Author:
Sheikhali Babaei, Mahsa

Title:
Multi-omics analyses to investigate molecular mechanisms underlying atrial fibrillation and stroke disease

Multi-omics analyses to investigate molecular mechanisms underlying atrial fibrillation and stroke disease

Mahsa Sheikhali Babaei

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Population Health Sciences

MRC Integrative Epidemiology Unit
Bristol Medical School
University of Bristol
UK

November 2021

Word count: 44,834

Abstract

Background: Genome-wide association studies (GWAS) in Europeans have robustly associated 111 loci with atrial fibrillation (AF) and 22 loci with stroke risk. However, the functional consequences of these associations have yet to be elucidated. Therefore, this thesis seeks to identify shared genetic effects between methylomic, transcriptomic and metabolomic traits to help improve our understanding of molecular mechanisms underlying stroke and AF.

Methods: To investigate this I developed and applied a multiple trait colocalization pipeline using the Bayesian “moloc” method. Molecular traits were considered to be colocalized with AF or stroke if they had a posterior probability of association (PPA)>80%. In addition, to demonstrate that there was evidence that genetic susceptibility for AF was linked to stroke, a two-sample Mendelian randomization (MR) study was conducted.

Results: In Phase I of the study, 23 AF and 11 stroke loci were found to colocalize between DNA methylation and circulating metabolites within the *cis* region. In Phase II of the study, eQTL data was integrated on the top findings from Phase I. Multiple CpG sites, gene expression and metabolites (mainly lipids and lipoproteins) were found to colocalize with AF and stroke, suggesting shared regulatory relationships between these intermediate phenotypes. Of the 34 prioritized loci, only the 16q22 region (harbouring the *HP* and *ZFHX3* genes) colocalized with both AF and stroke traits. MR analysis suggested that genetic predisposition to AF increases risk of stroke, although there was some evidence for a reverse MR effect.

Conclusions: This thesis demonstrates the application of multi-omics approach to discover genetic pathways linked to cardiovascular disease, and illustrates complexities around issues involving statistical power, directionalities of molecular effects, and tissue specificity. The moloc pipeline and framework developed could be applied to other diseases in the future and will become increasingly valuable as new molecular datasets are published.

Acknowledgements

I greatly appreciate the contributions of the following individuals, organizations and groups to the research work undertaken in this doctoral thesis:

My PhD supervisors: Prof Tom Gaunt, Dr Josine Min and Prof Caroline Relton, for their guidance, support and imparted wisdom over the last four years.

My father for the funding of my 4-year studentship.

The ALSPAC families and participants.

My colleagues in the MRC-IEU at the Bristol Medical School, in particular Dr Denis Baird, Dr Jie Zheng and Dr Thomas Richardson for their advice and support.

My parents and family, for their endless support and encouragement throughout the four years.

Declaration

I declare that the work in this thesis was carried out in accordance with the requirements of the University of Bristol's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of others, is indicated as such. Any views expressed in this thesis are those of the author.

Signed: ____________________ Date: ________________________

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Acronyms

AF	Atrial fibrillation
AFGen	Atrial Fibrillation Genetics
ALSPAC	Avon Longitudinal Study of Parents and Children
ARIES	Accessible Resource for Integrated Epigenomics Studies
CAD	Coronary artery disease
CES	Cardioembolic stroke
CHD	Coronary heart disease
CpG	Cytosine-phosphate-Guanine
CVD	Cardiovascular disease
DNAm	DNA methylation
eQTL	Expression QTL
eQTS	Expression quantitative trait score
EWAS	Epigenome-wide association studies
GCTA	Genome-wide complex trait analysis
GoDMC	Genetics of DNA Methylation Consortium
GTEx	Genotype-Tissue Expression Consortium
GWAS	Genome-wide association study
HRC	Haplotype Reference Consortium
HVH	Heart and Vascular Health
ICH	Intracerebral hemorrhage
IS	Ischemic Stroke
ISGC	International Stroke Genetics Consortium
IV	Instrumental variable

IVW	Inverse-variance weighted
LAS	Large-artery atherosclerotic stroke
LD	Linkage disequilibrium
LncRNA	Long non-coding RNAs
LOO	Leave-one-out
MAF	Minor allele frequency
MbQTL	Metabolite QTL
Moloc	Multiple-trait colocalization
mQTL	Methylation QTL
MR	Mendelian randomization
NINDS	The National Institute of Neurological Disorders and Stroke
NMR	Nuclear magnetic resonance
PC	Principal component
PP	Posterior probability
PPA	Posterior probability of association
PRS	Polygenic risk score
pQTL	Protein QTL
pQTS	Protein quantitative trait score
PWCoCo	Pairwise conditional and colocalization
QTL	Quantitative trait locus
SiGN	NINDS Stroke Genetics Network
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SVS	Small-vessel stroke

TWAS	Transcriptome-wide association studies
UCLEB	The University College-London School-Edinburgh-Bristol consortium
WR	Wald ratio

Chapter 1 Introduction

1.1 Background to the genetics of common diseases

Over evolutionary time the genome accumulates point mutations, or single-nucleotide variants (SNVs), at random to the genetic code. A SNV is when a nucleotide base pair (Adenine (A), Cytosine (C), Guanine, (G), Thymine (T)) in a person's DNA is substituted by a different base pair. If a SNV reaches an appreciable frequency in the population, for example a minor allele frequency of greater than 0.01 (MAF > 0.01), whereby different variations of the alleles become fixed in the population, then it becomes classed as a single-nucleotide polymorphism (SNP). The 1000 Genomes project[1] mapped 84.7 million SNPs with a MAF>0.01 in the human genome. Many of these SNPs have been demonstrated to be associated with risk of common, complex diseases, suggesting the genomic regions harbouring these SNPs have pathogenic consequences[1,2]. Other consortia such as the Encyclopaedia of DNA Elements (ENCODE) project[3,4] and the NIH Roadmap Epigenomics Mapping Consortium[5] have generated atlases of functional elements enabling these SNPs to be linked to downstream transcriptional activity.

It is well-established that disease-causing genetic variation contributes to pathophysiological phenotypic changes in traits and inherited predisposition to many diseases. Rare variants (MAF<0.01) conferring large effects ("penetrance") have been identified in families with extreme phenotypes, for example early-onset forms of cardiovascular diseases (CVDs) using linkage analysis[6,7]. Common variants (SNPs) have been associated with common CVDs in the general population which fit the common disease, common variant hypothesis where many common variants (MAF>0.01) each with a small effect contribute to susceptibility of the trait.

In this model, in contrast to highly penetrant monogenic or rare genetic variants, involvement of a specific individual SNP in causing a disease is neither essential nor adequate[8]. GWAS where millions of SNPs are interrogated for association with a human complex trait, have successfully mapped many variants associated with common diseases[8,9,10]. The GWAS design is based on linkage disequilibrium (LD) between SNPs that are genotyped and causal variants that are ungenotyped. The ungenotyped variants can be imputed with reference panels such as Haplotype Reference Consortium (HRC)[11] or 1000 Genomes Project[12].

SNPs which are associated with disease have the potential to identify molecular mechanisms for a disease, which will help to give a deeper understanding of molecular mechanisms and could help to identify potential drug targets for therapeutic intervention. However, interpretation of GWAS loci is often challenging mainly because of the genomic correlation structure known as linkage disequilibrium (LD)[13], which often causes ambiguity in identifying the true causal variants driving the association. A potential solution is to prioritize the SNPs in the region which can be mapped to the function of genes. However, this approach can also lack resolution if multiple genes are in LD within the genomic region[14].

A major challenge for finding the causal SNP is that many of the disease risk variants identified by GWASs fall inside non-coding regions so the mechanism of action is often unclear. Non-coding regions include introns, intergenic sequence and non-coding RNAs such as microRNAs and long non-coding RNAs[15], all which have functional roles in disease but through different molecular mechanisms[16]. Non-coding variants are significantly enriched within the functional elements such as enhancer and promoter elements, DNase I hypersensitive regions, chromatin marks and transcription factor binding sites[17,16] , leading to the conclusion that the SNP effects are likely to be mediated through regulation of gene expression[18,19]. However, for many SNPs it is unknown which genes they regulate, which cell types they act within, and what molecular pathways are involved. This has not been comprehensively examined due to limited availability of molecular datasets[20,21] to test this on, for example, to integrate GWAS loci with expression data requires expression data from disease-relevant cell types and tissues which often have limited sample size making drawing reliable conclusions challenging. Instinctively, genes situated in nearest proximity to the GWAS associated hit may seem to make the most promising causal genes, but this has been demonstrated often not to be the case[22,23].

1.2 Genetics of atrial fibrillation and stroke

Cardiovascular disease (CVD) affects more than 80 million people in the US. CVDs include a variety of disorders such as diseases of the circulatory vascular system, the heart's electrical system, the myocardium and congenital heart disease (CHD)[24,25]. The highly prevalent CVDs including atrial fibrillation (AF), coronary artery disease (CAD), hypertension, stroke and heart failure (HF), have complex pathologies, indicative of the interaction between genetic and environmental factors. In this PhD thesis, I focus on further understanding the molecular mechanisms explaining two CVDs: AF and stroke.

1.2.1 Common genetics of atrial fibrillation

AF arises as a result of chaotically firing of the electrical impulses from different places in the atria (the hearts top two chambers). AF is the most common cardiac arrythmia (sustained irregular heartbeat) affecting more than 33 million individuals globally[26]. The prevalence of AF was estimated to be from approximately 2.7 to 6.1 million in the US in 2010 and it is calculated to increase to 12.1 million in 2030 in the US[27]. AF is associated with an increased risk of major complications such as mortality[28,27], HF[29], myocardial infarction (MI)[30], CAD[31] and ischemic stroke (IS)[32,33,34,35]. Anticoagulation treatment can reduce the risk of death from stroke[36].

The heritability of AF, particularly early onset AF demonstrates a common, complex disease[37,38]. A study on Danish monozygotic and dizygotic twins estimated the twin heritability of AF to be 62%[39]. Individuals with first-degree relatives that suffered from AF had around 40% increased hazard (after adjusting for AF clinical risk factors)[40]. The contribution of genetic factors in AF pathogenesis has been revealed by the detection of both common and rare variants in individuals with AF[35,41,42,43,44]. A study conducted by Weng et al showed the overall estimation of narrow sense heritability for AF explained by both common and low frequency variants to be 22.1% with small proportion attributable to rare variants[35]. This study estimated 20.4% was explained by common variation (MAF>5%) whereas the rare variation (MAF 1-5%) only explained 1.7% of the AF heritability. Atrial remodelling, changes in atrial activity[45,46,47,48] and variable penetration caused by defects in rare genes[49] are the potential mechanisms suggested to be involved in risk of developing AF in carriers of common genetic variation. Candidate gene-based association analyses and genome-wide association studies (GWAS) have discovered a number of common genetic variants showing predisposition to the risk of AF[50,51]. A multi-ancestry GWAS meta-analysis of AF in cohorts of more than half million individuals detected 97 loci associated with AF[52]. To date, the largest meta-analysis of GWAS studies (n=1,030,836) for AF has discovered 142 independent risk variants associated with AF in 111 genomic regions[41]. The genome-wide heritability of AF explained by all common genetic variation captured in this study was estimated 11.2%. 4.6% of AF heritability was calculated to be explained by the AF-associated variants[41]. The majority of AF-associated

SNPs have been found within intronic or intergenic regions of the genome. For example, the *PITX2* locus on chromosome band 4q25 is the most strongly associated locus and associations at this locus have been replicated across multiple AF GWAS studies[53,41,52]. The risk variants located proximal to the *PITX2* gene increase the risk of AF up to seven-fold[54]. Studies in a mice model showed that insufficiency of *Pitx2*, the nearest proximal gene to the AF variant rs2200733, was linked to an atrial arrhythmogenesis[55]. A study using CRISPR/Cas9 editing technology identified functional evidence for the *PITX2* gene and revealed that long-range interaction of *Pitx2c* (which regulates transcriptional activity) with the gene's promoter decreases susceptibility to AF[56]. AF can also be caused by monogenic mutations, whereby some young patients with inherited diseases such as channelopathies or cardiomyopathies caused by monogenic mutations suffer from AF. Rare form of familial AF has been reported to be caused as a result of mutations in atrial natriuretic peptide[57], nuclear pore and potassium channel genes[42].

1.2.2 Common genetics of stroke

The brain needs to be supplied with oxygen-rich blood in order to be metabolically active. The functions of brain cells are severely influenced by decreased blood flow under 10ml/100g per minute while neurons do not survive long under 5ml/100g per minute level[58]. If the blood supply to the brain disrupts for a few minutes, the hypoglycaemia and hypoxia caused by this interruption results in brain tissue infarction or stroke.

Stroke is the second leading cause of lifelong disability and death among individuals over 60 years worldwide[59,60,61]. Stroke is a clinical condition characterized by a severe focal neurological injury. There are two types of stroke. Ischemic stroke (IS) which mainly happens as a result of infarction in the brain caused by inadequate cerebral blood supply due to arterial blockages[62]. IS (also referred to as brain infarction or cerebral ischemia) accounts for approximately 85% of all cases of stroke[63,64]. The etiological subtypes of IS include cardioembolic stroke (CES), large-artery atherosclerotic stroke (LAS) and small-vessel stroke (SVS)[65]. The other type of stroke is hemorrhagic stroke or intracerebral hemorrhage (ICH) which is caused by rupture of a blood vessel and intracerebral bleeding. Approximately 15% of stroke has been attributed to ICH. SVS also contributes to the cause of ICH.

IS and ICH share similar risk factors. Hypertension in particular is an important ICH risk factor which is also involved in atherosclerotic disease, a leading cause of IS[66]. Susceptibility to IS as a complex disease may be affected by several related cardiovascular traits or risk factors including AF, CAD, systolic blood pressure (SBP), diastolic blood pressure (DBP), and HF as well as metabolomic traits such as levels of high-density lipoprotein (HDL) and low-density lipoprotein (LDL) cholesterol which show evidence of shared genetic loci with IS[67,41]. Moreover, hyperlipidemia is an important IS risk factor which results from intracranial and extracranial vessel atherosclerosis[68]. AF has been identified as a risk factor for development of CES in particular, which is the severe subtype of IS[69]. About 20-30% of patients over age 80 with an IS have already been diagnosed with AF[70,71]. Furthermore, elevated inflammatory biomarkers have been causally linked to higher risk of AF by a multi-directional MR study[72]. This MR study also causally linked the genetic predisposition to AF to risk of CES[72].

Mendelian (single-gene) disorders have been found to be associated with stroke. Such disorders include CADASIL (Cerebral Autosomal Dominant Arteriopathy with Sub-cortical Infarcts and Leukoencephalopathy), an Autosomal Recessive equivalent (CARASIL)[73] and arterial tortuosity syndrome[74]. Missense mutations within the *NOTCH3* gene (located at the 19q12 locus) are associated with CADASIL, a vasculopathy of small vessel[75,76]. In individuals with parental and family history of IS, the risk of stroke increases by 30% in mutations carriers[77,78,79]. Monogenic mutations in genes which are involved in coagulation have also been found to be linked to IS[80,81,82].

GWASs have estimated the heritability of IS subtypes as 32.6% for CES, 40.3% for LAS, and 16.1% for SVS[83,84,85]. Previous GWASs in European ancestry samples have discovered genetic variants in 10 genomic regions associated with CES[86,87], LAS[88,89,90], SVS subtypes[91,92] and ICH[93] including some loci found to be linked to any stroke[91] or more than one subtype of IS[94,90,95,96,67]. For example, a study by the EuroCLOT consortium showed associations of the genetic variant, rs505922 (located in the *ABO* gene which defines blood group) with coagulation protein levels (such as factor VIII and von Willebrand factor) in LAS and CES but not the SVS subtype of IS[96]. Another study comprising METASTROKE cohorts identified associations between the genetic variants in *PITX2* and *ZFHX3* and CES subtype, and *HDAC9* and LAS subtype[97,98]. Common genetic variants within the 9p21 region have been linked to the risk of IS, specifically to the LAS subtype[99,98]. A large-scale multi-ancestry GWAS meta-analysis of stroke in a sample of 521,612 individuals detected 32 loci associated with all stroke and its etiologic subtypes[67]. In the same study on the GWAS conducted in 446,696 European individuals a total of 22 independent risk variants were detected. The risk variants explained up to approximately 2% of the phenotypic variation, which will help us understand the pathophysiology of this highly heterogeneous disease more deeply.

1.2.3 Shared genetic aetiology of atrial fibrillation and stroke

Evidence at the phenotypic level suggests that AF and IS may share a common etiology. Studies revealed that 20-30% of individuals with IS have already been diagnosed with AF[70,71].

Moreover, AF patients can also experience depression[100], reduced quality of life[101,102], cognitive impairment and lesions in the white matter of their brain[103]. CES occurs mainly as a result of AF, is the most severe IS subtype, represents 25% of IS cases and its risk increase by age. Patients who suffer from AF have been reported to have a 3 to 5-fold increased risk of developing IS and approximately 25-30% of all IS cases arise from cardioembolism[104,105]. Some of the AF risk variants have been also found to be associated with IS and CES, likely because of asymptomatic or silent AF[106,107,108].

AF and IS have previously been found to share two loci in common. The *PITX2* and *ZFHX3* loci are located on chromosome band 4q25 and 16q22 respectively and were originally identified as the two top loci for AF[97,87]. *PITX2* encodes a transcriptional regulator which contributes to sinoatrial node development[46] and regulation of genes involved in ion transport and intercalated disc structural remodelling[109]; *ZFHX3* encodes a transcriptional factor involved in atrial arrythmia and remodeling[110]. A GWAS study in an Icelandic cohort revealed that genetic variants conferring high risk for both AF and IS were particularly SNPs in these loci which also showed strong associations with the CES subtype[97]. Moreover, the associations of the *PITX2* and *ZFHX3* AF-associated gene variants with the CES subtype have been identified and replicated in GWASs[97,107,87,67].

Pulit et al[111] used AF Genetics (AFGen) consortium AF GWAS data[112] along with NINDS-Stroke Genetics Network (SiGN) consortium (AF in cardioembolic stroke) cohort data[90] and found that SNP associations between these two cohorts are highly correlated even when SNPs within *PITX2* and *ZFHX3* loci were removed. This study also demonstrated that 23.1% of the heritability in risk of CES is explained by genetic risk factors for AF. Polygenic risk score

(PRS) for AF was shown to be associated with cardioembolic IS after controlling for AF clinical risk factors[111]. Indeed, the genetic susceptibilities to the risk of both AF and CES are likely to contribute as part of a complex genetic predisposition[8].

Furthermore, genetic loci associated with subtypes of IS have also been implicated in its related vascular risk factors including AF, hypertension, CAD, formation of carotid plaque, venous thromboembolism (VTE) and lipid levels. Risk loci associated with AF (16q22 and 4q25) [97,87], blood pressure (6p21, 12q24, 7p21), CAD (9q34, 12q24, 7p21, 9p21, 19p13 and 4q31), LDL cholesterol levels (9q34 and 19p13), HDL cholesterol levels (12q24), formation of carotid plaque[113] (4q31 and 11q22) and neuro-inflammation (1p13) have been found to have genetic overlap with loci for stroke in a genetic risk score (GRS) association analysis conducted by Malik et al[67]. Two loci for white-matter hyperintensities (WMH)[114] (1q22 and *APOE*) have also been linked to ICH[115,116]. The genetic influences at the 9q34 (*ABO*) and 12q24 (*SH2B3*) loci have been found to be shared between IS subtypes[67]. It is possible that these loci might therefore be acting through a common biological mechanism, for example, by influencing atheroma, pathways of coagulation, and arterial thrombosis, by altering or increasing the risk of stroke risk factors such as AF (leading to blood clot) and hypertension[85].

1.3 The molecular aetiology of complex traits

Common diseases have a complex molecular aetiology involving multiple molecular phenotypes across a number of tissues in the human body. In this thesis I consider the interaction between the methylome, transcriptome and metabolome to better understand the molecular pathways underlying AF and stroke risk.

1.3.1 Epigenome: DNA methylation

Methylation of DNA is a process in which a methyl group added to the CpG dinucleotide sequence typically leads to a silencing of genetic information (gene transcription), which can occur by inhibition of transcription factors (TF) binding to gene promoters or recruitment of DNA binding proteins that affect chromatin remodeling[117,118]. DNA methylation is therefore an epigenetic mark that plays a role in controlling many cellular processes including the regulation of gene expression, differentiation and maintaining genomic stability[119]. Hence, DNA methylation in response to perturbations in the environment can change disease risk[120], and this in combination with broad role across biological systems makes DNA methylation important to consider when investigating the aetiology of common diseases. Genetic effects on cytosine methylation are typically estimated by comparing CpG methylation levels in individuals with different genotypes at a suspected regulatory locus, an approach termed methylation quantitative trait locus (mQTL) mapping. A number of mQTL studies have explored the correlation between genotype and DNA methylation levels mainly in blood[121,122,123,124]. A few studies have shown mapping of *cis* (local) or *trans* (distal) mQTLs across multiple tissues[125,126]. A number of mQTLs have also been demonstrated to be linked with changes in gene expression levels (in other words, they are also classified as expression QTLs or eQTLs)[121,127,128].

The effects of DNA methylation are assumed to be mediated by their impact on gene expression. Some studies have shown variation in the methylation of CpG sites located at the promoter regions is associated with higher expression of several genes[129]. More recent published studies have revealed a more complex situation, with both negative and positive associations between gene expression and DNA methylation traits in relation to cardiovascular diseases such as CAD and IS[130,131,121,127,132,125]. This is compatible with a probable shared genomic regulation of DNA methylation and gene expression, and indeed several studies have evaluated the overlap in genomic regulation of expression and methylation in humans[121,133,134].

Epigenome-wide association studies (EWAS) have identified the association of DNA methylation variation at thousands of CpG sites with complex traits[135]. Particularly relevant to the work in this thesis are a whole blood EWAS study of AF in the FHS cohort[136] and an EWAS of IS[137] which has identified differentially methylated CpG sites associated with IS within IS genetic risk loci (including 16q22 locus). The 16q22 locus is involved in angiogenesis, inflammation, glycolysis pathways and lipid metabolism[137]. A number of CVDs have been reported to be correlated with defects in the pattern of DNA methylation from peripheral blood[117,138,139,140,141,142]. For example, one study has revealed that the DNA methylation patterns of patients with CAD can distinguish between those with and without HF[143]. Moreover, variability of methylation CpG sites or differential methylation (i.e., hyper and hypo methylated CpG sites) has been observed between different cardiovascular tissues of CAD patients[144].

1.3.2 Transcriptome: gene expression

The transcription of genomic sequence into ribonucleic acid ("gene expression") is a fundamental part of all human systems and biological processes and is extremely important to the development and progression of disease. A large number of SNPs have been identified to be strongly associated with expression levels of protein-coding ribonucleic acid (RNA) (mature messenger RNA (mRNA)) and non-coding RNAs (e.g., long non-coding RNAs; lncRNAs). These expression-associated SNPs are known as expression quantitative trait loci (eQTLs), and have been identified in a number of different cell types and tissues[145,146,147,148,149]. It has been reported that mRNA abundance of almost all genes is associated with one or more genetic variants[146,149,150]. Transcription levels can be modified by either *cis* (SNPs proximal to the gene) or *trans* (SNPs distal to the gene, often on other chromosomes) eQTLs. A typical definition used in many studies identifies *cis* eQTLs as those located on the same chromosome and within 1Mb physical distance from either side of a gene transcript start site (TSS). By the same definition, *trans* eQTLs are located more than 1Mb distant from the regulated gene or on a different chromosome[151]. Studies have found eQTLs to be highly enriched among complex disease risk loci[149,146,152], suggesting that disease GWAS variants often drive their associations by influencing gene regulation.

Previous studies to explore the mechanisms underpinning eQTL associations uncovered that a large proportion of eQTLs influence the binding of transcription factors (TF) to control gene expression[153,154,155,156]. It has been reported that gene expression levels may be changed through alterations in chromatin function at promoters or enhancers[157,158,159,17] and some studies have identified genetic variants that influence the function of chromatin[3,4,160] by analyzing inter-individual variation in DNase I hypersensitivity, an indicator of chromatin accessibility[17]. Furthermore, disease-associated SNPs identified by GWASs are highly enriched and mapped within regions of active chromatin in relevant cell types[16,161].

1.3.3 Metabolome: metabolite levels

Circulating metabolites (including blood lipids) are another type of intermediate phenotype which is located more distal to the genome in pathophysiological pathways than traits such as DNA methylation and gene expression. Metabolites are the products of metabolic processes or metabolism[162,163] and may contribute to regulatory processes of gene expression and protein abundance and cellular activity[164,163,165]. Particular classes of metabolites can be quantified and identified using either high-throughput nuclear magnetic resonance (NMR) spectroscopy platform (a targeted method)[166] or mass spectrometry (MS)[167]. Although MS technology provides wide metabolite coverage, usage of NMR platform is more appropriate for large-scale studies as it is cheaper and can also quantify lipoproteins which MS cannot. In parallel with circulating metabolites (including plasma cholesterol, plasma triglycerides (TG), low-density lipoprotein (LDL), high-density lipoprotein cholesterol (HDL)), which have been recognized for their correlation with CVDs such as atherosclerosis, CAD and MI[168,169,170], inflammation is another known factor involved in pathogenesis. For example, mechanisms of lipid-induced inflammation response in atherosclerosis have also been connected to MI progression[171,172].

Genetic variants that influence circulating blood metabolite levels (also called metabolite quantitative trait loci or mbQTLs) have been identified in several studies[173,174,175] (including Avon Longitudinal Study of Parents and Children (ALSPAC) and The University College-London School-Edinburgh-Bristol (UCLEB) mbQTLs – unpublished data). Recently, genetic control of more than 249 metabolic biomarkers has been released and studied in blood samples of 130k from 500k participants in UK Biobank using Nightingale's biomarker profiling technology[176,177]. Several studies have linked elevated levels of lipoprotein(a) to increased risk of atherosclerosis and stroke[174,178,179,180,181].

1.3.4 Integrating QTLs to understand mechanisms of disease

1.3.4.1 DNA methylation and gene expression

eQTLs are frequently mapped to regulatory elements such as chromatin accessibility, DNase I hypersensitivity and histone modifications indicating coordinated epigenetic effects are likely to have a role in regulating gene expression[182,159,21,183,184]. For example, eQTLs are likely to act through the disruption of TF binding sites to control gene expression[182,159,183,184]. Information on genetic variation associated with levels of DNA methylation and gene expression have been incorporated with disease and complex traits to find overlaps between mQTLs, eQTLs and trait-associated variants[23,185,186,187,22,188]. For example, Richardson et al integrated epigenetic data with cardiovascular traits, and found a putative causal role of DNA methylation at CpG sites (enriched in gene promoter regions and histone marks) on cardiovascular disease risk[188]. Follow-up analysis combining these loci with genetic information on gene expression data also revealed evidence of an influence of these variants on gene expression levels.

1.3.4.2 Gene expression and metabolite

Integrating genetic information on genome, transcriptome and metabolome has mapped the interaction of multiple genes whose expression is involved in immune activity with serum circulating metabolites, such as amino acids, fatty acids, lipids and lipoprotein subclasses, providing evidence for a likely contribution of products of these genes in metabolic inflammation[189]. Studies have also linked levels of acylcarnitine and amino acids to metabolic diseases in adults[190,191]. Integrating complex trait-associated SNPs with expression QTL data in blood mononuclear cells along with genetic data on metabolite levels revealed evidence for a putative causal role for genes which may regulate metabolism[187]. Moreover, some studies have revealed that a large fraction of genetic variants associated with circulating metabolite concentrations (mbQTLs) were also *cis*-eQTLs[192,193].

1.3.4.3 DNA methylation and metabolite

There is considerable evidence that supports the role of epigenetic mechanisms of DNA methylation in the regulation of metabolic traits and diseases[194,195,196,197,198,199,200,201,202,203,204,205]. While, these studies have underlined the replication and overlap of CpG sites identified in cardiometabolic traits, the complete causal mechanisms between DNA methylation and circulating metabolites is unclear and needs to be elucidated. For example, some studies have shown that most of the putative causal variation in DNA methylation may be a consequence of alteration in circulating metabolites levels rather than a cause of it[203,206,207,208,209,210]. Studies which have assessed the correlation between DNA methylation at CpG sites and circulating metabolites, have detected several loci associated with complex diseases[208,211]. Recently a large EWAS of DNA methylation from leukocyte and 226 serum metabolites (mainly lipid-related metabolites) has identified 161 CpG-metabolite associations (including fatty acids and lipoproteins). This study also found a link between five metabolite-associated CpG sites and alteration of gene expression levels in adipose tissue and blood and showed that methylation of metabolite-associated CpG sites correlated with expression of genes, obesity and MI[212].

1.4 Statistical methods to investigate shared genetic effects between traits

Loci identified by GWAS are difficult to interpret due to LD and the fact that most complex trait-associated SNPs are located in non-coding regions of the genome. Performing lookups of GWAS SNPs in eQTL databases to find overlaps (where the GWAS SNP also associates with gene expression) is unreliable due to confounding by LD. It also appears that causal genes are not necessarily the closest genes[22,23]. Moreover, a LD “block” (region of high LD) in the genome might contain multiple genes and the locus of interest might have a cell type-specific or tissue-specific effect requiring cell type specific data (which might not have sufficient sample size). These challenges have led to the development of advanced statistical methods to help us better understand regulatory mechanisms, and potentially uncover new causal genes.

Several approaches have been developed to integrate eQTL and disease GWAS data. Description of the various integrating and colocalization approaches is shown in **Table 1**. For example, transcriptome-wide association studies (TWAS)[213,214,215] use genetically predicted gene expression to identify differentially expressed genes associated with the trait. PrediXcan[214] is a method for integration of eQTL and GWAS studies in order to prioritize candidate causal genes. This method detects the statistical associations between imputed expression of a gene (using eQTL) and the trait. GWAS individual-level data is used in the application of PrediXcan and GWAS summary-level data is employed in the application of S-PrediXcan[215]. TWAS approach is not used in my thesis, rather I am concerned with identification of colocalization between intermediate molecular QTL and GWAS variants at a genomic locus. Genetic colocalization methods (also detailed in **Table 1**) have been developed to test whether association signals for different traits map to the same causal variant and thereby can be postulated to have a functional connection. In the following sections, I describe in detail different approaches to colocalization and the "moloc" multi-trait colocalization method[216] I used in this thesis to determine shared regulatory effects at AF and stroke GWAS loci. I also describe the use of MR[217,218,219] to infer potential causal relationships between traits.

1.4.1 Method for colocalization of a pair of traits

Coloc is a Bayesian colocalization approach to elucidate if the observed overlap between a pair of traits is due to a shared causal regulatory effect[220,221]. This approach has three key assumptions: (i) that the causal association signal is included in the set of common SNPs between both datasets, either well imputed or directly genotyped. In the absence of the causal variant, the power of detecting a common SNP will be decreased depending on the LD between the causal variant and other variants included in the model. (ii) that there is at most one causal genetic variant for each trait in the genomic region. If multiple causal SNPs are present for each trait at the given locus, coloc is unable to detect colocalization of the primary association signal with other traits independent of secondary and additional signals. (iii) that the LD patterns and the allele frequencies are identical across independent studies[216]. Of note, the coloc approach is not sufficient to distinguish cases where apparent pleiotropic effects are causally related ('vertical pleiotropy') from those where the same causal association signal is influencing both molecular and complex trait via independent biological pathways ('horizontal pleiotropy') (**Figure 1**).

Coloc[220] computes posterior probabilities for each of 5 hypotheses: There is no association signal in the given region for either the GWAS phenotype or gene expression (H_0), there is only one association signal for gene expression in the given region (H_1), there is only one association signal for GWAS phenotype in the given region (H_2), there are two distinct causal association signals in the given region for gene expression and GWAS phenotype (H_3), There is one single (shared) causal association signal for both gene expression and GWAS phenotype in the given region (H_4). Hypothesis 4 gives a posterior probability (PP_4) of the presence of a single shared causal variant in a region which helps to prioritize the associated gene as potentially involved in the GWAS phenotype.

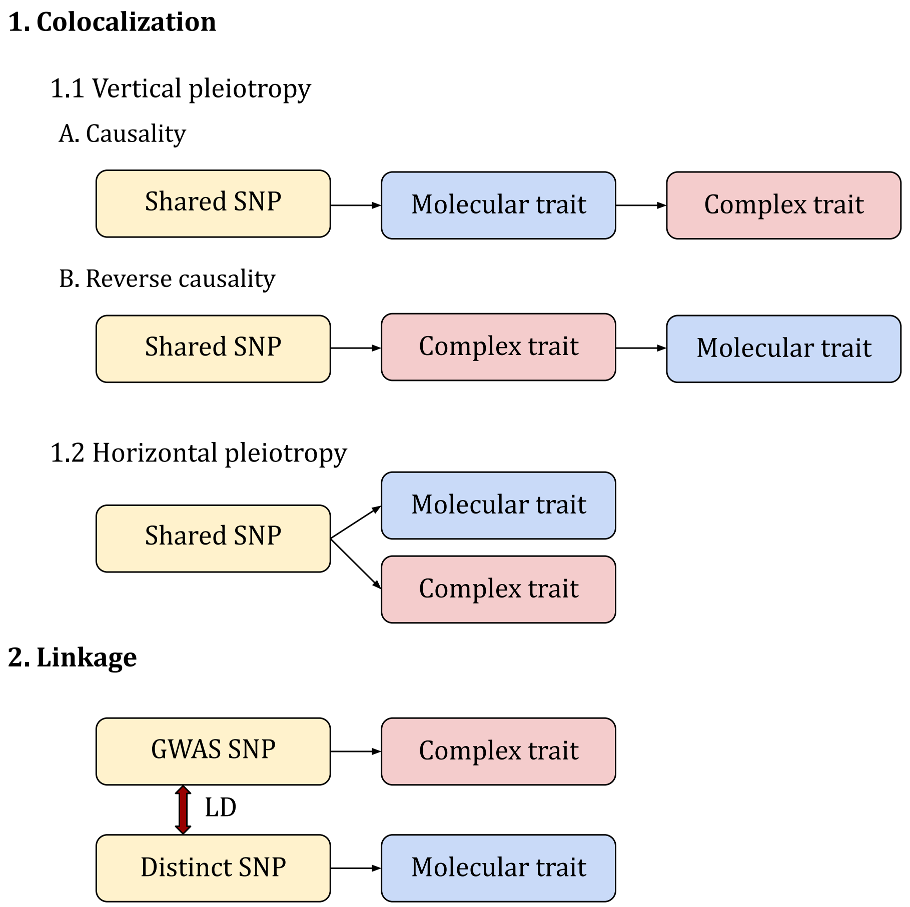

Figure 1. Colocalization explanations with regards to pleiotropy (i.e., either horizontal or vertical).

Explanation 1.1– Vertical pleiotropy (A) – Causality: A shared variant influences complex trait risk through changes in levels of the colocalized molecular phenotype at a locus; (B) – Reverse causality: A shared variant influences complex trait risk via biological pathways other than through the colocalized molecular phenotype (i.e., downstream effects of disease on intermediate phenotype levels). Explanation 1.2 – Horizontal pleiotropy: A shared variant affects both molecular phenotype levels and complex trait risk by two independent molecular pathways. Explanation 2 – A GWAS variant that affects the complex trait is in LD with another distinct variant which regulates changes in levels of the molecular trait (i.e., confounded by LD). Colocalization can distinguish explanation 2 from explanation 1.1 or 1.2 but, can't distinguish explanation 1.1 from 1.2.

1.4.2 Colocalization for multiple traits

Coloc can be applied to test the colocalization of just two traits, however, in 2018 a technique called Multiple-trait-colocalization (Moloc)[216] was reported which can investigate shared regulatory effects across multiple traits at GWAS risk loci. Moloc extends the coloc framework (posterior calculation) for integration of summary statistics of multiple traits to comprehensively estimate joint probabilities that a single association variant is causal across all traits tested. Moloc has more statistical power and better resolution for multiple traits, as it can evaluate all combinations jointly rather than carrying out a series of pairwise colocalization analyses. Moloc uses a Bayesian statistical framework to estimate the posterior probability of association (PPA) that multiple intermediate molecular traits and a complex trait share a single causal genetic variant. The PPA for each of 15 hypotheses (H0...H14), combining three traits is computed as a ratio:

$$\frac{P(H_h|D)}{P(H_0|D)} = \sum_{S \in S_h} \frac{P(D|S)}{P(D|S_0)} \times \frac{P(S)}{P(S_0)}$$

$$\frac{P(H_h|D)}{P(H_0|D)} = \prod_{s \in h} \pi_s \sum_{i=1}^{Q} BF_{i,s}$$

where $P(H_h|D)$ is the probability of the data D for each hypothesis h, $P(H_0|D)$ is the likelihood of baseline hypothesis of no association with any trait H_0, $\mathrm{BF}_{i,s}$ are the Bayes factor of a SNP among the traits of interest indexed in s and the prior probabilities that SNP i is the causal variant under a particular model are p.

1.4.3 Other colocalization methods

Other eQTL-GWAS colocalization approaches (**Table 1)** such as regulatory trait concordance (RTC)[222], Sherlock[223], QTLMatch[224], Coloc[220], eCAVIAR[225], enloc[226] can be applied to detect genes whose expression is under influence of the same causal genetic variants underlying GWAS association. Both RTC and Sherlock (a Bayesian statistical framework) adjust for the LD structure to detect the common causal signal in both studies. QTLMatch[224] is a method where the colocalized variant arises from co-occurrence or causal relationship between GWAS and eQTL variants. eCAVIAR[225] (eQTL and GWAS Causal Variant Identification in Associated Regions) models LD correlation between SNPs in the region to account for multiple causal variants and compute the colocalization posterior probability (CLPP) for the variant causal for both traits based on GWAS summary statistics. PWCoCo (developed locally at University of Bristol) carries out COJO (conditional and joint) analysis[227,228] to condition out the independent signals within a genomic region, and then conducts Bayesian pairwise coloc on any conditioned association peaks that remain for both traits (unpublished tool). SuSiE (Sum of Single Effects)[229] combines the resolution of fine-mapping with coloc approach. It first fine maps to identify credible sets belonging to each independent SNP in the region and then conducts colocalization on the credible intervals. For a single pair of traits, all of these methods exist, however, for exploring colocalization of multiple traits fewer methods are available. At the time of starting this PhD project the only tool available was moloc, but another method became available later on HyPrColoc[230] (Hypothesis Prioritization in multi-trait Colocalization). HyPrColoc is a Bayesian technique using GWAS summary statistics along with summary information of multiple molecular traits to simultaneously identify colocalization across many traits (i.e., jointly analysis of 3-100 traits) (**Table 1**).

Table 1. Summary of existing integrating and colocalization approaches.

Method name (publication date)	**input data**	**description**	**category**
PrediXcan (2015)	Individual-level data	TWAS approach which generates genetically predicted transcript expression levels imputed from eQTL data and tests this for association with GWAS trait.	TWAS
S-PrediXcan	Summary statistics and LD estimates of a reference population	PrediXcan algorithm adapted for use with GWAS summary levels statistics.	TWAS
QTLMatch (2009)	Summary statistics	Bayesian statistical method, performed on specific genes for GWAS top hits	Colocalization (single variant assumption)
Coloc (2012, 2014)	Summary statistics	Bayesian statistical framework	Colocalization (single variant assumption)
Moloc (2018)	Summary statistics	Extension of coloc method to estimate joint colocalization probabilities when using multiple traits. **Method used in my PhD.**	Colocalization (single variant assumption, multiple traits less than 5 at one time)
HyPrColoc (2021)	Summary statistics	Multiple trait colocalization methods that selects traits (hypothesis prioritization) based on a clustering algorithm to improve computational efficiency	Colocalization (single variant assumption, many traits)
RTC (2010)	Summary statistics from GWAS, individual level data of gene expression and LD estimates of a reference population	Conditional analysis technique which removes the effect of a GWAS SNP from eQTL to see if an effect still remains in a region	Colocalization (multiple independent SNPs)
Sherlock (2013)	Summary statistics and LD estimates of a reference population	Bayesian statistical method which adjusts for LD between SNPs in the region	Colocalization (multiple independent SNPs)
eCAVIAR (2016)	Summary statistics and LD estimates of a reference population	Bayesian statistical framework that models LD to account for multiple independent variants in a region	Colocalization (multiple independent SNPs)

Method name (publication date)	**input data**	**description**	**category**
enloc (2017)	Summary statistics and LD estimates of a reference population	Bayesian statistical method which incorporates fine-mapping and regulatory elements information	Colocalization (multiple independent SNPs)
SuSiE (2021)	Summary statistics and LD estimates of a reference population	Extension of the coloc method that fine maps region beforehand to provide colocalization probabilities per credible set.	Colocalization (multiple independent SNPs)
PWCoCo (unpublished)	Summary statistics	pairwise conditional analysis and colocalization analysis	Colocalization (multiple independent SNPs)

1.5 Mendelian randomization for causal inference

Whilst colocalization methods test for shared genetic variants between a molecular phenotype and disease, they do not test directionality (causality). Mendelian randomization is an approach which aims to address this challenge.

1.5.1 Mendelian Randomization

Mendelian randomization (MR) is a type of instrumental variable (IV) analysis in which genetic variants are used as instruments to make causal inferences in epidemiological research[217,218,219]. These instruments robustly associated with the exposure are employed to proxy the exposure of interest, owing to the fact that genetic variants are far less inclined to confounding and reverse causation. A confounder is a common cause of both exposure (X) and outcome (Y), which will result in a biased association if not adjusted for. Suppose we have unmeasured confounder (U) of exposure (X) and outcome (Y) relationship. MR attempts to find evidence of causality by removing this confounding through instrumenting on genetic variants (G). A schematic causal diagram is displayed in **Figure 2**.

The key set of IV assumptions for MR[231]

IV1 – Instrument (Gj) must be robustly associated with the exposure (X).

IV2 – Instrument (Gj) must not be associated with confounders (U) of the exposure-outcome relationship.

IV3 – Instrument (Gj) must only be associated with trait outcome (Y) through exposure (X) (i.e., horizontal pleiotropy is not allowed).

