liu Y, Loh M, Lohman KK, Louie T, Mägi R, Matsuda K, Meitinger T, Metspalu A, Milani L, Momozawa Y, Mosley TH Jr, Nalls MA, Nasri U, O'Connell JR, Ogunniyi A, Palmas WR, Palmer ND, Pankow JS, Pedersen NL, Peters A, Peyser PA, Polasek O, Porteous D, Raitakari OT, Renström F, Rice TK, Ridker PM, Robino A, Robinson JG, Rose LM, Rudan I, Sabanayagam C, Salako BL, Sandow K, Schmidt CO, Schreiner PJ, Scott WR, Sever P, Sims M, Sitlani CM, Smith BH, Smith JA, Snieder H, Starr JM, Strauch K, Tang H, Taylor KD, Teo YY, Tham YC, Uitterlinden AG, Waldenberger M, Wang L, Wang YX, Wei WB, Wilson G, Wojczynski MK, Xiang Y, Yao J, Yuan JM, Zonderman AB, Becker DM, Boehnke M, Bowden DW, Chambers JC, Chen YDI, Weir DR, de Faire U, Deary IJ, Esko T, Farrall M, Forrester T, Barry I Freedman, Froguel P, Gasparini P, Gieger C, Lesssa Horta B, Hung YJ, Jonas JB, Kato N, Kooner JS, Laakso M, Lehtimäki T, Liang KW, Magnusson PKE, Oldehinkel AJ, Pearls T, Pereira AC, Rauramaa R, Redline S, Rettig R, Samani NJ, Scott J, Shu XO, van der Harst P, Wagenknecht LE, Wareham NJ, Watkins H, Wickremasinghe AR, Wu T, Kamatani Y, Laurie

CC, Bouchard C, Cooper RS, Evans MK, Gudnason V, Hixson J, Kardia SLR, Kritchevsky SB, Psaty BM, van Dam RM, Arnett DK, Mook-Kanamori DO, Fornage M, Fox ER, Hayward C, van Duijn CM, Tai ES, Wong TY, Loos RJF, Reiner AP, Rotimi CN, Bierut LJ, Zhu X, Cupples LA, Province MA, Rotter JI, Franks PW, Rice K, Elliott P, Caulfield MJ, Gauderman WJ, Munroe PB, Rao DC, Morrison AC. Multi-ancestry genome-wide study incorporating gene-smoking interactions identifies multiple new loci for pulse pressure and mean arterial pressure. Submitted for publication.