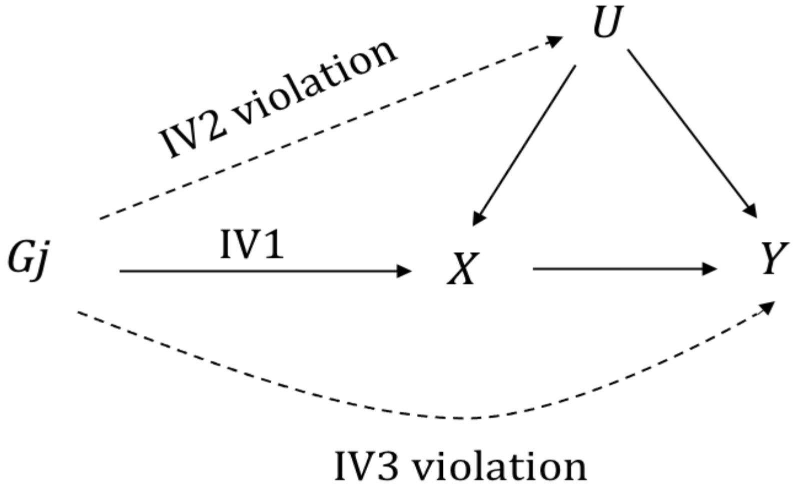

Figure 2. Schematic diagram displaying the causal relationship between instrumental variable (*Gj*), exposure (*X*) and trait outcome (*Y*).

Solid lines indicate *Gj* is a valid IV. Dashed lines show IV assumption violations.

1.5.2 Single SNP MR (Wald ratio) to link QTLs to disease SNPs

Acquiring a large sample with individual participant summary data is often impractical. Two-sample MR (2SMR) using summary statistics of molecular quantitative trait loci (molQTLs) and SNP-outcome associations from independent and large populations not only can overcome this issue but also increases the statistical power of the analysis[232,233,219]. Moreover, obtaining data with disease outcomes and all the molecular traits would be challenging, but the 2SMR framework allows us to integrate data from different sources[219]. The molQTLs are normally identified in separate non-overlapping studies to the outcome trait of interest, and are independent of each other (i.e., not in LD). Two-sample MR incorporates these SNPs as instrumental variables (IVs) to estimate the causal effect of the exposure on the trait outcome using two distinct study samples[233]. In the 2SMR model, molQTL is valid as an instrument if it is only associated with the outcome via its effect on the molecular trait or risk factor[231]. GWAS summary data for different traits are often publicly accessible and can be used for application of MR[10,234].

In molQTL analysis, the single SNP weight (Wald Ratio) from MR can be utilized to discover the causal role of intermediate phenotypes (such as gene expression and protein levels) on complex diseases by instrumenting on molQTLs with robust associations[232]. For example, for a particular gene all eQTLs (β_x) can be found and then SNPs can be looked up in outcome data (β_y) and the Wald ratio (β_{MR}) can be computed, which estimates the genetically predicted increase in disease risk per unit change in gene expression. Another method which conducts MR using gene expression as an exposure and complex trait as an outcome is SMR/HEIDI[22,235,236]. SMR finds shared genetic effects between expression and GWAS traits using a two-sample MR approach.

1.5.3 Combining MR and colocalization to find causal molecular phenotypes

Wald ratio analysis is augmented with colocalization to ensure that the MR is not due to the QTL and GWAS trait sharing coincidently due to LD patterns in the region. Using this approach, MR studies have led to the discovery of causal links between the methylome[188], transcriptome[237,238,239,22] and proteome[240] and human complex traits. Zhu et al conducted SMR followed by HEIDI to determine colocalization of signals, using summary statistics from five complex traits and blood gene expression and identified 126 candidate genes whose expression levels are likely to causally influence the traits. This included 25 newly discovered genes which were not previously reported in the literature and 77 genes (61%) which were not located/mapped to the closest physical proximity to the GWAS lead SNP in the region[22]. Phenome-wide MR analyses conducted by Richardson et al identified potential causal effects for 1,148 *cis*-mQTLs on 139 human complex traits. Moreover, integration of eQTL data showed that 306 of the mQTLs influenced both methylation levels at CpG sites and variation in 47 traits also affect gene expression[188].

1.5.4 Multi-SNP methods to evaluate causality between intermediate traits and disease

For QTL based MR analysis, the Wald ratio method is used by necessity due to the sparsity of instruments available, often only allowing MR to be conducted on a single SNP. However, MR is traditionally performed on multiple SNPs, allowing for more robust inference and improved evaluation of modelling assumptions (i.e., examination of bias due to pleiotropic factors). For many intermediate traits, such as circulating biomarkers, multi-SNP MR will be conducted. In a multi-SNP MR setting, Wald ratios are calculated between the instrument-exposure and instrument-outcome SNP effects, and then pooled together. Different methods for combining the Wald ratios have been developed each with different modelling assumptions regarding the influence of pleiotropy on the final MR estimate (see Methods (**2.2.2**) for more detail). It is common practice for a MR study to compare results across multiple MR methods, and then nominate findings that show consistency as these are likely to be robust against pleiotropic bias[234].

1.6 Overarching aims of this thesis

The main aims of this doctoral thesis are to employ multi-trait colocalization approach to gain insights into the molecular mechanisms of AF and stroke GWAS loci by uncovering the shared regulatory effects of intermediate molecular phenotypes and prioritizing candidate causal features (such as CpG sites, genes, and metabolites) for gene regulation. AF and stroke were selected to be studied for several reasons: first, these are well-defined common CVDs which create a significant health burden on society and have extensive published evidence of a relationship between them. Second, recent large GWAS were available for both diseases. Third, large-scale molecular QTL resources typically only exist for blood, and therefore probably best reflect the cardiovascular system which is likely to be a disease relevant tissue for both AF and stroke. In terms of molecular traits, I first conducted analysis on DNA methylome and the metabolome as I gained early access to these QTL resources: the Genetics of DNA Methylation Consortium mQTL (GoDMC)[124] and the UCL-LSHTM-Edinburgh-Bristol (UCLEB) Consortium and Avon Longitudinal Study of Parents and Children (ALSPAC) mbQTL summary statistics, which I call Phase I of the study. Later on in Phase II of my study I then integrated the top findings from Phase I with eQTL data from eQTLGen[241], when the public release of this dataset became available.

I applied the multi-trait colocalization pipeline to all recently identified AF and stroke loci from the most recent GWAS meta-analyses to explore regions which show evidence of colocalization with different combinations of molecular phenotypes and therefore identify the genes that may underpin AF and stroke associations. The aims of each Chapter are as follows: In **Chapter 3**, moloc is performed to explore whether the overlap between molecular traits and AF are explained by a single shared regulatory effect at each locus and identify the potential causal genes responsible for these shared effects at AF GWAS loci. In **Chapter 4**, moloc is conducted to explore whether the overlap between molecular traits and stroke are explained by a single shared regulatory effect at each locus and identify the genes involved at stroke GWAS loci. In **Chapter 5**, I conduct pairwise colocalization and pQTL analyses to relate AF and stroke GWAS variants to functional effects through *HP* expression at the 16q22 locus (this locus was colocalized with both AF and stroke in previous chapters). In this chapter, I hypothesized that AF might be an intermediate step before stroke and this pathway may be partly mediated by the *HP* gene, leading to the work in Chapter 6. In **Chapter 6**, to further investigate shared aetiology and determine whether a causal relationship exists between AF and stroke MR analyses are performed. Thesis workflow is shown in **Figure 3.**

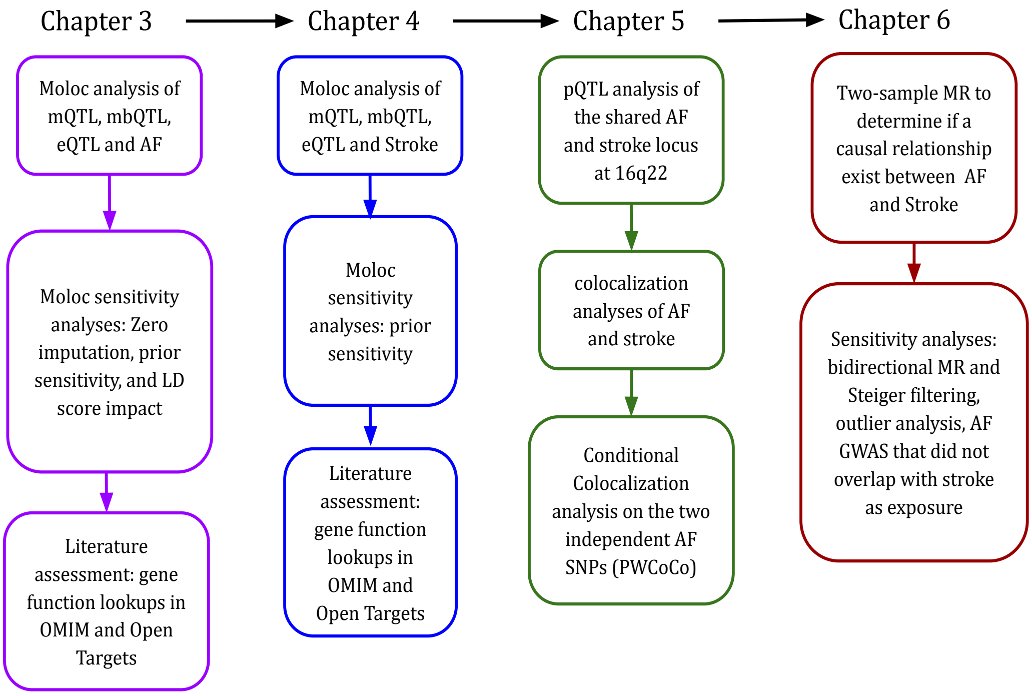
Chapter 3
Chapter 4
Chapter 5
Chapter 6
Moloc analysis of mQTL, mbQTL, eQTL and AF
Moloc sensitivity analyses: Zero imputation, prior sensitivity, and LD score impact
Literature assessment: gene function lookups in OMIM and Open Targets
Moloc analysis of mQTL, mbQTL, eQTL and Stroke
Moloc sensitivity analyses: prior sensitivity
Literature assessment: gene function lookups in OMIM and Open Targets
pQTL analysis of the shared AF and stroke locus at 16q22
colocalization analyses of AF and stroke
Conditional Colocalization analysis on the two independent AF SNPs (PWCoCo)
Two-sample MR to determine if a causal relationship exist between AF and Stroke
Sensitivity analyses: bidirectional MR and Steiger filtering, outlier analysis, AF GWAS that did not overlap with stroke as exposure

Figure 3. Flowchart of thesis analyses.

Chapter 2 Methods

2.1 Data sources

The colocalization and MR approaches used in this thesis are based on the analysis of summary level genetic association information generated in other studies (i.e., regression coefficients, standard errors and P values from GWA studies or molQTL analysis). The summary data used in this thesis are described in this section.

2.1.1 Atrial fibrillation GWAS summary statistics

I used summary statistics from the most recent and large-scale GWAS meta-analysis of AF[41] in European participants which were publicly accessible at http://csg.sph.umich.edu/willer/public/afib2018/ (date accessed: 3 August 2018). Nielsen et al performed the AF GWAS meta-analysis on over 1,000,000 individuals of European ancestry including 60,620 cases and 970,216 controls using Haplotype Reference Consortium (HRC) imputed genetic data.

2.1.2 Stroke GWAS summary statistics

I used the largest publicly available stroke GWAS meta-analysis dataset in Europeans for any type of stroke[67]. The full genome-wide summary association statistics were downloaded from the GWAS Catalog FTP website (https://www.ebi.ac.uk/gwas/studies/GCST006906) on the 5th May 2020. This stroke meta-analysis included 446,696 individuals of European ancestry comprising 40,585 cases and 406,111 controls (number of SNPs=8,211,693). The summary statistics were generated by the MEGASTROKE consortium using samples from 15 cohort studies and 2 consortia (the METASTROKE and CHARGE consortia) with genotypes imputed to 1000 Genomes Project (1000G) phase I haplotype panel[242]. This analysis used five stroke phenotypes; any stroke (stroke, n = 39,067 cases), any ischemic stroke (IS, n = 32,686 cases), IS subtypes: large artery stroke (LAS, n = 4,113 cases), cardioembolic stroke (CES, n = 6,820 cases), and small vessel stroke (SVS, n = 4,975). The genetic information at 22 stroke risk loci (11 loci for stroke, 8 loci for IS, 2 loci for CES subtype and 1 locus for LAS subtype) identified by this published GWAS were employed for the application of moloc in **Chapter 4**.

2.1.3 The Genetics of DNA Methylation Consortium (GoDMC) mQTL

I used GoDMC mQTL meta-analysis data[124] (http://mqtldb.godmc.org.uk/) which was generated by Min et al using a two-phase design across 36 datasets of European origin. Genetic and DNA methylation data were pre-processed in a standardized way and analyzed by each cohort analyst using the same pipeline (https://github.com/MRCIEU/godmc). The mQTL analysis was performed using 1000G imputed genetic data and DNA methylation profiles derived from blood samples of European ancestry and measured on Illumina Infinium HumanMethylation450k arrays. Summary statistics were available for *cis* and *trans* mQTL associations. An overview of the meta-analyses in GoDMC along with descriptive summary text can be found in "Supplementary Note" GoDMC publication[124].

2.1.4 Avon Longitudinal Study of Parents and Children (ALSPAC) mbQTL

I used ALSPAC mbQTL summary data for 230 metabolites which were provided on request from the lead analyst (Fotios Drenos, personal correspondence). The ALSPAC (http://www.bristol.ac.uk/alspac/) samples from children were genotyped by a customized Illumina 550 array. After quality control (QC) check, these data (MAF>0.01) were imputed into the combined UK10K/1000 Genomes reference panel of 28 million SNPs using IMPUTE2. Metabolic phenotypes were measured in serum samples of the ALSPAC cohort by

Nuclear Magnetic Resonance (NMR) spectroscopy (as previously described[243]) and rank inverse normal transformed. The metabolite QTL analysis was performed on all metabolic measures of plasma metabolites from the oldest time point available in ALSPAC children cohort[244] using SNPtest 2.5[245]. In regression model the metabolic data were adjusted for age, sex and the first 10 principal components (PCs) of the genetic data. The residuals which were rank inverse transformed to normality were then used in regression model to test for association with SNPs using an additive model which took uncertainty of imputation into account. The mbQTL data (n=5589 participants) provided to me for this project were restricted to top association signals (i.e., significant results) with P value less than 1x10^{-4}.

2.1.5 The University College-London School-Edinburgh-Bristol (UCLEB) mbQTL

I used mbQTL summary statistics for 288 metabolites generated by UCLEB consortium. The UCLEB samples were genotyped using the Illumina Cardio-Metabochip array. After quality control (QC) check, these data (MAF>0.001) were imputed into the 1000 Genomes Project phase I as previously described[246]. Metabolite profiles for UCLEB individuals were generated using serum samples taken at recruitment with age ranges from birth to 62 years. Levels of 288 metabolites were measured using a high-throughput serum nuclear magnetic resonance (NMR) platform following methods previously described elsewhere[167]. Metabolites were rank reverse transformed to normality. The metabolite QTL meta-analysis was conducted applying the fixed-effect inverse variance method implemented in the METAL software package[247]. The mbQTL meta-analysis used samples of European ancestry (UK-based) consisting of approximately 30,000 participants from 7 contributing studies. The analysis Early (pre-publication) access to these full mbQTL data were kindly provided by the UCLEB consortium[246] co-authors/analysts from UCL Institute of Cardiovascular Science, Population

Science and Experimental Medicine, Centre for Translational Genomics.

2.1.6 eQTLGen eQTL

The most recent eQTL summary statistics with the largest sample size were downloaded from eQTLGen (https://www.eqtlgen.org). The eQTLGen dataset includes association data for 19,942 genes conducted in a total of 31,684 samples (mostly Europeans) in blood meta-analyzed across 37 eQTL datasets[146]. The eQTLGen data contained full *cis*-eQTL association information on pseudogenes, non-coding and protein coding RNAs without a P value threshold cut-off.

2.2 Methods

This thesis used multi-omics colocalization technique to integrate statistical evidence for the potential involvement of different intermediate phenotypes in AF and stroke. The principal techniques used in this thesis are described in this section. The methods section of each analysis chapter provides additional details.

2.2.1 Moloc study design (overview of Phase I and II)

Moloc is a Bayesian statistical technique that integrates summary level information from GWAS studies with multiple molecular QTL data to identify a shared regulatory effect between molecular phenotypes and a complex trait[216]. The main reasons for selecting moloc was that it can be applied to up to 4 traits and because it works with summary level data. In these analyses, disease GWAS, DNA methylation, gene expression and circulating metabolite traits were designated as G, M, E, and Mb respectively for simplicity. Examples of four possible colocalization scenarios are illustrated in **Figure 4**. I was able to use publicly accessible data with the largest available sample size for both the whole blood eQTL and GWAS data.

Moloc analyses were conducted in two Phases (**Figure 5**). The main reason for performing Phase I analyses (integration of epigenetics and metabolomics data with GWAS) was early access to large QTL summary statistics datasets (i.e., GoDMC mQTLs and ALSPAC mbQTLs). eQTL were not included in Phase I as the eQTLGen data was not available at that time, and other resources (e.g., GTEx[148]) were based on smaller sample sizes. The reason for performing Phase II analyses was to integrate the new eQTLgen data to prioritize candidate causal genes which may mediate the phenotypic effects of shared variants at disease colocalized loci identified in Phase I. Moloc study pipeline (**Figure 5**) was developed to conduct multi trait colocalization analyses of all identified AF and stroke loci.

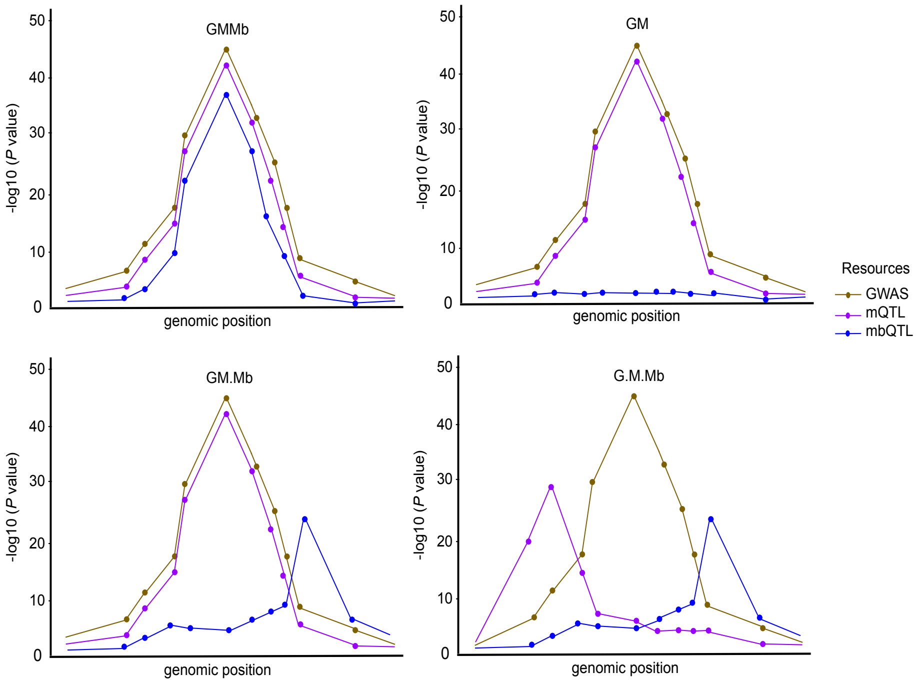

Figure 4. Schematic illustration of four plausible scenarios in a genomic region with 10 variants in common across GWAS, methylation and metabolite traits.

Scenarios where two or more traits have distinct causal variants are indicated by separating those traits with a "." (e.g. GM.Mb indicates colocalization of GWAS and DNA methylation, but a distinct non-colocalizing signal for metabolite).

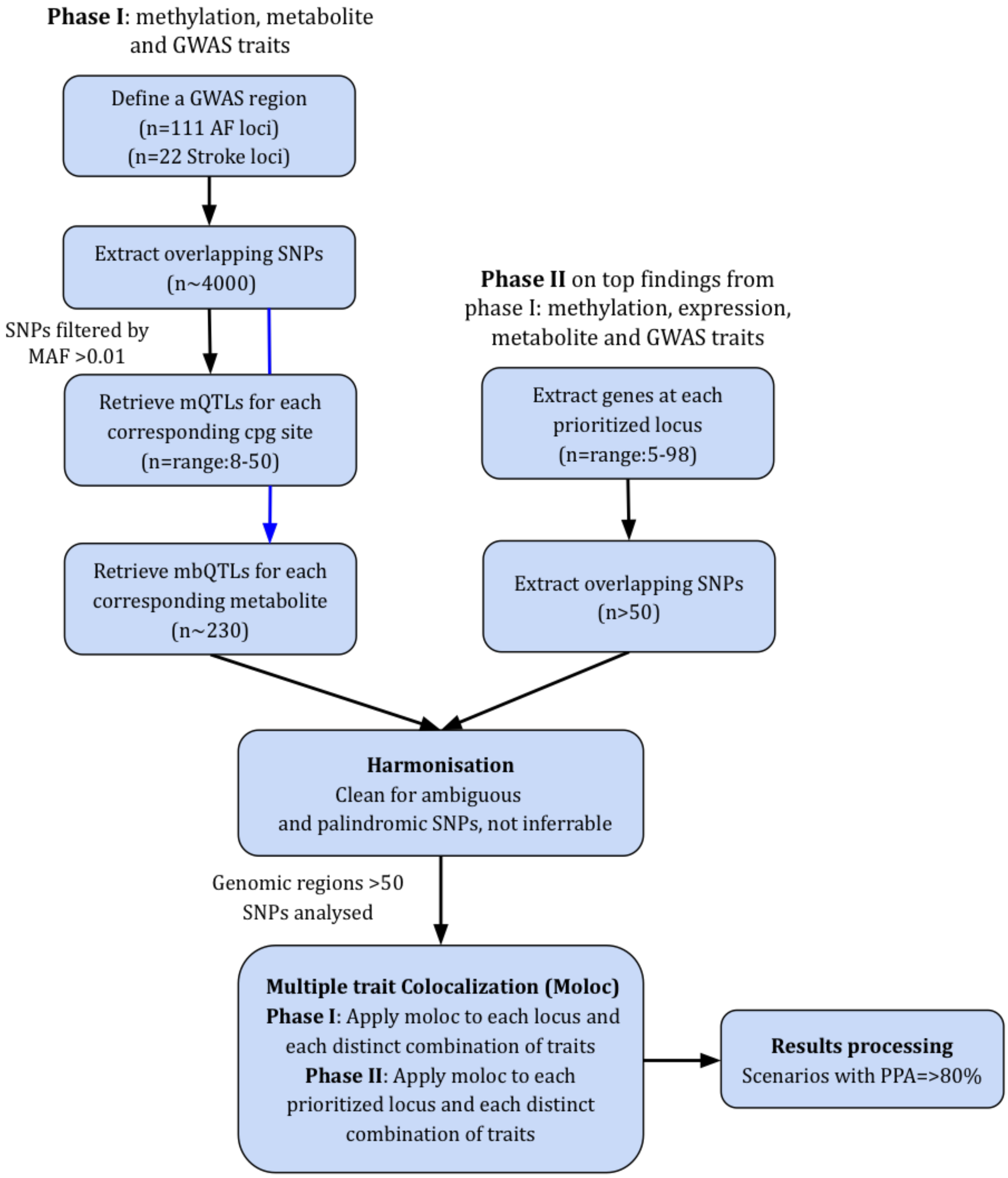
Phase I: methylation, metabolite and GWAS traits
Define a GWAS region (n=111 AF loci) (n=22 Stroke loci)
Extract overlapping SNPs (n~4000)
SNPs filtered by MAF >0.01
Retrieve mQTLs for each corresponding cpg site (n=range:8-50)
Retrieve mbQTLs for each corresponding metabolite (n~230)
Phase II on top findings from phase I: methylation, expression, metabolite and GWAS traits
Extract genes at each prioritized locus (n=range:5-98)
Extract overlapping SNPs (n>50)
Harmonisation Clean for ambiguous and palindromic SNPs, not inferrable
Genomic regions >50 SNPs analysed
Multiple trait Colocalization (Moloc) **Phase I**: Apply moloc to each locus and each distinct combination of traits **Phase II**: Apply moloc to each prioritized locus and each distinct combination of traits
Results processing Scenarios with PPA=>80%

Figure 5. Moloc analysis pipeline flowchart.

Phase I: My moloc analysis pipeline was applied to each GWAS locus (AF and stroke) and each distinct GWAS-CpG-metabolite combination of trait (E.g., 1 locus x n CpG sites x 230 metabolites). Phase II: moloc was then applied to each AF or stroke prioritized locus and each distinct GWAS-CpG-gene or GWAS-gene-metabolite or GWAS-CpG-gene-metabolite combination of trait. Phase II was only conducted on colocalized combinations found in Phase I.

2.2.1.1 Phase I: Identifying methylation CpG sites and metabolites colocalized at GWAS loci

In Phase I of these analyses, three datasets were employed containing summary level information for the genetic effects on methylation at CpG sites (mQTLs), circulating metabolites (mbQTLs), and complex trait (GWAS). Only common SNPs between all 3 datasets were kept for moloc analyses. Here I describe the analysis for AF, exactly the same approach was applied to stroke.

(i) Define GWAS region to conduct colocalization analysis on

A window size of 2Mb was defined in 111 AF genomic regions identified by Nielsen et al for colocalization analysis. The window was defined as 1Mb either side of the GWAS lead SNP at each of these genomic regions.

(ii) Extract the overlapping QTL and GWAS SNPs within each region

The summary statistics for all SNPs in common between the GWAS, mbQTL and mQTL datasets were extracted within the defined window (**Figure 6**). All three datasets were filtered by minor allele frequency (MAF) > 0.01 (1%).

(iii) Retrieve all mQTLs for each corresponding CpG site

All mQTLs (either *cis* or *trans* mQTLs) falling within the region were extracted. Every mQTL within the defined window was mapped to multiple CpG sites (ranging between 8 and 50 CpG sites per mQTL). The summary statistics for all the mapped mQTLs were extracted regardless of the CpG site location (i.e., CpG sites may fall outside our defined window).

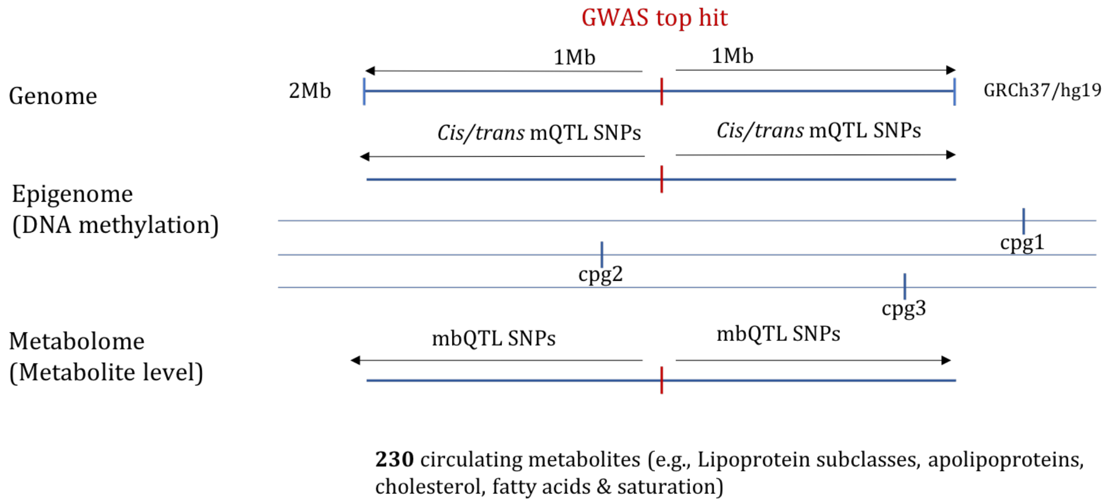

Figure 6. Illustrative diagram displaying the defined 2Mb window overlapping with mQTLs and mbQTLs in a GWAS region.

(iv) Retrieve all mbQTLs for each corresponding metabolite

All mbQTLs falling within the *cis* region were extracted. Every mbQTL within the defined window was mapped to multiple metabolites (approximately 230 per mbQTL). The summary statistics for all the mapped mbQTLs were extracted.

(v) Harmonisation of SNP effects

Data harmonisation was performed using the harmonisation functionality of the *TwoSampleMR*[234] (version 0.5.5) R package (version R-3.3.1-ATLAS) within our framework to ensure that the SNP effects of molecular traits and outcome trait correspond to the same strand and the same allele (i.e., the effect of a SNP on the DNA methylation and metabolite and the effect of that SNP on GWAS trait must each correspond to the same allele). Two sets of harmonisations were undertaken (treating the GWAS data as reference): harmonising the variant effects on DNA methylation to GWAS effect alleles, and separately harmonising the variant effects on metabolites to GWAS effect alleles. The two datasets were then merged together.

(vi) Run loci-wide moloc

To detect evidence of a shared causal variant across the mQTL, mbQTL and GWAS traits in a specific genomic region a multiple-trait colocalization implemented in the *moloc*[216] R package (v0.1.0) was applied. Moloc was run for each GWAS locus and each distinct combination of GWAS-CpG-metabolite (E.g., 1 locus x n CpG sites x 230 metabolites). In total, posterior probabilities of 15 potential scenarios (summarizing how a methylation CpG site, a metabolite and GWAS trait share a single causal variant in a given risk region) were calculated. The scenarios are listed in **Table 2**. Moloc analysis was only conducted on windows that had greater than 50 SNPs within them. The default priors of $1x10^{-4}$, $1x10^{-6}$ and $1x10^{-7}$ were used for the prior probability for any one layer of association ($p1$), for any two layers of associations ($p2$) and for colocalization of all three layers of associations ($p3$) across the traits. In these analyses, scenarios that support a possible sharing of the association signal for more than one of the phenotypes (i.e., GM, GM.Mb, GMb, GMb.M for two and GMMb for three phenotypes) were prioritized. A posterior probability of association (PPA) equal to, or greater than 0.8 (PPA>=80%) for these scenarios was defined as evidence of colocalization in the test region (**Figure 7**). Moloc analyses were conducted in R (version 3.3.1), using *moloc* downloaded as an R package (v0.1.0), which is available at Github (https://github.com/clagiamba/moloc).

For example, the code below was used to execute moloc of 3 traits.

```
ABF <- adjust_bfs_overlap(moloc_input_data, overlap=F, prior_var=0.15^2,
       from_p=F)
lkl <- config_coloc(ABF, n_files=3, priors=c(1e-04,1e-06,1e-07))
```

moloc_input_data contains R data frames of all three trait datasets (GWAS=G, mQTL=M, mbQTL=Mb) that are listed to be tested for colocalization in moloc. For example, a data frame for a GWAS trait contains “SNP”, “CHR”, “POS”, “A1”, “A2”, “EAF”, “BETA”, “SE”, “PVAL”, "N", which are vectors of the SNP-GWAS chromosome number, position, effect allele, alternative allele, effect allele frequency, coefficients, standard errors, p value, and sample size respectively.

`ABF` (Adjusted Bayes factors) is an array containing the log adjusted Bayes Factors (moloc logBFs) for each SNP and each configuration combination was returned using *adjust_bfs_overlap* function.

`lkl` (Likelihood frame) is a data frame containing 15 likelihoods and posterior probabilities (PPA) values for each of the 15 scenarios or combination of traits (e.g., AF, a specific CpG site, a specific metabolite) in the input using *config_coloc* function.

For my main moloc analyses priors were set to $p1$=1x10^{-4}, $p2$=1x10^{-6} $p3$=1x10^{-7}. For sensitivity moloc analyses priors were set to $p1$=1x10^{-5}, $p2$=1x10^{-7}, $p3$=1x10^{-8}.

Table 2. 15 plausible scenarios summarising sharing or not sharing of a likely causal variant among GWAS, methylation and metabolite traits.

Hypotheses for association of each trait with a genetic variant in a region. The "." separates traits with independent (non-colocalizing) causal variants.

Scenario	Hypothesis	Colocalization of variant
Null	H0- no association to any trait	no associations
G	H1- association for GWAS only	no colocalization
M	H2- association for methylation only	no colocalization
Mb	H3- association for metabolite only	no colocalization
GM	H4- association for GWAS and methylation	GWAS and mQTL
MMb	H5- association for methylation and metabolite	mQTL and mbQTL
GMb	H6- association for GWAS and metabolite	GWAS and mbQTL
G.M	H7- association for GWAS and methylation - two distinct causal variants	no colocalization
M.Mb	H8- association for methylation and metabolite - two distinct causal variants	no colocalization
G.Mb	H9- association for GWAS and metabolite - two distinct causal variants	no colocalization
G.MMb	H10- association for all traits - two distinct causal variants	mQTL and mbQTL not GWAS
GM.Mb	H11- association for all traits - two distinct causal variants	GWAS and mQTL not mbQTL
GMb.M	H12- association for all traits - two distinct causal variants	GWAS and mbQTL not mQTL
G.M.Mb	H13- association for all traits - three distinct causal variants	no colocalization
GMMb	H14- one single association for all 3 traits - one causal varaint	GWAS and mQTL and mbQTL

2.2.1.2 Overview of Phase II: Mapping to the putative causal gene(s) at prioritized loci

The pipeline described in **Figure 5** was extended further to integrate eQTL data from the eQTLGen consortium to map the likely causal genes responsible for the shared GWAS signals. Phase II was conducted as a follow-up analysis on the loci prioritized from Phase I (**Figure 7**) as data only became available after phase I had been completed. The same steps described for Phase I were followed, but with the addition of eQTLs as a third layer for scenarios where two traits colocalized or as a fourth layer for scenarios where three traits colocalized. These analyses were conducted on all *cis*-genes that lay within the defined window. The *trans* eQTL effects were not considered as all the prioritized CpG sites were *cis*-acting. The full eQTL summary statistics (with no p-value cut-off) were used in this analysis.

(i) Extract gene transcripts

A 2Mb window around each GWAS signal (GWAS SNP -/+ 1Mb) was taken (i.e., either GWAS top hit or additional independent variant identified as a shared variant in phase I) to select all *cis*-Genes at each locus. These *cis*-genes were extracted for all colocalized loci using the GENCODE.v19 annotation file based on overlap with the start or end point of the gene (gene annotation file mapped to GRCh37 genome build reference). The Ensembl gene IDs extracted from the Gencode annotation file were then merged with the same gene IDs provided with the eQTLGen summary statistics. Summary statistics for the *cis*-acting eQTLs within each window were then harmonized and merged with the summary statistics from the colocalized traits in Phase I. Moloc analysis was then conducted on each of these harmonized regions to estimate the probability of the *cis*-gene expression sharing the same likely causal variant as the other traits. A posterior probability (PPA) ≥ 80% for colocalization of a gene was considered as evidence for a potential causal gene.

If considering colocalization of 3 traits (GWAS, mQTL, and eQTL) the moloc technique calculates the evidence underpinning the 15 possible scenarios or combinations of traits (H0...H14) (**Table 3**), of sharing of a causal signal at the specific risk locus across all traits. In the case of integrating 4 traits, moloc computes the 52 possible scenarios (H0...H51). The scenarios considering 4 traits are listed in **Table 4**. For example, the code below was used to execute moloc integrating eQTL data as a fourth layer of trait.

```
ABF <- adjust_bfs_overlap(moloc_input_data, overlap=F, prior_var=0.15^2,
       from_p=F)
lkl <- config_coloc(ABF, n_files=4, priors=c(1e-04,1e-06,1e-07,1e-08))
```

`moloc_input_data` contains data frames of all four trait datasets (GWAS=G, mQTL=M, eQTL=E, mbQTL=Mb) that are listed to be tested for colocalization in moloc.

`ABF` (Adjusted Bayes factors) is an array containing the log adjusted Bayes Factors (moloc logBFs) for each SNP and each configuration combination was returned using *adjust_bfs_overlap* function.

`lkl` (Likelihood frame), a data frame containing 52 likelihoods and posterior probabilities (PPA) for each of the 52 scenarios or combination of traits (e.g., AF, a specific CpG site, a specific metabolite and a particular gene) in the input using *config_coloc* function.

For my main moloc analyses of four traits, priors were set to $p1=1\text{x}10^{-4}$, $p2=1\text{x}10^{-6}$ $p3=1\text{x}10^{-7}$, $p4=1\text{x}10^{-8}$. For sensitivity moloc analyses of four traits, priors were set to $p1=1\text{x}10^{-5}$, $p2=1\text{x}10^{-7}$, $p3=1\text{x}10^{-8}$, $p4=1\text{x}10^{-9}$.

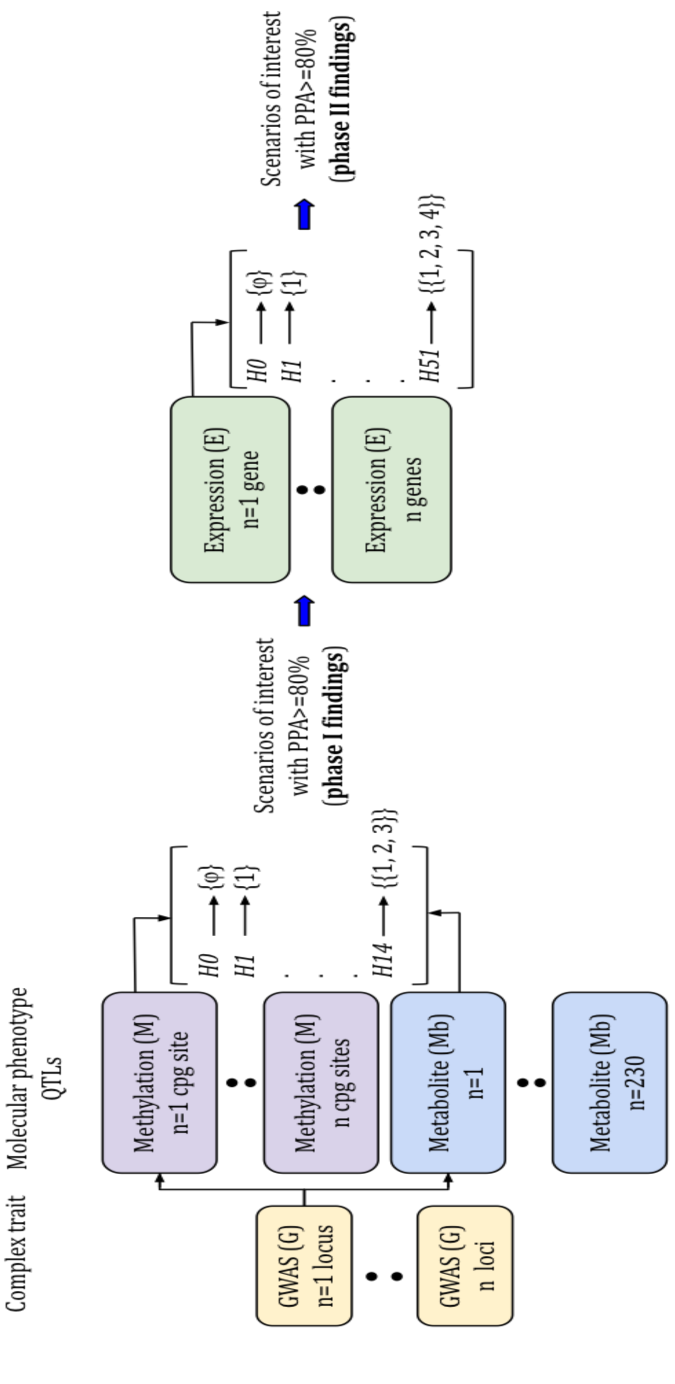

Figure 7. Moloc study design diagram.

Each GWAS locus was tested for colocalization with different CpG sites and different circulating metabolites at phase I. The 15 hypotheses/scenarios computed by moloc represent the possible combinations of the three traits in phase I. Each locus which colocalized in phase I, was then tested for colocalization with gene expression at phase II. The 52 hypotheses/scenarios computed by moloc represent the possible combinations of the four tested traits in phase II. Scenarios of interest were those that passed a posterior probability of colocalization threshold of 80% (PPA>=80%).

Table 3. 15 plausible scenarios summarising sharing or not sharing of a likely causal variant among GWAS, methylation and expression traits.

Colocalization hypotheses for association of each QTL with GWAS trait in a region. The "." separates traits with independent (non-colocalizing) causal variants.

Scenario	Hypothesis	Colocalization of variant
Null	H0- no association to any trait	no associations
G	H1- association for GWAS only	no colocalization
M	H2- association for methylation only	no colocalization
E	H3- association for expression only	no colocalization
GM	H4- association for GWAS and methylation	GWAS and mQTL
ME	H5- association for methylation and expression	mQTL and eQTL
GE	H6- association for GWAS and expression	GWAS and eQTL
G.M	H7- association for GWAS and methylation - two distinct causal variants	no colocalization
M.E	H8- association for methylation and expression - two distinct causal variants	no colocalization
G.E	H9- association for GWAS and expression - two distinct causal variants	no colocalization
G.ME	H10- association for all traits - two distinct causal variants	mQTL and eQTL not GWAS
GM.E	H11- association for all traits - two distinct causal variants	GWAS and mQTL not eQTL
GE.M	H12- association for all traits - two distinct causal variants	GWAS and eQTL not mQTL
G.M.E	H13- association for all traits - three distinct causal variants	no colocalization
GME	H14- one single association for all 3 traits - one causal varaint	GWAS and mQTL and eQTL

Table 4. 52 plausible scenarios summarising sharing or not sharing of a likely causal variant among GWAS, methylation, expression and metabolite traits.

Colocalization hypotheses for association of each QTL with GWAS trait in a region. The "." separates traits with independent (non-colocalizing) causal variants.

Scenario	Hypothesis	Colocalization of variant
Null	H0	no associations
G	H1	no colocalization
G.M	H2	no colocalization
G.E	H3	no colocalization
G.Mb	H4	no colocalization
G.ME	H5	no colocalization
G.MMb	H6	mQTL and mbQTL not GWAS
G.EMb	H7	eQTL and mbQTL not GWAS
G.MEMb	H8	mQTL and eQTL and mbQTL not GWAS
G.M.E	H9	no colocalization
G.M.Mb	H10	no colocalization
G.M.EMb	H11	eQTL and mbQTL not mQTL not GWAS
G.E.Mb	H12	no colocalization
G.MMb.E	H13	mQTL and mbQTL not eQTL not GWAS
G.ME.Mb	H14	mQTL and eQTL not mbQTL not GWAS
G.M.E.Mb	H15	no colocalization
M	H16	no colocalization
M.E	H17	no colocalization
M.Mb	H18	no colocalization
GE.M	H19	GWAS and eQTL not mQTL
GMb.M	H20	GWAS and mbQTL not mQTL
M.EMb	H21	eQTL and mbQTL not mQTL
GEMb.M	H22	GWAS and eQTL and mbQTL not mQTL
M.E.Mb	H23	no colocalization
GMb.M.E	H24	GWAS and mbQTL not mQTL not eQTL
GE.M.Mb	H25	GWAS and eQTL not mQTL not mbQTL
E	H26	no colocalization
E.Mb	H27	no colocalization
GM.E	H28	GWAS and mQTL not eQTL
GMb.E	H29	GWAS and mbQTL not eQTL
MMb.E	H30	mQTL and mbQTL not eQTL
GMMb.E	H31	GWAS and mQTL not mbQTL not eQTL
GM.E.Mb	H32	GWAS and mQTL not eQTL not mbQTL
Mb	H33	no colocalization
GM.Mb	H34	GWAS and mQTL not mbQTL
GE.Mb	H35	GWAS and eQTL not mbQTL
ME.Mb	H36	mQTL and eQTL not mbQTL
GME.Mb	H37	GWAS and mQTL and eQTL not mbQTL
GM	H38	GWAS and mQTL only
GM.EMb	H39	GWAS and mQTL, eQTL and mbQTL
GE	H40	GWAS and eQTL only
GE.MMb	H41	GWAS and eQTL, mQTL and mbQTL
GMb	H42	GWAS and mbQTL only

Scenario	Hypothesis	Colocalization of variant
GMb.ME	H43	GWAS and mbQTL, mQTL and eQTL
ME	H44	mQTL and eQTL only
MMb	H45	mQTL and mbQTL only
EMb	H46	eQTL and mbQTL only
GME	H47	GWAS and mQTL and eQTL only
GMMb	H48	GWAS and mQTL and mbQTL only
GEMb	H49	GWAS and eQTL and mbQTL only
MEMb	H50	mQTL and eQTL and mbQTL only
GMEMb	H51	GWAS and mQTL and eQTL and mbQTL

2.2.2 Two-sample Mendelian randomisation

I used 2SMR analysis in my thesis to determine whether there was evidence for a relationship between genetically predicted AF and stroke liability using the SNP summary statistics available from AF and stroke GWASs (see section **2.1**). To conduct the MR analysis, I used the mr function provided by the *TwoSampleMR* R package (version 0.5.5) maintained by MR-Base[234] (https://www.mrbase.org/). Detailed description of how MR was applied provided in **Chapter 6**. Below is a brief description of the different MR methods I used:

2.2.2.1 Wald ratio (WR)

Wald ratio (WR) is a very basic method to perform two-sample MR where only a single genetic instrument is required to estimate the magnitude of the causal effect by dividing the SNP-outcome coefficient by the SNP-exposure coefficient[248,218,249]:

$$\beta_{MR} = \frac{\beta_y}{\beta_x}$$

WR MR was performed using the mr-singlesnp function provided by the *TwoSampleMR* R package (version 0.5.5), and these effect estimates pooled using the following different methods.

2.2.2.2 Inverse variance weighted (IVW) MR

The Inverse Variance Weighted (IVW)[250] method uses a fixed effect meta-analysis to calculate an unbiased estimate of the overall causal effect in which each instrumental variant contribution is weighted by the inverse of the variance of the instrument-outcome effect. The IVW method has the most statistical power but does not adjust for invalid instruments and assumes that all genetic variants are valid IVs, whilst other methods have different assumptions to adjust for invalid instruments. If the IV1-IV3 assumptions hold, the causal estimate obtained from IVW linear regression model is unbiased[231,234].

2.2.2.3 MR-Egger

In the presence of directional horizontal pleiotropic effects, where the effect of the genetic instrument on the disease outcome is mediated through pathways other than the exposure, the MR-Egger regression technique[251] returns an unbiased effect estimate by allowing pleiotropic effects to exist, provided that there is no correlation between the size of these pleiotropic effects and the size of the instrument-exposure effects, known as the InSIDE (Instrument Strength Independent of Direct Effect) assumption[251,252].

2.2.2.4 Weighted median

The weighted median approach provides an unbiased causal effect estimate in the presence of a number of invalid instruments (that is in the case when the greater number of the instruments are valid) by taking all SNPs median effect with stronger SNPs contributing more to the estimate[253].

2.2.2.5 Weighted mode

If a number of SNPs have a pleiotropic effect the weighted mode groups SNPs by considering their causal effects and evaluates an unbiased causal estimate based on the biggest cluster of SNPs with valid instruments[254].

Chapter 3 Dissecting the molecular aetiology of atrial fibrillation

3.1 Introduction

AF is the most common sustained arrythmia which is associated with an increased risk of major complications such as mortality[28], heart failure[29], myocardial infarction[30] and stroke[33,34,35]. A large GWAS meta-analysis by Nielsen et al. has identified 111 genetic loci with increased susceptibility to AF[41]. As most of the identified SNPs mapped to non-coding regions of the genome, it is challenging to map these genetic loci to the causal gene(s) and underlying molecular mechanisms. These AF SNPs are strongly enriched in regions of open chromatin and active enhancers[41], supporting the hypothesis that most common genetic risk signals affect transcriptional regulation rather than influencing the coding regions of genes directly[18]. In Nielsen et al[41] different methods were applied to identify potential functional effects of SNPs. The candidate genes in each of the regions were prioritized based on several criteria which included (i) physical proximity of the gene to the AF variant (ii) overlap of the gene's eQTLs with an AF associated region (iii) the gene containing a non-synonymous SNP which is in LD with the AF variant and (iv) the gene has tissue specific expression consistent with AF phenotype. The identified candidate genes perform a wide range of functions: they encoded transcription factors with a potential role in mediation of events related to development (e.g. *GATA4*, *ZFHX3* and *PITX2*), cell structural proteins likely to be involved in skeletal and cardiac muscle integrity and function (e.g. *MYOZ1*, *RBM20*, and *SYNPO2L*), and genes likely to regulate endocrine function (e.g. *ESR2* and *CGA*) and be involved in intracellular signal transduction process in the heart (e.g. *CAMK2D* and *CALU*) and function of cardiac ion channels (e.g. *HCN4* and *KCNN3*)[41]. This study also prioritized AF candidate genes in several heart-related tissues using GTEx datasets. However, for most of the loci Nielsen et al detected the molecular mechanism underpinning the genetic associations remains unknown.

A common approach to map genetic association signals to molecular mechanism is to identify shared genetic variants between the GWAS signal and potential molecular mediators[225,188,255]. Several studies have found genetic variants associated with circulating metabolite concentrations (mbQTLs) were also *cis*-eQTLs[174,193]. Other studies have shown that DNA methylation is also associated with metabolite levels[200,201,202,203,204,205], therefore an integrative multi-omics investigation of these two molQTLs should be informative. Indeed, previous studies demonstrated that integration of multi-omics data (DNA methylation, metabolite levels and gene expression) can be usefully applied to expand our knowledge of molecular mechanisms at several loci[237,238,239,22,255]. For example, multi-omics studies of AF have identified candidate genes which might have a functional mediatory role along the pathway to AF susceptibility[256,257]. Assum et al integrated atrial tissue-specific genetic information on transcriptome (*cis*-eQTLs), proteome (*cis*-pQTLs) along with AF GWAS and found potential relationships between gene expression and protein abundance controlled by the same variant at several loci. Assum et al also constructed eQTSs (expression quantitative trait rank scores) based on the AF PRS association with transcript expression and trans pQTSs (protein quantitative trait rank scores) based on the AF PRS association with protein abundance to explore the trans-acting regulatory mechanisms at AF loci. A follow-up analysis using trans eQTSs and trans pQTSs identified a regulatory role for transcription factor (TF) NKX2-5 as a mediator along the pathway from AF GWAS signal (rs9481842) located at the 6q22 locus to AF susceptibility[257]. Interestingly, the AF-relevant pathways reported involved in metabolism and contractile function of the heart[258,259,260]. Wang et al used multi-omics approaches on whole blood data from AFGen GWAS of AF[261], EWAS of AF[136] and transcriptome-wide association study (TWAS) of AF[262] and found 1931 genes relevant to AF. Follow-up analysis incorporating gene set findings with cardiac-specific gene interaction network, uncovered cardiac related molecular pathways linked to AF susceptibility[256].

In this chapter I employ a multi-omics colocalization technique to further understand the genetic pathways contributing to the risk of AF. Herein, a large-scale two-phased multi-omics colocalization study of AF loci in Europeans was performed by employing GWAS of AF (n=1,030,836), summary statistics of DNA methylation (GoDMC, n=27,750) and circulating metabolite levels (ALSPAC, n=5,589), to identify shared regulatory effects between DNA methylation, metabolome and AF (**Figure 8**). In a second phase, moloc analysis was further extended to integrate eQTL data from eQTLGen (n=31,684), to map the potential causal genes responsible for these shared effects. I used moloc to explore whether the observed overlap between molecular traits and AF may be due to a single variant driving the shared regulatory effect at each AF risk locus.

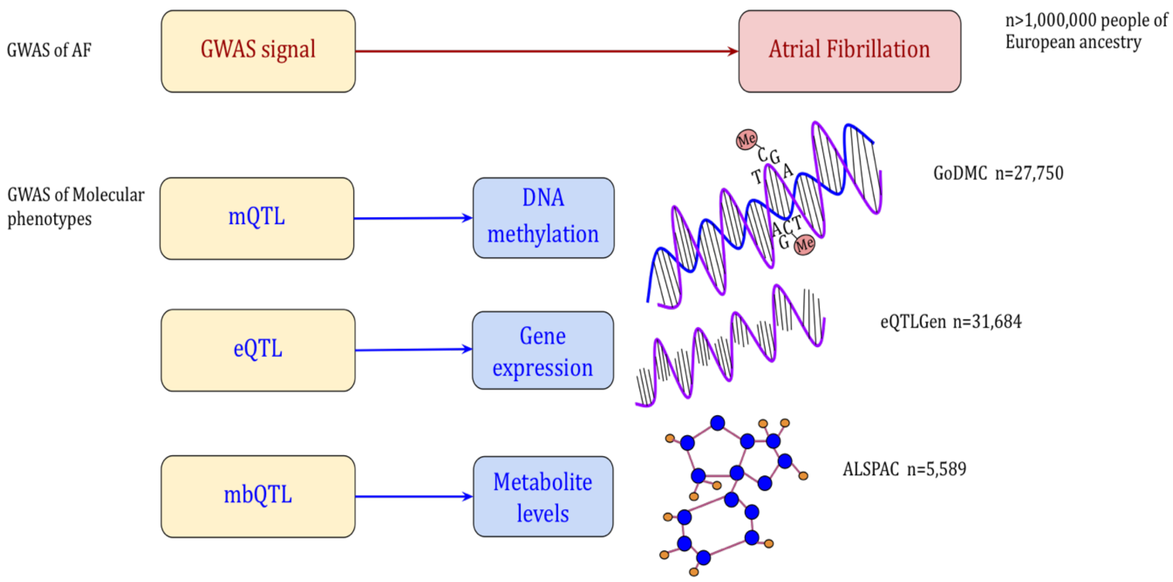

Figure 8. Schematic diagram of SNP-AF and SNP-molecular phenotype associations used in moloc analyses.

3.2 Methods

I downloaded AF GWAS summary statistics[41] (number of SNPs = 34,459,399) from: http://csg.sph.umich.edu/willer/public/afib2018/ (date accessed: 3 August 2018). Quality control (QC) check was conducted by removing SNPs with non rsIDs (identifiers), indels and multi allelic SNPs. SNPs with MAF>0.01 were kept, leaving 22,760,461 associations for moloc analyses. SNP associations were further filtered based on P value>0.05, keeping 5,538,471 SNP associations to generate a Manhattan plot. I used the SNP2GENE function of FUMA (https://fuma.ctglab.nl/)[263] to provide the Manhattan plot for genetic associations of AF GWAS (**Figure 9**).

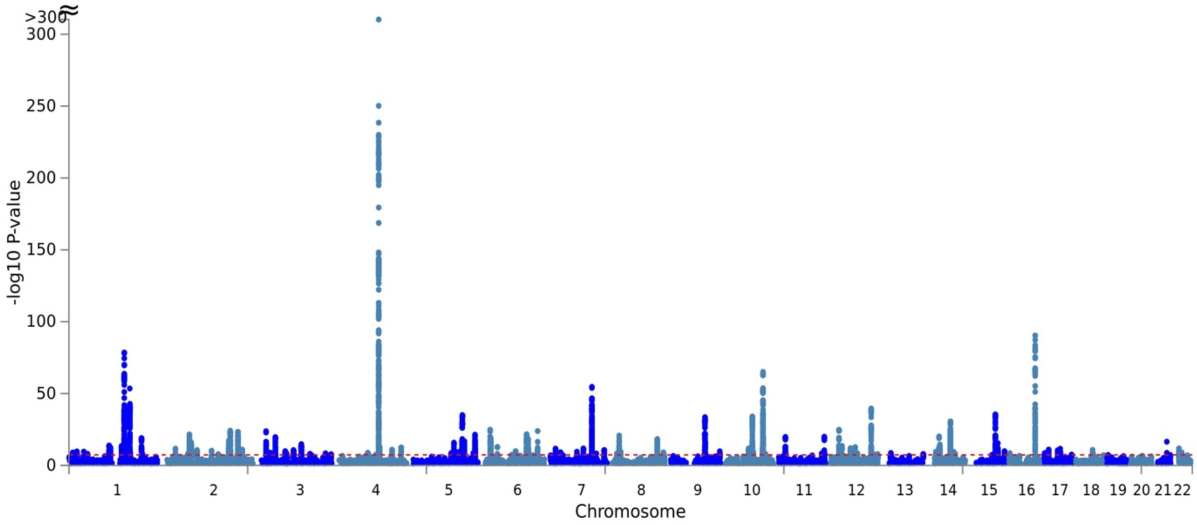

Figure 9. Manhattan plot displaying genetic risk loci associated with AF.

3.2.1 Phase I – moloc of methylation, metabolite and AF

A large-scale multi-omics colocalization analysis was conducted on 111 AF-associated loci by employing GWAS of AF, mQTL summary statistics of DNA methylation sites profiled with the Illumina Infinium HumanMethylation 450k array (GoDMC) and mbQTLs of plasma metabolite levels profiled with NMR platform (ALSPAC). Analyses were performed by defining a window size of +/–1Mb around each of the top AF risk SNPs at 111 loci. For each AF risk locus, the moloc analysis pipeline was applied to identify statistical evidence of a shared regulatory effect between AF, each methylation CpG site (n=ranging from 8 to 50) and each metabolite (n~230). Detailed description of my pipeline and how it was applied at each locus can be found in Chapter 2 **2.2.1**). Across all analyses, a scenario (i.e., combination of traits) was considered "colocalized" or with "evidence of colocalization" if its posterior probability of association (PPA) was equal to or greater than 80%. Overall, 23 loci were prioritized at this phase. CpG sites for which I found shared genetic effects with metabolite and AF were annotated to CpG site physical position in the genome and the gene they are annotated to, using the *meffil* (v1.0.0) R package[264] (https://github.com/perishky/meffil/).

3.2.2 Phase II – moloc of methylation, gene expression, metabolite and AF at prioritized loci

In the second phase, the moloc analysis pipeline was extended to integrate eQTL data from the eQTLGen consortium to identify the potential causal gene(s) responsible for the shared AF signal at each of the 23 prioritized loci. The pipeline is described in Chapter 2 **2.2.1**). At each locus, colocalization of all *cis*-genes (ranging from 5 to 98 genes) with each combination of traits (i.e., GM (colocalized AF-CpG site), GMb.M (colocalized AF-metabolite but not CpG site), GMMb (colocalized AF-CpG site-metabolite) was tested.

3.2.2.1 Assessment of current state of knowledge to evaluate putative drug targets

The colocalized genes were looked up in two publicly available data resources to evaluate the potential druggability and functional evidence for involvement in related CVDs specially AF pathology. The following online resources were used: (i) Open Targets[265] to detect if the gene is used or known as pharmaceutical drug target for any indications/diseases. The Ensemble ENSG gene identifier was used in the Open Targets Genetics database (https://genetics.opentargets.org) to view information on the gene. In addition, the "Mouse Phenotypes" section in the Open Targets platform was searched for evidence that gene knock down/out experiments recapitulated/demonstrated cardiovascular system phenotype, (ii) Online Mendelian Inheritance in Man (OMIM)[266] was used to look for relevant human monogenic cardiovascular disorders caused by mutation(s) in the gene. The "Allelic Variants" field of the OMIM website (https://omim.org) was manually reviewed for the specific gene and any evidence of phenotypes related to CVDs caused by a single-gene mutation was recorded. The literature was also searched for evidence in animal models if there was no data available in the Open Targets Mouse Phenotypes annotations section.

3.2.3 Sensitivity analyses on moloc probabilities

3.2.3.1 Selection of more stringent priors

In phase I and II analyses, the priors in moloc (v0.1.0) were set to the default settings (i.e., $p1=1x10^{-4}$, $p2=1x10^{-6}$, $p3=1x10^{-7}$ and $p4=1x10^{-8}$ when the fourth layer was included). *p*2 is the prior for one genetic variant associated with two traits and *p*3 and *p*4 are the prior probabilities for association with three or four traits. To assess whether the results were robust to these default settings, moloc analyses were re-run applying more stringent prior probabilities for *p*1, *p*2, *p*3 and *p*4 ($p2=1x10^{-5}$, $p2=1x10^{-7}$, $p3=1x10^{-8}$ and $p4=1x10^{-9}$) for AF regions (n=23 in phase I and n=10 in phase II) with evidence of colocalization.

3.2.3.2 LD score analysis

Colocalized findings might be confounded by the LD pattern and number of SNPs in the region. The LD score is defined as the sum of the pairwise r^2 between the lead variant and all the variants residing within 500kb of a region. It provides a representation of the extent of the variant correlation with nearby SNPs, capturing both quantity of correlated SNPs and strength of LD. To estimate LD scores[267], all the overlapping SNPs (MAF>0.01) in each colocalized region (with no r^2 cut-off) were used and LD data from HRC-imputed unrelated ALSPAC individuals was employed as an LD reference panel. The *ld-score* command implemented in GCTA software (gcta_1.91.1beta)[227] was used. LD scores of regions were then compared to the PPA values.

3.2.3.3 Zero imputation of missing GoDMC mQTL data (P=1, Beta=0)

For computational reasons, the GoDMC mQTL data has been generated using a two-phase design (see Methods **2.1.3**) across 36 datasets and therefore doesn't include summary statistics on all SNPs. Instead, it has summary statistics on all SNPs across all 36 datasets that were P value<1x10^{-5} in at least one dataset for *cis*-mQTLs and two datasets for *trans*-mQTLs. As I only used overlapping SNPs across all 3 datasets for colocalization, many SNPs were removed from the tested region. Therefore, as a sensitivity analysis, the mQTLs were zero imputed in the 17q12 region (where evidence of colocalization (PPA.GMMb=80.4%) was detected with 149 SNPs in the region) to test if this would affect the moloc PPA value. mQTLs for cg22833065 CpG site were imputed using the 1000G SNP file (nSNPs=10,085,072; MAF>0.01) (which had all SNPs tested in GoDMC mQTL analysis) to map all missing mQTLs in the 17q12 region. This imputation was performed by replacing the missing mQTLs with null effects, using a completely flat effect size (β coefficient) =0.00 and P value=1.00. The minor allele for each of the missing variants was selected as the effect allele.

3.3 Results

3.3.1 Phase I – Identification of AF loci that colocalized with DNA methylation and metabolite traits.

In the first phase of this study, colocalization analyses were performed on 111 AF loci to map shared genetic effects linking AF loci to two types of intermediate molecular phenotypes: DNA methylation and metabolite levels. Of the 111 loci for AF, evidence for colocalization was detected at 3 loci across all three traits, 23 loci with a CpG site only and 1 locus with metabolite traits only comprising a total of 23 loci with evidence of colocalization. Multiple scenarios were found at 3 loci (**Table 5**).

3.3.1.1 Regions with evidence of colocalization between methylation and AF

While we were most interested in finding evidence of colocalization where the same single SNP is shared among all three traits (GWAS, mQTLs and mbQTLs (GMMb)) colocalization was detected at 23 loci where AF risk variants are shared with a significant *cis* mQTL only ($P<1x10^{-8}$) for at least one CpG site (PPA.GM >=80%, 41 unique SNP-CpG combinations), (**Table 5**). Of the 23 loci, 9 (39.1%) showed evidence of colocalization with genetic effects of methylation at multiple different CpG sites whereas 14 (60.9%) loci colocalized with mQTLs that were associated with a single CpG methylation site. For example, the AF risk SNPs at 1q32 and at 12p12 shared a *cis*-acting genetic variant associated with DNA methylation at three and four CpG sites respectively (**Table 5**). The finding that a substantial number of shared variants might influence the risk of AF through DNA methylation at multiple CpG sites is consistent with the existence of co-methylated blocks in the genome that are partly genetically regulated[268]. However, horizontal pleiotropy should be considered as another potential interpretation (**Figure 1**, explanation 1.2). All the mQTLs that colocalized with AF risk signals were found to be associated in *cis* with methylation CpG sites (i.e., *cis*-acting influences with distance <1Mb between mQTL and CpG site). In addition, among the shared AF-mQTL SNPs, both mQTL effect size and significance tend to increase as the distance between CpG site and the shared variant decreases, suggesting that colocalized CpGs closer to the shared AF-mQTL signal are more likely to be strongly affected by the shared AF variant.

Table 5. Colocalized CpG sites and metabolites identified in the phase I of moloc analysis.

Number of SNPs in the given region (nSNPs) with scenarios represents sharing of variant between AF and methylation CpG site (GM), sharing of variant between AF and circulating metabolite but not CpG site (i.e.,2 causal variants) (GMb.M), sharing of variant between AF, methylation CpG site and metabolite (GMMb), posterior probability (PPA) of colocalization between molecular traits and AF in the main moloc and sensitivity moloc analysis. Variant with * indicates the additional independent AF risk variant at the locus. High density lipoprotein (HDL), Very low-density lipoprotein (VLDL). Lipoprotein subclasses and their ratio : Total cholesterol to total lipids ratio in medium VLDL (M_VLDL_C_PC), Cholesterol esters to total lipids ratio in small HDL (S_HDL_CE_PC), Cholesterol esters to total lipids ratio in small VLDL (S_VLDL_CE_PC), Total cholesterol to total lipids ratio in small VLDL (S_VLDL_C_PC), Triglycerides to total lipids ratio in small VLDL (S_VLDL_TG_PC), Total cholesterol in very small VLDL (XS_VLDL_C), Free cholesterol in very small VLDL (XS_VLDL_FC), Fatty acids and saturation : 22:6, docosahexaenoic acid (DHA), Ratio of 22:6 docosahexaenoic acid to total fatty acids (DHA_FA), Amino acids: Alanine (ALA), Apolipoproteins: Ratio of apolipoprotein B to apolipoprotein A-I (APOB_APOA1). Each row in the locus column represents a single AF hit locus, with sub-rows representing the different colocalized molecular traits.

locus	risk SNP	scenario	nSNPs	CpG site	metabolite	main moloc	Sensitivity moloc
1p36	rs7529220	GM	297	cg16583536		96.6	90.5
1q21	rs11264280	GM	123	cg19233405		96.8	96.2
1q24	rs577676*	GM	165	cg22693806		96.3	86.3
1q32	rs10753933	GM	92	cg03900565		91.4	81.0
			62	cg11656175		88.6	49.2
			115	cg23098069		89.3	51.3
3p25	rs7650482	GM	504	cg24848339		95.6	92.7
3p14	rs34080181	GM	69	cg15724417		89.4	48.9
4q34	rs10520260*	GM	208	cg24950233		97.5	95.4
			92	cg18575740		96.9	87.0
			57	cg13935962		96.8	86.9
5q35	rs6891790	GM	105	cg13004182		89.0	83.8
			416	cg12825773		83.9	78.1
			186	cg18839504		88.4	83.6
	rs28439930*		370	cg06889108		90.3	63.2
7q21	rs56201652	GM	74	cg10481072		85.4	39.6
7q32	rs55985730	GM	200	cg18693656		95.5	94.0
			308	cg13951589		95.3	95.1
			410	cg10826733		94.9	93.6
8q24	rs6994744	GM	610	cg26291848		93.6	89.0
			612	cg14396066		87.9	50.3
			632	cg10996527		82.3	36.4
9q34	rs2274115	GM	207	cg04455058		97.2	96.5
10q21	rs12245149	GM	879	cg01631684		80.1	50.1
		GMMb	879	cg01631684	M_VLDL_C_PC	80.5	55.9
					S_HDL_CE_PC	82.5	58.5
					S_VLDL_CE_PC	80.1	52.7
					S_VLDL_C_PC	84.0	61.9
					S_VLDL_TG_PC	84.8	63.3
					XS_VLDL_C	83.5	60.5
					XS_VLDL_FC	84.4	63.0
		GMb.M	197		DHA	82.0	31.4
					DHA_FA	82.0	31.3

locus	risk SNP	scenario	nSNPs	CpG site	metabolite	main moloc	Sensitivity moloc
10q22	rs60212594	GM	75	cg16228286		90.1	62.3
			73	cg24637261		86.2	60.1
		GMMb	75	cg16228286	ALA	85.7	80.1
			73	cg24637261		80.8	62.1
10q24	rs11598047	GM	147	cg17426192		97.2	91.6
12p12	rs17380837	GM	247	cg22232504		94	69.2
			127	cg11332519		92.3	59.7
			229	cg07725355		94.6	72.1
			129	cg02593205		93.7	66.3
14q24	rs74884082	GM	58	cg25949241		87.6	44.2
15q24	rs74022964	GM	74	cg10576051		84.5	37.7
			78	cg06071033		88.1	58.9
			239	cg01796676		93.6	81.8
15q25	rs2759301*	GM	60	cg12292492		85.3	38.9
			1018	cg13148921		85.8	46.3
16q22	rs2359171	GM	82	cg03463523		92.2	60.9
17p13	rs9899183	GM	665	cg01557754		89.6	55.6
17q12	rs11658278	GM	149	cg22833065		81.3	32
		GMMb			APOB_APOA1	80.4	63.4
17q25	rs12604076	GM	343	cg23834688		84.6	40.1

3.3.1.2 Regions with evidence of colocalization between methylation, metabolite and AF

Three loci were identified with PPA above 80% for the GMMb scenario supporting the evidence of sharing the same variants between DNA methylation, metabolite and AF traits (**Table 5**). These colocalizations corresponded to 10 unique AF-CpG-metabolite combinations (PPA.GMMb) and 2 unique AF-metabolite pairs (PPA.GMb.M). The top AF hit rs12245149 at the chromosome 10q21 locus was found to be shared with DNA methylation at cg01631684 and circulating metabolites (multiple lipoprotein subclasses). Analysis of colocalization in the 10q22 region demonstrated that the AF risk variant, rs60212594 was shared between AF, two CpG sites (cg16228286 and cg24637261) and Ala (Alanine) amino acid. Moreover, rs11658278 at the 17q12 locus was shared across all three traits (PPA.GMMb=80.4%). At this locus, cg22833065 CpG site and ApoB_ApoA1 (Ratio of apolipoprotein B to apolipoprotein

A-I) colocalized with AF-associated signal, rs11658278. The shared variant resides within the large LD block spanning multiple genes including *ZPBP2*, *GSDMB* and *ORMDL3*. These colocalized loci were further studied in Phase II to identify which gene(s) may be prioritized as a putative causal gene(s) responsible for the shared genetic effect of AF.

3.3.2 Phase II – Mapping Phase I methylome and metabolome findings to the potential causal gene(s) for atrial fibrillation

3.3.2.1 Regions with evidence of colocalization between AF, methylation, metabolites and/or gene expression

Multi-trait colocalization analyses were expanded by integrating eQTL data at 23 colocalized loci found in phase I. The number of loci and their unique combinations of traits at PPA>80% in these analyses are reported in **Table 6**. Ten loci were identified with probability of a shared variant (PPA>80%) for at least one of the 3 possible molecular colocalization scenarios. Of these loci, I identified 9 loci with 23 GME scenarios (unique AF-CpG-gene configurations), 2 loci with 3 GMEMb scenarios (unique AF-CpG-gene-metabolite configurations) and 1 locus with 4 GEMb.M scenarios (unique AF-gene-metabolite configurations and not CpG site). At the 17q12 locus, evidence of colocalization was found between *ERBB2* gene expression, cg22833065 methylation, ratio of apolipoprotein B to apolipoprotein A-I (APOB_APOA1, an apolipoprotein) and AF (PPA.GMEMb=81.8%). The C allele of rs11658278 was significantly associated with decreased expression of *ERBB2* gene in the eQTLGen data (ß=-0.092, P=7.21x10^{-15} [i.e., *cis*-eQTL]). These findings suggest that the genetic variant driving the observed effect on AF might be responsible for alterations in cg22833065 methylation, *ERBB2* gene expression and APOB_APOA1 metabolite levels at this locus. Of the 10 genes tested for colocalization with AF-cg03463523 pair (PPA.GM=92.2%) at the 16q22 locus, only *HP* (haptoglobin) had evidence of colocalization. This locus encompasses two independent AF variants (rs2359171 and rs876727) with *ZFHX3* gene residing in close physical proximity to both independent variants. The top AF risk variant, rs2359171-T (ß=-0.175, P=4.65x10^{-91}) at this locus associated with DNA methylation at cg03463523 CpG site (*cis*-mQTL; ß=0.143, P=1.87x10^{-36}) colocalized with the *cis*-eQTL SNP that was associated with decreased expression level of the *HP* gene (ß=-0.095, P=7.45x10^{-10}) in this analysis (PPA.GME=96.7% rs2359171-cg03463523-*HP* combination) (**Figure 10**). No evidence of colocalization was identified between the nearby *ZFHX3* gene (PPA.GM.E=93.2%) and the colocalized AF-cg03463523 pair found in the phase I analysis. Rs2359171 was not associated with the expression of *ZFHX3* (P=0.24). The secondary AF risk variant, rs876727 at this locus (ß=-0.084, P=1.97x10^{-23}) was showed only weak evidence of association with the expression of the *HP* gene in whole blood (ß=0.045, P=1.62x10^{-3}). This locus is studied in more detail in **Chapter 5**.

Table 6. Colocalized combination of traits identified in the phase II of moloc analysis of AF. Number of SNPs in the given region (nSNPs) with scenarios represents sharing of variant between AF, methylation, and expression (GME), sharing of variant between AF, expression, metabolite but not methylation (i.e.,2 causal variants) (GEMb.M), sharing of variant between AF, methylation, expression and metabolite (GMEMb), posterior probability of colocalization between molecular traits and AF in the main moloc and sensitivity moloc analysis. Each row in the locus column represents a single AF hit locus, with sub-rows representing the different colocalized molecular traits.

locus	risk SNP	scenario	nSNPs	CpG site	gene	metabolite	main moloc	Sensitivity moloc
1q21	rs11264280	GME	106	cg19233405	*DAP3*		84.7	83.4
					SYT11		90.9	90.1
					YY1AP1		98.0	97.9
					MSTO2P		98.2	98.2
1q32	rs10753933	GME	62	cg11656175	*CHI3L1*		89.8	83.3
			109	cg23098069	*KLHL12*		84.9	76.1
7q21	rs56201652	GME	72	cg10481072	*GATAD1*		96.2	94.5
					KRIT1		93.3	89.4
7q32	rs55985730	GME	190	cg18693656	*RBM28*		99.9	99.9
			296	cg13951589			99.9	99.8
10q21	rs12245149	GEMb.M	197		*NRBF2*	DHA	87.7	77.8
					JMJD1C	DHA	93.4	89.6
					NRBF2	DHA_FA	87.7	77.8
					JMJD1C	DHA_FA	93.4	89.6
10q22	rs60212594	GME	70	cg16228286	*CAMK2G*		94.6	92.0
					P4HA1		91.0	87.2
			68	cg24637261	*BMS1P4-AGAP5*		93.0	87.9
					P4HA1		89.0	78.9
					MRPS16		83.7	66.4
		GMEMb	70	cg16228286	*CAMK2G*	ALA	82.6	82.7
		GMEMb	68	cg24637261	*BMS1P4-AGAP5*		82.5	82.3
14q24	rs74884082	GME	58	cg25949241	*PSEN1*		98.1	97.7
					ACOT4		87.4	83.8
					AC004846.1		95.5	92.9
16q22	rs2359171	GME	80	cg03463523	*HP*		96.7	95.2
17q12	rs11658278	GME	147	cg22833065	*ERBB2*		95.4	92.0
					IGFBP4		90.1	81.7
					RP11-94L15.2		84.5	72.1
		GMEMb			*ERBB2*	APOB_APOA1	81.8	81.6
17p13	rs9899183	GME	629	cg01557754	*RP11-186B7.4*		93.4	91.8

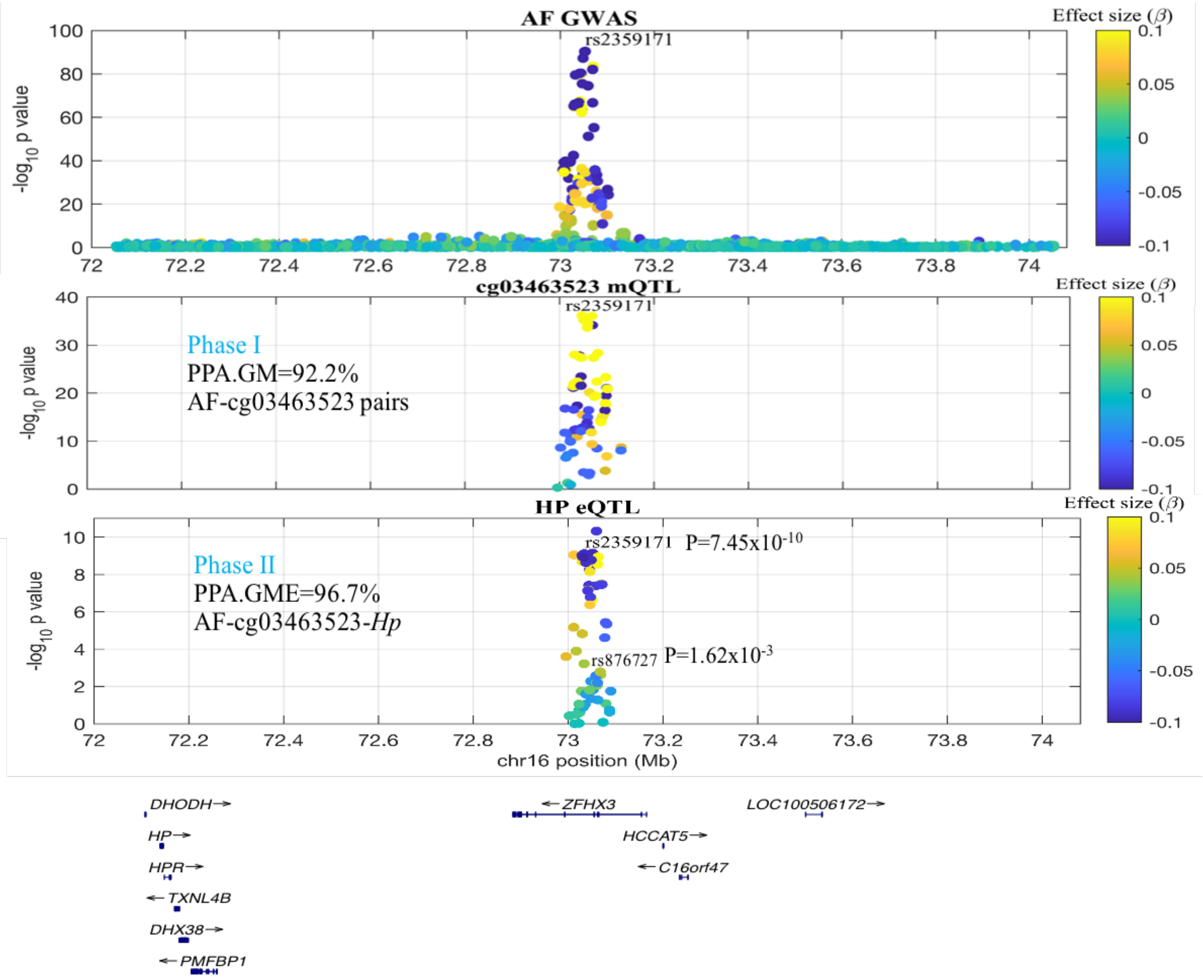

Figure 10. Locus association plot displaying a single association peak for AF, cg03463523 CpG site and *HP* gene in the *cis* region.

The top hit for AF (rs2359171) colocalized with the independent hit (mQTL) for cg03463523 and the independent hit (eQTL) for *HP* gene at the 16q22 locus (PPA.GME=96.7%). The secondary hit for AF (rs876727) is weakly associated with the expression of the *HP* gene.

3.3.2.2 Evidence of molecular pleiotropy

Among 10 loci with evidence of colocalization with multiple traits, eight loci were found with shared genetic effects between AF and multiple molecular phenotypes of the same kind (i.e., different CpG sites, different genes, and different metabolites) whose regulation is under the influence of a shared *cis*-regulatory signal. Two loci (10q21 and 10q22) were identified with multiple AF-CpG-metabolite combinations (GMMb) (**Table 5**), 7 loci (1q21, 1q32, 7q21, 7q32, 10q22, 14q24, and 17q12) with multiple AF-CpG-gene combinations (GME), 1 locus (10q21) with multiple AF-gene-metabolite combinations (GEMb.M), and 1 locus (10q22) with multiple AF-CpG-gene-metabolite combinations (GMEMb) (**Table 6**). These loci may act as master regulators (a SNP that controls groups of intermediate phenotypes) on the methylation of multiple CpG sites or may co-regulate the expression of multiple genes (including causal and non-causal genes) and/or multiple metabolites (**Figure 11**).

3.3.2.3 Direction of effects for colocalized variants for SNP-CpG-gene and SNP-CpG-metabolite combinations

Among 23 unique AF-CpG-gene combinations with strong evidence of colocalization (PPA.GME>80%), the direction of association of the variant with methylation and expression was in the opposite direction for 52.2% of the associations and in the same direction for 47.8% in the same directions (i.e., roughly equal). For the 10 colocalized unique AF-CpG-metabolite combinations, the direction of the effect of the variant on methylation and metabolite level was more frequently in the opposite direction (80%) than in the same direction (20%). Two examples of mQTLs that colocalized with an eQTL that has the same or opposite effect on nearby genes are shown in **Figure 12.** This is consistent with the hypothesis that the decrease in methylation of CpG site located in the promoter region is indicative of a more open chromatin state and/or upregulation of gene expression or transcriptional activity and that increased promoter CpG methylation is indicative of induced transcriptional repression.

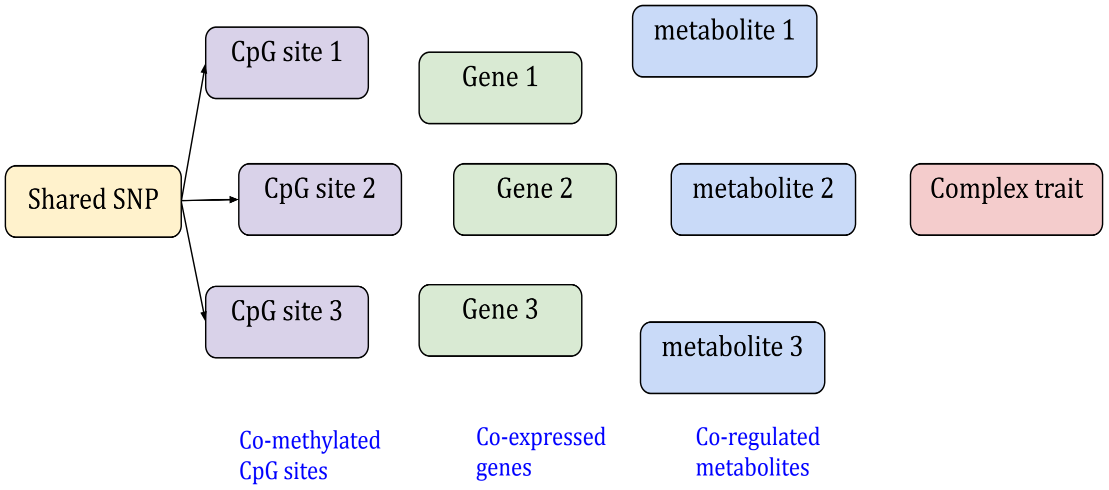

Figure 11. An explanation for colocalization between multiple molecular traits and a complex trait.

Potential scenarios that can explain colocalization of multiple traits of the same kind. A shared genetic variant influences a complex trait through multiple CpG sites or/and genes or/and circulating metabolites which are co- methylated, expressed or regulated with one another. I.e., colocalization scenarios: GM, GMb, GMMb, GME, GEMb, GMEMb. Colocalization can pinpoint potential causal gene(s) and prioritize biological candidate molecular traits which might have relationship with one another. Any of the CpG sites/genes/metabolites might be causal and the other CpG sites/genes/metabolites might be non-causal (as a bystander) but, are co-regulated together. Alternatively, multiple molecular phenotypes of the same kind might be co-regulated by a single shared variant and all might be involved in the same pathway to the GWAS trait.