de Vries PS*, Brown MR*, Bentley AR*, Yun J Sung YJ*, Winkler TW*, Ntalla I*, Schwander K, Kraja AT, Guo X, Franceschini N, Cheng CY, Sim X, **Vojinovic D**, Huffman JE, Musani SK, Li C, Feitosa MF, Richard MA, Noordam R, Aschard H, Bartz TM, Bielak LF, Deng X, Dorajoo R, Lohman KK, Manning AK, Rankinen T, Smith AV, Tajuddin SM, Evangelou E, Graff M, Alver M, Boissel M, Chai JF, Chen X, Divers J, Gandin I, Gao C, Goel A, Hagemeijer Y, Harris SE, Hartwig FP, He M, Horimoto ARVR, Hsu FC, Jackson AU, Kasturiratne A, Komulainen P, Kühnel B, Laguzzi F, Lee JH, Luan J, Lyytikäinen LP, Matoba N, Nolte IM, Pietzner M, Riaz M, Said MA, Scott RA, Sofer T, Stančáková A, Takeuchi F, Tayo BO, Van der Most PJ, Varga TV, Wang Y, Ware EB, Wen W, Xiang YB, Yanek LR, Zhang W, Zhao JH, Afaq S, Amin N, Amini M, Arking DE, Aung T, Ballantyne C, Boerwinkle E, Broeckel U, Campbell A, Canouil M, Charumathi S, Chen YDI, Connell JM, de Faire U, de las Fuentes L, de Mutsert R, de Silva HJ, Ding J, Dominiczak AF, Duan Q, Eaton CB, Eppinga RN, Faul JD, Fisher V, Forouhi NG, Forrester T, Franco OH, Friedlander Y, Ghanbari M, Giulianini F, Grabe HJ, Grove ML, Gu CC, Hallmans G, Harris TB, Heikkinen S, Heng CK, Hirata M, Hixson JE, Howard BV, Ikram MA, InterAct Consortium, Jacobs Jr DR, Johnson C, Jonas JB, Kammerer CM, Katsuya T, Khor CC, O Kilpeläinen TO, Koh WP, Koistinen HA, Kolcic I, Kooperberg C, Krieger JE, Kritchevsky SB, Kubo M, Kuusisto J, Lakka TA, Langefeld CD, Langenberg CL, Launer LJ, Lehne BC, Lemaitre RN, Li Y, Liang J, Liu J, Liu K, Loh M, Louie T, Mägi R, Manichaikul AW, McKenzie CA, Meitinger T, Metspalu A, Milaneschi Y, Milani L, Mohlke KL, Momozawa Y, Mosley Jr TH, Mukamal KJ, Nalls MA, Nauck M, Nelson CP, Nona S, O'Connell JR, Palmer ND, Pazoki R, Pedersen NL, Peters A, Peyser PA, Polasek O, Poulter N, Raffel LJ, Raitakari OT, Reiner AP, Rice TK, Rich SS, Robino A, Robinson JG, Rose LM, Rudan I, Schmidt CO, Schreiner PJ, Scott WR, Sever P, Shi Y, Sidney S, Sims M, Smith BH, Smith JA, Snieder H, Starr JM, Strauch K, Tan N, Taylor KD, Teo YY, Tham YC, Uitterlinden AG, van Heemst D, Vuckovic D, Waldenberger M, Wang L, Wang Y, Wang Z, Wei WB, Williams C, Wilson Sr G, Wojczynski MK, Yao J, Yu B, Yu C, Yuan JM, Zhao W, Zonderman AB, Becker DM, Boehnke M, Bowden DW, Chambers JC, Cooper RS, Deary IJ, Esko T, Farrall M, Franks PW, Freedman BI, Froguel P, Gasparini P, Gieger C, Horta BL, Kamatani Y, Kato N, Kooner JS, Laakso M, Laurie CC, Leander K, Lehtimäki T, Lifelines Cohort Study, Magnusson PKE, Oldehinkel AJ, Pearls T, Penninx B, Pereira AC, Rauramaa R, Samani NJ, Scott J, Shu XO, van der Harst P, Wagenknecht LE, Wang YX, Wareham NJ, Watkins H, Weir DR, Wickremasinghe AR, Wu T,

Zheng W, Elliott P, North KE, Bouchard C, Evans MK, Gudnason V, Liu CT, Liu Y, Psaty BM, Ridker PM, van Dam RM, Kardia SLR, Zhu X, Rotimi CN, Mook-Kanamori DO, Fornage M, Kelly TN, Fox ER, Hayward C, van Duijn CM, Tai ES, Wong TY, Liu J, Rotter JI, Gauderman WJ, Provinc MA, Munroe PB**, Rice K**, Chasman DI**, Cupples LA**, Rao DC**, Morrison AC**. Multi-ancestry genome-wide association study incorporating gene-alcohol interactions identifies new lipid loci. Submitted for publication.

Bentley AR*, Sung YJ*, Brown MR*, Kraja AT*, Winkler TW*, Ntalla I*, Schwander K, Lim E, Deng X, Guo X, Liu J, Lu Y, Cheng CY, Sim X, **Vojinovic D**, Huffman JE, Musani SK, Li C, Feitosa MF, Richard MA, Noordam R, Baker J, Aschard H, Bartz TM, Chasman DI, Ding J, Dorajoo R, Manning AK, Rankinen T, Smith AV, Tajuddin SM, Zhao W, Alver M, Boissel M, Chai JF, Chen X, Divers J, Evangelou E, Gao C, Goel A, Hagemeijer Y, Harris SE, Hartwig FP, He M, Horimoto ARVR, Hsu FC, Hung YJ, Jackson AU, Kasturiratne A, Komulainen P, Kühnel B, Leander K, Lin KH, Luan J, Lyytikäinen LP, Matoba N, Nolte IM, Pietzner M, Prins B, Riaz M, Robino A, Said MA, Schupf N, Scott RA, Sofer T, Stančáková A, Takeuchi F, Tayo BO, Van der Most PJ, Varga TV, Wang TD, Wang Y, Ware EB, Wen W, Xiang YB, Yanek LR, Zhang W, Zhao JH, Adeyemo A, Afaq S, Amin N, Amini M, Arking DE, Arzumanyan Z, Aung T, Ballantyne C, Barr GR, Bielak LF, Boerwinkle E, Bottinger EP, Broeckel U, Brown M, Cade RE, Campbell A, Canouil M, Charumathi S, Chen G, Chen YDI, Christensen K, COGENT-Kidney Consortium, Concas MP, Connell JM, de las Fuentes L, de Silva HJ, de Vries PS, Doumatey A, Duan Q, Eaton CB, Eppinga RN, Faul JD, Floyd JS, Forouhi NG, Forrester T, Franco OH, Friedlander Y, Gandin I, Gao H, Gharib SA, The GIANT Consortium, Gigante B, Giulianini F, Grabe HJ, Graff M, Gu CC, Harris TB, Heikkinen S, Heng CK, Hirata M, Hixson JE, Ikram MA, InterAct Consortium, Jacobs Jr DR, Jia Y, Joehanes R, Johnson C, Jonas JB, Justice AE, Katsuya T, Khor CC, Kilpeläinen TO, Koh WP, Kolcic I, Kooperberg C, Krieger JE, Kritchevsky SB, Kubo M, Kuusisto J, Lakka TA, Langefeld CD, Langenberg C, Launer LJ, Lehne BC, Lewis CE, Li Y, Liang J, Lin S, Liu CT, Liu J, Liu K, Loh M, Lohman KK, Louie T, Luzzi A, Mägi R, Mahajan A, Manichaikul AW, McKenzie CA, Meitinger T, Metspalu A, Milaneschi Y, Milani L, Mohlke KL, Momozawa Y, Morris AP, Murray AD, Nalls MA, Nauck M, Nelson CP, North K, O'Connell JR, Palmer ND, Papanicolau GJ, Pedersen NL, Peters A, Peyser PA, Polasek O, Poulter N, Raitakari OT, Reiner AP, Renström F, Rice TK, Rich SS, Robinson JG, Rose LM, Rosendaal FR, Rudan I, Schmidt CO, Schreiner PJ, Scott WR, Sever P, Shi Y, Sidney S, Sims M, Smith JA, Snieder H, Starr JM, Strauch K, Stringham HM, Tan NYQ, Tang H, Taylor KD, Teo YY, Tham YC, Tiemeier H, Turner ST, Uitterlinden AG, Understanding Society Scientific Group, van Heemst D, Waldenberger M, Wang H, Wang L, Wang L, Wei WB, Williams C, Wilson Sr G, Wojczynski MK, Yao J, Young K, Yu C, Yuan JM, Zhou J, Zonderman AB, Becker DM, Boehnke M, Bowden DW, Chambers JC, Cooper RS, de Faire U, Deary IJ, Eleftheria Z, Elliot P, Esko T, Farrall M, Franks PW, Freedman BI, Froguel P, Gasparini P, Gieger C, Horta BL, Juang JMJ, Kamatani Y, Kammerer CM, Kato N, Kooner JS, Laakso M, Laurie CC, Lee

IT, Lehtimäki T, Lifelines Cohort Study, Magnusson PKE, Oldehinkel AJ, Penninx B, Pereira AC, Rauramaa R, Redline S, Samani NJ, Scott J, Shu XO, van der Harst P, Wagenknecht LE, Wang JS, Wang YX, Wareham NJ, Watkins H, Weir DR, Wickremasinghe AR, Wu T, Zheng W, Bouchard C, Evans MK, Gudnason V, Kardia SLR, Liu Y, Psaty BM, Ridker PM, van Dam RM, Mook-Kanamori DO, Fornage M, Province MA, Kelly TN, Fox ER, Hayward C, van Duijn CM, Tai ES, Wong TY, Loos RJF, Franceschini N, Rotter JI, Zhu X, LJ Bierut, Gauderman WJ, Rice K**, Munroe PB**, Morrison AC**, Rao DC**, Rotimi CN**, Cupples LA**. Multi-ancestry Genome-wide Association Study incorporating Gene × Smoking Interactions Identifies Novel Lipid Loci. Submitted for publication.