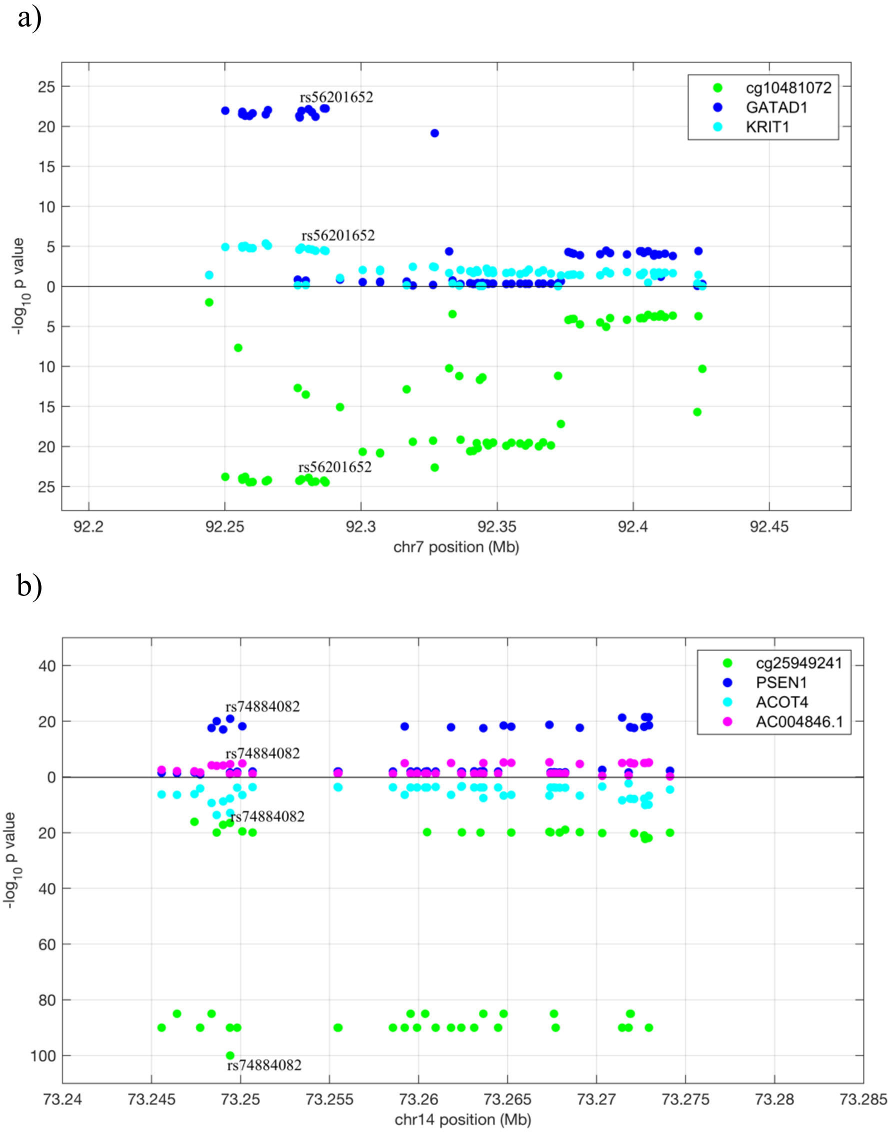

Figure 12. Examples of mQTLs that colocalize with an eQTL that has the same or opposite effects on nearby genes.

Results for the CpG selected for colocalization analysis (green) are shown on the bottom of each plot (descending), overlaid with the eQTL results (dark/light blue and pink). For the 1st scenario (a), the methylation-decreasing allele (rs56201652, G) at the 7q21 locus is associated with increased expression of the *GATAD1* and *KRIT1* genes. For the 2nd scenario (b), the methylation-decreasing allele (rs74884082, C) at the 14q24 locus is associated with increased expression of the *PSEN1* and *AC004846.1* (lncRNA) but decreased expression of *ACOT4* (shown (descending) in light blue).

3.3.2.4 Colocalized genes inform drug discovery

To evaluate if the genes prioritized by moloc might be candidate targets for drug discovery or be druggable genes, existing literature was reviewed. 18 protein coding genes were assessed for prior biological evidence of a causal link to cardiovascular diseases. Of the 18 genes, two were identified as potential drug targets (*ERBB2* and *PSEN1*), two (*GATAD1* and *YY1AP1*) were identified as containing monogenic mutations and five genes (*CHI3L1, KRIT1, CAMK2G, IGFBP4* and *JMJD1C*) had experimental evidence from animal models (**Table 7**). As these genes have a demonstrated biological role that can be observed at the phenotypic level, they make strong candidates for genes to be taken forward as potential drug targets to treat disease in the early drug discovery process. However, it should be noted that these genes were identified using GWAS and therefore are susceptibility loci, and it is not certain whether this will have a similar role in disease progression.

Two of the 18 genes (*ERBB2* and *PSEN1*) showed evidence for being druggable targets for approved pharmaceutical drugs that have been clinically tested and used commercially. A member of the epidermal growth factor (EGF) receptor family is encoded by *ERBB2* gene. *ERBB2* targeted drugs, such as AFATINIB (inhibitor) and trastuzumab (inhibitor) have been approved and used for non-small cell lung carcinoma and breast cancer respectively (Open Targets annotations). *ERBB2* signaling has been found to play a crucial role in adult heart function according to adverse side effects such as cardiac dysfunction revealed by the anticancer drug (trastuzumab, cyclophosphamide and anthracycline) that targets this gene[269,270]. A mutant mouse model deficient in *ERBB2* showed various dilated cardiomyopathy related physiological phenotypes including reduced contractility, wall thinning, chamber dilation, and cardiac dysfunction (OMIM 164870, Animal Model)[271,272]. Negro et al identified impairment of cardiac contractility in the hearts of *Erbb2* knockout mice[273]. *ERBB2* has been shown to be essential for electrical function of the atria and proliferation of cardiomyocytes during development[274,275,276]. Therapeutic drugs such as SEMAGACESTAT-inhibitor for Alzheimer's disease; TARENFLURBIL-modulator for dementia; RG-4733-inhibitor for non-small cell lung carcinoma, ovarian carcinoma, fallopian tube carcinoma, Different body organs carcinoma and cancers have been also approved to target *PSEN1* gene (Open Targets annotations). *PSEN1* plays an important role in cardiac development and its morphogenesis[277,278] and has been reported to be involved in smooth endoplasmic reticulum calcium ion homeostasis[279]. Missense mutations in *PSEN* genes were identified to cause dilated cardiomyopathy and heart failure (*PSEN1* (Asp333Gly) and *PSEN2* (Ser130Leu)) (OMIM 104311)[277]. Genetically modified mice with *Psen1* knocked out or deleted have reduced muscles fibers, sarcomere lengths in cardiomyocytes and diastolic dysfunction of the heart[280]. In addition, abnormal development of blood vessel and capillary morphology was reported in a mouse model with a defect in the *PSEN1* gene (Open Targets, Mouse Phenotype annotations)[281].

Two genes (*GATAD1* and *YY1AP1*) had monogenic mutations that resulted in cardiovascular disease. *GATAD1* encodes a protein containing a zinc finger domain which controls gene expression. Missense mutations in the *GATAD1* gene cause an autosomal recessive dilated cardiomyopathy (OMIM 614518)[282,283]. An *in vivo* model of *Gatad1* knockout adult zebrafish showed phenotypes similar to HF[284]. Loss-of *YY1AP1* function results in Grange syndrome which affects the blood vessels with internal carotid artery stenosis and hypertension phenotypes (OMIM 607860)[285,286,287]. Moreover, homozygous missense variants in *YYA1P1* leads to Grange syndrome with a later onset cardiovascular disease such as ischemic stroke and hypertension[288]. A novel biallelic mutation in the *YY1AP1* gene leads to an early onset hemorrhagic stroke and/or hypertension with Grange syndrome[289].

Five of the 18 genes (*CHI3L1*, *KRIT1*, *CAMK2G*, *IGFBP4* and *JMJD1C*) were detected with only related CVD phenotypes in mouse models. *CHI3L1* encodes a glycoprotein involved in the biological process of inflammation response and is secreted by cells of the innate immune system. *CHI3L1* originates at the inflammation site, for example in the myocardium in contrast to other inflammatory biomarkers such as HP and C-reactive protein (CRP), which originate in the liver. A study in mice revealed inhibition of *CHI3L1* expression and increased microRNA levels enhanced the influence of dexmedetomidine (DEX) preconditioning to protect myocardial ischemia–reperfusion injury[290]. Furthermore, murine models have shown that *CHI3L1* and miR-24 are two important biomarkers for vascular inflammation and abdominal aortic aneurysm[291]. Studies of the *KRIT1* gene in a mouse model have been reported to result in abnormal heart development, abnormal dorsal aorta morphology, abnormal pericardium morphology; aorta dilation; abnormal blood vessel morphology; aortic endothelial dysfunction[292,293,294,295,296]. *CAMK2G* is part of the Ca(2+) and calmodulin-dependent protein kinase subfamily involved in the regulation of calcium ion transport, skeletal muscle adaptation and muscle contraction[297]. Chelu et al identified increased predisposition to AF caused by enhanced ryanodine receptors (*RyR2s*) phosphorylation by *CAMK2* due to swift atrial pacing in mice with gain-of *RyR2s* function mutation[298]. The increased activity of *CAMK2* was also reported in patients diagnosed with chronic AF[298,299,300]. Moreover, increased AF was detected in diabetic mice as a result of increased oxidized *CAMK2*[301]. *IGFBP4* is part of the insulin-like growth factor binding protein (IGFBP) family contribute to negatively regulation of canonical Wnt signaling pathway[302]. Canonical Wnt signaling is essential for cardiogenesis and both in *vitro* and in *vivo* experiments showed that knockdown of Igfbp4 cause attenuation of cardiomyogenesis[302]. Upregulation of *JMJD1C* has been detected in human and mice with hypertrophic heart[303]. In cardiomyocytes, knockdown of *Jmjd1c* displayed decrease in expression of hypertrophic genes and suppressed the mediation effect of angiotensin II on increasing cardiomyocyte size[303].

Table 7. Known functional evidence for genes prioritized in moloc analysis of AF.

GeneCards (http://www.genecards.org/) was used to look-up the gene name. The Open Targets Genetics database (https://genetics.opentargets.org) was used to look-up information linking the gene to the CVD related phenotype based on experimental work in animal model. The Online Mendelian Inheritance in Man (OMIM) platform (https://www.omim.org/) was used for evidence on CVD caused by defects in the gene and knock down/out in animal model. The publication reporting the evidence for the candidate gene (PMID).

colocalized gene	gene name	gene direction for increased AF risk (my study)	cardio human phenotype (PMID) (OMIM)	cardio animal model
ERBB2	ERB-B2 receptor tyrosine kniase 2	increased expression	no	yes (PMID: 27390088, 11984589) (OMIM 164870)
RBM28	RNA Binding Motif Protein 28	reduced expression	no	no
SYT11	Synaptotagmin 11	reduced expression	no	no
DAP3	Death Associated Protein 3	reduced expression	no	no
YY1AP1	yin yang 1-associated protein 1	reduced expression	yes (PMID: 27939641, 31633303, 31270375) (OMIM 607860)	no
CHI3L1	Chitinase 3 Like 1	reduced expression	no	yes (PMID: 33461162, 25358394)
KLHL12	Kelch Like Family Member 12	increased expression	no	no
GATAD1	GATA Zinc Finger Domain Containing 1	increased expression	yes (PMID: 21965549) (OMIM 614518)	yes (PMID: 28955713)
KRIT1	KRIT1 Ankyrin Repeat Containing	increased expression	no	yes (PMID: 24990152, 20668652, 14993192, 25625206, 31590384)
CAMK2G	Ca2+/calmodulin dependent protein kinase II Gamma	reduced expression	no	yes (PMID: 19603549, 24030498, 33151911, 29903013)
P4HA1	collagen enzyme P4HA1, protein is higher in atrioventricular valves	reduced expression	no	no
MRPS16	Mitochondrial Ribosomal Protein S16	reduced expression	no	no

colocalized gene	gene name	gene direction for increased AF risk (my study)	cardio human phenotype (PMID) (OMIM)	cardio animal model
PSEN1	Presenilin 1	increased expression	yes (PMID: 17186461) (OMIM 104311)	yes (PMID: 28617969, 12834865)
ACOT4	Acyl-CoA Thioesterase 4	reduced expression	no	no
HP	Haptoglobin	increased expression	no	no
IGFBP4	Insulin Like Growth Factor Binding Protein 4	increased expression	no	yes (PMID: 18528331, 24610529, 32597006)
NRBF2	Nuclear Receptor Binding Factor 2	increased expression	no	no
JMJD1C	Jumonji Domain Containing 1C	reduced expression	no	yes (PMID: 32625104)

3.3.3 Sensitivity analyses on moloc probabilities

3.3.3.1 Comparison of main moloc findings against more stringent priors

It has been reported that the colocalization posterior probability is affected by selecting different prior values[304]. Therefore, sensitivity analyses were conducted where main moloc results using the default prior settings, $p1$=1x10^{-4}, $p2$=1x10^{-6}, $p3$=1x10^{-7}, $p4$=1x10^{-8} (see section **2.2.1.1** for definition of prior probabilities) were compared against more stringent prior settings for both phase I ($p1$=1x10^{-5}, $p2$=1x10^{-7}, $p3$=1x10^{-8}) and phase II ($p1$=1x10^{-5}, $p2$=1x10^{-7}, $p3$=1x10^{-8}, $p4$=1x10^{-9}). These analyses showed that evidence of colocalization (PPA>80%) remained for only half of the findings from the main moloc analyses using the more stringent prior thresholds. The number of colocalized findings dropped from 53 to 18 (34%) when $p1$=1x10^{-5}, $p2$=1x10^{-7} and $p3$=1x10^{-8} priors were used in phase I (**Table 5)** and from 30 to 24 (80%) in phase II sensitivity moloc (**Table 6**). This was not only due to a thresholding effect as 35 (40%) of the colocalized results showed large differences of greater than 0.2 in PPA values between the different prior settings. These results suggest that moloc findings at some loci are sensitive to the priors specified and should therefore be treated with caution. However, approximately 50% of all moloc results were robust to stringent priors with 28.6% of this attributed to phase 2 colocalization evidence. An inverse correlation (Pearson r^2=-0.35, 95%CI [-0.564 -0.0850], P=1.08x10^{-2}) was observed between the number of SNPs in the region and the posterior probability (PPA) of the shared causal variant (**Figure 13**). Colocalized combinations with small numbers of SNPs were less sensitive to priors (*p2* or *p3*) than those with large numbers of SNPs. These results demonstrate that having less SNP density in the region will not necessarily decrease the chance of finding colocalization.

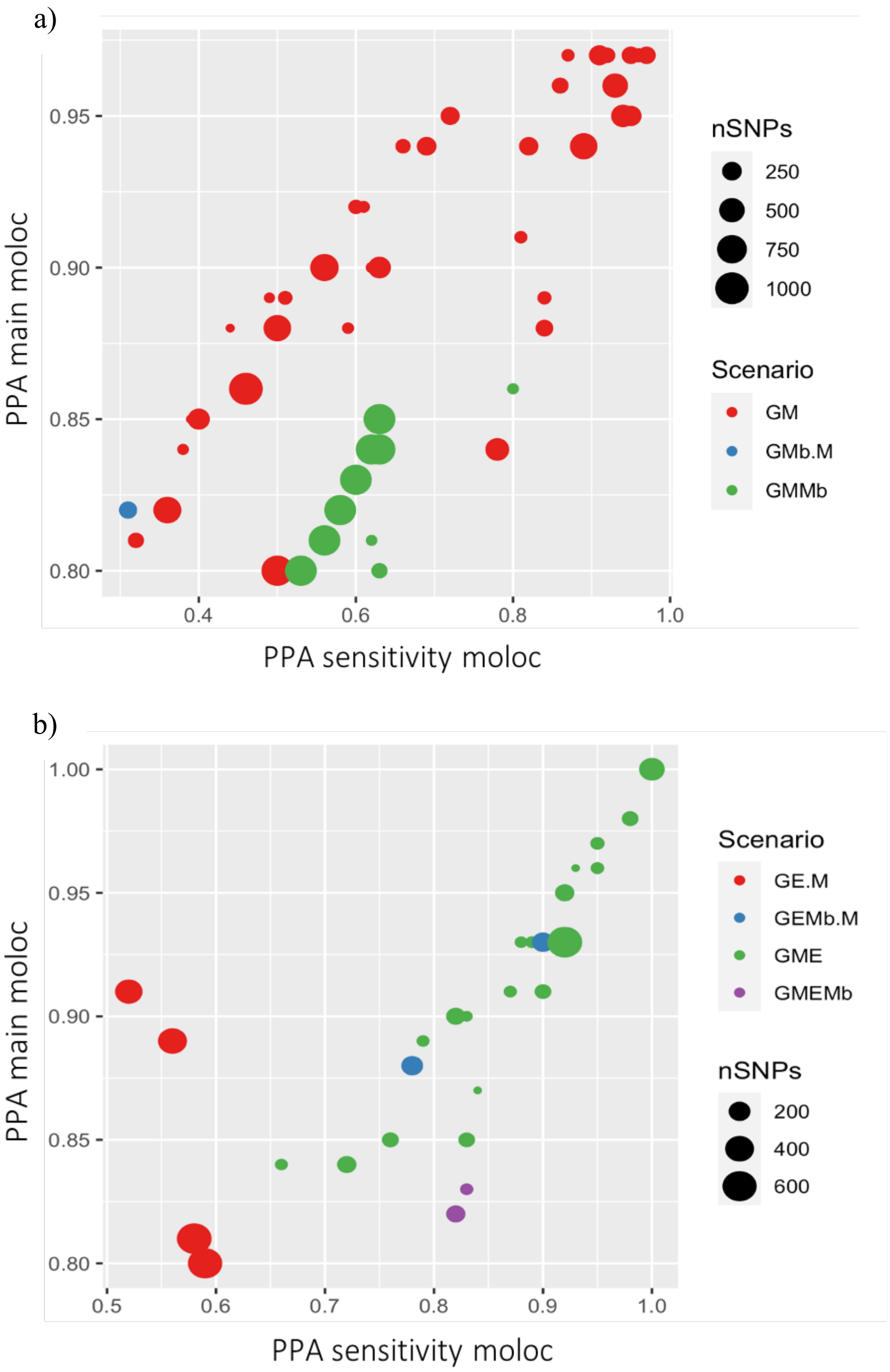

Figure 13. Main moloc results compared to sensitivity moloc results in phase I and II.

a) Comparison including all results from moloc phase I. (b) Same plot but restricting to all results from moloc phase II. Main moloc analysis used p_1=1x10^{-4}, p_2=1x10^{-6} and p_3=1x10^{-7} (in phase I) and p_4=1x10^{-8} (in phase II when expression trait was added) for prior values. Sensitivity moloc analysis used p_1=1x10^{-5}, p_2=1x10^{-7} and p_3=1x10^{-8} (in phase I) and p_4=1x10^{-9} (in phase II when expression trait was added) for prior values. Size of coloured points reflect number of SNPs in a region for different scenarios (i.e., trait combinations).

3.3.3.2 LD pattern of region and PPA

The correlation structure of the genomic region (i.e., local LD pattern) could influence the posterior probability (PPA) of finding a shared causal variant[216]. Therefore, the relationship between the LD score for each AF lead variant and all possible moloc probabilities was tested for the SNP-CpG (PPA.GM), SNP-metabolite (PPA.GMb) and SNP-CpG-metabolite (PPA.GMMb) combinations using phase I colocalization results. Pearson's correlations between LD score of regions versus PPA of trait combinations in phase I of main moloc analysis showed inverse association between these two factors (r^2=-0.44, 95% CI [-0.64 -0.20], P=9.30x10^{-4}). Moreover, testing for correlation coefficient between numbers of SNPs and LD scores of the regions showed positive significant association (Pearson's correlation: r^2=0.73, [95% CI=0.58 0.84], P=4.16x10^{-10}) between these two factors that affect PPA values (**Figure 14**). A total of 24 of the 53 (45%) combinations of traits in the main moloc analysis in phase I had a relatively high LD score of greater than 50. In these regions the moloc probabilities may be biased towards lower probabilities (i.e., increased risk of missing a moloc finding in these regions of high LD). These findings are consistent with the findings of a previous study which examined the impact of LD structure of the regions on colocalization probabilities[304]. This relationship demonstrates that there is more uncertainty in determining sharing of a causal signal in genomic regions with large numbers of highly correlated variants (i.e., when many variants are in high LD).

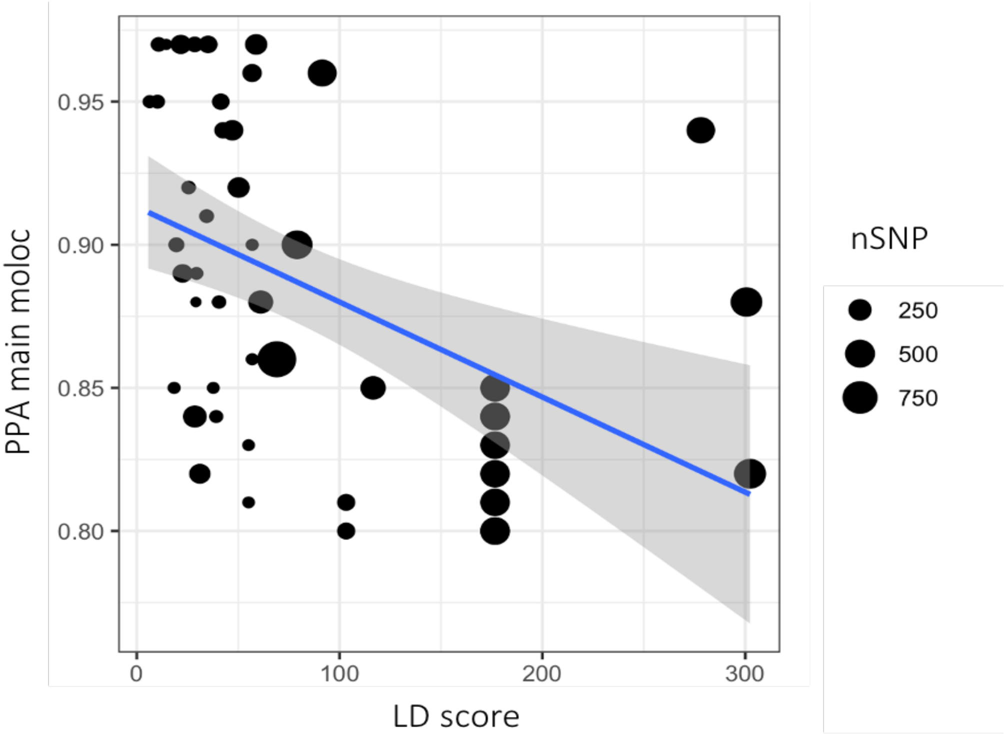

Figure 14. Scatter plot illustrating correlation between posterior probability of association (PPA) of main moloc results and LD score of regions in Phase I.

Size of black points reflect number of SNPs in a region for different scenarios (i.e., GM, GMb.M and GMMb trait combinations) found in Phase I.

3.3.3.3 Zero imputation of missing GoDMC mQTL data

Because of the absence of complete mQTL summary statistics in the GoDMC mQTL data, one region was selected to include missing SNPs and to test the impact of the number of SNPs in the region on the PPA value. This sensitivity moloc analysis was performed on the 17q12 region where 4419 SNPs were in common between AF, cg22833065 CpG site and APOB_APOA1 metabolite datasets. This analysis showed that posterior probability only slightly attenuated to PPA.GMMb=78.0% when the missing SNPs in the region where zero imputed as compared to a PPA of 80.4% where only 149 SNPs were available for moloc analysis. This result is in concordance with the earlier results found in this chapter.

3.4 Discussion

3.4.1 Main findings

In this study, a moloc analysis pipeline was developed and applied to evaluate the shared genetic influences of molecular traits at AF-associated loci by colocalization to elucidate molecular mechanisms in atrial fibrillation. 23 loci with evidence of colocalization between methylation and metabolites were identified. Of the 23 loci, 10 found to be also colocalized with expression of a gene. The gene lookups in two public databases showed that of the 18 genes, *ERBB2* and *PSEN1* are drug targets and *GATAD1* and *YY1AP1* have human monogenic cardiovascular phenotypes, highlighting an important role in disease.

Colocalization revealed evidence of molecular pleiotropy for 8 loci (i.e., region colocalized with multiple molecular phenotypes). For example, for a set of highly correlated metabolites (very small VLDL lipoprotein subclasses) the genetic effects on circulating metabolite levels were found to be shared with the AF lead variant, rs12245149 at the 10q21 locus which also colocalized with methylation of CpG site, cg01631684. It is hypothesized that the colocalization of multiple metabolites with one CpG at this locus may be due to the shared genetic variant co-regulating levels of different circulating metabolites through changes in methylation levels of CpG site. In addition, colocalization of *JMJD1C* (PPA.GEMb.M=93.4%) and *NRBF2* (PPA.GEMb.M=87.7%) genes with DHA and DHA_FA (22:6, docosahexaenoic acid, an omega-3 fatty acid, and its ratio to total fatty acids), was found at this locus. Notably, the rs12245149 variant is in LD (r^2=0.71) with the rs10740118 variant (P=8.1x10^{-9}) near *NRBF2* which has been associated with the n6 fatty acids linoleic acid (LA) in the CHARGE cohorts[305]. *NRBF2* encodes for nuclear receptor binding factor 2 which upon interaction with a transcription factor, PPAR-α (which is highly expressed in the skeletal muscle) upregulate oxidation of fatty acid and lipoprotein lipase activity[306,307]. *JMJD1C* encodes jumonji domain containing 1C, a histone demethylase and has been shown to be involved in demethylation of histones and chromatin modification[308]. Inhibition of *JMJD1C* has been correlated with the attenuation of cardiac hypertrophy[309,303]. Genetic variation, rs10761741 in *JMJD1C* (in LD with the AF shared variant r^2=0.71) has been reported to affect pathways that are involved in platelet reactivity as well as modulation of platelet development[310]. Increased serum testosterone genetic predictors in the *JMJD1C* gene region have been linked to lipids, CAD and IS[311]. MR studies have found a causal relationship between plasma phospholipid arachidonic acid (previously linked to coagulation and inflammation) and IS, ischaemic heart disease (IHD), venous thromboembolism (VTE) and peripheral artery disease[312]. This evidence taken together with the moloc findings, suggests pleiotropic effects of the shared genetic variant on AF through regulation of likely causal genes, *JMJD1C, NRBF2* and fatty acids at this region.

The moloc findings showed shared genetic effects on multiple CpG sites and multiple genes at 1q32, 10q22 and 7q32 loci. Among these regions, pairs of methylation CpG sites were found to have a shared mQTL signal that is enriched for opposite directions of effect (CpG pairs with inverse associations - cg11656175, cg23098069 at 1q32 and cg16228286, cg24637261 pairs at 10q22) and concordant directions of effect (cg18693656, cg13951589 pair at 7q32 with negative associations) in GoDMC. These findings are consistent with previous reports by Bonder et al[128] that have linked multiple CpG sites with each gene and also revealed the existence of both positive and negative correlation between methylation of DNA and gene expression. Gene annotations for these colocalized CpG sites and their genomic locations showed that the colocalized CpG pairs reside comparatively in close relation with each other and were annotated to the same gene at a locus. Evidence of colocalization for GM (phase I) and GME (phase II) scenarios showed that colocalized pairs of CpG sites are more likely to be co-regulated in a region. These results are consistent with findings that the levels of methylation for proximally located CpG sites are often correlated[313,268], equivalent to the correlation structure of the genome (LD). For example, two neighboring CpG sites (cg16228286, cg24637261) located within 3kb of each other were found to be influenced by the shared signal in the 10q22 region. These two CpG sites colocalized with multiple genes (such as *CAMK2G*, *P4HA1*, *MRPS16* (protein coding genes) and *BMS1P4-AGAP5* (lncRNA)), suggesting that a shared variant may co-regulate the expression of *cis*-genes through *cis*-effects on the methylation of adjacent CpG sites. However, the extent to which unique pairs or combinations of intermediate molecular traits implicated by the colocalized association signals, are causal for AF or represent genuine biological mechanisms is further complicated by overlap of multiple molecular traits in each region and by potential horizontal pleiotropy. Taken together, it is likely that many of these regions represent a mixture of vertical and horizontal pleiotropy.

Several sensitivity analyses were performed to evaluate the robustness of the results. First, using more stringent prior thresholds in sensitivity moloc analyses, which although reducing the type I error rate (i.e., false positive finding), is likely to also increase the type II error rate (i.e., false negative finding). Secondly, examining the potential effects of LD score of the region on PPA identified that power to detect the true causal variant is lower when LD score of the lead variant is higher. Third, zero imputation of P values and effects for missing mQTLs slightly attenuated the posterior probability of colocalization for GMMb combination. In fact, zero imputation the region with missing genetic information down-weighted the colocalization probabilities which may be due to this conservative approach assuming missing mQTLs were null, which is unlikely to be true. This result is consistent with the simulation study shown in the GoDMC study where sparse summary data performed comparably to complete summary data with regards to power or FDR[124].

3.4.2 Strength and limitations

A major strength of this study is the systematic application of a moloc analysis pipeline to integrate multiple molecular QTL data with AF which allowed fuller understanding of molecular pleiotropy and master regulators to be identified. Moloc was performed integrating multiple omics data in a single analysis (i.e., either 3 or 4 traits), which captured colocalization of unique molecular phenotypes with shared genetic effects in the *cis* region. While many of the shared variants exert a small effect in risk of AF individually through different molecular phenotypes, targets based on evidence from colocalization are more likely to be therapeutically well grounded compared to those that are not. Moreover, the use of summary statistics from the largest mQTL, eQTL and AF GWAS studies ensured that the moloc analyses were highly powered to detect share regulatory signals between the traits.

Another strength is that moloc requires only information on genetic associations (i.e., summary statistics) and there is no requirement for LD estimates of the datasets when in-population LD is not available, although the statistics (posterior probabilities of combinations) depend on the pattern of association (LD) and the number of correlated SNPs in the region[216].

However, moloc studies in Phase I were of limited genomic coverage and lack of power for the mbQTL dataset since the ALSPAC mbQTL study was limited by small sample size and therefore, failed to investigate the shared genetic regulatory effects between AF and metabolomic traits in regions which otherwise could improve detection of circulating metabolites with potential roles in pathways towards pathogenesis of AF. Moreover, moloc analysess were limited by the lack of full summary datasets for methylation and metabolite traits (i.e., missing of mQTLs and mbQTLs for some CpG sites and metabolites, respectively). Therefore, only mQTLs and mbQTLs at the threshold of $P<1x10^{-5}$ and 10^{-4} were analyzed in moloc. However, restricting molQTLs to genetic variants that are significant ensured that the SNPs robustly associated with DNA methylation and metabolite traits are present in the model, but this approach might be conservative as moloc requires >50 SNPs in the model.

Blood is thought to be a good proxy as a high degree of consistency in overlap between mQTLs and eQTLs and risk variants has been observed across blood and relevant tissues[239,128,314,315]. This study detected shared genetic influences in peripheral blood, which may not be the most suitable tissue in which to look for genetic overlaps between intermediate phenotypes and AF and more conspicuous shared genetic effects may be detected in human cardiac tissue (such as left and right atrial appendages). Moreover, blood has multiple cell types which may all have different methylation levels. Therefore, studies using large-scale QTL mapping of DNA methylation and gene expression in relevant tissues or cell types to identify whether the shared genetic variants are tissue or cell type specific and also infer whether another candidate gene is responsible for a shared regulatory effect, are needed. However, distinguishing between tissue-specific and subject-specific effects of variants is not possible without having summary statistics data calculated from tissue samples of the same individuals which are often not easily accessible.

3.4.3 Future directions

Colocalization of summary statistics of genetic, epigenetic, transcriptomic and metabolomic profiles from multiple studies provides an opportunity to characterize the relationship between hierarchies of molecular phenotype regulation and untangle how these regulatory associations affect a vast array of cardiovascular diseases through shared genetic variation of molecular phenotypes and trait. Detailed understanding of this relationship helps us to correctly interpret the contribution of genetic, epigenetic, transcriptomic and metabolomic variations to AF. The availability of multi-omic statistics from GoDMC, eQTLGen and ALSPAC studies enabled evaluation of statistical evidence supporting a common causal variant at AF risk loci across multiple traits using moloc. Moloc findings suggested that the shared variant might be influencing the risk of AF through changes in DNA methylation, gene expression and circulating metabolites, although horizontal pleiotropy should be considered as another potential interpretation.

Further studies such as multivariable Mendelian randomization (MVMR)[316,317] or two step mediation MR[318] analysis to unravel the independent contributions of each individual molecular phenotype found to be colocalized or co-regulated (e.g., multiple CpG sites, multiple genes) to AF risk should facilitate investigation of causality (mediation) and horizontal pleiotropy.

Extending these analyses beyond DNA methylation, gene expression and metabolite levels to other molecular phenotypes such as variable histone modification and protein levels can greatly intensify the mechanistic and functional interpretation of AF genetic associations. Furthermore, inclusion of pQTL summary statistics as a fifth layer would facilitate discovery of proteins influencing the risk of AF through methylation at a particular CpG site and expression of a specific gene. Access to large-scale molecular QTL datasets is important for application of the moloc approach, and so these other analyses will become possible as other molQTL sample sizes grow in the future and future studies establish better coverage of the genome (e.g., the 450k array captures only 2% of the genome[124].

3.4.4 Conclusions

A comprehensive multiple-trait colocalization analysis was carried out to explore the molecular mechanisms underpinning AF. Combining the largest mQTL dataset available with metabolomic and gene expression data identified evidence of shared genetic etiology with molecular traits at 23 loci, including 20 with at least two molecular traits involved. Whilst this approach does not distinguish horizontal from vertical pleiotropy it does provide more insights into the pathways which are influenced by genetic predisposition to AF, which could be of value in drug target prioritization or in identifying biomarkers for early detection of disease.

Chapter 4 Dissecting the molecular aetiology of stroke

4.1 Introduction

Stroke is the second leading cause of lifelong disability and death among individuals over 60 years worldwide[59,60,61]. IS accounts for approximately 85% of all cases of stroke[63,64] in contrast to ICH which accounts for around 15% of stroke[65]. GWA meta-analyses in European ancestry have identified several risk loci associated with IS subtypes[85,67]. IS has been estimated to have a high heritability (h^2=16-40%)[85] and may be affected by several related cardiovascular traits or risk factors including AF, CAD, blood pressure (systolic and diastolic BP), and HF as well as metabolomic traits such as levels of HDL and LDL cholesterol which showed evidence of overlapping genetic loci with IS[41,67]. Moreover, 4q25 and 16q22 loci identified to be associated with CES by Malik et al are concordant with genetic pathways involved in regulating the cardiac mechanisms associated with AF[97,87]. Among loci identified to be associated with IS, two loci showed associations with more than one IS subtype including 9q34 (*ABO*) locus with shared genetic effects between LAS and CES and 12q24 (*SH2B3*) locus with shared genetic influences between LAS and SVS as analysed by GWAS-pairwise analyses[67].

Variants which influence methylation, gene expression and metabolite levels have been linked to stroke risk in the literature. One recent study has integrated IS GWAS and brain eQTL and mQTL datasets to link functional genetic variants to IS through methylation and gene expression and found IS susceptibility loci overlapped with loci influencing DNA methylation and gene expression[319]. Circulating lipoprotein lipids have previously been identified as risk factors for IS[320]. Moreover, triglyceride-rich lipoproteins have been linked to the risk of IS[321] and CHD[322,323]. MR studies have found causal relationships between circulating lipoprotein lipids and apolipoproteins with risk of IS[324,325,326] and its risk factor CHD[327]. It has been revealed that elevated levels of APOB, triglycerides, and LDL cholesterol increase the risk of IS[324,326,328]. This contrasts with the MR studies that found a relationship between high apolipoprotein A-I (APOAI) and HDL levels and lower risk of IS which indicates that APOAI and HDL levels may have protective roles[324,326,329]. The 9q34 (*ABO*) and 12q24 (*SH2B3*) loci have been shown to be associated with CAD through multiple circulating proteins[330] and have been previously linked to both CAD and IS[331].

This chapter aims to improve the statistical power of detecting evidence of colocalization for different intermediate molecular phenotypes at stroke loci by incorporating a larger blood metabolite QTL resource from the UCLEB consortium[246] meta-analysis (n ~ 30,000) of circulating metabolites (n~230) using the Illumina Cardio-Metabochip platform. Here, moloc was applied to identify shared genetic aetiology between stroke and intermediate molecular phenotypes such as DNA methylation, gene expression and circulating metabolites (including lipid and lipoproteins). By investigating the shared genetic aetiology of multi-omic phenotypes I aim to provide evidence to prioritize potential causal genes, candidate CpG sites and metabolites at susceptibility loci and improve our understanding of the molecular mechanisms implicated in any stroke pathophysiology.

4.2 Methods

4.2.1 Data sources

In this analysis summary statistics from the largest publicly available GWAS meta-analysis for any type of stroke (described in section **2.1.2**) was used. The genetic information at 22 stroke-associated loci (11 loci (1p36, 1p13, 6p21, 9p21, 10q24, 13q14, 19p13 (*SMARCA4-LDLR*), 1q22, 6p22, 6p25, 16q24) for stroke, 8 loci (4q25 (*ANK2*), 4q31, 5q23, 7q21, 19p13 (*ILF3-SLC44A2*), 9q34, 12q24, 11q22) for IS, 2 loci (4q25 (*PITX2*) and 16q22) for CES subtype and 1 locus (7p21) for LAS subtype) identified by this published GWAS were employed for the application of moloc in this chapter.

GoDMC mQTL, UCLEB mbQTL and eQTLGen eQTL data used in this moloc study described in Chapter 2, section **2.1**.

The same study design for stroke was used as AF, where moloc was performed between mQTL, mbQTL in Phase I followed by integrating eQTL in Phase II. (see section **2.2**).

4.2.2 Phase I – moloc analysis of methylation, metabolite and stroke

I extracted all SNPs in common between stroke GWAS, DNA methylation (GoDMC consortium) and metabolite (UCLEB consortium) datasets that were within a 2Mb window centred on each of the 22 stroke GWAS index SNP (see Chapter 2, section **2.2.1** “Define a GWAS region” and “get overlapping SNPs”). A locus was kept for further analysis if methylation QTLs or metabolite QTLs ($P<5x10^{-5}$) were present in the given region and excluded if there was no signal for any CpG site or metabolite in that window. 2 out of 22 loci were not tested for colocalization with multiple traits due to the stroke GWAS hit not being present at the locus for any CpG sites (i.e., no GWAS-mQTL for any CpGs with P<5x10^{-5} at the specific locus). More details of the data preparation process including harmonisation of molecular QTLs with GWAS and moloc analysis Phase I and II can be found in Chapter 2 (**2.2.1**). In this Phase, 11 loci with evidence of colocalization with any of the molecular traits were prioritized for analyses in Phase II.

4.2.3 Phase II – moloc analysis of methylation, gene expression, metabolite and stroke at prioritized loci

Eleven of the 20 regions were prioritized and tested for colocalization with gene expression. At each locus, moloc was applied to detect a potential causal gene(s) and investigate whether gene expression, DNA methylation, metabolite and stroke share a single association signal. Colocalization analysis of SNPs with eQTLs associated with each gene was performed at each locus using summary statistics from all eQTL-gene pairs within a 2Mb window centred on the GWAS hit. The number of genes tested for colocalization at each particular locus varied from 5 to 74 genes.

4.2.4 Sensitivity moloc Phase I and II analyses

Sensitivity analyses for Phase I and Phase II moloc were conducted to compare the main moloc results (using the default prior setting) against sensitivity moloc results where a more stringent prior setting was used in Phase I ($p1$=1x10^{-5}, $p2$=1x10^{-7}, $p3$=1x10^{-8}) and Phase II ($p1$=1x10^{-5}, $p2$=1x10^{-7}, $p3$=1x10^{-8}, $p4$=1x10^{-9}). This is explained in more detail in Chapter 3 (**3.2.3.1**).

4.3 Results

4.3.1 Phase I – results of methylation, metabolite and stroke moloc analysis

Table 8 presents the results of this analysis. Of the 22 reported loci, 20 loci that overlapped with mQTLs and mbQTLs were tested for evidence of colocalization. Eleven of the 20 loci were found with evidence of colocalization, including: the GM scenario (stroke-CpG pair) at 8 loci (10q24, 16q22, 16q24, 1p32, 4q25 (*ANK2*), 4q25 (*PITX2*), 4q31 and 6p21), the GMb.M scenario (stroke-metabolite pair without CpG site) at 3 loci (9q34, 12q24 and 19p13), and the GMMb scenario (stroke-CpG-metabolite combination) at 1 locus (12q24). At most of the 11 loci multiple combinations of molecular traits were found to be colocalized, which led to a total of 127 scenarios; GM=14, GMb.M=111 and GMMb=2 combinations. At the 12q24 locus, two scenarios of interest, GMMb and GMb.M were detected. The rs3184504-cg18714086 mQTL showed evidence of a shared genetic influence with cholesterol esters in very large HDL (XL_HDL_CE) (PPA.GMMb=90.6%), total cholesterol in very large HDL (XL_HDL_C) (PPA.GMMb=85.7%) and stroke. Evidence of colocalization was also found for apolipoprotein A-I (APOAI) and very large HDL (XL_HDL) lipoprotein subclasses, and HDL cholesterol (HDL_C) (PPA.GMb.M>95%) (**Table 8**). Three loci, 4q25 (*PITX2*) for CES subtype, 6p21 and 10q24 for stroke, showed evidence of colocalization with multiple DNA methylation sites. The T and A alleles of the shared variants (rs13143308 and rs16896398) at the 4q25 and 6p21 loci affect multiple CpG sites in the same directions in contrast to the A allele of the shared stroke variant, rs2295786 which has a negative influence on methylation of two CpG sites and a positive effect on methylation of 3 different CpG sites at the 10q24 locus. A complete list of metabolites colocalized in this moloc analysis are shown in **Table 9**.

Table 8. CpG sites and circulating metabolites identified in the moloc analysis Phase I.