Kilpeläinen TO, Bentley AR, Noordam R, Sung YJ, Schwander K, Winkler TW, Jakupović H, Manning A, Aschard H, Ntalla I, Brown MR, de las Fuentes L, Franceschini N, Guo X, **Vojinovic D**, Aslibekyan S, Feitosa MF, Kho M, Musani SK, Richard M, Wang H, Wang Z, Bartz TM, Bielak LF, Campbell A, Chasman DI, Dorajoo R, Fisher V, Hartwig FP, Horimoto AR, Li C, Lohman KK, Marten J, Sim X, Smith AV, Tajuddin SM, Alver M, Amini M, Boissel M, Chai JF, Chen X, Divers J, Evangelou E, Gao C, Graff M, Harris SE, He M, Hsu FC, Jackson AU, Hua ZJ, Kraja AT, Kühnel B, Laguzzi F, Lyytikäinen LP, Nolte IM, Rauramaa R, Riaz M, Robino A, Rueedi R, Stringham HM, Takeuchi F, van der Most PJ, Varga TV, Verweij N, Ware EB, Wen W, Xiaoyin L, Yanek LR, Amin N, Arnett DK, Boerwinkle E, Brumat M, Cade B, Canouil M, Chen YDI, Concas MP, Connell J, de Mutsert R, de Silva HJ, de Vries PS, Demirkan A, Ding J, Eaton CB, Faul JD, Friedlander Y, Gabriel KP, Ghanbari M, Giulianini F, Gu CC, Gu D, Harris TB, He J, Heikkinen S, Heng CK, Hunt SC, Ikram MA, Jonas JB, Koh WP, Komulainen P, Krieger JE, Kritchevsky SB, Kutalik Z, Kuusisto J, Langefeld CD, Langenberg C, Launer LJ, Leander K, Lemaitre RN, Lewis CE, Lifelines Cohort Study, Liu J, Mägi R, Manichaikul AW, Meitinger T, Metspalu A, Milaneschi Y, Mohlke KL, Mosley Jr TH, Murray AD, Nalls MA, Nang EEK, Nelson CP, Nona S, Norris JM, Nwuba CV, O'Connell J, Palmer ND, Papanicolau GJ, Pazoki R, Pedersen NL, Peters A, Peyser PA, Polasek O, Porteous DJ, Poveda A, Raitakari OT, Rich SS, Risch N, Robinson JG, Rose LM, Rudan I, Schreiner PJ, Scott RA, Sidney SS, Sims M, Jennifer A. Smith,, Harold Snieder, Tamar Sofer,, John M. Starr,, Barbara Sternfeld, Strauch K, Tang H, Taylor K, Tsai M, Tuomilehto J, Uitterlinden AG, van de Ende YM, van Heemst D, Voortman T, Waldenberger M, Wilson G, Xiang YB, Yao J, Yu C, Yuan JM, Zhao W, Zonderman AB, Becker DM, Boehnke M, Bowden DW, de Faire U, Deary IJ, Elliott P, Esko T, Freedman BI, Froguel P, Gasparini P, Gieger C, Kato N, Laakso M, Lakka TA, Lehtimäki T, Magnusson PKE, Oldehinkel AJ, Penninx BWJH, Samani NJ, Shu XO, van der Harst P, Van Vliet-Ostapchouk JV, Vollenweider P, Wagenknecht LE, Wang YX, Wareham NJ, Weir DR, Wu T, Zheng W, Zhu X, Evans MK, Franks PW, Gudnason V, Hayward C, Horta BL, Kelly TN, Liu Y, North KE, Pereira AC, Ridker PM, Tai ES, van Dam RM, Fox ER, Kardia SLR, Liu CT, Mook-Kanamori DO, Province MA, Redline S, van Duijn CM, Rotter JI, Kooperberg CB, Gauderman WJ, Psaty BM, Rice K, Munroe PB, Fornage M,

Cupples LA, Rotimi CN, Morrison AC, Rao DC, Loos RJF. Multi-Ancestry Genome-Wide Study of Blood Lipid Levels Identifies Four Loci Interacting with Physical Activity. Submitted for publication.

Kunkle BW*, Grenier-Boley*, Sims R, Bis JC, Damotte V, Naj AC, Boland A, Vronskaya M, van der Lee SJ, Amlie-Wolf A, Bellenguez C, Frizatti A, Chouraki V, Martin ER, Sleegers K, Badarinarayan N, Jakobsdottir J, Hamilton-Nelson KL, Aloso R, Raybould R, Chen Y, Kuzma AB, Hiltunen M, Morgan T, Ahmad S, Vardarajan BN, Epelbaum J, Hoffmann P, Boada M, Beecham GW, Garnier JG, Harold D, Fitzpatrick AL, Valladares O, Moutet ML, Gerrish A, Smith AV, Qu L, Bacq D, Denning N, Jian X, Zhao Y, Zompo MD, Fox NC, Grove ML, Choi SH, Mateo I, Hughes JT, Adams HH, Malamon J, Garcia FS, Patel Y, Brody JA, Dombroski B, Naranjo MCD, Daniilidou M, Eiriksdottir G, Mukherjee S, Wallon D, Uphill J, Aspelund T, Cantwell LB, Garzia F, Galimberti D, Hofer E, Butkiewics M, Fin B, Scarpini E, Sarnowski C, Bush W, Meslage S, Kornhuber J, White CC, Song Y, Barber RC, Engelborghs S, Pichler S, **Vojinovic D**, Adams PM, Vandenberghe R, Mayhaus M, Cupples LA, Albert MS, De Deyn PP, Gu W, Himali JJ, Beekly D, Squassina A, Hartmann AM, Orellana A, Blacker D, Rodriguez-Rodriguez E, Lovestone S, Garcia ME, Doody RS, Fernadez CM, Sussams R, Lin H, Fairchild TJ, Benito YA, Holmes C, Comic H, Frosch MP, Thonberg H, Maier W, Roschupkin G, Ghetti B, Giedraitis V, Kawalia A, Li S, Huebinger RM, Kilander L, Moebus S, Hernández I, Kamboh MI, Brundin R, Turton J, Yang Q, Katz MJ, Concari L, Lord J, Beiser AS, Keene CD, Helisalmi S, Kloszewska I, Kukull WA, Koivisto AM, Lynch ATarraga L, Larson EB, Haapasalo A, Lawlor B, Mosley TH, Lipton RB, Solfrizzi V, Gill M, Longstreth WT Jr, Montine TJ, Frisardi V, Ortega-Cubero S, Rivadeneira F, Petersen RC, Deramecourt V, Ciaramella A, Boerwinkle E, Reiman EM, Fievet N, Caltagirone C, Rotter JI, Reisch JS, Hanon O, Cupidi C, Uitterlinden AG, Royall DR, Dufouil C, Maletta RG, Moreno-Grau S, Sano M, Brice A, Cecchetti R, St George-Hyslop P, Ritchie K, Tsolaki M, Tsuang DW, Dubois B, Craig D, Wu CK, Soininen H, Avramidou D, Albin RL, Fratiglioni L, Germanou A, Apostolova LG, Keller L, Koutroumani M, Arnold SE, Panza F, Gkatzima O, Asthana S, Hannequin D, Whitehead P, Atwood CS, Caffarra P, Hampel H, Baldwin CT, Lannfelt L, Rubinsztein DC, Barnes LL, Pasquier F, Frölich L, Barral S, McGuinness B, Beach TG, Johnston J, Becker JT, Passmore P, Bigio EH, Schott JM, Bird TD, Warren JD, Boeve BF, Lupton MK, Bowen JD, Proitsi P, Boxer A, Powell JF, Burke JR, Kauwe JK, Burns JM, Mancuso M, Buxbaum JD, Bonuccelli U, Cairns NJ, McQuillin A, Cao C, Livingston G, Carlson CS, Bass NJ, Carlsson CM, Hardy J, Carney RM, Bras J, Carrasquillo MM, Guerreiro R, Allen M, Chui HC, Fisher E, Cribbs DH, Masullo C, Crocco EA, DeCarli C, Bisceglio G, Dick M, Ma L, Duara R, Graff-Radford NR, Evans DA, Hodges A, Faber KM, Scherer M, Fallon KB, Riemenschneider M, Fardo DW, Heun R, Farlow MR, Ferris S, Leber M, Foroud TM, Heuser I, Galasko DR, Giegling I, Gearing M, Hüll M, Geschwind DH, Gilbert JR, Morris J, Green RC, Mayo K, Growdon JH, Feulner T, Hamilton RL, Harrell LE, Drichel D, Honig LS, Cushion TD, Huentelman MJ,