Linked vascular traits were related by Malik et al[67] via lookups in published GWASs from EBI GWAS catalog: Systolic blood pressure (SBP), Diastolic blood pressure (DBP), Atrial fibrillation (AF), Venous thromboembolism (VTE), Coronary artery disease (CAD), White-matter hyperintensities on brain MRI (WHB MRI), Low density lipoprotein (LDL), High density lipoprotein (HDL). *SNP location* (the IS risk SNP Location in relation to the nearest gene), *scenario* (combination of traits with evidence of colocalization at the given locus), *moloc main* (posterior probability (PPA) of colocalization between methylation, metabolite and stroke phenotype), *moloc sensitivity* (PPA of colocalization where more stringent priors; p_1=1x10^{-5}, p_2=1x10^{-7}, and p_3=1x10^{-8} was used in the moloc analysis). Each row in the locus column represents a single stroke hit locus, with sub-rows representing the different colocalized molecular traits.

locus	linked vascular trait(s)	risk SNP	SNP location	scenario	CpG site	metabolite	main moloc	sensitivity moloc
1p36	SBP, DBP	rs880315	Intronic	GM	cg05396182		96.6	82.3
4q25	AF	rs13143308	Intergenic	GM	cg03587884		94.8	80.0
					cg06126494		85.6	40.2
4q25		rs34311906	Intergenic	GM	cg11289139		85.2	45.2
4q31	VTE	rs6825454	Intergenic	GM	cg22130008		88.2	47.9
6p21	SBP, DBP	rs16896398	Intergenic	GM	cg13033722		83.8	39.4
					cg00084398		88.3	45.5
9q34	CAD, VTE, LDL levels	rs635634	Intergenic	GMb.M		APOB	97.6	80.2
						ESTC	94.1	61.4
						FAW6	93.2	57.7
						GLY	89.9	47.2
						IDL_CE	95.0	65.7
						IDL_CE_PC	95.6	68.4
						IDL_C	97.3	80.0
						IDL_FC	96.8	75.5
						IDL_C_PC	95.2	66.6
						IDL_L	97.7	80.9
						IDL_PL	96.5	81.7
						IDL_P	97.8	81.7
						IDL_TG	93.7	60.0
						IDL_TG_PC	92.2	54.3
						LA_FA	91.6	52.3
						LA	93.2	57.8
						LDL_C	97.3	80.3
						L_LDL_CE_PC	95.7	69.1
						L_LDL_C	97.4	77.0
						L_LDL_CE	97.1	77.0
						L_LDL_C_PC	95.3	66.9
						L_LDL_FC	97.5	79.7
						L_LDL_L	97.4	79.1
						L_LDL_PL	97.8	81.3
						L_LDL_P	97.6	80.3
						L_LDL_TG_PC	93.1	57.5
						L_VLDL_CE_PC	84.6	35.5
						L_VLDL_C_PC	94.3	62.4
						L_VLDL_TG_PC	92.2	54.2
						M_LDL_CE	97.1	76.8
						M_LDL_CE_PC	86.4	38.9
						M_LDL_C	97.4	79.1
						M_LDL_C_PC	94.7	63.9
						M_LDL_FC	94.6	63.8
						M_LDL_L	97.2	77.6
						M_LDL_PL	97.4	79.1
						M_LDL_P	97.4	79.2

locus	linked vascular trait(s)	risk SNP	SNP location	scenario	CpG site	metabolite	main moloc	sensitivity moloc
9q34	CAD, VTE, LDL levels	rs635634	Intergenic	GMb.M		M_LDL_TG_PC	91.3	51.4
						M_VLDL_CE_PC	94.5	63.0
						M_VLDL_C_PC	94.9	65.1
						M_VLDL_TG_PC	93.7	59.8
						S_LDL_CE	94.9	65.0
						S_LDL_CE_PC	82.2	31.6
						S_LDL_C	97.3	78.2
						S_LDL_C_PC	95.5	67.8
						S_LDL_L	97.2	77.8
						S_LDL_PL	94.6	63.6
						S_LDL_P	97.2	77.8
						S_LDL_PL_PC	83.2	33.2
						S_VLDL_CE	92.9	56.6
						S_VLDL_CE_PC	95.2	66.3
						S_VLDL_C	95.2	66.5
						S_VLDL_C_PC	94.8	64.8
						S_VLDL_FC	96.7	74.3
						S_VLDL_L	93.8	60.2
						S_VLDL_PL	96.8	75.1
						S_VLDL_TG_PC	94.1	61.6
						VLDL_C	89.7	46.6
						XS_VLDL_CE	91.7	52.4
						XS_VLDL_C	91.9	53.0
						XS_VLDL_C_PC	92.4	55.1
						XS_VLDL_FC	92.1	53.9
						XS_VLDL_L	97.7	80.6
						XS_VLDL_PL	97.8	81.9
						XS_VLDL_P	97.7	80.9
						XS_VLDL_TG_PC	91.1	50.6
						SERUM_C	95.4	67.5
						REMNANT_C	95.6	68.3
10q24	WHB MRI	rs2295786	Intergenic	GM	cg25866173		83.8	35.4
					cg25181684		83.3	34.3
					cg24911198		84.9	37
					cg11819799		84.7	37.9
					cg07671776		84.6	36.8
12q24	CAD, DBP, SBP, HDL levels	rs3184504	Exonic; nonsynonymous	GMMb	cg18714086	XL_HDL_CE	90.6	87.9
						XL_HDL_C	85.7	82.1
				GMb.M		APOA1	97.2	77.7
						HDL_C	95.7	69.1
						XL_HDL_CE	98.7	88.2
						XL_HDL_C	98.8	89.2
						XL_HDL_FC	98.9	90.4
						XL_HDL_L	98.1	83.7
						XL_HDL_PL	83.0	32.9
						XL_HDL_P	97.7	81.0
16q24		rs12445022	Intergenic	GM	cg04245248		97.4	94.3
16q22	AF	rs12932445	Intronic	GM	cg03463523		91.2	53.8
19p13	CAD, LDL levels	rs8103309	Intergenic	GMb.M		APOB	98.6	87.6
						IDL_CE	95.2	87.6
						IDL_C	98.6	87.6
						IDL_FC	98.6	87.6
						IDL_L	98.6	87.6
						IDL_PL	98.6	87.6
						IDL_P	98.6	87.6
						IDL_TG	98.6	87.4
						LDL_C	98.6	87.6
						L_LDL_CE	98.6	87.6
						L_LDL_C	98.6	87.6

locus	linked vascular trait(s)	risk SNP	SNP location	scenario	CpG site	metabolite	main moloc	sensitivity moloc
19p13	CAD, LDL levels	rs8103309	Intergenic	GMb.M		L_LDL_FC	98.6	87.6
						L_LDL_L	98.6	87.6
						L_LDL_PL	98.6	87.5
						L_LDL_P	98.6	87.6
						M_LDL_CE	98.6	87.6
						M_LDL_C	98.6	87.6
						M_LDL_L	98.6	87.6
						M_LDL_PL	98.6	87.5
						M_LDL_P	98.6	87.6
						REMNANT_C	87.6	41.4
						SERUM_C	98.6	87.6
						S_LDL_C	98.6	87.6
						S_LDL_L	98.6	87.6
						S_LDL_P	98.6	87.6
						S_VLDL_C	98.6	87.6
						S_VLDL_FC	98.6	87.4
						S_VLDL_L	97.6	80.5
						S_VLDL_PL	85.0	36.2
						S_VLDL_P	94.0	61.2
						XS_VLDL_C	84.4	35.2
						XS_VLDL_L	98.6	87.6
						XS_VLDL_PL	98.6	87.6
						XS_VLDL_P	98.6	87.6
						XS_VLDL_TG	98.1	83.8

Table 9. Metabolites categories and names identified in the moloc analyses.

Intermediate density lipoprotein (IDL), High density lipoprotein (HDL), Low density lipoprotein (LDL), Very low-density lipoprotein (VLDL).

category	metabolite	name/subclass
Apolipoproteins	APOB	Apolipoprotein B
	APOA1	Apolipoprotein A-I
IDL	IDL_CE	Cholesterol esters
	IDL_CE_PC	Cholesterol esters to total lipids ratio
	IDL_C	Total cholesterol
	IDL_C_PC	Total cholesterol to total lipids ratio
	IDL_FC	Free cholesterol
	IDL_L	Total lipids
	IDL_PL	Phospholipids
	IDL_P	Particles
	IDL_TG	Triglycerides
	IDL_TG_PC	Triglycerides to total lipids ratio
Very large HDL	XL_HDL_CE	Cholesterol esters
	XL_HDL_C	Total cholesterol
	XL_HDL_FC	Free cholesterol
	XL_HDL_L	Total lipids
	XL_HDL_PL	Phospholipids
	XL_HDL_P	Particles
Large LDL	L_LDL_CE	Cholesterol esters
	L_LDL_CE_PC	Cholesterol esters to total lipids ratio
	L_LDL_C	Total cholesterol
	L_LDL_C_PC	Total cholesterol to total lipids ratio
	L_LDL_FC	Free cholesterol
	L_LDL_L	Total lipids
	L_LDL_PL	Phospholipids
	L_LDL_P	Particles
	L_LDL_TG_PC	Triglycerides to total lipids ratio
Medium LDL	M_LDL_CE	Cholesterol esters
	M_LDL_CE_PC	Cholesterol esters to total lipids ratio
	M_LDL_C	Total cholesterol
	M_LDL_C_PC	Total cholesterol to total lipids ratio
	M_LDL_FC	Free cholesterol
	M_LDL_L	Total lipids
	M_LDL_PL	Phospholipids
	M_LDL_P	Particles
	M_LDL_TG_PC	Triglycerides to total lipids ratio
	M_VLDL_CE_PC	Cholesterol esters to total lipids ratio
	M_VLDL_C_PC	Total cholesterol to total lipids ratio
	M_VLDL_TG_PC	Triglycerides to total lipids ratio

category	metabolite	name/subclass
Small LDL	S_LDL_CE	Cholesterol esters
	S_LDL_CE_PC	Cholesterol esters to total lipids ratio
	S_LDL_C	Total cholesterol
	S_LDL_C_PC	Total cholesterol to total lipids ratio
	S_LDL_L	Total lipids
	S_LDL_PL	Phospholipids
	S_LDL_P	Particles
	S_LDL_PL_PC	Phospholipids to total lipids ratio
Large VLDL	L_VLDL_CE_PC	Cholesterol esters to total lipids ratio
	L_VLDL_C_PC	Total cholesterol to total lipids ratio
	L_VLDL_TG_PC	Triglycerides to total lipids ratio
Small VLDL	S_VLDL_CE	Cholesterol esters
	S_VLDL_CE_PC	Cholesterol esters to total lipids ratio
	S_VLDL_C	Total cholesterol
	S_VLDL_C_PC	Total cholesterol to total lipids ratio
	S_VLDL_FC	Free cholesterol
	S_VLDL_L	Total lipids
	S_VLDL_PL	Phospholipids
	S_VLDL_P	Particles
	S_VLDL_TG_PC	Triglycerides to total lipids ratio
Very small VLDL	XS_VLDL_CE	Cholesterol esters
	XS_VLDL_C	Total cholesterol
	XS_VLDL_C_PC	Total cholesterol to total lipids ratio
	XS_VLDL_FC	Free cholesterol
	XS_VLDL_L	Total lipids
	XS_VLDL_PL	Phospholipids
	XS_VLDL_P	Particles
	XS_VLDL_TG	Triglycerides
	XS_VLDL_TG_PC	Triglycerides to total lipids ratio
Cholesterol	LDL_C	LDL cholesterol
	HDL_C	HDL cholesterol
	VLDL_C	VLDL cholesterol
	ESTC	Esterified cholesterol
	SERUM_C	Serum total cholesterol
	REMNANT_C	Remnant cholesterol (non-HDL, -LDL -cholesterol)
Fatty acids & saturation	FAW6	Omega-6 fatty acids
	LA	Linoleic acid
	LA_FA	Linoleic acid to total fatty acids ratio
Amino acids	GLY	Glycine

4.3.2 Phase II – Mapping Phase I findings to the potential causal gene(s) for stroke

To identify the potential causal gene at loci from Phase I that have evidence of shared genetic effects between stroke, CpG site and/or metabolite, moloc Phase II was conducted as previously described (**2.2.1.2**). Of the 11 colocalized loci detected in Phase I, five (6p21 (stroke), 9q34 (IS), 12q24 (IS), 16q22 (CES), 19p13 (stroke)) were found with evidence of colocalization (PPA ≥ 80%) with gene expression (**Table 10**). In total, 215 scenarios suggesting a single shared association signal for a combination of traits at a locus were identified. These colocalizations correspond to the following combinations: 5 unique SNP-CpG-gene (GME) colocalizations at the 6p21 and 16q22 loci, 204 unique SNP-gene-metabolite (without CpG site) (GEMb.M) colocalizations at 9q34, 12q24 and 19p13 loci, 6 unique SNP-CpG-gene-metabolite (GMEMb) colocalizations at 12q24 locus (**Table 10**).

The greatest number of colocalized phenotypes was detected at the 9q34 locus (GEMb.M scenario (n=122 phenotypes)), suggesting broad pleiotropic effects (marked by SNP rs635634) on many molecular traits. Of the 36 genes tested for colocalization at the 9q34 locus, *ABO, CACFD1* and *GBGT1* genes showed evidence of colocalization with multiple clusters of circulating metabolites and stroke in blood tissue. *ABO* and *CACFD1* genes had the highest posterior probability (PPA>95%) for colocalization with stroke and multiple circulating lipoprotein lipids and apolipoprotein B (APOB) compared to *GBGT1* gene with PPA ≥ 80%. The stroke risk variant rs635634, also an eQTL signal associated with expression of *ABO* (the closest gene), was found to be colocalized with fatty acids and saturation (e.g., omega-6 fatty acids (FAW6), linoleic acid to total fatty acids ratio (LAFA)), triglycerides in IDL lipoprotein, VLDL cholesterol, Medium/small LDL lipoprotein subclasses, small VLDL lipoprotein subclasses. *GBGT1* an important paralog of the *ABO* gene was also detected to colocalize with apolipoprotein B (APOB), LDL cholesterol, IDL lipoprotein subclasses, large to small LDL lipoprotein subclasses, small and very small VLDL, non-HDL, non-LDL -cholesterol (remnant cholesterol), serum total cholesterol and stroke. *CACFD1* was also identified to share a single variant with apolipoprotein B (APOB), IDL lipoprotein subclasses, large to small LDL lipoprotein subclasses, LDL cholesterol, VLDL cholesterol, fatty acids and saturation (e.g., omega-6 fatty acids (FAW6) and linoleic acid (LA), small to very small VLDL lipoprotein subclasses, large VLDL lipoprotein subclasses ratios, non-HDL, non-LDL -cholesterol (remnant cholesterol), serum total cholesterol and stroke (**Table 10**).

Colocalization of stroke, methylation and expression (GME scenario) were identified at two loci, 6p21 and 16q22. At the 6p21 locus, evidence of colocalization was found between multiple genes, including two protein coding genes (*SRF* and *ABCC10*), 1 pseudogene (*RPL34P14*), cg13033722 CpG site and stroke. At the CES-associated locus, 16q22, the T allele of rs12932445 (ß=-0.054, P=4.99x10^{-6}) has an opposite direction of effect on methylation of cg03463523 CpG site (ß=0.149, P=6.95x10^{-35}) in GoDMC and expression of *HP* gene (ß=-0.084, P=3.20x10^{-8}) in eQTLGen whole blood. At the 12q24 locus, colocalization evidence was identified for multiple combinations of traits. Six GMEMb scenarios indicating a shared association signal, rs3184504 between one single CpG site (cg18714086), three protein coding gene (*TCTN1*, *HVCN1*, and *GPN3*), and two subclasses of very large HDL (XL HDL) lipoprotein (total cholesterol and cholesterol esters) was detected. *TCTN1* showed the strongest probability of colocalization with both XL HDL lipoprotein subtypes (PPA.GMEMb=99.9%). Evidence of colocalization was also found for a protein coding gene, *TRAFD1* and a long intergenic non-coding RNA (lincRNA) transcript, *RP3-473L9.4* and various XL HDL lipoprotein subclasses with PPA.GEMb.M of 99.7% and 88% respectively. At the 19p13 risk locus, of the 74 genes tested for colocalization with APOB, different circulating IDL/LDL/VLDL lipoprotein subclasses, LDL cholesterol, remnant cholesterol and total cholesterol, only two neighbouring genes, *SMARCA4* and *C19orf52* (also named *TIMM29*) were colocalized (PPA.GEMb.M>99%) (**Table 10**). The T allele of rs8103309 risk variant at this locus is strongly associated with increased expression of *SMARCA4* (ß=0.202, P=1.48x10^{-58}) and *C19orf52* gene (ß=0.092, P=2.64x10^{-13}). In addition, the T allele of this variant is associated with increased circulating concentrations of all highly correlated metabolites colocalized in this region (ß=0.052).

Table 10. Genes, CpG sites and metabolites identified in the phase II of moloc analysis of stroke.

Locus (locus found with evidence of colocalization (stroke phenotype involved in each colocalization))*, Scenario* (combination of traits with evidence of colocalization/a shared signal at the given locus), *main moloc* (posterior probability (PPA) of colocalization), *sensitivity moloc* (PPA of colocalization where more stringent priors; p_1=1x10^{-5}, p_2=1x10^{-7}, p_3=1x10^{-8}, and p_4=1x10^{-9} was used in the moloc analysis). Each row in the locus column represents a single stroke hit locus, with sub-rows representing the different colocalized molecular traits.

locus	risk SNP	scenario	CpG site	metabolite	gene	main moloc	sensitivity moloc
6p21 (Stroke)	rs16896398	GME	cg13033722		*SRF*	98.5	98.4
					ABCC10	97.6	96.6
					RPL34P14	91.2	84.1
			cg00084398		*AL035587.1*	89.0	78.6
9q34 (IS)	rs635634	GEMb.M		APOB	*GBGT1*	83.1	80.4
				IDL_CE_PC		80.1	75.9
				IDL_C		83.9	81.0
				IDL_FC		85.1	81.9
				IDL_L		83.8	81.3
				IDL_PL		84.3	80.8
				IDL_P		83.8	81.3
				IDL_TG		82.7	76.1
				LDL_C		83.6	80.7
				L_LDL_CE_PC		80.1	77.3
				L_LDL_CE		83.6	80.5
				L_LDL_FC		83.8	81.1
				L_LDL_L		83.6	80.8
				L_LDL_PL		83.6	81.1
				L_LDL_P		83.6	81.0
				M_LDL_CE		83.4	80.3
				M_LDL_C		83.6	80.8
				M_LDL_L		83.4	80.4
				M_LDL_PL		83.4	80.6
				M_LDL_P		83.4	80.6
				M_VLDL_CE_PC		80.2	69.4
				REMNANT_C		80.0	80.2
				SERUM_C		84.4	77.2
				S_LDL_C_PC		80.2	77.2
				S_LDL_C		83.5	80.5
				S_LDL_L		83.2	80.2
				S_LDL_P		83.2	80.2
				S_VLDL_C		84.2	79.8
				S_VLDL_FC		83.4	80.0
				S_VLDL_L		82.1	75.9
				S_VLDL_PL		83.3	80.0
				XS_VLDL_L		83.9	81.3
				XS_VLDL_PL		83.9	81.5
				XS_VLDL_P		83.8	81.3
9q34 (IS)	rs635634	GEMb.M		APOB	*CACFD1*	99.0	98.9
				FAW6		98.1	97.9
				IDL_CE_PC		98.7	98.5
				IDL_CE		98.5	98.3
				IDL_C_PC		98.5	98.4
				IDL_C		99.0	98.9
				IDL_FC		98.9	98.8
				IDL_L		99.1	99.0
				IDL_PL		98.8	98.7

locus	risk SNP	scenario	CpG site	metabolite	gene	main moloc	sensitivity moloc
9q34 (IS)	rs635634	GEMb.M		IDL_P	*CACFD1*	99.1	99.0
				IDL_TG_PC		98.2	97.9
				IDL_TG		98.6	98.4
				LA_FA		98.1	97.8
				LA		98.1	97.8
				LDL_C		99.0	98.9
				L_LDL_CE_PC		98.7	98.5
				L_LDL_CE		98.9	98.8
				L_LDL_C_PC		98.6	98.4
				L_LDL_FC		99.0	98.9
				L_LDL_L		99.0	98.9
				L_LDL_PL		99.1	99.0
				L_LDL_P		99.0	98.9
				L_LDL_TG_PC		98.4	98.1
				L_VLDL_CE_PC		97.2	96.5
				L_VLDL_C_PC		98.4	98.2
				L_VLDL_TG_PC		98.2	97.9
				M_LDL_CE_PC		97.6	97.0
				M_LDL_CE		98.9	98.8
				M_LDL_C_PC		98.6	98.3
				M_LDL_C		99.0	98.9
				M_LDL_FC		98.4	98.2
				M_LDL_L		98.9	98.8
				M_LDL_PL		99.0	98.9
				M_LDL_P		99.0	98.9
				M_LDL_TG_PC		98.2	97.8
				M_VLDL_CE_PC		98.4	98.2
				M_VLDL_C_PC		98.5	98.3
				M_VLDL_TG_PC		98.2	98.0
				REMNANT_C		98.6	98.4
				SERUM_C		98.5	98.3
				S_LDL_CE_PC		97.1	96.3
				S_LDL_CE		98.6	98.4
				S_LDL_C_PC		98.6	98.5
				S_LDL_C		99.0	98.9
				S_LDL_L		98.9	98.8
				S_LDL_PL_PC		97.2	96.4
				S_LDL_PL		98.5	98.3
				S_LDL_P		98.9	98.8
				S_VLDL_CE_PC		98.5	98.3
				S_VLDL_CE		98.0	97.7
				S_VLDL_C_PC		98.5	98.2
				S_VLDL_C		98.5	98.3
				S_VLDL_FC		98.8	98.7
				S_VLDL_L		98.6	98.3
				S_VLDL_PL		98.9	98.8
				S_VLDL_TG_PC		98.3	98.1
				VLDL_C		98.0	97.5
				XS_VLDL_CE		97.8	97.4
				XS_VLDL_C_PC		98.1	97.8
				XS_VLDL_C		97.8	97.4
				XS_VLDL_FC		97.9	97.5
				XS_VLDL_L		99.1	99.0
				XS_VLDL_PL		99.1	99.0
				XS_VLDL_P		99.1	99.0
				XS_VLDL_TG_PC		98.0	97.7

locus	risk SNP	scenario	CpG site	metabolite	gene	main moloc	sensitivity moloc
9q34 (IS)	rs635634	GEMb.M		FAW6	*ABO*	97.7	97
				IDL_CE_PC		97.7	97.2
				IDL_TG_PC		97.4	96.4
				IDL_TG		98.8	98.6
				LA_FA		97.5	96.5
				L_LDL_TG_PC		97.5	96.5
				L_VLDL_CE_PC		98.4	98.1
				L_VLDL_C_PC		97.8	97.3
				L_VLDL_TG_PC		97.5	96.6
				M_LDL_CE_PC		97.9	97.4
				M_LDL_C_PC		97.6	97
				M_LDL_TG_PC		98.8	98.6
				S_LDL_CE_PC		97.6	96.9
				S_LDL_CE		97.7	97.2
				S_LDL_C_PC		97.7	97.2
				S_LDL_PL_PC		97.7	97.1
				S_LDL_PL		97.7	97.1
				S_VLDL_FC		93.6	88.5
				S_VLDL_L		98.8	98.6
				S_VLDL_PL		94.8	90.8
				VLDL_C		98.6	98.3
				XS_VLDL_C_PC		97.6	96.8
				XS_VLDL_TG_PC		97.4	96.2
12q24 (IS)	rs3184504	GMEMb	cg18714086	XL_HDL_CE	*TCTN1*	99.9	99.9
					GPN3	81.6	81.9
					HVCN1	89.8	90.5
				XL_HDL_C	*TCTN1*	99.9	99.9
					GPN3	85.9	86.3
					HVCN1	84.4	84.7
		GEMb.M		XL_HDL_CE	*TRAFD1*	99.7	99.7
					RP3-473L9.4	88.2	89.9
				XL_HDL_C	*TRAFD1*	99.7	99.7
					RP3-473L9.4	88.2	90.0
				XL_HDL_FC	*TRAFD1*	99.7	99.7
					RP3-473L9.4	88.2	90.1
				XL_HDL_L	*TRAFD1*	99.7	99.7
					RP3-473L9.4	88.2	89.4
				XL_HDL_PL	*TRAFD1*	97.8	97.6
					RP3-473L9.4	84.9	75.9
				XL_HDL_P	*TRAFD1*	99.7	99.7
					RP3-473L9.4	88.2	89.0
16q22 (CES)	rs12932445	GME	cg03463523		*HP*	81.1	60.1
19p13 (Stroke)	rs8103309	GEMb.M		APOB	*SMARCA4*	99.7	99.6
				IDL_CE		99.4	99.4
				IDL_C		99.7	99.6
				IDL_FC		99.7	99.6
				IDL_L		99.7	99.6
				IDL_PL		99.7	99.6
				IDL_P		99.7	99.6
				IDL_TG		99.7	99.6
				LDL_C		99.7	99.6
				L_LDL_CE		99.7	99.6
				L_LDL_C		99.7	99.6
				L_LDL_FC		99.7	99.6
				L_LDL_L		99.7	99.6
				L_LDL_PL		99.7	99.6

locus	risk SNP	scenario	CpG site	metabolite	gene	main moloc	sensitivity moloc
19p13 (Stroke)	rs8103309	GEMb.M		L_LDL_P	*SMARCA4*	99.7	99.6
				M_LDL_CE		99.7	99.6
				M_LDL_C		99.7	99.6
				M_LDL_L		99.7	99.6
				M_LDL_PL		99.7	99.6
				M_LDL_P		99.7	99.6
				REMNANT_C		98.8	98.6
				SERUM_C		99.7	99.6
				S_LDL_C		99.7	99.6
				S_LDL_L		99.7	99.6
				S_LDL_P		99.7	99.6
				S_VLDL_C		99.7	99.6
				S_VLDL_FC		99.7	99.6
				S_VLDL_L		99.6	99.6
				S_VLDL_PL		98.5	98.4
				S_VLDL_P		99.3	99.3
				XS_VLDL_C		98.5	98.3
				XS_VLDL_L		99.7	99.6
				XS_VLDL_PL		99.7	99.6
				XS_VLDL_P		99.7	99.6
				XS_VLDL_TG		99.6	99.6
19p13 (Stroke)	rs8103309	GEMb.M		APOB	*C19orf52*	99.7	99.6
				IDL_CE		99.4	99.4
				IDL_C		99.7	99.6
				IDL_FC		99.7	99.6
				IDL_L		99.7	99.6
				IDL_PL		99.7	99.6
				IDL_P		99.7	99.6
				IDL_TG		99.7	99.6
				LDL_C		99.7	99.6
				L_LDL_CE		99.7	99.6
				L_LDL_C		99.7	99.6
				L_LDL_FC		99.7	99.6
				L_LDL_L		99.7	99.6
				L_LDL_PL		99.7	99.6
				L_LDL_P		99.7	99.6
				M_LDL_CE		99.7	99.6
				M_LDL_C		99.7	99.6
				M_LDL_L		99.7	99.6
				M_LDL_PL		99.7	99.6
				M_LDL_P		99.7	99.6
				REMNANT_C		98.8	98.6
				SERUM_C		99.7	99.6
				S_LDL_C		99.7	99.6
				S_LDL_L		99.7	99.6
				S_LDL_P		99.7	99.6
				S_VLDL_C		99.7	99.6
				S_VLDL_FC		99.7	99.6
				S_VLDL_L		99.6	99.6
				S_VLDL_PL		98.5	98.4
				S_VLDL_P		99.3	99.3
				XS_VLDL_C		98.5	98.3
				XS_VLDL_L		99.7	99.6
				XS_VLDL_PL		99.7	99.6
				XS_VLDL_P		99.7	99.6
				XS_VLDL_TG		99.6	99.6

4.3.3 Sensitivity moloc analyses

To investigate whether the main moloc findings at Phase I and Phase II would be sensitive to more stringent setting of Bayesian priors, sensitivity moloc analyses were performed for both phases. Detailed information, including prior selection, can be found in Chapter 3 (**3.2.3.1**). Sensitivity moloc analysis showed that evidence of colocalization (PPA ≥ 80%) remained for 52 (41%) of 127 scenarios (i.e., GM, GMb.M, GMMb) which were identified in the main moloc analysis Phase I (**Table 8**). Among 75 scenarios which showed attenuated colocalization probability and failed to pass the threshold of 0.8, 16 had moderate colocalization evidence (PPA ≥ 70%). These results suggest that moloc findings at some loci are sensitive to the priors specified and should therefore be treated with caution. In the sensitivity analysis for phase II, 204 (95%) of 215 scenarios (i.e., GEMb.M, GME, GMEMb) found in Phase II of the main moloc analysis showed evidence of colocalization at or greater than 80% PPA (**Table 10**). These results suggest that moloc findings where a third, or at some loci a fourth, trait with high statistical power (such as gene expression from eQTLGen data) was colocalized, were not sensitive to the stringent priors and were robust to different priors. Comparison of the results of the main moloc analysis with sensitivity moloc analysis for Phase I and II is shown in **Figure 15.** Large differences between PPA of these results occurred for GM (4q25, 4q31, 6p21, and 10q24 loci), with attenuation of PPA (50-79%) for GMb.M (9q34) and GM (16q22) scenarios (**Figure 15a**). Apolipoprotein B and LDL cholesterol retained a robust colocalization evidence (GMb.M.PPA>80%) at 9q34 locus (**Table 8**).

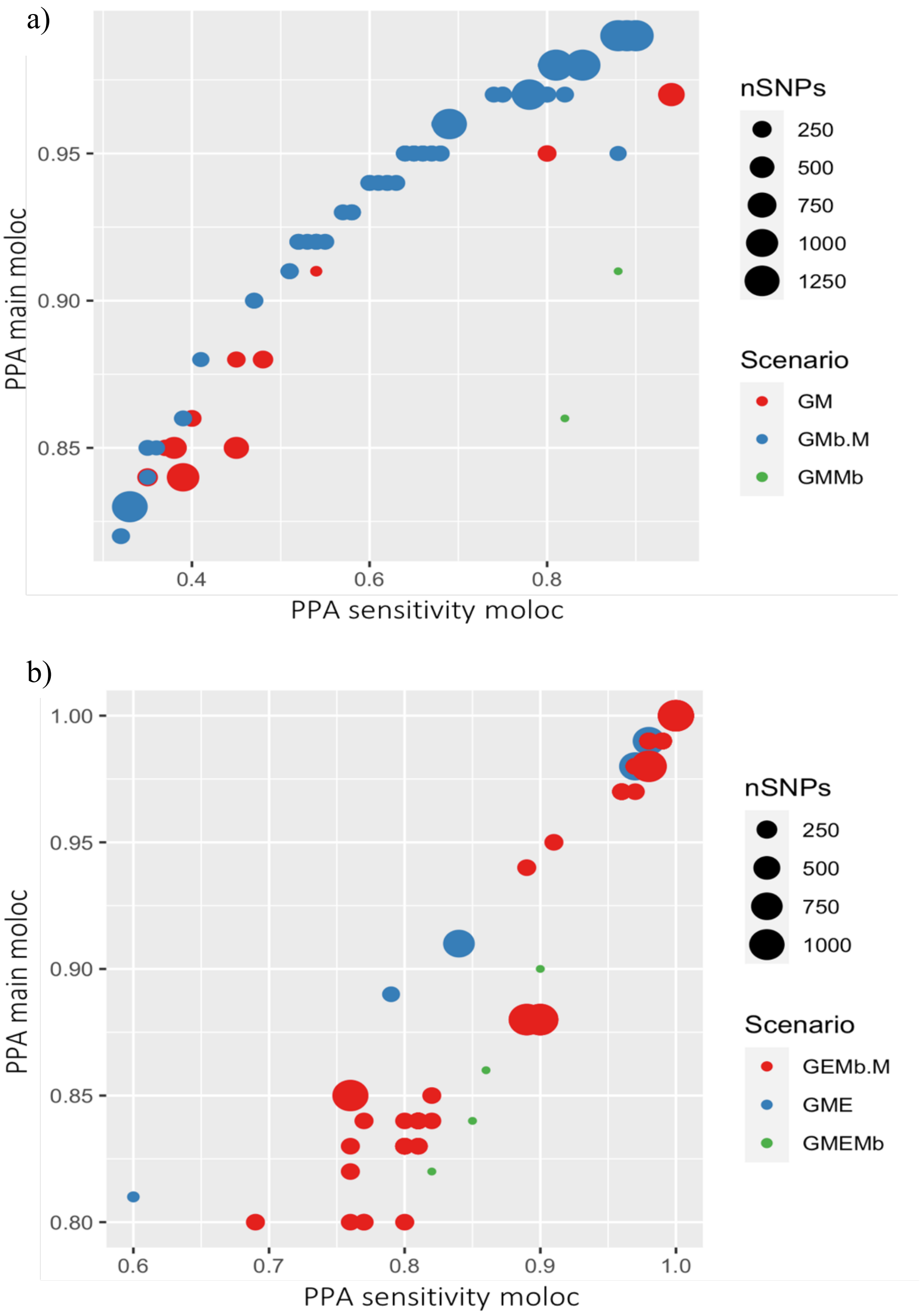

Figure 15. Main Moloc results compared to sensitivity moloc results (imposing more stringent priors) in phase I and II.

(a) Comparison including all results from moloc phase I. (b) Same plot as (a) but restricted to all results from moloc phase II. Main moloc analysis used p_1=1x10^{-4}, p_2=1x10^{-6} and p_3=1x10^{-7} (in phase I) and p_4=1x10^{-8} (in phase II when additional trait was added) for prior values. Sensitivity moloc analysis used p_1=1x10^{-5}, p_2=1x10^{-7} and p_3=1x10^{-8} (in phase I) and p_4=1x10^{-9} (in phase II when additional trait was added) for prior values. Size of coloured points reflect number of SNPs in a region for different scenarios (i.e., trait combinations).

4.3.4 Colocalized genes inform drug discovery

Genes colocalized with stroke were assessed to evaluate potential drug targets. Each of the 11 protein coding genes were assessed for functional evidence for involvement in stroke pathology by reviewing the existing literature and performing lookups in two open resources described in **3.2.2.1**. All 11 genes were found with no evidence for being the biological target of any pharmaceutical drugs and no evidence for protein coding variants that cause Mendelian disorders of CVDs. Of the 11 genes, five (*SRF, CACFD1, HVCN1*, *TRAFD1* and *SMARCA4*) showed evidence for related vascular phenotypes in mouse models caused by knockdown or knockout of the genes (**Table 11**).

SRF (located at the stroke-associated locus, 6p21) encodes a ubiquitous nuclear protein which is involved in regulation of both cell population proliferation and differentiation including regulation of vascular smooth muscle cell differentiation and their contraction[332]. It is also involved in related vascular biological process including platelet activation, angiogenesis and response to cytokines (such as immune response IL-6 signaling pathway)[333,334,335]. Expression of the *SRF* gene was detected to be greatly limited to skeletal and cardiac muscle tissues during development of mouse embryos[336]. Niu et al revealed that mouse embryonic stem cells lacking the *SRF* gene showed failure in expression of myocardin and cardiac myogenic alpha-actins and failed to form beating cardiac myocytes[337]. In mutant mouse embryos, deletion of endothelial cell-specific SRF resulted in haemorrhages, reduced density of capillaries, defects in migration of endothelial cells of small vessels and subsequently embryonic lethality was observed[335].

Mouse lines carrying homozygous mutations in *CACFD1* (located at the IS-associated locus, 9q34), exhibited disruption of homeostasis and metabolism such as elevated levels of circulating cholesterol and triglycerides (Open Targets, Mouse Phenotypes annotations). Furthermore, downregulation of *CACFD1* gene in *Drosophila* was shown to result in impaired Ca(2+) handling[338].

HVCN1 **(**located at the IS-associated locus, 12q24), encodes a voltage-gated protein channel protein which is primarily expressed in specific cells of the immune system[339]. The biological processes linked to this gene include voltage-gated proton transmembrane transport[340] and regulation of ion transmembrane transport. *HVCN1* functions in immune related pathways including NOX/ROS and RNS production in phagocytes[341]. *HVCN1* has been revealed to exacerbate brain damage through the modulation of NOX-dependent ROS production and activity in the mouse model of IS[342]. Furthermore, excessive expression of *HVCN1* along with other oxidative stress genes were reported to have higher expression levels in the ischemic hemisphere of old mice relative to young mice using Illumina cDNA microarrays[343]. In a mouse model of photothrombotic stroke, smaller infarction of brain with little motor deficiency was observed in *Hvcn1* knockout mice compared to the wild-type group due to increase in polarization of microglia from anti-inflammatory state (M2)[344]. Thus, developing a drug as an inhibitor to target HVCN1 could be promising for providing neuroprotection in CNS injuries like stroke.

TRAFD1 **(**located at the IS-associated locus, 12q24), has been shown to play a role in negative control of intemperate innate immune response by its involvement in the Toll-like Receptor (TLR) signaling pathway as a negative regulator[345]. Sanada et al identified that *TRAFD1* mutant mice displayed elevated Il-6 secretion and circulating Il-6 levels compared to the wild-type group (Open Targets, Mouse Phenotypes annotations)[346]. Study of TLR4 in ischemia models of TLR4 knockout mice have demonstrated that suppression of IL-6 as a pro-inflammatory cytokine results in decreased tissue injury[347]. Furthermore, TLR4 deficiency in mice has been reported to show inhibition of IL-6 and TNFα cytokines as well as reduced cerebral ischemia-reperfusion injury[348,349]. Mice deficient in TLR2 displayed smaller size of cerebral infarction relative to wild-type mice in response to cerebral artery occlusion (ischemia)[350,351]. These findings suggest that developing a drug for promotion of *TRAFD1* gene as a negative regulator of TLR signalling may attenuate stroke risk or its pathophysiology.

SMARCA4 (located at the stroke-associated locus, 19p13), encodes a protein which is part of the large ATP-dependent SNF/SWI chromatin remodeling complex and is involved in positively regulation of Wnt signaling pathway[352]. Mice lacking *SMARCA4* expression in smooth muscle (*SMARCA4* (*BRG1*) knockout mice) exhibited cardiopulmonary defects such as patent ductus arteriosus and complete atrioventricular septal defect specifying a key role of *SMARCA4* gene in controlling the development of smooth muscle[353]. Moreover, *SMARCA4* deficiency in the endocardial lineage of the mouse embryos resulted in disorganized and thickened semilunar valve (SLV) cusps and ventricular septal defect (Open Targets, Mouse Phenotypes annotations)[353,354].

Table 11. Known functional evidence for genes prioritized in moloc analysis of stroke.

GeneCards (http://www.genecards.org/) was used to look-up the gene name. The Open Targets Genetics database (https://genetics.opentargets.org) was used to look-up information linking the gene to the CVD related phenotype based on experimental work in animal model. The Online Mendelian Inheritance in Man (OMIM) platform (https://www.omim.org/) was used for evidence on CVD caused by defects in the gene and knock down/out in animal model. The publication reporting the evidence for the candidate gene (PMID).

Colocalized gene	gene name	gene direction for increased stroke risk (my study)	cardio human phenotype (OMIM)	Related-vascular animal model
SRF	Serum Response Factor	reduced expression	no	yes (PMID: 15929941, 15647354, 18804439)
ABCC10	ATP Binding Cassette Subfamily C Member 10	reduced expression	no	no
ABO	Alpha 1-3-N-Acetylgalactosaminyltransferase And Alpha 1-3-Galactosyltransferase	reduced expression	no	no
GBGT1	Globoside Alpha-1,3-N-Acetylgalactosaminyltransferase 1 (FORS Blood Group)	reduced expression	no	no
CACFD1	Calcium Channel Flower Domain Containing 1	reduced expression	no	yes (Open Target, Mouse Phenotypes annotations)
TCTN1	Tectonic Family Member 1	reduced expression	no	no
HVCN1	Hydrogen Voltage Gated Channel 1	reduced expression	no	yes (PMID: 22388960, 27470181, 28774948)
GPN3	GPN-Loop GTPase 3	increased expression	no	no
TRAFD1	TRAF-Type Zinc Finger Domain Containing 1	increased expression	no	yes (PMID: 18849341)
HP	Haptoglobin	increased expression	no	no
SMARCA4	SWI/SNF Related, Matrix Associated, Actin Dependent Regulator Of Chromatin, Subfamily A, Member 4	increased expression	no	yes (PMID: 21518954, 26100917)
C19orf52	Chromosome 19 Open Reading Frame 52	increased expression	no	no

4.4 Discussion

4.4.1 Main findings

In this study, summary statistics from DNA methylation, gene expression, circulating metabolite and stroke datasets were integrated by conducting multiple trait colocalization in two phases to identify statistical evidence of shared genetic effects at stroke risk loci and pinpoint prioritized candidate intermediate phenotypes likely to be involved in development of stroke. Here, Phase I moloc analyses at 20 stroke-associated loci identified 11 loci with evidence of colocalization with methylation CpG sites and multiple clusters of lipoprotein lipids. In Phase II, integration with eQTL data pinpointed 12 candidate genes (protein coding) in which changes in gene expression may contribute to the risk of stroke. Moreover, 3 loci (9q34 (IS), 12q24 (IS), 19p13 (stroke)) showed evidence of shared regulatory effects between multiple genes, multiple metabolites and stroke in these moloc analyses, suggesting a functional role for these genes and lipoprotein lipids in the pathogenesis of stroke.