Hollingworth P, Hulette CM, Hyman BT, Marshall R, Jarvik GP, Meggy A, Abner E, Menzies G, Jin LW, Leonenko G, Jun G, Grozeva D, Karydas A, Russo G, Kaye JA, Kim R, Jessen F, Kowall NW, Vellas B, Kramer JH, Vardy E, LaFerla FM, Jöckel KH, Lah JJ, Dichgans M, Leverenz JB, Mann D, Levey AI, Pickering-Brown S, Lieberman AP, Klopp N, Lunetta KL, Wichmann HE, Lyketsos CG, Morgan K, Marson DC, Brown K, Martiniuk F, Medway C, Mash DC, Nöthen MM, Masliah E, Hooper NM, McCormick WC, Daniele A, McCurry SM, Bayer A, McDavid AN, Gallacher J, McKee AC, van den Bussche H, Mesulam M, Brayne C, Miller BL, Riedel-Heller S, Miller CA, Miller JW, Al-Chalabi A, Morris JC, Shaw CE, Myers AJ, Wiltfang J, O'Bryant S, , Olichney JM, Alvarez V, Parisi JE, Singleton AB, Paulson HL, Collinge J, Perry W, Mead S, Peskind E, Rosser M, Pierce A,Ryan N, Poon WW, Nacmias B, Potter H, Sorbi S, Quinn JF, Sacchinelli E, Raj A, Spalletta G, Raskind M, Bossù P, Reisberg B, Clarke R, Reitz C, Smith AD, Ringman JM, Warden D, Roberson ED, Wilcock G, Rogaeva E, Bruni AC, Rosen HJ, Gallo M, Rosenberg RN, Ben-Shlomo Y, Sager MA, Mecocci P, Saykin AJ, Pastor P, Cuccaro ML, Vance JM, Schneider JA, Schneider LS, Seeley WW, Smith AG, Sonnen JA, Spina S, Stern RA, Swerdlow RH, Tanzi RE, Trojanowski JQ, Troncoso JC, Van Deerlin VM, Van Eldik LJ, Vinters HV, Vonsattel JP, Weintraub S, Welsh-Bohmer KA, Wilhelmsen KC, Williamson J, Wingo TS, Woltjer RL, Wright CB, Yu CE, Yu L, Alzheimer Disease Genetics Consortium (ADGC), The European Alzheimer's Disease Initiative (EADI), Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium (CHARGE), Genetic and Environmental Risk in AD/Defining Genetic, Polygenic and Environmental Risk for Alzheimer's Disease Consortium (GERAD/PERADES), Pilotto A, Bullido M.J., Peters O, Crane PK, Bennett DA, Bosco P, Coto E, Boccardi V, De Jager PL, Lleo A, Warner N, Lopez OL, McDonough S, Ingelsson M, Deloukas P, Cruchaga C, Graff C, Gwilliam R, Fornage M, Goate AM, Sanchez-Juan P, Kehoe PG, Amin N, Ertekin-Taner N, Berr C, Debette S, Love S, Launer LJ, Younkin SG, Dartigues JF, Corcoran C, Ikram MA, Dickson DW, Campion D, Tschanz J, Schmidt H, Hakonarson H, Clarimon J, Munger R, Schmidt R, Farrer LA, Van Broeckhoven C, O'Donovan MC, DeStefano AL, Jones L, Haines JL, Deleuze JF, Owen MJ, Gudnason V, Mayeux R, Escott-Price V, Psaty BM, Ramirez A, Wang LS, Ruiz A**, van Duijn CM**, Holmans PA**, Seshadri S**, Williams J** Amouyel P**, Schellenberg GD**, Lambert JC**, Pericak-Vance MA**. Meta-analysis of genetic association with diagnosed Alzheimer's disease identifies novel risk loci and implicates Abeta, Tau, immunity and lipid processing. Submitted for publication.

Mishra A, Chauhan G, ViolleauMH, **Vojinovic D**, Jian X, Bis J, Li S, Saba Y, Yang Q, Bartz TM, Hofer E, Soumare A, Peng F, Duperron MG, Mosley TH, Schmidt R, Psaty BM, Launer LJ, Boerwinkle E, Zhu Y, Mazoyer B, Lathrop M, van Duijn CM, Ikram MA, Schmidt H, Longstreth WT, Fornage M, Seshadri S, Joutel A, Tzourio C, Debette S. Association of variants in HTRA1 and NOTCH3 with MRI-defined extremes of small vessel disease of brain in older community persons: an exome sequencing study. In preparation.