At the 9q34 locus, moloc identified a shared common association signal between 3 likely causal genes (*ABO*, *GBGT1, CACFD1),* a cluster of lipoprotein lipids and stroke. The T allele of the shared SNP, rs635634 shows evidence of effect on expression of *ABO* (ß=-0.510, P=5.22x10^{-279}), *GBGT1* (ß=-0.153, P=8.93x10^{-25}), and *CACFD1* (ß=-0.153, P=1.16x10^{-9}) genes in eQTLGen data[146], suggesting that these genes might be influencing stroke through regulation of lipoprotein lipids metabolisms. The intergenic risk variant (rs635634) at this locus is located upstream of the nearest gene, *ABO* which determines blood group and has previously been reported to be linked to ischemic heart disease (IHD)[355,356] and IS[357]. The shared causal variant, rs635634 was strongly associated with expression of *ABO* gene (P=5.22x10^{-279}) in eQTLGen whole blood and the eQTL showed strong evidence of colocalization (PPA.GEMb.M=97%) with a cluster of circulating medium/small dense LDL/VLDL (a large triglyceride-rich lipoprotein) lipoprotein subclasses, VLDL cholesterol, triglycerides in IDL and stroke. At this locus, no evidence of colocalization was found for rs8176719 SNP, which is the most common genetic variant linked to blood group O through a frameshift mutation that results in *ABO* gene inactivation[358]. This suggests that rs635634 has a primary prominent role in regulation of expression of the *ABO* gene and circulating lipoprotein metabolites rather than just tagging the SNP associated with the O blood group type. Moreover, the rs507666-GMP140 pQTL has previously been found to overlap with the CHD-associated SNP at this this locus[330], which is in high LD with the IS shared variant, rs635634 (r^2=0.99). GMP140 protein has been found to play a role as a mediator between leukocytes and endothelial cells or platelets involved in inflammatory pathways. A MR study performed by Yao et al[330] also identified the causal effect of *ABO* gene on circulating levels of GMP140 protein in heart atrial appendage. Glycosyltransferase protein encoded by *ABO* gene (along with its paralog, *GBGT1*) is involved in the metabolism of lipids and lipoproteins and the glycosphingolipid biosynthetic pathway. ABO blood group has been previously reported to be linked to IS[359,360] and CVD in the Framingham Heart Study (FHS) cohort[361]. In addition, studies revealed the correlation of ABO blood type with CVD through effects on coagulation pathway and mechanisms[362,363]. Taken together, these findings suggest that the association is likely to be driven by the IS causal variant in high LD with the CHD-associated SNP and likely to be mediated via *ABO* gene expression and concentrations of APOB and multiple LDL/VLDL lipoprotein subclasses. However, the possibility of horizontal pleiotropy cannot be ruled out at this locus using moloc.

This lead SNP, rs635634 is also in LD with the rs495828 SNP associated with venous thromboembolism (VTE)[364] (P=3x10^{-16}) (r^2=0.84 in the 1000 Genomes EUR population). This VTE-associated SNP is also strongly associated with blood metabolite ratios[175] (P=6x10^{-34}). Furthermore, rs635634 is in moderate LD with another VTE variant, rs9411377 (P=1x10^{-224}) in a study published by Klarin et al[365] (r^2=0.43). There are two other VTE-associated SNPs at this locus, rs8176719 (P=6x10^{-12}) (r^2=0.34 with the IS variant) and rs2519093 (P=8x10^{-16}) (r^2=0.98 with the IS variant). Notably, the rs2519093 variant is also associated with triglyceride levels[366] and CAD (P=1x10^{-11}) at this locus[367]. Richardson et al[327] performed a multivariable MR between circulating lipoprotein lipids, apolipoproteins and CHD using 440 triglyceride-associated variants and provided further evidence for a potential causal effect of triglyceride levels on high risk of CHD at this locus. Hence, the rs635634 signal is likely to be a shared causal regulatory effect behind VTE, CHD and IS phenotypes.

IS variant (rs635634) is in moderate LD with the intronic HF-associated SNP, rs9411378 (r^2=0.53). PheWAS studies of rs9411378 SNP on traits in UK Biobank and GWAS traits from the GWAS Atlas[368] have reported on its association with VTE, metabolic and hematologic traits (e.g., blood cell counts and hemoglobin levels)[369]. However, this study revealed that the *ABO* locus is associated with HF independent of any of these traits after conditioning on their effects[369]. This study along with colocalization results, suggests that the effect of the IS colocalized variant on regulation of circulating lipoprotein metabolites and IS might be independent of the HF variant and its pathway, however it needs further analysis to confirm that its regulatory role is not owing to LD with the HF variant.

In this chapter the 12q24 locus with exonic variant rs3184504 was identified with evidence of colocalization between methylation cg18714086, expression of 3 genes (*TCTN1*, *HVCN1*, *GPN3*) and very large HDL lipoprotein subclasses with stroke. A previous study has shown rs3184504 association with CAD[367,370]. *HVCN1* (*HV1*) encodes for a voltage-gated proton channel which controls NOX enzymes activity and is expressed in various cell types such as leukocytes[371] and microglial cells[342]. *HVCN1* hyperactivity has previously been shown to increase brain damage in IS in a mouse model[342,344]. MR analyses by Yuan et al provide evidence for decreased levels of HDL cholesterol causally linked to enhanced risk of IS[326]. Richardson et al[327] performed MR between circulating lipoprotein lipids/apolipoproteins and CHD using 534 HDL cholesterol-associated variants and found that individual MR analyses showed a 1-standard-deviation-higher HDL cholesterol (OR 0.80; 95% CI: 0.75–0.86; $P < 0.001$) and apolipoprotein A-I (OR 0.83; 95% CI: 0.77–0.89; $P < 0.001$) to lower the risk of CHD (i.e., increased HDL levels potentially causing lower CHD risk) however, the HDL effect attenuated in multivariable MR[327]. These findings support the hypothesis for an influence of a shared variant on stroke via regulation of HDL lipoprotein levels and *HVCN1* expression. However, the possibility of a "reverse" mechanism between intermediate molecular phenotypes (e.g., that the alteration in levels of HDL lipoprotein could cause changes in expression levels) cannot be excluded.

At the 19p13 locus, evidence of colocalization with shared genetic effects was found between two genes (*SMARCA4* and *C19orf52*) and circulating metabolites including apolipoprotein B, multiple IDL/LDL/VLDL lipoprotein subclasses, LDL cholesterol, remnant cholesterol, total cholesterol levels and stroke. The shared intergenic variant, rs8103309 is located upstream of *SMARCA4* (or *BRG1*)*,* a gene that encodes for a protein member of SWI/SNF complex which regulates chromatin configuration around certain genes. rs8103309 is associated with expression of the *SMARCA4* gene in heart-atrial appendage tissue in GTEx ($P=6.30x10^{-5}$, ß=-0.12). The stroke variant in LD with an intronic SNP, rs1122608 ($r^2=0.569$), has been reported as a shared genetic factor for CAD and IS[331] and LDL cholesterol levels [372]. A case control study in the Chinese population showed that the rs1122608 variant is associated with upregulation of total cholesterol levels and downregulation of the *SFRS3* gene in relation to stroke[373]. rs1122608 was found not to be associated with *SMARCA4* gene expression in this region[373], which suggests that association signals for CAD and stroke may be acting independently in regulating different genes. Of note, differentially expressed exons of *SMARCA4* gene have been reported in blood of patients with small vessel disease (cause of SVS) compared to controls using whole transcriptome microarrays[374]. MR studies have shown that increased levels of apolipoprotein B, triglycerides, and LDL cholesterol are causally linked to higher risk of CHD[327] and IS[326]. Multivariable MR analyses in these studies provided evidence for a predominant role of APOB in controlling the effect of LDL cholesterol and triglycerides on CHD and IS developments. These results are consistent with the moloc findings of colocalization of APOB, multiple lipoprotein subclasses and LDL cholesterol at different stroke loci. Wang et al performed multivariate metaCCA analysis on 7 related risk factors for ischemic stroke and identified association of pleiotropic genes such as the *SMARCA4* gene with CAD and levels of total cholesterol which may influence IS via these phenotypes[375]. These results are in agreement with moloc findings of evidence for colocalization between *SMARCA4* and total cholesterol levels at this locus in this chapter and support the hypothesis that the shared genetic variant rs8103309 exerts its effect on stroke through regulation of multiple downstream intermediate phenotypes and or risk factors for stroke.

The GME scenario of colocalization indicating a shared genetic aetiology between cg03463523 CpG site, *HP* gene and CES, was identified at the 16q22 locus. This is in agreement with GME colocalization evidence found in the previous chapter for AF (**3.3.2.1**) at this locus, thus suggesting a shared causal mechanism influences both of these medical conditions at this locus. This locus is investigated in more detail in **Chapter 5**.

4.4.2 Strengths and limitations

A major challenge for colocalization of multiple traits is to obtain association datasets with large enough sample sizes to be powered for detecting colocalized combinations of traits. In this moloc analysis, a large-scale GWAS of metabolites was employed which increased power for detection of potential candidate metabolites due to the large sample size and strength of mbQTL associations.

Regions with less than 50 SNPs in common between 3 phenotypes were not tested for colocalization (i.e., combination of traits with <50 overlapped SNPs were removed from analysis). This issue arose due to the lack of full summary statistics from the GoDMC consortium[124], because association signals with P value $>1x10^{-5}$ are not included in the published GoDMC mQTL dataset or incomplete SNP coverage of a region. Therefore, availability of full summary level statistics from association studies is a necessity to enable us to effectively detect molecular phenotypes sharing a signal with a disease trait. Also, this study was limited to whole blood (a mixture of different cell types), which is available in large sample sizes. Other tissues might be informative too, although might be limited by the sample size and availability post-mortem.

Chapter 5 Follow-up analyses of the 16q22 locus

5.1 Introduction

One of the key results found in moloc analysis of AF (**Chapter 3**) and stroke (**Chapter 4**) was the 16q22 locus. In this chapter, I describe detailed follow-up analyses of this locus to investigate whether AF and stroke share the same genetic aetiology at this locus.

5.1.1 *HP* gene

Haptoglobin (HP) protein is encoded by the *HP* gene on chromosome 16q22 locus. HP protein is a main plasma acute-phase glycoprotein which binds free hemoglobin (HB) to form an irreversible HP-HB complex facilitating its removal from circulation[376] while preventing the loss of iron and kidney damage resulting from the accumulation of hemoglobin following hemolysis[377,378,379]. The HP protein also binds to a variety of lipid molecules[380,381,382]. The *HP* gene contains internal copy number variation of a pair of exons which produce two co-dominant alleles (HP1 and HP2). Based on the genomic architecture for these alleles of the two-exon segments, the copy number of a multimerization domain is encoded. A dimer is formed by single-copy *HP1* allele encoding two subunits α1 and ß, but multimers form as a result of α2 and a portion of the ß subunit encoded by two-copies of the *HP2* allele[383]. A common copy number variation (CNV) within this gene influences the structure of haptoglobin molecular phenotypes: HP1-1, HP2-1, HP2-2[383]. The binding affinities of each HP phenotype for free hemoglobin varies[384] with each phenotypic functional variation contributing differently to biological processes[385]. These HP phenotypes have been linked to cardiovascular diseases such as myocardial infarction[386,387] and infections[388] with HP2 genotype more susceptible to the risk of diseases. In addition to this major functional variant several other mutations have been described in the *HP* gene. A missense mutation (I24T) in *HP2* associated with downregulated serum haptoglobin protein has been linked to anhaptoglobinemia[389]. A rare splice donor mutation in HP1 genotype has been found to be associated with higher levels of non-HDL cholesterol and risk of coronary artery disease[390]. However, the contribution of each of these functional variations to human phenotype is not yet clear.

The relationship of the *HP* CNV with GWAS association signals close to the *HP* gene has not been clear[391] due to not being included in imputation panels as it is a copy number variation, and while it can possibly be imputed, it is not routinely[392]. A GWAS of plasma cholesterol levels identified an association signal at markers close to *HP*[393], however, the causal signal explaining this association at most GWAS risk loci are not easily pinpointed. Kazmi et al showed that the rs2000999 SNP located downstream of the *HP* gene is associated with levels of circulating HP protein independently of the *HP* CNV effect on HP levels and HP structure[394]. The genetic variant rs2000999 has also been linked to LDL cholesterol levels and total cholesterol levels[393]. The existence of two distinct mechanisms for the CNV (influencing both protein levels and structure) mean that disease associations observed only for the CNV are likely to represent protein structure effects, whilst those also observed for rs2000999 are likely to represent effects of protein quantity.

5.1.2 Previous colocalization studies between *HP*, AF and stroke

There has been limited formal colocalization studies conducted between the 16q22 risk locus and stroke and no colocalization studies between 16q22 locus and AF to map the causal genes involved. However, previous literature has linked the *ZFHX3* gene, which belongs to the same genomic region as *HP* gene, with AF. The *HP* gene was identified to be colocalized with both AF and stroke in Chapters 3 and 4 respectively. *ZFHX3* has been previously prioritized as a candidate gene at this AF-associated locus only based on its closest proximity to the index AF SNP. No evidence of colocalization of *ZFHX3* expression level with AF was found in Genotype-Tissue Expression (GTEx) consortium eQTL data with 54 tissues[148,147]. This is consistent with my finding of no colocalization evidence for *ZFHX3* in blood with any of the traits tested in moloc studies in this thesis using eQTLGen data (PPA.GM.E=93.2%, (**3.3.2.1**)). The *ZFHX3* (Zinc Finger Homeobox 3) gene encodes a transcription factor (TF) (DNA-binding protein) containing multiple zinc finger motifs and multiple homeodomains which play fundamental roles in adult tissues and development[395,396]. A knockdown of the *ZFHX3* gene was found to disrupt regulation of Ca^{2+} homeostasis and to be associated with high risk of AF in a mouse model[397]. A study in mouse atrial cardiomyocytes with *ZFHX3* knockdown showed changes in expression of miRNAs and promoted susceptibility to arrythmia[110]. Van ouwerkerk et al[398] constructed a mutant mouse with a deletion in the regulatory region of the *ZFHX3* first intron where AF risk variants are mostly located and found no *in vivo* changes in expression levels of any genes in atria and ventricles including *ZFHX3*, suggesting that AF risk variants might not be acting through regulation of *ZFHX3* expression to affect AF.

Furthermore, Data-driven Expression Prioritized Integration for Complex Traits (DEPICT)[399] enrichment analysis by Nielsen et al[41] identified the *HP* gene to be listed in a gene set (P-value<0.05) involved in molecular mechanisms which might be related to pathogenesis of AF, including increased sensitivity to induced mortality and morbidity, hemoperitoneum, heparin binding, reactome platelet degranulation. These findings are consistent with recent reports from a MR study in UK Biobank[176] that detected direct causal effects of hemoglobin concentration on cardiac arrythmia as well as platelet structure involving platelet count and blood volume occupied by platelets in the blood (plateletcrit)[400]. In addition, one recent population-based study has linked hemoglobin concentration with AF[401]. GWASs have not associated any SNPs in the 16q22 with blood cell counts or hemoglobin levels.

The AF top hit (rs2359171) in Nielsen et al is located in an intronic region of the *ZFHX3* gene and is approximately 950 kb upstream of the *HP* gene (see **Figure 10**). The 16q22 risk locus harboring genetic predisposition to CES has been reported to act through regulation of mechanisms underlying AF. Chauhan et al postulated in a 2016 review of the genetics of ischemic stroke, that genetic risk factors influencing both CES and AF may be linked through molecular processes involved in cardioembolism[85].

5.1.3 Questions to be addressed in this chapter

Identification of whether or not a common association signal is causal for both a disease trait and an intermediate molecular trait is challenging due to the uncertainty instigated by the genome correlation structure known as linkage disequilibrium (LD)[402] and the presence of multiple causal variants in some genomic regions. However, in Chapters 3 and 4, where the moloc approach was applied to multi-omics data, I assumed that a single causal variant within the region was underlying the AF association signal. However, this assumption does not hold for the 16q22 locus as two independent SNPs are associated with AF at this locus. The focus of this chapter is the colocalized *16q22* locus. This locus was found to be colocalized with both

AF and stroke in the moloc analyses performed for each trait separately (i.e., for each trait the GME scenario at this locus passed the PPA of 80% and the *HP* gene was identified as the colocalized gene (the potential candidate gene)). The 16q22 locus (with two AF independent association signals) was found to be colocalized with AF, methylation at the cg03463523 CpG site and expression of the *HP* gene in the moloc analysis in Chapter 3 and with stroke, methylation at the cg03463523 CpG site and expression of the *HP* gene in Chapter 4. The main objectives for the analyses in this chapter were to explore the relationship between the *HP* gene, AF and stroke disease. First, a set of pairwise coloc analyses of AF and stroke were conducted to examine whether a single common genetic effect (i.e., AF primary variant) is shared between AF and stroke at this locus. Second, a newly developed colocalization method, PWCoCo was then applied to conduct conditional coloc analysis in order to account for colocalization of more than one causal variant in the region.

5.2 Methods

5.2.1 Conditional analysis to find the independent mQTLs

Conditional analysis was undertaken using COJO (conditional and joint association analysis) in GCTA[227,228] (version 1.91.1beta) (https://cnsgenomics.com/software/gcta/#Overview) at the 16q22 locus harbouring two independent AF signals identified with the colocalization evidence in Phase I (**Table 5**) (**3.3.1.1**). COJO uses summary data from GWAS and linkage disequilibrium (LD) estimated from the genotype data of a reference panel. COJO was used with summary statistics to estimate single-SNP association conditional on the AF lead SNP at the 16q22 locus to identify whether there is a peak for mQTL that might be colocalized with the secondary AF variant and influence methylation independently (i.e., also represents an independent *cis*-mQTL). Individual-level HRC-imputed genotype data was used from ALSPAC with a sample size of 8890 as a reference sample to match the mQTL meta-analysis results to the genotype data and estimate the LD structure of the region.

5.2.2 Association between the 16q22 region and haptoglobin protein abundance in ALSPAC

To test whether the shared primary AF variant, rs2359171 (**3.3.2.1**) is associated with haptoglobin protein levels in ALSPAC, protein QTL (pQTL) analysis was conducted. pQTL analysis was performed to estimate the effect of genetic variants at the 16q22 locus on circulating haptoglobin levels in ALSPAC. Detailed information of all data is provided in the ALSPAC study website (http://www.bristol.ac.uk/alspac/researchers/our-data/) through a completely searchable data dictionary and variable search tool. Haptoglobin measurements (individual level data) from ALSPAC at the teen focus 4 clinic (variable: Hapt_TF4) was extracted using ALSPAC package version 0.6.1 (https://github.com/explodecomputer/alspac/). Plasma haptoglobin levels from blood samples of approximately 3,250 children at one time point (age 17 years), were previously assayed using high-throughput proton NMR spectroscopy (following a detailed experimental protocol[403,404]). Participants who had withdrawn consent were removed from this analysis. In this analysis, extreme outliers (3SD of the mean) were filtered out and normalization of circulating haptoglobin concentration values was performed across samples by an inverse rank-based transform to the standard normal distribution. Using HRC-imputed genotype data, SNPs within 1Mb on either side of AF top variant (rs2359171) were extracted at the 16q22 locus. Only variants (n=4,715) with MAF >0.01 (common genetic variants) were included in the analysis. A total of 2,707 unrelated individuals (1404 females and 1303 males) were used in pQTL analysis. The genetic variants were tested for association with plasma HP levels using linear regression, including sex and the top 20 genetic principal components (PCs) of ancestry as covariates using the software PLINK[405] version 1.9. Sex was previously reported to have an effect on circulating haptoglobin levels[379] so, it was included as a covariate in this analysis.

5.2.3 Conditional analysis to find independent pQTLs within the 16q22 region

To identify independent variants for haptoglobin levels, conditional analysis was undertaken using a stepwise regression approach. Conditional association analysis was performed using an additive genetic model with PLINK[405] (version 1.9) (http://pngu.mgh.harvard.edu/purcell/plink/), adjusting for sex and 20 genetic PCs. This analysis conditioned on the haptoglobin lead SNP (top pQTL signal that had the highest strength of the association) to identify a secondary signal associated with levels of circulating HP at the 16q22 locus. This procedure was repeated (i.e., conditioning on the next most significant SNP; finding a new conditional top SNP; and then conditioning on this *new* top SNP) until the p-value showed little evidence of association (P>0.05). R^2 between the top haptoglobin pQTLs and the AF lead variant was looked up in the European (EUR) population from the 1000 Genomes (1000G) Project using the LDpair tool from LDlink (https://analysistools.nci.nih.gov/LDlink/).

5.2.4 Pairwise colocalization analyses between AF and stroke at 16q22 locus

Pairwise colocalization was tested between AF and stroke GWAS results from meta-analysis conducted by Nielsen et al[41] and Malik et al[67] respectively. GWAS summary statistics of SNPs within a 2Mb window around the AF top hit (rs2359171) at the 16q22 locus were used for this analysis. Summary statistics of each trait within this region were mapped to the variant call format (VCF) with harmonisation to the 1000G reference FASTA GRCH37/hg19 using gwas2vcf harmonisation tool[406] (version 1.2.1), implemented in Python (version 3.8.0). The gwas2vcf software was downloaded from https://github.com/mrcieu/gwas2vcf. Harmonisation was performed to align the non-effect allele and alleles coded on the forward strand to the human genome reference sequence build 37 to ensure consistency of the data. Any SNPs in the dataset which did not map to the reference genome were removed and the sign of the regression coefficient (beta, the effect size estimate) was switched if the effect allele needed to be switched. The rsID field of AF and stroke harmonized summary statistics within the 16q22 region was updated to the latest version of dbSNP identifier (v153) build 37. Pairwise colocalization analysis between AF and stroke traits was conducted using harmonised VCF datasets (nSNPs=4923) using coloc implemented in the gwasglue R package (version 0.0.0.9000)[10] available at Github (https://github.com/MRCIEU/gwasglue). Analyses were carried out in R (version 3.6.2). The analysis was repeated using a smaller window size (1Mb rather than 2Mb around the lead SNP).

5.2.5 Pairwise colocalization analysis between AF and stroke on SNPs in the AF top hit LD block

To compare the colocalization results of previous analyses with the results where the SNPs in the region are limited to the LD block, a pairwise colocalization analysis was performed on SNPs with $r^2>0.1$ within the LD block for the AF top SNP using gwasglue[10]. The LD threshold of $r^2=0.1$ was used to filter on all SNPs present in the 1000 Genomes EUR population with correlation coefficient greater than 0.1 within a 500 kilobases (kb) window of the AF top hit in order to exclude secondary AF SNP in this block. The AF lead SNP was not in LD ($r^2=0.064$) with the secondary SNP in the 1000G EUR population (i.e., truly independent SNP). Ensembl version 87 and human Genome reference assembly GRCh37 were selected for genome annotations. Correlation coefficients between the AF top hit and proxies was looked up using the proxy_search function in the SNiPA[407] (v3.3) (data accessed: 12 October 2020) online platform (https://snipa.helmholtz-muenchen.de/snipa3/). A LD block is defined as a genomic region where the target variant is not in significant LD with the other variants located outside this region[408]. Therefore, based on this definition, in this analysis, the LD block was constructed for each AF independent SNP using --ld-wind option provided by the GCTA-LDF[408], GCTA software[227] (version 1.91.1beta) where correlation coefficient (r^2) of SNPs in LD with each AF SNP within a 1Mb window (i.e., 500Kb either direction) of the AF hit were estimated by regression test.

5.2.6 Pairwise Conditional and Colocalization (PWCoCo) analysis of AF and stroke

Loci harbouring multiple independent association signals make it challenging to identify the shared causal variant between GWASs as this violates the standard assumptions of the coloc tool and affects the performance of the method. Recently a Pairwise Conditional and Colocalization analysis (PWCoCo) pipeline implemented in C++ has been developed within the MRC-IEU by Jamie Robinson (https://ieugit-scmv-d0.epi.bris.ac.uk/jr18055/pwcoco) which integrates conditional and colocalisation methods from GCTA-COJO software tool[227] (https://cnsgenomics.com/software/gcta/#Overview) and the coloc[220,221] (https://chr1swallace.github.io/coloc/index.html) R package for regions with more than one independent variant.

PWCoCo analyses were performed between two conditionally independent AF hits and conditionally independent signals for stroke in the 16q22 region. This stepwise conditional colocalisation analysis conditioned out the SNP effect of the AF primary top hit, rs2359171 and the secondary hit, rs876727 one at a time before conducting pairwise colocalization analysis between AF and stroke traits. In the presence of only one peak in the region, PWCoCo output results for unconditioned coloc. The conditional analysis was carried out using the GCTA-COJO package (version gcc-9.1.0) using HRC imputed genotype data from mothers in ALSPAC as the LD reference panel[244,409]. ALSPAC data is described in Chapter 2 (**2.1.4**). SNPs with MAF > 0.01 were filtered and included in this analysis. AF-stroke association signals with posterior probability (PP_4) greater than 80% were considered as evidence of colocalization.

5.3 Results

5.3.1 Conditional analysis to find the independent mQTLs in the 16q22 region

In Chapter 3 phase I (**3.3.1.1**), evidence of colocalization was identified for the *16q22* locus (PPA.GM=0.92) encompassing two independent association signals for AF, rs2359171 ($P=4.65x10^{-91}$) and rs876727 ($P=1.97x10^{-23}$). These two variants are not in LD with each other (1000 Genomes European population; $r^2=0.06$). Both primary, (rs2359171) and secondary (rs876727) SNPs located within the intronic region of the *ZFHX3* gene are significant mQTLs with genetic effects on methylation at the same CpG site, cg03463523 (ß=0.143, $P=1.87x10^{-36}$ and ß=0.086, $P=1.50x10^{-12}$) in GoDMC data. Here, in this chapter, conditional analysis at this locus confirmed that the two mQTLs (found in the colocalized 16q22 region) influence methylation at the same CpG site independently of each other (**Table 12**). Therefore, it is likely that cg03463523 and AF share two causal variants in this genomic region (i.e., the primary AF SNP colocalizes with primary mQTL and secondary AF SNP colocalizes with secondary mQTL). In fact, multiple independent mQTLs associated with cg03463523 (rs739414 (secondary), rs56311231 (tertiary)) were detected in this region (**Table 12**). Of note, rs739414 mQTL is not an AF variant. The secondary AF signal, rs876727 did not have the lowest P value after removing the effect of the top mQTLs, (rs2359171 and rs739414) and ranking the association results by P value. However, rs56311231 mQTL was in high LD with rs876727 mQTL (r^2=0.77) (in 1000G EUR population). Taken together, there might be two independent shared association signals between cg03463523 and AF. However, in Chapter 3 Phase II, moloc showed only the AF lead variant which was significantly associated with expression of *HP* gene (rs2359171-*HP cis*-eQTL), also colocalized with cg03463523, *HP* gene and AF (PPA.GME=96.7%) (**3.3.2.1**) in this region. Of note, the rs879324 variant, previously found to be associated with cardioembolic stroke[98], was also identified as a significant mQTL in high LD with the primary AF variant (r^2=0.93) at this locus.

Table 12. Genetic variants detected as independent mQTLs for cg03463523 by GCTA-COJO.
Single-SNP association analysis results, correlation coefficient (Beta), SE and P-value of mQTL effect conditioned on the set of SNPs (conditional on the given SNP(s)). LD (r^2) between Top conditionally independent mQTL and secondary AF/mQTL SNP, rs876727 in 1000G EUR population.

Top mQTL	Beta	SE	P	conditional on the given SNP(s)	r^2 with secondary AF/mQTL SNP
rs2359171	0.127	0.011	4.01×10^{-29}	rs876727	0.064
rs739414	0.087	0.011	6.15×10^{-15}	rs2359171	0.023
rs56311231	-0.077	0.012	5.75×10^{-11}	rs2359171, rs739414	0.766

5.3.2 Identification of independent SNPs associated with haptoglobin plasma levels in ALSPAC

pQTL analysis was performed to investigated whether the shared AF causal variant, rs2359171 identified at the 16q22 locus could potentially influence haptoglobin (HP) plasma levels. The association testing of 4,715 variants in this region against levels of plasma haptoglobin in 2,707

ALSPAC individuals was performed. Strong evidence of association ($P<5x10^{-8}$) was found between circulating haptoglobin plasma levels and 353 variants at this locus (**Figure 16**). The strongest association with haptoglobin was detected for SNP rs217184 (ß=0.639, $P=2.57x10^{-77}$), located on chromosome 16, which was in high LD (r^2 >0.95 in the 1000 Genomes European population) with rs77303550 and rs217181 variants in this region. There were multiple independent pQTLs (mostly intronic variants) strongly associated with haptoglobin protein levels in this region. Using stepwise conditional analysis, evidence of four conditionally independent associations with haptoglobin plasma levels was found (**Table 13**).

Table 13. Conditionally independent pQTLs identified in the 16q22 region by conditional analysis.
Top conditionally independent pQTL identified after adjusting for a set of covariates: (genetic principal components (PCs) and conditioning on variant(s) (Top SNP), P-value of the AF primary SNP, rs2359171 on haptoglobin level after conditional analysis, LD (r2) between Top pQTL and primary, rs2359171 and secondary, rs876727 AF SNP in 1000G EUR population.

Top SNP	Beta	SE	P	covariates	primary AF SNP P	r^2 with primary AF SNP	r^2 with secondary AF SNP
rs217184	0.636	0.033	$2.57x10^{-77}$	sex, top 20 PCs	0.698	0.002	0.009
rs9302635	0.679	0.033	$2.61x10^{-86}$	sex, top 20 PCs, rs217184	0.277	0.001	0.001
rs9941087	-0.366	0.031	$5.07x10^{-32}$	sex, top 20 PCs, rs217184, rs9302635	0.208	0.002	0.002
rs2336601	-0.203	0.054	$2.02x10^{-4}$	sex, top 20 PCs, rs217184, rs9302635, rs9941087	0.586	0.006	0.006

5.3.3 No relationship between AF risk variants and haptoglobin plasma levels

The shared AF risk variant, rs2359171 is not in LD with the primary *cis*-pQTLs that influence haptoglobin concentrations (r^2 <0.002). This AF variant (located 947Kb upstream of the top *cis*-pQTL, rs217184) showed little evidence of association with haptoglobin protein levels in

ALSPAC (i.e., not a significant pQTL; ß=0.015, P=0.698). **(Figure 16**). The secondary AF independent signal, rs876727 (located approximately 962Kb upstream of the top *cis*-pQTL) showed modest evidence of association with haptoglobin levels in this region (ß=-0.105, P=4.24x10^{-3}). Conditioning on the top HP-associated SNPs did not find AF risk variants (rs2359171 and rs876727) as significant pQTLs at this region (**Table 13**). In addition, none of the four independent pQTLs was in LD with the AF shared variant, rs2359171 (r^2<0.007) suggesting that changes in plasma haptoglobin levels might not be on the causal pathway to AF. The lead AF SNP at this locus affects methylation at cg03463523 CpG site and *HP* gene transcript levels but not plasma protein levels of haptoglobin. Sex showed strong evidence of association with haptoglobin levels in this analysis (ß=0.318, P=5.25x10^{-17}). In addition, none of the four independent pQTLs was in LD with the primary AF shared variant, rs2359171 (r^2<0.007) suggesting that changes in plasma haptoglobin levels does not colocalize with AF in this region, therefore SNP regulation is only affecting methylation and *HP* gene transcript levels, but not protein expression.

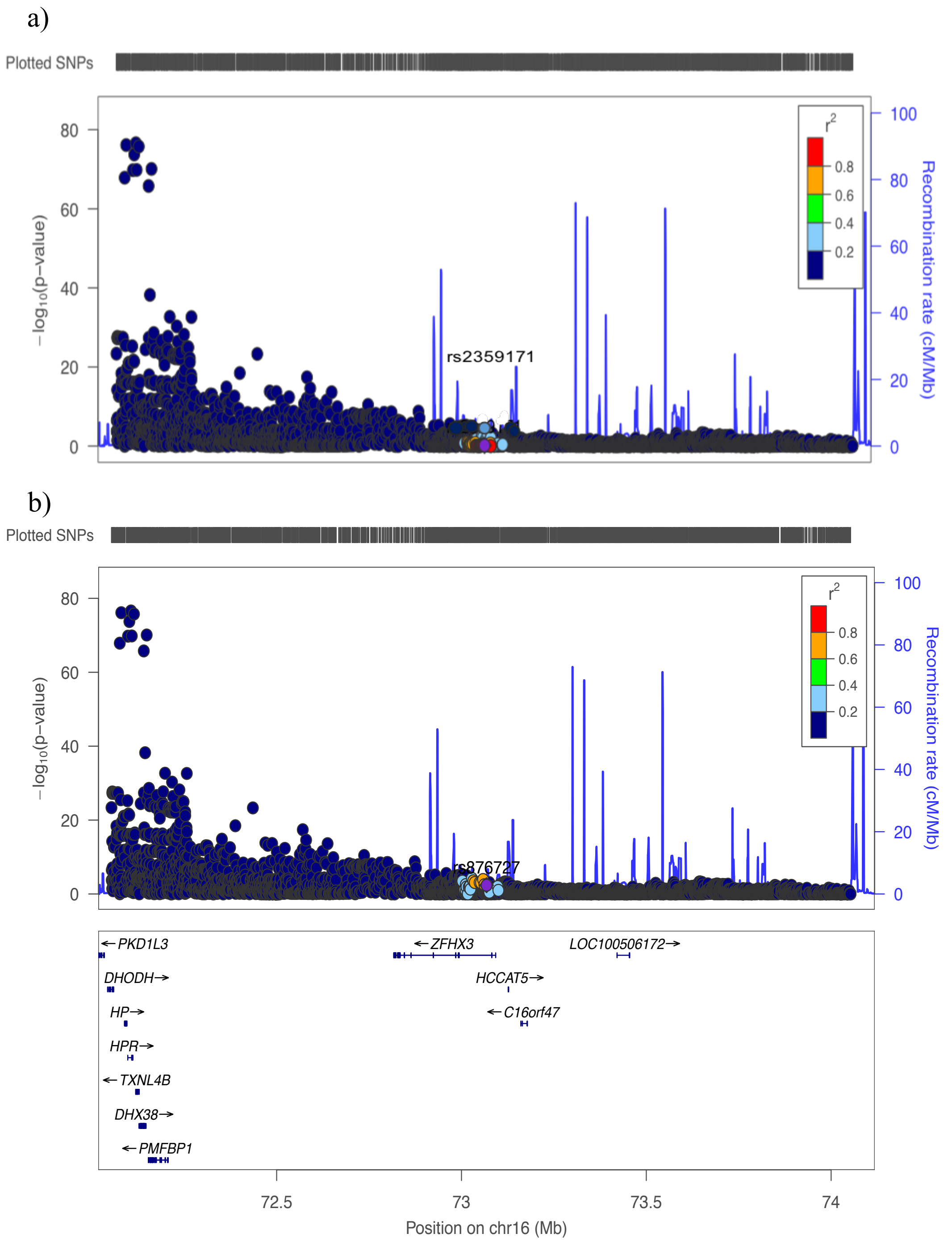

Figure 16. Regional association plots displaying the pQTL association peaks for two independent AF signals in the 16q22 *cis* region.

a) The primary SNP (rs2359171) and b) the secondary SNP (rs876727) for AF are not associated with circulating HP protein in this region. SNPs tested for association are plotted on the *x*-axis by their regional position in megabases. The -log10 P values of pQTL associations are plotted on the *y*-axis. The second *y*-axis shows recombination rate from 1000 Genomes. LD (r^2) is displayed based on the 1000 Genomes EUR population reference panel.

5.3.4 Pairwise coloc between AF and stroke at the 16q22 locus

To identify whether the genetic associations with both AF and stroke at this locus shared a common causal variant, pairwise coloc was conducted using harmonised VCF summary statistics of SNPs (n=4923) for AF and stroke at the 16q22 locus. Moderate evidence of colocalization (posterior probability (PP_4) =74.6%) was detected between the AF- and stroke-associated signals. This analysis was repeated using SNPs (n=2520) within a 1Mb window around the AF top hit (rs2359171) to test the effect of the number of SNPs in the region on the probability of a shared variant (PP_4) between AF and stroke. Evidence of colocalization slightly improved from PP_4 =74.6% to 74.8%. Of note, the AF secondary variant, rs876727 (located 14,739 bp upstream of the primary variant) shows very little evidence of association with stroke at this locus (Beta=-0.002, P=0.891).

5.3.5 Pairwise coloc between AF and stroke on SNPs in the AF top hit LD block

To avoid unreliable colocalization findings of a shared causal variant due to the presence of two independent AF signals in the region, pairwise coloc was performed on SNPs within the LD block of the top AF hit with the AF secondary SNP excluded from the region. Genetic variants (n=96) in LD (r^2>0.1) with the AF top hit within a 500Kb region in either direction (i.e., LD block for AF lead SNP, rs2359171 with length of 106,248 bp in chr16:72,995,996-73,102,243 genomic region) were extracted from the VCF file summary statistics of AF and stroke. After integrating AF and stroke summary statistics at the 16q22 locus to map SNPs in common between both datasets, 88 SNPs were left in the region for pairwise coloc analysis. Strong evidence of sharing a causal variant (PP_4>80%) (**Table 14; Figure 17**) behind both traits were identified by gwasglue, suggesting that reducing the noise of additional signals increased the specificity of colocalization to provide more reliable inference for a shared causal variant.

Table 14. Pairwise colocalization evidence for AF and stroke in the 16q22 region.

Colocalization evidence identified in different sized windows around the AF top hit, rs2359171, each containing a different number of SNPs (nSNPs). In the first two colocalization analyses (**5.3.4**) (first two rows), both AF primary and secondary signals (rs2359171 and rs876727) were present in the 16q22 region. In the last analysis on SNPs (r^2 >0.1 with the lead SNP) within +/- 500Kb LD block for the independent AF top hit, only rs2359171 variant was present and the rs876727 SNP was filtered out. Posterior probability of colocalization between AF and stroke (PP.H4.abf), probability of other coloc scenarios evaluated (PP.H0 – H3).

window size	nSNPs	PP.H0.abf	PP.H1.abf	PP.H2.abf	PP.H3.abf	PP.H4.abf
1Mb	4923	0.00%	10.90%	0.00%	14.50%	74.6%
500Kb	2520	0.00%	11.00%	0.00%	14.20%	74.8%
500Kb LD block	88	0.00%	12.00%	0.00%	4.90%	83.1%

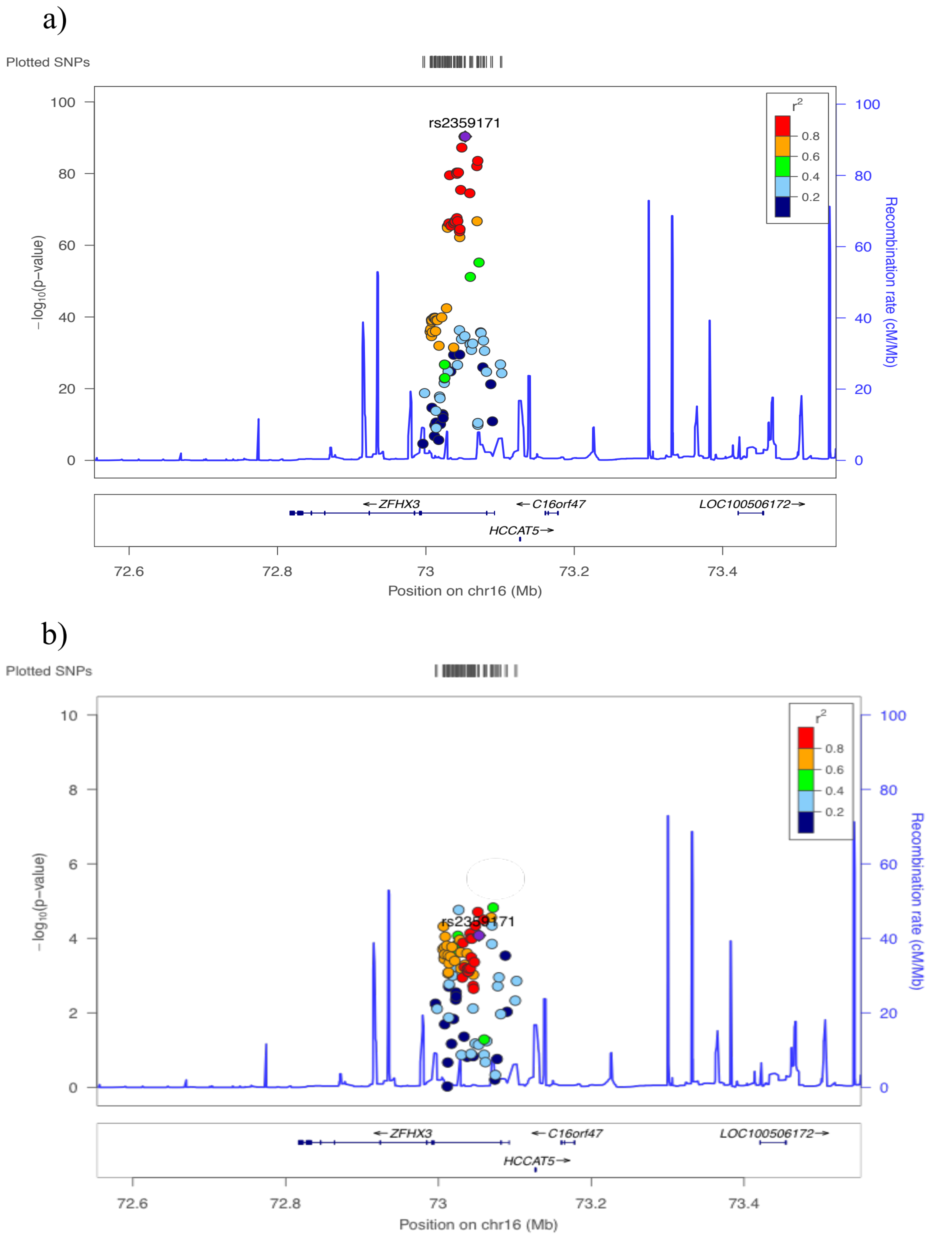

Figure 17. Regional association plot depicting a shared single association peak for AF and Stroke in the 16q22 region.

a) The top hit for AF (rs2359171) colocalized with b) the hit for stroke. 88 SNPs (r^2 >0.1) common between two traits within +/- 500Kb LD block for rs2359171 variant.

5.3.6 PWCoCo analysis to confirm that AF colocalizes with stroke at the 16q22 locus

In the previous chapters (**3.3.2.1** and **4.3.2**), the moloc results at the 16q22 locus showed that expression of the *HP* gene was colocalized with AF and stroke. However, colocalization analysis assumes a single causal variant within the region. As the 16q22 region had two weakly independent SNPs, rs2359171 and rs876727 in pairwise LD of r^2=0.064, (**Figure 18**) the assumptions of standard coloc have been violated. Here, I conducted PWCoCo analysis to assess whether there is evidence of a colocalization signal remain after conditioning on each of the independently associated SNPs (see methods for more detailed description (**5.2.6**)). Application of PWCoCo confirmed that the AF and stroke traits colocalized at the single primary signal for AF, rs2359171 in this region (nSNPs=4907; PP_4=81.4% unconditioned coloc). The second association signal for AF did not colocalize between AF and stroke after PWCoCo analysis.

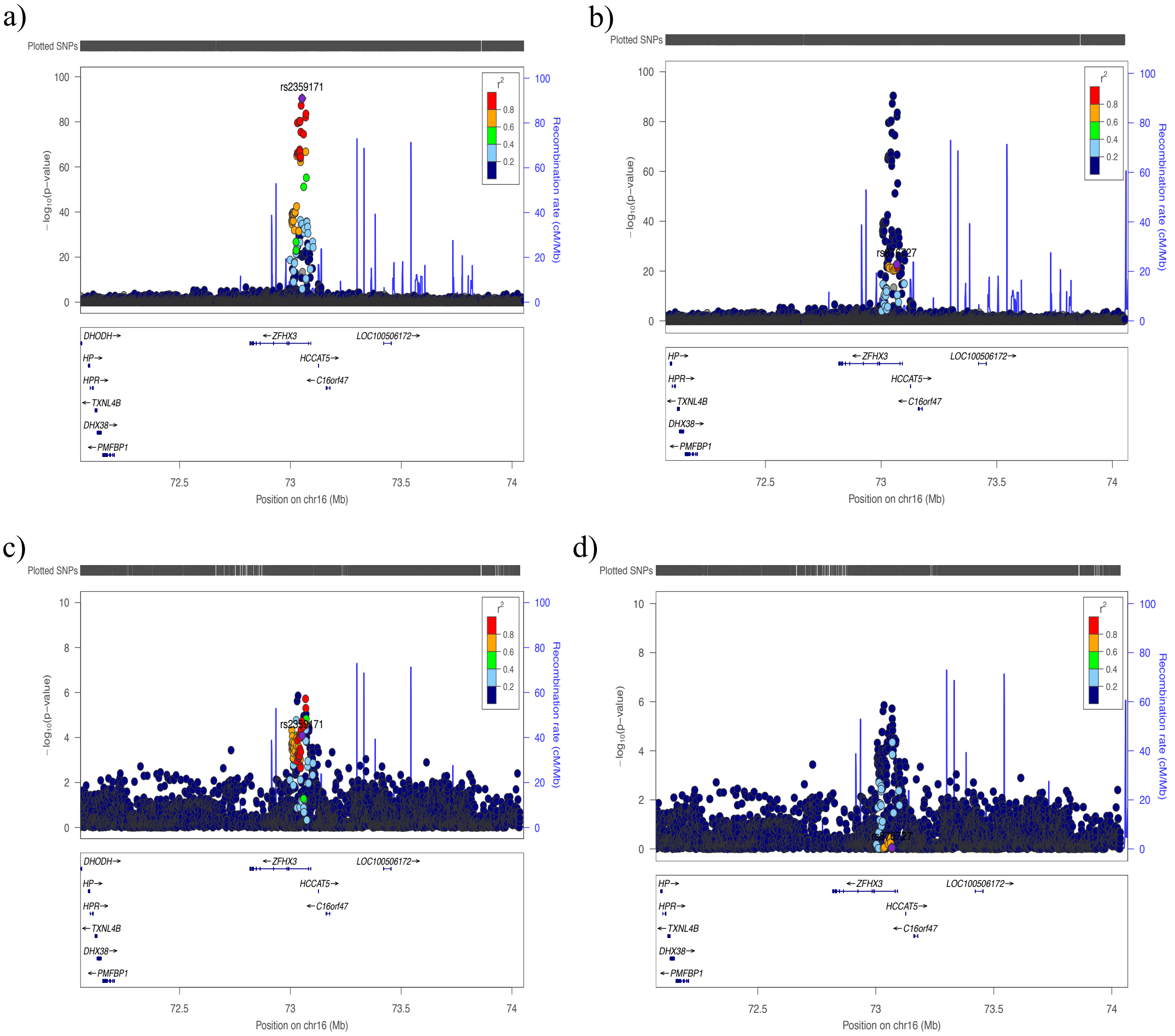

Figure 18. Regional association plots displaying the association peaks for AF and stroke GWASs in the 16q22 region.

a-d, Regional plots of the 16q22 locus. Regional plots of AF showing (a) the primary variant rs2359171 and (b) the secondary variant rs876727 without conditional analysis. Regional plots of stroke showing (c) AF variant, rs2359171 (r^2=0.92 with the top stroke hit rs12932445 in the 1000 Genomes EUR population) and (d) the second AF variant rs876727 for stroke without conditional analysis. LD (r^2) is displayed based on the 1000 Genomes EUR population reference panel.

5.4 Discussion

5.4.1 Main findings

In this study, statistical analyses were used to explore genetic mechanisms underlying the potential relationship between AF and stroke at the 16q22 locus. In chapters 3 and 4 methylation of cg03463523 and expression of *HP* gene was identified to colocalize with association signals for both AF and stroke at the 16q22 locus. The risk allele of the shared genetic risk for AF, rs2359171-T (ß=-0.175, P=4.65x10^{-91}) was positively correlated with DNA methylation of the cg03463523 CpG site (ß=0.143, P=1.87x10^{-36}) in the GoDMC study. Additionally, expression of the *HP* gene (encoding haptoglobin, a plasma glycoprotein) was downregulated (ß=-0.095, P=7.45x10^{-10}) in eQTLGen data. This gene showed a positive direction of effect in relation to AF and CES (**Table 7** and **Table 11**), suggesting a relationship between upregulated expression of *HP* gene and increased risk of AF and CES.

Here, in this chapter no evidence of the effect being through plasma HP levels was found and no effect of the AF lead SNP, rs2359171 on plasma HP protein levels persisted when conditioning on the top haptoglobin pQTLs and rs2000999. The AF risk variant and the identified independent pQTLs of circulating haptoglobin levels (including rs2000999 reported by Kazmi et al[394] and SNP haplotypes found by Boettger et al[392] for tagging the different HP structural features) were independent of each other and were not in LD (r^2<0.01 in 1000G EUR population). Previous studies between eQTLs and pQTLs performed by Sun et al showed that not all eQTLs colocalized with pQTLs[410], which suggests that there might be other plausible ways through which the genetic variants could affect the risk or development of atrial fibrillation. An explanation for this might be that many circulating proteins originate in other tissues such as haptoglobin which is expressed primarily in the liver before entering the circulatory system, and therefore blood eQTLs might not be that relevant to circulating protein levels. Moreover, there are many other biological mechanisms between transcription and circulating protein levels (post-transcriptional mechanisms: such as protein clearance from the bloodstream, degradation, secretion and binding) and technological reasons that can alter both measured quantity of transcript and measured quantity of protein[411,412]. This is consistent with the findings by Sun et al where based on the high LD ($r^2 \geq 0.8$) between *cis* pQTLs and eQTLs, 40% of the pQTLs were reported to be overlapped with eQTLs for the same gene in at least one tissue or cell type. Colocalization test conducted on these overlapping eQTLs-pQTLs by this group showed enrichment of approximately 78% pQTLs with eQTLs ($P<1x10^{-4}$) for the same gene in at least one tissue or cell type. However, only 12.2% retained significant (PP>80%) after filtering *cis* eQTLs for the most significant ones in a well-powered eQTL study in whole blood[410].

PWCoCo was used to identify if AF and stroke share two independent causal signals at the 16q22 locus (i.e., if the two genetic variants for both traits colocalize). The AF secondary signal, rs876727 showed little evidence of association with stroke in GWAS[67], however, conditioning on the primary variant using PWCoCo may help identification of an additional independent peak for colocalization which also substantiates findings from GWAS. In addition, PWCoCo was performed to ensure that the colocalization evidence between AF and stroke at this locus was not due to alternative causal variants in LD. PWCoCo provided strong evidence of a shared genetic effect of rs2359171 (colocalized *cis*-eQTL-mQTL) with both AF and stroke, suggesting a potential functional role of the *HP* gene (but *not* the circulating protein) in pathogenesis of AF and stroke, with gene expression controlled by the shared variant. Taken together, these findings highlight the evidence of potentially a single shared causal variant between AF and stroke in the 16q22 region supported by PWCoCo and the elimination of the rs2359171 peak from being confounded by the LD.

The AF primary lead SNP (located in the *ZFHX3* gene) was not associated with *ZFHX3* expression in eQTLGen data[241] and Martin et al did not identify any association between any of the AF-associated variants and *ZFHX3* expression in both atrial tissue and peripheral blood[413]. Animal models have previously shown that the *ZFHX3* gene might be involved in AF[110]. No evidence of colocalization of the *ZFHX3* gene with AF or stroke was found in the analyses performed in Chapter 3 and 4, respectively. This suggests that AF variant might act through *ZFHX3* expression and protein levels in another tissue, which needs further investigation. In contrast, colocalization analyses at this locus revealed evidence of sharing genetic effects on methylation at cg03463523 CpG site, expression of the *HP* gene and risk of AF and stroke, indicating that the *HP* gene could be a functionally relevant or link to these diseases. However, the finding that there is no evidence from this analyses that circulating haptoglobin levels (through which gene expression would presumably act) colocalize with AF risk support that it is possible for gene expression to functionally affect disease risk without detecting evidence in circulating HP levels. It is also very plausible that there might be pleiotropy at the genetic variant level. Another possible hypothesis might be timing of the measurement of plasma haptoglobin level, which may result in different pQTL effects across time. However, when I compared the HP pQTLs measured in this study (ALSPAC 17 year olds) to the HP blood plasma pQTLs in Sun et al in adults[410] I found consistency between the effect sizes so timing of pQTL effect is unlikely to be explaining the lack of overlap between HP eQTL and pQTL. A system under stress (age or disease) will behave differently than one that is not.

Studies in patients with AF revealed that inflammation is associated with perpetuation of AF[414]. Plasma haptoglobin is one of the inflammation-sensitive proteins which have been linked to stroke and it is a well-established biomarker for stroke[415]. A population-based study showed that its increased levels is a risk factor for IS[416]. A previous study has also reported that high cholesterol is linked to increased levels of inflammation-sensitive plasma proteins (ISP) (including haptoglobin) associated with higher risk of IS[417].

The findings in this chapter suggested that AF might mediate the genetic effect exerted by a single causal variant at the 16q22 locus on CES subtype through changes in methylation at cg03463523 CpG site and expression of the *HP* gene. However, the lack of evidence for colocalization of HP protein levels and AF suggest that the mechanism may be complex, or that these results may reflect pleiotropy at some level.

5.4.2 Strengths and limitations

The major strength of this chapter is the detailed information presented in regard to the AF-CES shared locus.

This study has a number of key limitations. One of the limitations of this study is the smaller sample size (n=2,707) (compared to eQTLGen study sample size n=31,684) was used for haptoglobin pQTL analysis in whole blood. In addition, different technological platform used to measure protein levels relative to corresponding gene expression levels. Another limitation is the fact that no data on ZFHX3 protein levels was available in whole blood to compare the output of HP pQTL with that of ZFHX3 pQTL. Furthermore, a lack of well-powered tissue-specific data for both genes limited the interpretation of biological mechanisms in terms of uncovering potentially tissue-specific function of these genes in AF and CES pathogenesis.

5.4.3 Future directions

Further studies based on HP protein levels measured in CVD patients could clarify the functional consequences of the AF-CES-related gene expression (*HP*). Considering the tissue-specific nature of some molQTL effects, further studies in AF and stroke -relevant tissues and cell types with big sample sizes should be conducted to confirm or add on to these findings.

Chapter 6 Mendelian Randomization analysis between atrial fibrillation and stroke

6.1 Introduction

Previous studies have shown links between AF and stroke. In my previous chapters I have explored the shared molecular aetiology of these two diseases. In this chapter, I use Mendelian randomization to explore the potential causal relationships between AF and stroke to better understand the basis for this shared aetiology.

6.1.1 Current knowledge of AF and stroke relationship

AF remains a leading contributor to mortality and morbidity of cardiovascular diseases such as stroke and heart failure[418,419]. Population studies such as Framingham heart study (FHS) have shown that AF is associated with an increased risk of stroke with its attributable risk significantly increasing with age[105]. Furthermore, contemporary studies reported that 20-30% of IS cases are attributed to pre-existing AF[70]. Stroke patients that have undergone 72 hours prolonged electrocardiogram (ECG) monitoring have enhanced detection of AF[71]. Observational studies have shown that genetic risk factors for AF are highly associated with the CES subtype of stroke[111]. Pulit et al reported that 23.1% heritability in risk of CES is explained by genetic risk factors for AF. AF PRS comprising 934 SNPs have also been found to associate with cardioembolic stroke after adjusting for the clinical risk factors of AF[111]. However, although observational studies can provide evidence regarding etiology of the disease, they are prone to reverse causation, confounding factors and measurement error which may bias the results and limit ability to appraise causality[420]. Nielsen et al[41] constructed a polygenic risk score (PRS) for AF (111 AF risk variants and 31 conditionally additional independent risk variants), and tested for association with a range of multiple disease groups in UK biobank participants[176]. This study identified associations with further cardiovascular complications in addition to strong associations with AF, including stroke, ischemic heart disease and heart failure. Their AF PRS was very specific for AF (i.e., finding no associations with any CVDs after exclusion of participants with any type of cardiac arrythmia), suggesting that AF might be acting as a mediator linking genetic variants to other related vascular diseases such as stroke[41].

MR attempts to overcome the limitations of observational studies to find evidence of causality by removing unmeasured confounding through instrumenting on genetic variants that proxy for the risk factor or exposure of interest. Three MR studies have been performed between AF and stroke to date, on different datasets which came to conflicting conclusions. Wang et al[72] and Fill et al found MR evidence for a causal effect of AF on stroke. Fill et al[421] performed two sample MR between AF, stroke, IS and subtypes of IS using GWAS data from AFGen consortium (n=133,073) and Nielsen et al for AF and MEGASTROKE consortium (n=521,612) for stroke and its subtypes and found that genetic predisposition to AF is causally linked to a higher risk of all stroke, IS and CES but not the other two IS subtypes (LAS and SVS). Wang et al[72] conducted MR on a larger phenome-wide scale and identified independent potential causal roles of AF risk factors including height, adiposity, SBP and CAD on AF and causal effects for several proteins (including decreased levels of interleukin 6 receptor (IL-6R)) with roles in inflammation on increasing risk of AF. In addition, they also found that genetic predisposition to AF associated with higher risk of stroke, IS and only the CES subtype in both univariable and MVMR model. In contradiction with these findings, Hou et al[422] conducted a network MR analysis using IS summary data from ISGC consortium (n=29,633) and AF statistics from AFGen consortium (n=588,190) and found evidence for the reverse direction of causal effect, from IS to AF[422].

6.1.2 Motivation for a two-sample MR

The primary motivation for this chapter is to further assess the relationship between AF and stroke which will lead to a better understanding of how these two diseases may be linked. In Chapters 3 (**3.3.2.1**) and 4 **(4.3.2)**, the moloc results at the 16q22 locus showed that the *HP* gene was colocalized with both these traits. Furthermore, Chapter 5 (**5.3.6**) pairwise colocalization analyses demonstrated that AF and stroke share a single causal variant at this locus. Therefore, it is possible that the *HP* gene may be a shared risk factor for AF and stroke. To assess the strength of evidence for a causal relationship between AF and stroke a two-sample Mendelian Randomization (MR) analysis was conducted using the genome-wide summary statistics from the Nielsen et al study for AF[41] and Malik et al study for stroke[67]. MR is a statistical technique called instrumental variable analysis, which makes use of the random allocation of SNP genotypes at birth as genetic instruments[217,219] in order to infer causality between traits. MR methods and assumptions are explained in more detail in section **1.5**.

6.1.3 Aims

The aim of this result chapter was to conduct two-sample MR analysis between AF and stroke to assess evidence for AF causing stroke. A set of sensitivity analyses were then performed to assess validity of MR findings including heterogeneity analysis[231] on the genetic instruments to investigate horizontal pleiotropy, reverse MR[219,423] and Steiger filtering[424,425] to determine if the causal hypothesis is orientated in the correct direction (i.e., from AF to stroke and not stroke to AF), and leave-one-out (LOO) analysis[234] to determine robustness of MR estimate to outliers. Validation analyses were undertaken to determine if the MR findings are replicated in an independent cohort with little overlap between AF and stroke cohorts.

6.2 Methods

6.2.1 Instrument selection and preparation

The publicly available GWAS summary statistics curated by MRC Integrative Epidemiology Unit (IEU) OpenGWAS database[10] (https://gwas.mrcieu.ac.uk) was used to access the GWAS summary data. The genetic instruments (i.e., genetic variants associated with the exposure) were acquired from the AF meta-analysis conducted by Nielsen et al[41] on European cohorts from six studies (UK Biobank, deCODE, HUNT, DiscovEHR, MGI, and the AFGen Consortium) (**2.1.1**). This summary dataset comprises a total of 33,519,037 SNP associations measured in 1,030,836 participants (OpenGWAS ID: ebi-a-GCST006414). I filtered results based on $P<5x10^{-8}$ and $P<5x10^{-5}$ as *P* value significance thresholds and then performed linkage disequilibrium (LD) clumping on the GWAS SNPs to obtain the top independent instruments for AF (i.e., the strongest AF-associated variants). To conduct the LD clumping the default settings were used from the tophits function provided by the *ieugwasr* R package (version 0.1.5)[10] (https://mrcieu.github.io/ieugwasr/) (maintained by IEU OpenGWAS database) which used a 10,000kb window and r^2 cutoff of 0.001 to select out the independent SNPs ($r^2<0.001$ based on 1000 Genomes EUR population LD reference panel). This very strict LD clumping procedure was used to ensure the independent instrument assumption of MR was satisfied. This procedure selected 111 independent AF top hit SNPs (between SNP LD $r^2<0.001$) identified in the Nielsen et al[41] paper as the genetic instrument set but did not select out the additional 31 conditionally independent SNPs that were reported in this paper (between SNP LD $r^2>0.001$).

6.2.2 Outcome lookup and harmonisation of data

To look up instruments in outcome GWAS the associations function provided by the *ieugwasr* R package (version 0.1.5)[10] was used. Instrument coverage was checked, and non-proxy search was used only as not many SNPs were missing. (i.e., the LD proxies parameter was set to 0 to search for the exact rsid present in the specific outcome GWAS summary dataset). The summary statistics for the selected instruments were extracted from the outcome GWAS, and then the instrument-exposure and instrument-outcome associations were harmonised to reflect the same effect allele using the *TwoSampleMR* R package (version 0.5.5)[234]. After harmonisation, instrument strength and validation tests were performed on the remaining instruments. Two-sample MR analyses (see section **6.2.4** describing 2SMR in more detail) were then conducted on the harmonised summary statistics to estimate the genetically predicted causal effect of AF disease as the exposure (or the risk factor) on the stroke disease as the outcome using multiple SNP instruments.

6.2.3 Instrument strength

In MR studies, weak instrumental variables can cause bias if the proportion of the variation in the exposure or the risk factor explained by those variants is small (exacerbated by a small sample size)[426,427]. To assess instrument strength and to ensure unbiased causal estimates the F-statistic for each instrument was calculated using the Cragg-Donald statistics[428,429]:

$$PVE = \beta^2/(\beta^2 + se^2 * n)$$

where PVE is the proportion of variance explained by the exposure instrumental variable, β and se are the effect estimate/size and standard error for the instrument, n is the sample size for the instrument. The PVE can then be used to calculate the F statistic:

$$F\ statistic = PVE * (n - 1 - k)/(1 - PVE) * k$$

where k is the number of instruments used in the 2SMR estimate (for single variant MR $k = 1$) and n is the sample size for the instrument when $k = 1$. If $k > 1$ an average of sample size for k selected instruments is used. F statistics were calculated for the individual SNP instruments as well as the pooled MR estimates.

6.2.4 Two-sample MR

Two-sample MR analyses were performed to investigate the causal relationship between AF (OpenGWAS ID: ebi-a-GCST006414) and stroke (OpenGWAS ID: ebi-a-GCST006906). MR methods which pooled effect estimates across multiple SNP instruments were used. These consisted of the five default methods available in the mr function in the *TwoSampleMR* R package[234] (i.e., IVW, Weighted mode, Weighted median, Simple mode, MR-Egger). For a full description of these methods see Chapter 2 (**2.2.2**). These steps were implemented using the *TwoSampleMR* R package (version 0.5.5) maintained by MR-Base[234] (https://www.mrbase.org/). The MR results were plotted as forest and scatter plots using the *ggplot2* R package (v3.3.2) (https://ggplot2.tidyverse.org/authors.html) in R (version 4.0.3). The following sections describe the sensitivity analyses conducted to explore possible violations of the assumptions of MR.

6.2.5 Reverse MR and Steiger analysis to evaluate the directionality of MR effect

A potential concern with the MR analysis could be reverse causality (i.e., preclinical phenotypes of the disease affect the risk factor or changes in risk factor is a consequence of the genetic liability to disease rather than a cause of the disease). In order to determine whether the

AF to stroke MR relationship was orientated in the correct direction reverse MR analysis was performed, instrumenting on the outcome GWAS (stroke) instead of the exposure (AF) using $P<5x10^{-8}$ and $P<5x10^{-5}$ thresholds. The stroke instruments were then looked up in the AF GWAS and the effect estimates for both traits harmonised. The same MR methods as detailed previously were then used to conduct the MR. Evidence of an MR effect in both the forward and reverse directions would indicate ambiguity in direction of causation.

Steiger filtering[424,425] of instruments also helps to reduce the likelihood of directional mis-inference. The Steiger filtering approach tests if the variance explained by the instruments is greater in the exposure than the outcome (r^2 exposure > r^2 outcome). If this is not true, then the MR effect could be orientated in the wrong direction (i.e., from outcome to exposure rather than exposure to outcome). A Steiger directionality test was performed on all the instruments to check whether the overall IVW effect was orientated in the correct direction. Steiger filtering was also conducted on the individual SNP instruments to identify outliers that were driving the reverse orientation of MR effect. Instruments which failed the Steiger test were removed, and MR analysis re-performed to determine if the MR estimates were altered. Steiger filtering on the individual SNP instruments and MR Steiger test of directionality were conducted using the steiger_filtering function and the directionality_test function on harmonised data respectively, provided by *TwoSampleMR* R package (version 0.5.5).

6.2.6 Validation analyses to replicate AF and stroke MR effect in another cohort

To replicate MR estimates in another cohort (i.e., with less sample overlap between cohorts used in both AF exposure and stroke outcome GWAS studies) validation analyses were conducted. The AF GWAS published by Nielsen et al[41] was performed across six studies including the deCODE study which introduces sample overlap with stroke in Malik et al, who performed stroke GWAS using European ancestry studies with the deCODE cohort included (n=5,520 cases, n=255,213 controls, total n=260,733). In order to evaluate whether sample overlap between the two GWAS studies used in the main MR analyses biased the causal estimate, validation analyses were conducted using a recently published AF GWAS by Roselli et al[52] (GWAS ID: ebi-a-GCST006061) which has little sample overlap with the stroke GWAS. The European ancestry sample used in this AF GWAS (total n = 537,409) consisted of 55,114 AF cases and 482,295 controls with the majority from two consortia: the Atrial Fibrillation Genetics (AFGen) and the Broad AF study (Broad AF) consortia. Samples from the Heart and Vascular Health (HVH) study (n=681 cases, n=1,331 controls) and the NINDS Stroke Genetics Network (SIGN) study[90] (n=7,743 cases, n=17,970 controls) part of the AFGen consortium were used in the MEGASTROKE stroke GWAS analysis.

6.2.7 Statistical tests for the evidence of instrument heterogeneity and directional pleiotropy

In MR analyses the presence of horizontal pleiotropy can violate the IV2 assumption (**1.5.1**) which is one of the fundamental assumptions for MR. IV2 assumes that all instruments are valid IVs and their effect on the outcome of interest is not through any pathway other than the instrumented exposure[251,252]. However, it is possible that the SNPs in question might not meet this assumption due to the large number of instruments incorporated in MR analysis and inadequate knowledge regarding the functional role of these genetic variants (i.e., they might be pleiotropic variants)[231]. Use of pleiotropic SNPs as an IV can invalidate their use in MR analysis and bias the MR estimate[231]. To explore whether there was statistical evidence for pleiotropic instruments included in the MR analyses, a heterogeneity test was performed by calculating Cochran's Q statistic[430], which follows a chi-squared distribution with the number of instruments minus 1 degree of freedom[231]. Heterogeneity was calculated for the MR-Egger and IVW regression estimates.

Directional pleiotropy happens in the case where the mean of the pleiotropy distribution, referred to as α is deviated from zero[251]. In other words, all pleiotropic instruments have biased effects in one direction (i.e., either increasing or decreasing their apparent effect on the outcome). The MR-Egger intercept test for directional pleiotropy was conducted to detect evidence of directional pleiotropy influencing MR results. The MR-Egger test provides an estimate for the intercept of the MR-Egger regression, where a non-null intercept would indicate evidence of directional pleiotropy[231]. As a note, it is possible to have balanced pleiotropy (i.e., no evidence of directional pleiotropy) but with high heterogeneity indicating the potential presence of horizontal pleiotropy.

Statistical tests for heterogeneity across multiple instruments and MR-Egger intercept were conducted using the mr_heterogeneity function on harmonised data and the mr_pleiotropy_test function respectively, provided by *TwoSampleMR* R package (version 0.5.5).

6.2.8 Leave-one-out sensitivity analysis

The leave-one-out (LOO) analysis is used as a tool to identify outliers. If more than one of the SNPs were identified as outliers (i.e., having a large influence on the MR estimate) in the LOO analysis, the MR analysis was repeated, dropping all the outlier SNPs together. LOO analysis was carried out to assess if a single SNP which might have a large horizontal pleiotropic effect is biasing the MR estimate or driving the association (between stroke instruments and AF outcome in main and validation reverse MR). Using this tool on the default setting where the method applied is the IVW method, the effect was re-estimated by dropping an individual SNP at a time in sequence.

6.2.9 Phenome-wide association studies (PheWAS) of outliers

Horizontal pleiotropy can sometimes be detected by carrying out phenome-wide association studies (PheWAS) on potential pleiotropic variants. Inspection of the scatter and forest plots (from the main reverse MR analysis – stroke=>AF at 5x10^{-8} and 5x10^{-5} (**Figure 20a** and **c**), and main and validation reverse LOO analyses (**Figure 24a** and **b**), highlighted two outlier SNPs (rs2634074 and rs6838973) as being possibly pleiotropic. To identify if these outliers affect multiple phenotypes or have pleiotropic effects on the AF outcome, PheWAS of these SNP outliers was performed against all available trait GWAS summary association datasets (n=39,603) in the MRC-IEU OpenGWAS database (v.3.5.1) (https://gwas.mrcieu.ac.uk/)[10] with a P value threshold of 1x10^{-5} using the *ieugwasr* R package (version 0.1.5).

6.3 Results

6.3.1 Instrument selection and outcome lookup

To select the AF instruments two different p-value thresholds (P<5x10^{-8} and P<5x10^{-5}) were used, and the SNPs were LD clumped (r^2<0.001) to obtain independent SNPs. The effect estimates for all AF instruments were extracted from the stroke outcome GWAS using a direct lookup of the original rsid (i.e., no proxy search was used), harmonised to ensure that the instrument-outcome associations reflected the same effect allele before conducting MR analyses. Instrument strength for all the MR analyses was sufficient to avoid weak instrument bias (F>10) (**Table 15)**. For the main MR analysis (AF=>stroke), 111 SNPs at the P<5x10^{-8}

threshold and 354 SNPs at the P<5x10^{-5} threshold were present after LD clumping. After harmonisation, 111 (100%) instruments at the P<5x10^{-8} threshold and 337 of the 354 (95.2%) instruments at the P<5x10^{-5} threshold remained for the MR analyses. F statistic for each individual SNP weight was greater than 10 (F>10).

Table 15. Counts of instruments at each step of the instrument selection process at two instrument thresholds in the main and validation MR analyses.

Number of instruments available after LD clumping (n instrument), number of instruments found in the outcome GWAS (after outcome lookup) percentage looked up (%), number of instruments available after harmonisation step for the MR analysis (after harmonisation) percentage harmonised (%), overall F statistic for all the SNP instruments for the MR analysis (F statistic).

				after outcome lookup		after harmonisation		
Study	**MR analysis**	**instrument threshold**	**n instrument**	**n**	**%**	**n**	**%**	**F statistic**
Main	AF to stroke	5x10^{-8}	111	111	100.0	111	100.0	90.6
	stroke to AF (reverse)	5x10^{-8}	8	8	100.0	8	100.0	41.8
	AF to stroke	5x10^{-5}	354	344	97.2	337	95.2	44.2
	stroke to AF (reverse)	5x10^{-5}	164	163	99.4	162	98.8	18.1
Validation	AF to stroke	5x10^{-8}	103	103	100.0	90	87.4	77.1
	stroke to AF (reverse)	5x10^{-8}	8	8	100.0	7	87.5	40.0
	AF to stroke	5x10^{-5}	298	296	99.3	248	83.2	41.7
	stroke to AF (reverse)	5x10^{-5}	164	152	92.7	135	82.3	20.9

6.3.2 MR relationship between AF exposure and stroke outcome

Two-sample MR between AF and stroke was conducted based on instruments selected genome-wide at P<5x10^{-8} (n=111 IVs) and P<5x10^{-5} (n=337 IVs) thresholds. The MR analysis provided evidence of a causal relationship between genetic predisposition to AF and increased risk of stroke at both instrument cut-offs. The IVW direction of effect (ß=0.200, P=1.16x10^{-27}, at P<5x10^{-8} instrument cut-off; ß=0.186, P=3.87x10^{-44} at the P<5x10^{-5} instrument cut-off) was consistent with the other MR estimates derived across different MR methods (**Table 16a**; **Figure 19a** and **b**). This suggests that the IVW estimates are robust to violation of different modelling assumptions regarding pleiotropy. MR was also conducted on the IS subtype to check the MR results were similar to any stroke which was used as the outcome in this study. As expected, due to the close agreement in the instrumenting SNP effects between the two measures similar MR effects were also observed (**Appendix B**). To validate these results, MR using an independent AF GWAS dataset (Roselli et al) which only has a small sample overlap with the stroke GWAS (Malik et al) was performed (**Figure 19c** and **d**). MR results between the main and this validation dataset were concordant (**Table 16b**) (**Figure 22a**), therefore the MR effect of AF on stroke is robust to bias caused by sample overlap between AF and stroke.

Table 16. MR estimates for five different methods at two instrument cut-offs in the main and validation study.

The number of instruments used in MR analysis (nSNP), Inverse Variance Weighted (IVW), MR estimates with effect size (Beta), standard error (SE) and p-value (P).

a) main MR

exposure	outcome	cut-off	nSNP	MR method	Beta	SE	P
AF	Stroke	$5x10^{-8}$	111	MR Egger	0.188	0.036	$6.81x10^{-7}$
				Weighted median	0.225	0.024	$2.04x10^{-20}$
				IVW	0.200	0.018	$1.16x10^{-27}$
				Simple mode	0.237	0.052	$1.45x10^{-5}$
				Weighted mode	0.218	0.024	$4.12x10^{-15}$
		$5x10^{-5}$	337	MR Egger	0.201	0.026	$1.34x10^{-13}$
				Weighted median	0.206	0.023	$3.44x10^{-19}$
				IVW	0.186	0.013	$3.87x10^{-44}$
				Simple mode	0.184	0.055	$8.81x10^{-4}$
				Weighted mode	0.211	0.025	$3.13x10^{-16}$

b) validation MR

exposure	outcome	cut-off	nSNP	MR method	Beta	SE	P
AF	Stroke	$5x10^{-8}$	90	MR Egger	0.087	0.050	$8.32x10^{-2}$
				Weighted median	0.156	0.024	$4.11x10^{-11}$
				IVW	0.180	0.020	$1.86x10^{-19}$
				Simple mode	0.137	0.052	$1.03x10^{-2}$
				Weighted mode	0.151	0.037	$9.85x10^{-5}$
		$5x10^{-5}$	248	MR Egger	0.126	0.035	$4.56x10^{-4}$
				Weighted median	0.156	0.021	$2.85x10^{-13}$
				IVW	0.181	0.016	$8.24x10^{-31}$
				Simple mode	0.151	0.050	$2.75x10^{-3}$
				Weighted mode	0.151	0.029	$2.52x10^{-7}$

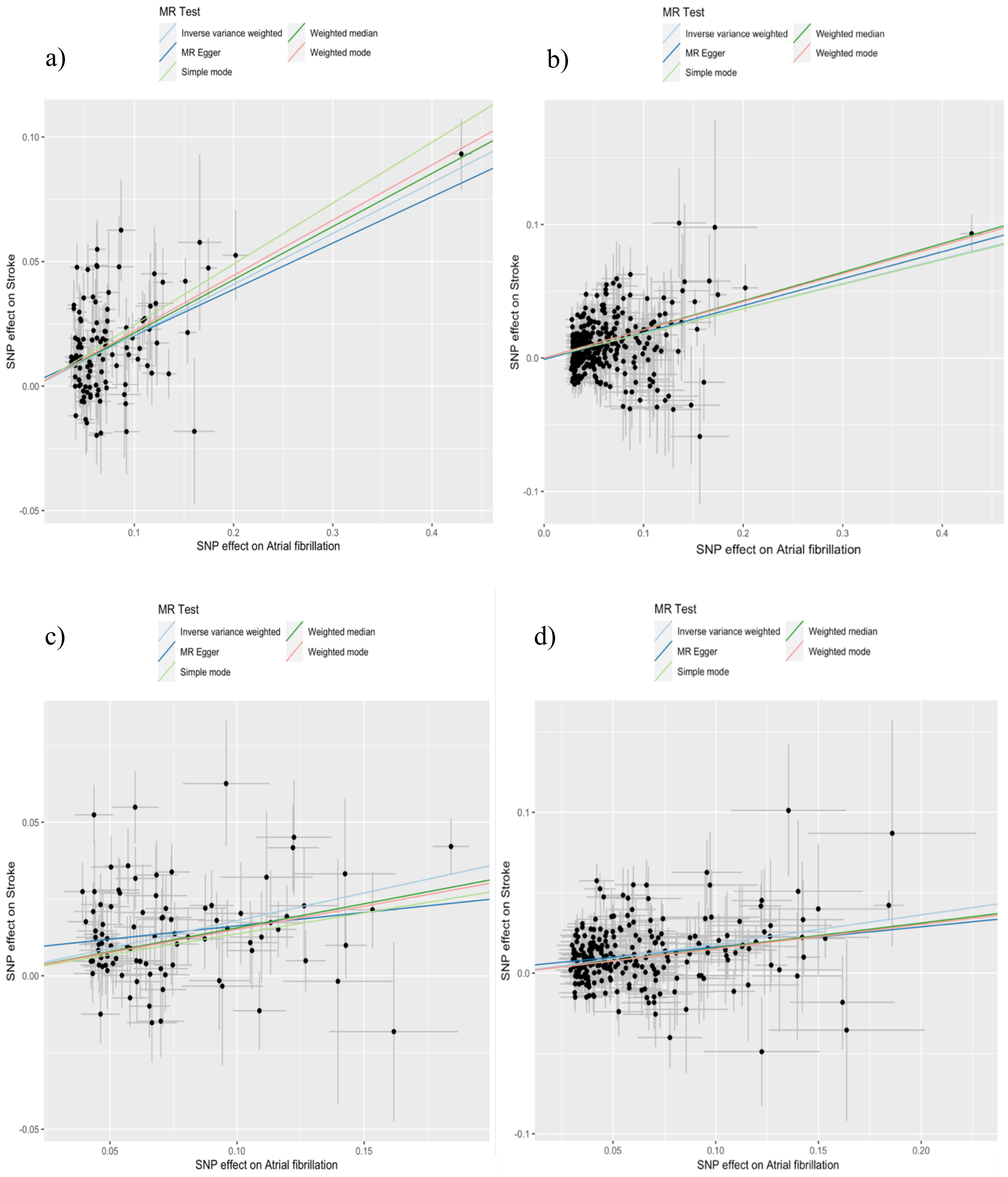

Figure 19. Scatterplot showing the MR slopes for the AF versus stroke relationship in the main and validation study.

(a-d) Scatterplots representing the MR relationships between instrument-AF associations and the instrument-stroke associations. (a (main) and c (validation)) all AF IVs at $p<5x10^{-8}$ cut-off (b (main) and d (validation)) all AF IVs at $p<5x10^{-5}$ cut-off. On the x-axis is the SNP effects for the AF associations and on the y-axis is the SNP effects for the outcome associations (on the log odds scale). Each point represents the individual SNP weights with the 95 CIs intervals plotted. The different coloured lines correspond to the slopes through these points calculated using each of the five MR approaches.

6.3.3 Reverse MR relationship using stroke as exposure and AF as outcome

To determine whether the MR effect was orientated in the correct direction reverse MR analysis was conducted, where stroke was instrumented and the causal effect of genetic liability to stroke on risk of AF was estimated. Two-sample MR between stroke and AF was conducted based on instruments selected genome-wide at $P<5x10^{-8}$ (n=8 IVs) and $P<5x10^{-5}$ (n=162 IVs) thresholds. There was some evidence for a causal effect of genetic liability to stroke on AF using reverse MR (**Table 17**) (**Figure 20a** and **b**), with the IVW analysis showing clear evidence of reverse effect between stroke and AF at $5x10^{-5}$ instrument cut-off for the validation analysis (ß=0.182,P=3.70x10^{-7}), which was also observed to a lesser extent for the Weighted median method (ß=0.123; P=2.64x10^{-8}) (**Table 18**). The other three methods did not show statistical evidence at P<0.05 of a reverse MR effect (**Table 18**) (**Figure 21a** and **b**). MR results from the Nielsen et al study (AF GWAS used in the main reverse MR analysis) were in concordance with MR results from the Roselli et al study (AF GWAS used in the validation reverse MR analysis) (**Figure 22b**), therefore this estimate of the effect of stroke on AF is robust to bias caused by sample overlap between AF and stroke. Steiger analysis showed a lack of evidence for MR in the reverse direction from stroke to AF (**Table 19**). All MR analysis results before and after removing instruments which failed the Steiger test are shown in the **Appendix D.** MR estimates remained similar indicating the outlier SNPs had little influence on the slopes for both the forward and reverse directions.

Table 17. Reverse MR estimates for different methods at two instrument cut-offs in the main study.

The number of instruments used in MR analysis (nSNP), Inverse Variance Weighted (IVW), MR estimates with effect size (Beta), standard error (SE) and p-value (P).

exposure	outcome	cut-off	nSNP	MR method	Beta	SE	P
Stroke	AF	5x10^{-8}	8	MR Egger	6.425	3.812	0.143
				Weighted median	0.101	0.064	0.111
				IVW	0.854	0.586	0.145
				Simple mode	0.120	0.080	0.177
				Weighted mode	0.089	0.070	0.244
		5x10^{-5}	162	MR Egger	0.081	0.126	0.519
				Weighted median	0.060	0.019	1.98x10^{-3}
				IVW	0.212	0.049	1.83x10^{-5}
				Simple mode	0.062	0.051	0.223
				Weighted mode	0.053	0.044	0.239

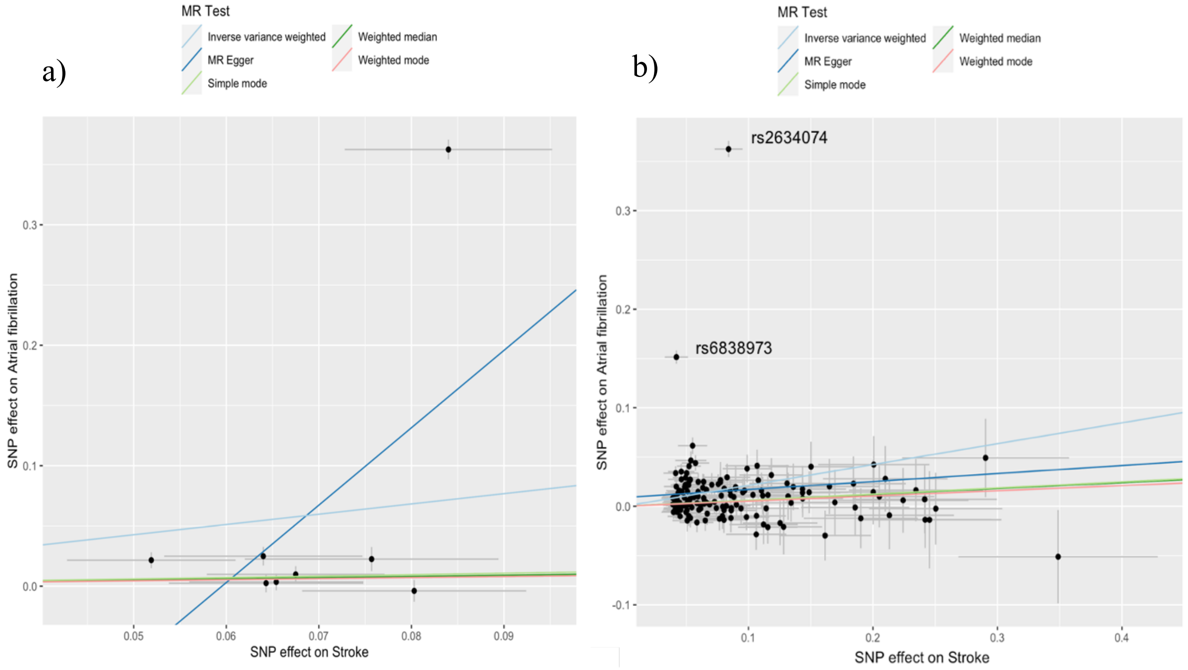

Figure 20. Main MR study of the effect of stroke on AF.

(a and b) Scatter plots representing the MR relationships between instrument-stroke associations and the instrument-AF associations. (a) all stroke IVs at p<5x10^{-8} and (b) at p<5x10^{-5} cut-offs. On the x-axis is the SNP effects for the stroke associations and on the y-axis is the SNP effects for the outcome associations (on the log odds scale). Each point represents the individual SNP weights with the 95 CIs intervals plotted. The different coloured lines correspond to the slopes through these points calculated using each of the five MR approaches.

Table 18. Reverse MR estimates for different methods at two instrument cut-offs in the validation study.