Bressler J, Satizabal CL, Bis JC, **Vojinovic D**, Amin N, DeStefano AL, Fitzpatrick AL, Lopez OL, Psaty BM, van Duijn CM, Fornage M, Mosley TH, Boerwinkle E, Seshadri S. Whole Exome Sequencing Identifies Novel Genes Implicating DNA Methylation and Branched Chain Amino Acid Pathways Associated with Memory Function in the CHARGE Consortium. In preparation.

Vojinovic D*, Radjabzadeh D*, Kurilshikov A*, Amin N, Wijmenga C, Franke L, Uitterlinden AG, Zhernakova A, Fu J**, Kraaij R**, van Dujin CM**. Relationship between gut microbiota and circulating metabolites in population-based cohorts. In preparation.

Vojinovic D, Kalaoja M, Trompet S, Fischer K, Shipley MJ, Li S, Havulinna AS, Perola M, Salomaa V, Yang Q, Sattar N, Jousilahti P, Amin N, Vasan RS, Ikram MA, Ala-Korpela M, Jukema JW, Seshadri S, Kettunen J, Kivimaki M, Esko T, van Duijn CM. Circulating metabolites and risk of stroke in seven population-based cohorts. In preparation.

Vojinovic D, Hayward C, Smith JA, Zhao W, Bressler J, Trompet S, Sarnowski C, Sargurupremraj M, Yang J, Timmers PRHJ, Hansell NK, Ahola-Olli A, Krapohl E, Bis JC, Gustavson DE, Palviainen T, Saba Y, Thalamuthu A, Giddaluru S, Weinhold L, Amin N, Armstrong N, Bielak LF, Böhmer AC, Boyle PA, Brodaty H, Campbell H, Clark DW, Couvy-Duchesne B, De Jager PL, Elman JA, Espeseth T, Faul JD, Fitzpatrick A, Gordon SD, Hankemeier T, Hofer E, Ikram MA, Joshi PK, Kaddurah-Daouk R, Kaprio J, Kardia SLR, Kentistou KA, Kleineidam L, Kochan N, Kwok J, Leber M, Lee T, Lehtimäki T, Loukola A, Lundquist A, Lyytikäinen LP, Mather K, Montgomery GW, Reppermund S, Rose RJ, Rovio S, Sachdev P, Schmid M, Schmidt H, Uitterlinden AG, Vuoksimaa E, Wagner M, Wagner H, Weir DR, Wright MJ, Yu M, Nyberg L, Ramirez A, Le Hellard S, Ames D, Schofield P, Schmidt R, Dick D, Porteous D, Kremen WS, Psaty BM, Raitakari O, Martin NG, Wilson JF, Bennett DA, Debette S, Jukema JW, Mosley TH, Jr, Seshadri S, van Dujin CM. Genetic determinants of general cognitive function and their association to circulating metabolites: a cross-omics study. In preparation.

Chapter 7.4

About the author

Dina Vojinović was born in Čačak, Serbia, on March 21th 1985. In 2004 she completed her Gymnasium in Čačak, Serbia and started her medical studies at the Faculty of Medicine, University of Belgrade, Belgrade, Serbia, where she obtained her medical degree in 2010. The same year she started working as a medical doctor at the Clinic for Neurology and Psychiatry for Children and Youth in Belgrade, Serbia and as a researcher on a scientific project on neuromuscular diseases funded by the Ministry of Science and Technological Development, Belgrade, Serbia. In 2012, she received ERAWEB scholarship and moved to the Netherlands where she continued her education and received Master of Science and Doctor of Science degrees in Genetic Epidemiology at Netherlands Institute of Health Sciences (NIHES). In 2014 she started her PhD at the Genetic Epidemiology Unit, Department of Epidemiology, Erasmus Medical Center, Rotterdam, the Netherlands under the supervision of Prof. Cornelia M. van Duijn. In 2016, as a part of Marie Curie Research and Innovation Staff Exchange, Dina spent three months at Better Value Healthcare, Oxford, United Kingdom and Linkcare Health Services, Barcelona, Spain where she worked on personalized prevention of chronic diseases. After obtaining her PhD degree, Dina will continue working at the Department of Epidemiology.