The number of instruments used in the MR analysis (nSNP), Inverse Variance Weighted (IVW), MR estimates with effect size (Beta), standard error (SE) and p-value (P).

exposure	outcome	cut-off	nSNP	MR method	Beta	SE	P
Stroke	AF	$5x10^{-8}$	7	MR Egger	-0.192	0.585	0.756
				Weighted median	0.129	0.063	0.040
				IVW	0.175	0.070	0.012
				Simple mode	0.142	0.077	0.117
				Weighted mode	0.140	0.076	0.116
		$5x10^{-5}$	135	MR Egger	0.020	0.089	0.824
				Weighted median	0.123	0.022	$2.64x10^{-8}$
				IVW	0.182	0.036	$3.70x10^{-7}$
				Simple mode	0.065	0.060	0.280
				Weighted mode	0.065	0.058	0.264

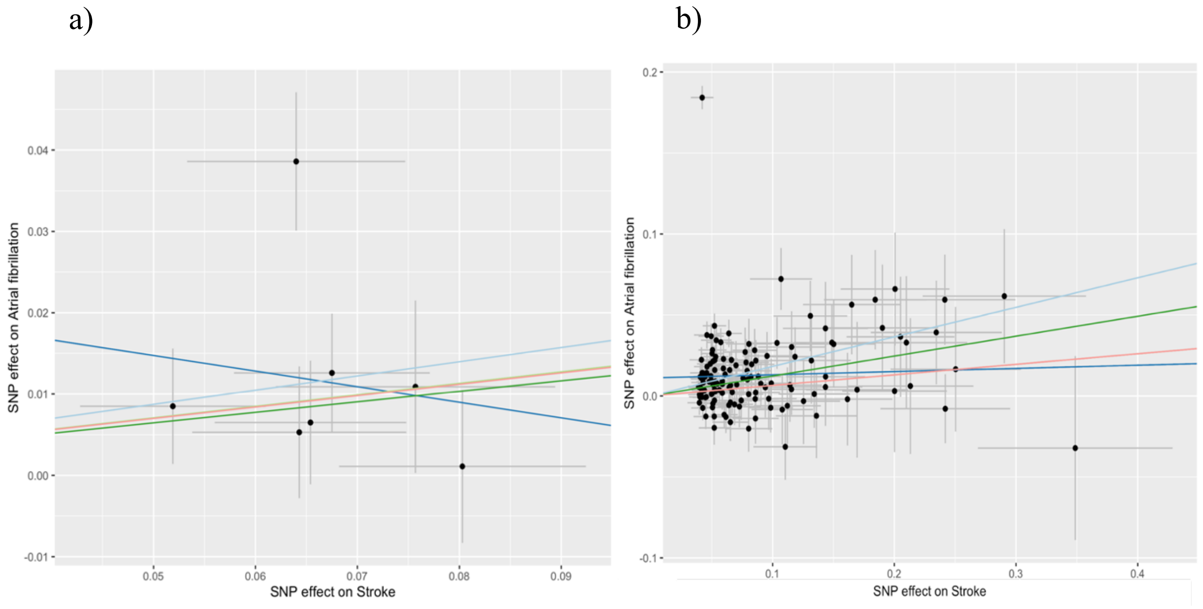

Figure 21. Validation MR study of the effect of stroke on AF.

(a and b) Scatter plots representing the MR relationships between instrument-stroke associations and the instrument-AF associations. (a) all stroke IVs at $p<5x10^{-8}$ and (b) at $p<5x10^{-5}$ cut-offs. On the x-axis is the SNP effects for the stroke associations and on the y-axis is the SNP effects for the outcome associations (on the log odds scale). Each point represents the individual SNP weights with the 95 CIs intervals plotted. The different coloured lines correspond to the slopes through these points calculated using each of the five MR approaches.

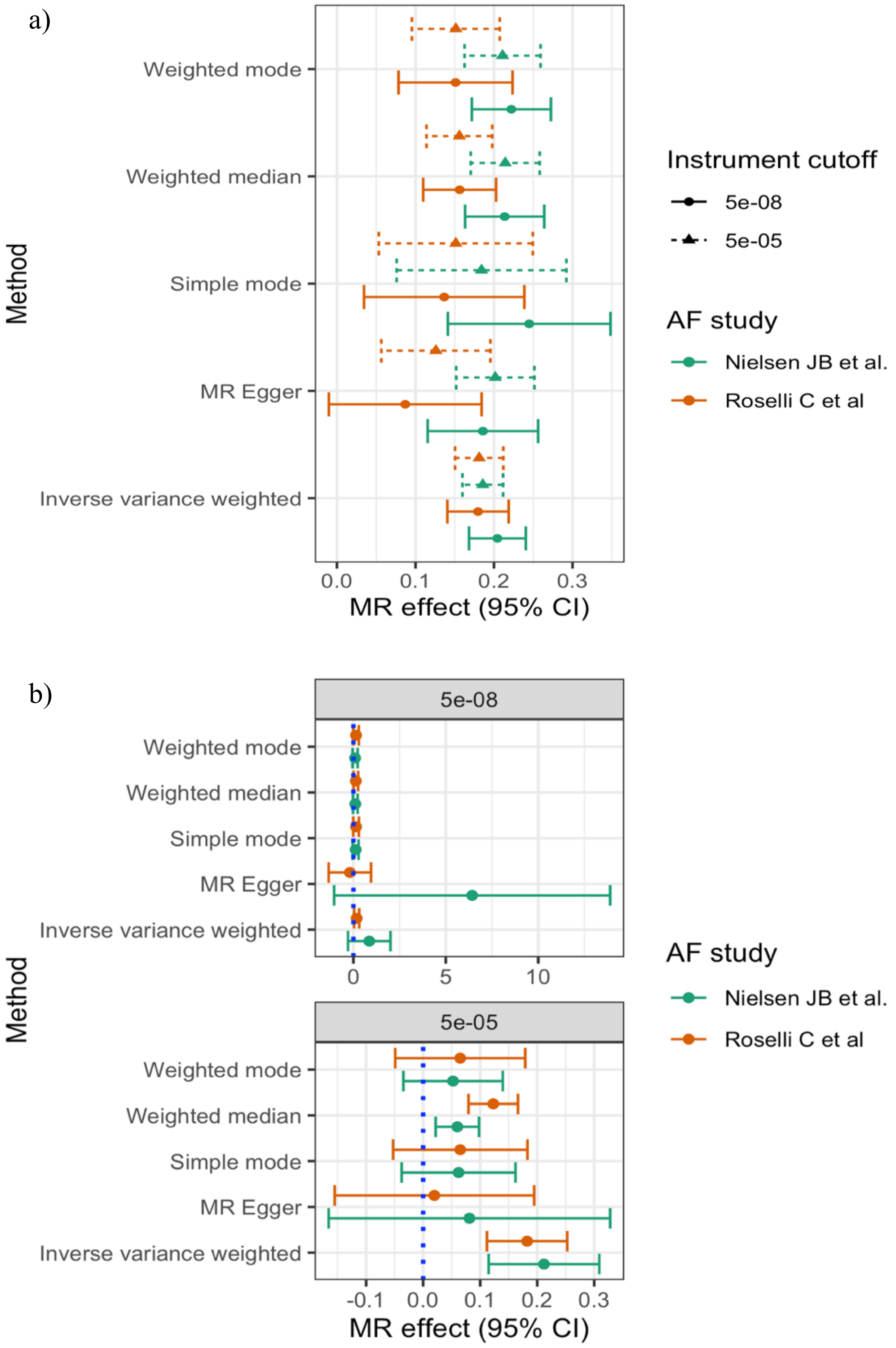

Figure 22. Error plots of results from two-sample MR studies.

(a) Error plot depicting two-sample MR estimates derived from different methods for the relationship between genetic predisposition for AF and stroke. (b) Error plot depicting reverse MR estimates for the relationship between genetic predisposition for stroke and AF. Green line represents main MR study (where AF GWAS from Nielsen et al was used) and orange line shows validation MR study (where AF GWAS from Roselli et al was used).

Table 19. MR Steiger results from steiger directionality test.

MR steiger estimates using two instrument cut-offs, P<5x10^{-8} and P<5x10^{-5}. Estimated r-squared for the exposure (i.e., the proportion of the variance in the exposure explained by the instruments) (SNP r^2 exposure) and the outcome (i.e., the proportion of the variance in the outcome explained by the instruments) (SNP r^2 outcome); MR analysis failed (FALSE) and passed (TRUE) the directionality test (correct causal direction) with p-value and outlier SNPs removed (TRUE) or not removed (FALSE). Steiger P represents the strength of evidence for a non-zero difference between the r^2 in the exposure and outcome.

Study	MR analysis	instrument threshold	SNP r^2 exposure	SNP r^2 outcome	correct causal direction	steiger P	outlier removed
Main	AF to stroke	5x10^{-8}	9.10x10^{-3}	9.36x10^{-4}	TRUE	4.20x10^{-289}	FALSE
	AF to stroke		9.07x10^{-3}	8.83x10^{-4}	TRUE	2.52x10^{-295}	TRUE
	stroke to AF (reverse)		7.48x10^{-4}	9.14x10^{-4}	FALSE	0.108	FALSE
	stroke to AF (reverse)		6.22x10^{-4}	2.80x10^{-5}	TRUE	5.14x10^{-28}	TRUE
	AF to stroke	5x10^{-5}	1.40x10^{-2}	1.65x10^{-3}	TRUE	0	FALSE
	AF to stroke		1.39x10^{-2}	1.55x10^{-3}	TRUE	0	TRUE
	stroke to AF (reverse)		7.64x10^{-3}	1.83x10^{-3}	TRUE	2.15x10^{-138}	FALSE
	stroke to AF (reverse)		7.42x10^{-3}	4.01x10^{-4}	TRUE	4.71x10^{-300}	TRUE
Validation	AF to stroke	5x10^{-8}	1.67x10^{-2}	8.39x10^{-4}	TRUE	0	FALSE
	AF to stroke		1.66x10^{-2}	7.74x10^{-4}	TRUE	0	TRUE
	stroke to AF (reverse)		7.48x10^{-4}	1.75x10^{-3}	FALSE	8.48x10^{-13}	FALSE
	stroke to AF (reverse)		6.22x10^{-4}	5.08x10^{-5}	TRUE	1.36x10^{-18}	TRUE
	AF to stroke	5x10^{-5}	2.43x10^{-2}	1.52x10^{-3}	TRUE	0	FALSE
	AF to stroke		2.41x10^{-2}	1.39x10^{-3}	TRUE	0	TRUE
	stroke to AF (reverse)		7.17x10^{-3}	3.66x10^{-3}	TRUE	3.23x10^{-33}	FALSE
	stroke to AF (reverse)		6.95x10^{-3}	6.83x10^{-4}	TRUE	9.23x10^{-177}	TRUE

6.3.4 Assessment of instrument heterogeneity and pleiotropy

A heterogeneity test was conducted on instruments to evaluate the potential pleiotropy of instruments used in MR analyses in this chapter. Instruments (n=8 and 162 at P<5x10^{-8} and P<5x10^{-5} thresholds, respectively) used in the main reverse MR analyses (stroke=>AF) exhibited strong evidence for heterogeneity in SNP effects (n=8 IVs, Cochran 's Q=1.115.98, P=7.28x10^{-238}; n=162 IVs, Cochran 's Q=2.609.84, P=1.00x10^{-200}), (**Table 20**) (**Figure 20a** and **b**), indicating that one or more of the instruments for stroke exposure might not be a valid IV, due to horizontal pleiotropy. There was also evidence of high heterogeneity amongst instruments in the main and validation MR analyses (AF=>stroke) (**Table 20**), suggesting that pleiotropy might be present, which could be an issue due to potential violation of MR assumption 2 and 3.

The MR-Egger intercept test was performed to identify directional pleiotropy. There was no evidence of directional pleiotropy in the main MR analyses (AF=>stroke) using the MR-Egger intercept test (α=0.002 [SE=0.003], P=0.549). The MR-Egger intercept did not differ from zero and the P value was very high, indicating little evidence for a non-zero intercept and therefore little evidence of directional pleiotropic effects of the SNPs on stroke. The MR-Egger intercept showed some weak evidence for a non-zero intercept (α=0.008 [SE=0.004], P=0.045) at IV threshold P<5x10^{-8} and (α=0.004 [SE=0.002], P=0.084) IV threshold P<5x10^{-5} in the validation MR analyses (AF=>stroke) (**Table 20**), indicating that the MR estimates may partly be biased by directional horizontal pleiotropy.

Table 20. MR-Egger intercept results and Q statistics from heterogeneity test on instruments.

MR-Egger intercept with standard error (SE) and p-value (P). Cochran's Q test statistics with Q value and P value for the MR-Egger analyses.

						MR-Egger	
Study	**MR analysis**	**instrument threshold**	**Egger intercept**	**SE**	**P**	**Q**	**Q P**
Main	AF to stroke	$5x10^{-8}$	0.002	0.003	0.549	201.24	$1.86x10^{-7}$
	stroke to AF (reverse)	$5x10^{-8}$	-0.382	0.259	0.190	1,115.98	$7.28x10^{-238}$
	AF to stroke	$5x10^{-5}$	-0.001	0.001	0.463	460.69	$5.95x10^{-6}$
	stroke to AF (reverse)	$5x10^{-5}$	0.009	0.008	0.261	2,609.84	$<1.00x10^{-200}$
Validation	AF to stroke	$5x10^{-8}$	0.008	0.004	0.045	145.00	$1.26x10^{-4}$
	stroke to AF (reverse)	$5x10^{-8}$	0.024	0.039	0.556	12.34	0.030
	AF to stroke	$5x10^{-5}$	0.004	0.002	0.084	363.42	$2.40x10^{-6}$
	stroke to AF (reverse)	$5x10^{-5}$	0.011	0.006	0.049	823.43	$5.02x10^{-100}$

6.3.5 Outlier exclusion and leave-one-out sensitivity analyses for the reverse MR

Two outliers (rs2634074 and rs6838973) which might be exhibiting horizontal pleiotropy were initially identified by manually inspecting the MR scatterplot from performing genome-wide reverse MR selecting stroke IVs at a $p<5x10^{-8}$ (**Figure 23a**) and $p<5x10^{-5}$ threshold (**Figure 23c**). These SNPs were then removed manually, and the main reverse MR (stroke=>AF) was repeated with the remaining stroke instruments across all the methods (**Figure 23b** and **d**). Outlier exclusion analysis was not performed in the forward direction from AF to stroke. Both SNPs when removed showed little evidence of changing the MR estimates other than for the IVW (**Table 21**). For the IVW when both SNPs were removed, the causal effect evidence attenuated in magnitude but was more precise (IVW beta=0.212, SE=0.049, $P=1.83x10^{-5}$ to IVW beta=0.126, SE=0.019, $P=1.79x10^{-11}$). In **Appendix E**, MR estimates for all the methods with the outliers removed are provided. In addition, LOO analysis was performed on the IVW method across all SNPs. On visual inspection of the LOO plot (**Figure 24a** and **b)**, the MR estimate with the rs2634074 and rs6838973 SNPs removed showed the clearest deviation from the overall IVW estimate, with no other additional outliers being highlighted.

PheWAS analyses were then conducted on rs2634074 and rs6838973 against all traits in the IEU OpenGWAS database[10] (using P=1x10^{-5} as a p-value threshold) to find if they are associated with any traits which could explain horizontal pleiotropy. The stroke-increasing alleles of these variants, rs2634074 (effect allele (T)) and rs6838973 (effect allele (C)) (located on chromosome 4) were strongly associated with increased risk of AF. PheWAS confirmed that these two SNPs are associated with cardiovascular traits (such as Arrhythmia, AF, stroke and its subtypes). No evidence of horizontal pleiotropy was detected. Results from PheWAS on outlier SNPs are shown in **Appendix F**. These SNPs were also identified as outliers in the Steiger analysis between stroke and AF (**Appendix C**), confirming the strong AF effect for these SNPs relative to stroke.

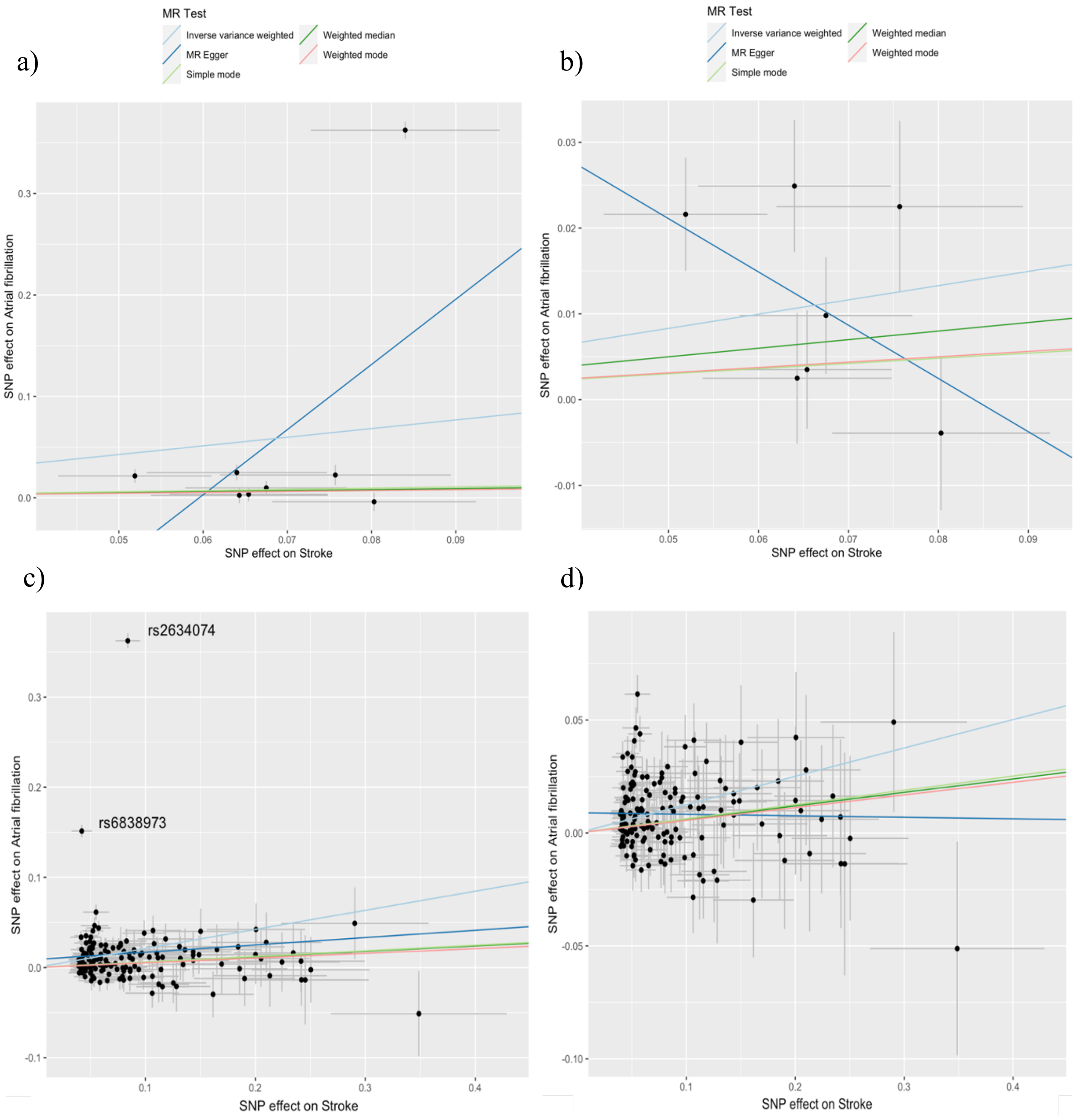

Figure 23. Main MR study and outlier exclusion analysis of the effect of stroke on AF.

(a-d) Plots showing the relationship between the effect magnitude of the instrument-stroke associations (on the x-axis, Beta coefficient) and the instrument-AF associations (on the y-axis, Beta coefficient) with standard error bars. The lines with slopes correspond to estimates of the causal effects calculated using each of the five MR approaches using stroke IVs at $P<5x10^{-8}$ (a and b) and at $P<5x10^{-5}$ (c and d). Reverse MR plots with outlier represented (a) and with outlier removed (b). Plots with 2 outliers shown (c) and exclusion of both outliers from reverse MR analysis (d).

Table 21. MR estimates for SNPs excluded in the leave-one-out sensitivity analysis.

Two outlier SNPs with influence on the IVW estimate detected in the main reverse MR analysis (rs2634074 and rs6838973). rs6838973 outlier detected only in the validation reverse analysis (rs2634074 outlier SNP was not present in this MR analysis due to being excluded in the harmonisation step). Inverse Variance Weighted (IVW) and Wald ratio (WR) (statistic) with MR estimate: effect size (Beta), standard error (SE) and p-value (P). IVW represents overall effect of stroke on AF. WR represents the ratio effect estimate of the specific outlier SNP on AF.

			MR estimate	
MR analysis	**statistic**	**Beta**	**SE**	**P**
stroke to AF (main reverse MR)	IVW (all SNPs)	0.212	0.049	1.83x10^{-5}
	rs6838973 WR	3.596	0.159	1.03x10^{-111}
	leave out rs6838973 IVW	0.192	0.045	2.26x10^{-5}
	rs2634074 WR	4.315	0.098	1.00x10^{-200}
	leave out rs2634074 IVW	0.146	0.028	2.04x10^{-7}
	leave both out IVW	0.126	0.019	1.79x10^{-11}
stroke to AF (validation reverse MR)	IVW (all SNPs)	0.182	0.036	3.70x10^{-7}
	rs6838973 WR	4.375	0.171	1.35x10^{-142}
	leave out rs6838973 IVW	0.153	0.019	2.44x10^{-15}

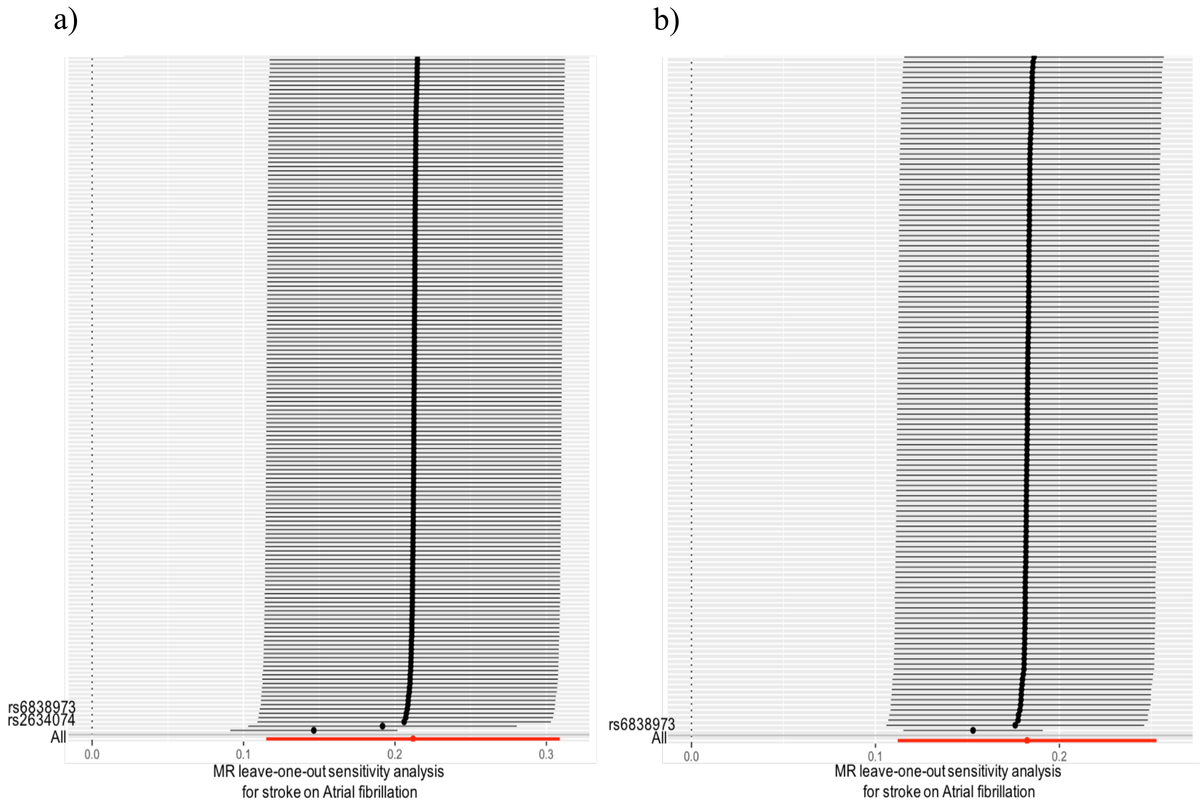

Figure 24. Leave-one-out sensitivity analysis of the effect of stroke on AF.

Each black point in the forest plots depicts the IVW estimate for the causal effect of stroke on AF excluding the specific SNP from the analysis (at IVs threshold of $P=5x10^{-5}$). The red point represents the IVW estimate combining all variants. The removal of the specific SNPs, rs2634074 and rs6838973 from main reverse MR analysis (a) and rs6838973 from validation reverse MR (b) resulted in attenuation of the overall IVW estimated causal effect depicted as the black point further from the centre).

6.4 Discussion

6.4.1 Main findings

In this chapter two-sample MR analysis was conducted to identify if there is evidence of a causal relationship between AF and stroke. MR was utilized to infer causality by mitigating confounders of the exposure-outcome association. A range of sensitivity analyses were conducted to determine the robustness and validity of this conclusion. A rigorous outlier analysis was performed which included visual inspection of scatter and LOO plots, Steiger filtering, and PheWAS of the outliers. These analyses confirmed that removal of the outliers was justified (i.e., they are highly likely to be true outliers) and when the outliers were removed it had little influence on overall MR inference. Therefore, the MR analysis will be relatively robust to horizontal pleiotropy strongly acting through specific SNP instruments. The MR findings suggested that increased genetic predisposition to AF was associated with an increased risk of stroke. MR analyses of AF and IS (which account for approximately 85% of all cases of stroke[64]) also showed a similar pattern to AF and stroke results. This result is consistent with previous findings that genetic susceptibility to AF is linked to increased risk of stroke, IS and CES subtype showed by two-sample MR and phenome-wide multi-directional MR studies of AF[421,72]. The two-sample MR study by Fill et al has identified genetically determined AF as causally linked to higher risk of stroke, IS and CES subtype but not the other two IS subtypes (LAS and SVS)[421].

This result was in agreement with the findings that revealed causal role of AF in stroke and IS by Wang et al in the univariable and MVMR[72]. Furthermore, they have performed univariable and MVMR to assess independent causal role of not only AF but also AF risk factors such as

CAD and SBP on risk of three IS subtypes and found that genetic predisposition to AF is strongly associated with higher risk of CES in both models. This group also found MR evidence for an independent causal effect of AF on CES in MVMR adjusted out for these risk factors. No evidence of genetic predisposition to AF association with LAS and SVS subtypes was found in any of the MR models[72]. Hou et al conducted a network MR analysis and found a bidirectional causal relationship between AF and IS[422]. In agreement with the other two studies, Hou et al also identified a CES-specific effect of AF after performing MR on all three subtypes of IS. This study also reported on a potential presence of the causal role of IS on AF mediated through blood pressure (BP) (i.e., IS=>BP=>AF)[422].

The results of the sensitivity analyses did not identify any major issue which would substantially affect the causal inference. The directionality of MR effects was assessed by bidirectional MR[219] and Steiger filtering[425,424] conducted for all MR findings. MR estimates for both the forward and reverse direction remained similar after removing the SNP outliers identified through Steiger analysis. MR findings in the main analysis were consistent with the validation MR conducted on the Roselli et al AF GWAS[52] and Malik et al stroke GWAS[67] (which had small sample overlap), confirming that this AF relationship is robust to confounding from sample overlap between AF and stroke.

Reverse MR results at $5x10^{-8}$ did not show any evidence of effect but this might be due to the small number of SNPs, however at $P=5x10^{-5}$ with more data points there was evidence for IVW and Weighted median. This result is in line with Hou et al two-sample MR study, which found a bidirectional causal relationship between AF and IS[422]. Therefore, there is some ambiguity regarding the direction of causal effect. However, the more relaxed $P=5x10^{-5}$ analysis does come with the drawback of potentially being less directionally robust (as you relax the p-value threshold more AF hits passing will pass the threshold for stroke). Moreover, after Steiger filtering was conducted in this thesis, which removed the stroke instruments driving the reverse effect, the AF to stroke relationship persisted, which does indicate more strength of evidence for the AF=>stroke direction.

The two variants, rs2634074 and rs6838973 identified in the LOO analyses (stroke=>AF) are AF SNPs that have such a large effect that they appeared in the stroke GWAS[37,431]. The rs2634074 SNP located in the 4q25 region upstream of the *PITX2* gene is in high LD with rs67249485 (r^2=0.98 in the 1000G EUR population), the strongest AF-associated risk variant (P=7.32x10^{-443}) identified by Nielsen et al[41]. The rs2634074 SNP have been previously reported by Pulit SL et al[111] as a risk variant associated with IS (P=3.00x10^{-14}) and its subtypes in a IS GWAS conducted in SiGN cohort[90] (a multi-ancestry cohort). This SNP is also in high LD with rs13143308, the stroke-associated variant (P=1.61x10^{-13})[67] at this locus (r^2=0.93). The rs6838973 variant is an independent AF variant identified by Lubitz et al after a conditional analysis conducted on the 4q25 locus[37]. This SNP is in moderate LD with an independent AF risk variant, rs3853445 (r^2=0.40) reported in the AF GWAS by Nielsen et al[41] (P=3.73x10^{-52}) and by Roselli et al[52] (P=6.00x10^{-113}). The effect alleles, T and C for rs2634074 and rs6838973 SNPs were identified to be associated with an increased risk of AF by PheWAS conducted in this Chapter. This is consistent with the previous findings by Lubitz et al[37] and Kiliszek et al[431], confirming the protective role of these variants. In addition to these two SNPs, the rs4151702 variant also failed the Steiger directionality test. The rs4151702 variant is in high LD with the AF SNP, rs3176326 (intronic to *CDKN1A*) (r^2=0.97) identified by Nielsen et al (P=1.42x10^{-13}) in the 6p21 region. rs4151702 is in perfect LD (r^2=1) with the rs730506 variant highly associated with the electrocardiographic PR interval correlated with AF[432,433].

6.4.2 Strengths and limitations

The selected instruments for AF exposure (n=111) explained 4.6 % of the variation in atrial fibrillation (Nielesen et al), corresponding to a F statistic of 90.6 in the main MR analysis. The instrument strength for all the MR analyses was determined with F statistic >10, indicating that instruments are well enough powered and unlikely to be vulnerable to weak instrument bias or bias due to overlap of samples[434]. In addition, validation analysis in the less overlapping datasets was performed to avoid induction of bias due to sample overlapping between AF and stroke GWAS[233]. To ensure the robustness and validity of the findings to avoid violation of MR estimate and influence of pleiotropy on MR estimates, a range of sensitivity analyses were performed to reduce heterogeneity. Moreover, findings were compared with validation results and across five different MR methods using two instrument cut-offs to check their consistency with each other.

6.4.3 Future directions

Further analyses would be beneficial in the availability of larger scale GWASs with non-overlapping datasets or participants.

6.4.4 Conclusions

In conclusion, the two-sample MR results in this chapter demonstrate a relationship between genetic susceptibility to AF and stroke risk. This result is consistent with the three other MRs conducted between AF and stroke in the literature, which indicates that AF is a risk factor for stroke. In Chapter 3 and 4 5 *HP* gene expression colocalized with both AF and stroke, and in Chapter 5, AF and stroke colocalized at one single causal variant at the 16q22 locus, therefore MR supports the possibility that *HP* acts downstream to influence the AF=>stroke pathway.

Chapter 7 Discussion

This doctoral thesis aimed to explore the shared molecular aetiology of two cardiovascular diseases, AF and stroke, using a multi-omics approach and a range of existing datasets. The approach taken here is generalizable, so could be applied to a range of other disease areas and is likely to yield more informative results as new large-scale molecular datasets become available.

7.1 Main findings

7.1.1 Multiple trait colocalization on AF

In **Chapter 3,** a multi-trait colocalization analysis pipeline was developed and applied to multiple molecular QTL and AF summary statistics. 23 loci with evidence of colocalization between AF and one or more molecular traits (CpG sites, genes and metabolites) were identified. Of these, 3 loci were found to share genetic effects across all three traits (PPA.GMMb $\geq$ 80%). In phase II, ten of these 23 loci were further prioritized by showing colocalization evidence with gene expression, including 23 GME scenarios at 9 loci, 3 GMEMb scenarios at 2 loci, and 4 GEMb.M scenarios at 1 locus. The multi-trait colocalization analysis on AF demonstrated shared genetic regulatory effects between multiple CpG sites, genes and circulating metabolites implying a complex relationship of changes in DNA methylation, gene expression and metabolite levels underlying AF susceptibility. This is consistent with a 2SMR study which revealed that genetic effects on 47 of the 139 tested complex traits (including CVDs) may be mediated through changes in DNA methylation[188]. A follow-up study showed that 306 of the 348 DNA methylation-trait associations were shared with eQTLs. MR mitigates reverse causation and this study did not find any evidence of a direction of effect from complex traits to methylation levels of a CpG[188]. This suggests that levels of methylation at CpG sites are unlikely to be influenced by complex traits. However, this conclusion could not really be evaluated comprehensively by Richardson et al[188] due to the small sample sizes[207] and hence low statistical power when using the mQTL data as an outcome in their reverse MR. Colocalized intermediate phenotypes may have coordinated regulatory roles in the pathway from genetic variant to AF. In coordinated regulation (a causal model), methylation at CpG sites influences gene expression which consequently affects metabolite levels and disease risk. Of note, colocalization does not establish the direction of molecular effects and is unable to examine the potential impact of horizontal pleiotropy. Some of these prioritized genes were enriched for biological processes/pathways relevant to CVDs and AF which makes them promising candidates for follow-up experimental analysis and drug development. Biological relevance of these genes to AF are discussed in section **7.3**.

My moloc analysis also indicated that the prioritized genes are not necessarily the ones in closest physical proximity to the AF lead variant in the associated regions, consistent with previously published studies[22,23], and highlighting the importance of this type of integrative multi-omics analysis.

7.1.2 Multiple trait colocalization on Stroke

Having established an effective multi-trait colocalization pipeline, this was then applied to stroke in **Chapter 4**, as evidence from the literature suggests that this disease is related to AF and may therefore share molecular pathways. Evidence for sharing of regulatory effects between methylation CpG sites and multiple clusters of lipoprotein lipids was detected at eleven of the 20 loci in phase I. In Phase II, of the 11 loci, five (6p21 (stroke), 9q34 (IS), 12q24

(IS), 16q22 (CES), 19p13 (stroke)) also showed evidence of colocalization with gene expression, suggesting that these intermediate phenotypes are likely to be involved in risk of stroke.

7.1.3 Follow-up analyses of the 16q22 locus

Multi-trait colocalization analysis on AF and stroke in **Chapter 3** and **Chapter 4** showed that DNA methylation at the cg03463523 CpG site and *HP* gene expression colocalized with both AF and stroke in the 16q22 region. In **Chapter 5**, pairwise coloc analyses of AF and stroke at the 16q22 locus detected that AF and stroke showed moderate evidence of colocalization in this region (PP_4=75%). Interestingly, PWCoCo (which accounts for multiple independent causal variants) confirmed strong evidence of sharing the same causal variant (rs2359171, PP_4=81.4%) for AF and stroke at this locus. The secondary genetic variant for AF (rs876727) at this locus was not associated with stroke and was only weakly associated with expression of the *HP* gene (P=0.0016) and was not identified to be colocalized between AF and stroke. In contrast, the rs876727 variant was associated with *ZFHX3* expression ($P=2.57x10^{-7}$) in the eQTLGen data, however no colocalization evidence was detected for this gene and AF or stroke suggesting that the link to *ZFHX3* is likely to be due to linkage disequilibrium rather than a causal role in AF. Furthermore, the pQTL mapping of haptoglobin revealed no association between the top AF colocalizing SNP (rs2359171) and plasma haptoglobin levels in ALSPAC, which indicates uncertainty that the circulating levels of the HP protein contribute to the risk of AF and highlights the complexity of biological mechanisms underlying 16q22 colocalizations. This is consistent with findings revealed by Sun et al and Assum et al that blood eQTLs and pQTLs for plasma proteins do not always overlap[410,257]. In addition, a recent study have found about 50% overlap between protein and expression effects[435].

7.1.4 Two-sample MR between AF and stroke

In **Chapter 6**, MR analysis suggested that the genetic predisposition to AF was linked to increased risk of stroke. These MR results were largely in agreement across all 5 MR methods indicating findings are robust to bias via potential horizontal pleiotropic mechanisms. However, there was some evidence for a bi-directional MR effect, demonstrating that the AF effect is partly driven by stroke variants making orientation of the causal direction difficult to discern. Validation analyses showed consistent results with the main analyses. These results are in line with previous findings, where genetic liability to AF causes stroke, IS and CES, but not SVS and LAS[421,72]. Wang et al[72] found that genetic predisposition to risk factors of AF (such as CAD and SBP) and AF associated with risk of stroke, IS and CES in univariable analyses, and with risk of stroke and IS in a multi-variable MR model. However, only the genetic predisposition to AF retained an independent causal effect for CES in the MVMR adjusted for the AF risk factors. These findings may indicate a potential mediatory role of AF in the causal pathway for the CES subtype (i.e., SBP/CAD=>AF=>CES). In contradiction with these results is a bidirectional causal relationship between AF and IS that has been reported by Hou et al[422].

7.2 Using a multi-omics approach to identify causal genes in AF and stroke

Previous studies have shown that colocalization can be informative in identifying potential mechanisms of disease[436,150,437]. In this thesis, I show that this approach could pinpoint potential mechanisms but that it also indicates complex relationships between gene expression and DNA methylation due to co-expression or co-methylation of multiple genes or DNA methylation sites.

In **Chapter 3**, mQTLs and eQTLs were integrated with the aim to identify potential causal gene(s) at a genomic locus. Of the 41 CpG sites associated with AF risk, 12 also colocalized with at least one gene. Of the 10 AF loci, two (7q32 and 16q22) had a single gene connected to disease risk. When multiple genes colocalize with a single CpG site it is difficult to determine which gene might be influenced by DNA methylation of that CpG site. For example, in the 1q21 region, 4 genes (*DAP3, SYT11, YY1AP1* and *MSTO2P*) were colocalized with one single CpG site (cg19233405), which may be due to co-expression of genes in the same pathway. This was the case for 6 out of the 12 CpG sites. Similarly, multiple CpG sites colocalized with a single gene which might be due to co-methylation of blocks of correlated CpG sites which might be under genetic control[268]. For example, the *RBM28* gene at the 7q32 locus colocalized with two correlated CpGs (cg18693656 and cg13951589) which are both negatively associated with mQTL. These findings are consistent with expression quantitative trait methylation (eQTM) analysis in other studies[185,128]. These studies showed that DNA methylation at multiple CpG sites correlated with gene expression levels of a single gene and that DNA methylation and gene expression were both negatively and positively correlated[128]. However, it has also been shown that not all variation at CpG sites affects gene expression[185,128].

Furthermore, most CpG site(s) colocalized in this thesis were found to be located distal from the gene body of the colocalized gene(s) (i.e., they are not CpG-annotated genes (**Appendix A)**), which is in disagreement with the findings previously reported by Richardson et al and Taylor et al where CpGs were enriched in promoter regions and histone mark peaks in their colocalization analysis[188,255]. However, these studies used tissue-specific DNA methylation and gene expression data. Taken together, although most CpG sites which colocalize with gene expression are likely to be located within the gene body or promoter regions and enhancers, there are many biological mechanisms and relationships between intermediate phenotypes which might explain colocalization in this thesis and challenge the use of genomic annotations based on physical proximity.

Molecular pleiotropy is often an issue with QTL studies, which is when *cis*-eQTLs for different genes within the same genomic region can be mapped to the same disease endpoint, challenging attribution of the causal gene. For example, at the 10q22 locus, four genes (*CAMK2G, P4HA1, MRPS16* and *BMS1P4-AGAP5*) and at the 6p21 locus, four genes (*SRF, ABCC10, RPL34P14, AL035587.1*) colocalized with AF. In my study, this was an issue for 7 of the 10 AF loci and 4 of the 5 stroke loci which had multiple genes connected to AF and stroke risk respectively. This demonstrates that combining mQTL and eQTL information does not completely help to resolve the gene at a locus.

In addition, the QTL findings were also combined with metabolite data, to triangulate the genes with the metabolic processes which are important to stroke and AF. In fact, combining mbQTL with eQTL data helped to narrow down the likely causal gene at the locus. For instance, of the 8 AF loci found with multiple genes colocalized (**3.3.2**), two (10q22 and 17q12) also showed GMEMb scenario where a single gene linked to a single metabolite and AF risk. A limitation of this analysis is that the ALSPAC metabolite QTLs had a relatively small sample size, which limited the ability for the AF moloc analyses to detect colocalized metabolic measures that are likely to contribute to the risk of AF.

In contrast, using UCLEB mbQTL data facilitated moloc analyses with substantial power to detect potential candidate metabolites including lipids and lipoprotein subclasses likely to contribute to the risk of stroke through influencing metabolism. Of the 11 colocalized loci, three (9q34, 12q24 and 19p13) colocalized with various cluster of lipoproteins with likely roles in stroke susceptibility. These three loci also colocalized with potentially causal genes likely to play a functional role and be involved in pathways relevant to stroke pathogenesis (see section **7.3**). For example, in the 9q34 region, multiple genes (*GBGT1, CACFD1, ABO*) colocalized with multiple metabolites (APOB, LDL cholesterol, different lipoprotein subclasses) which support findings of correlated metabolites detected by previous studies[327]. Furthermore, metabolites can be difficult to interpret as many are strongly correlated with each other making it harder to resolve the causal metabolite. For example, MVMR has been applied to determine the independent causal effect of each lipid component on CVDs, and found that the adverse LDL-C effect is explained through APOB levels[327].

7.3 Biological function of the prioritized genes

14 of the 29 genes highlighted by moloc were implicated in biological pathways related to cardiovascular phenotypes which consists of 9 genes with a potential role in AF (*GATAD1*, *PSEN1*, *YY1AP1, ERBB2*, *CHI3L1*, *KRIT1*, *CAMK2G*, *IGFBP4* and *JMJD1C)* and 5 genes in stroke (*SRF*, *CACFD1*, *HVCN1*, *TRAFD1* and *SMARCA4*), where there is published data of being a drug target (for any indication) or evidence for having a monogenic mutation in humans or mouse knockdown/knockout with a cardiovascular phenotype. For AF, the *ERBB2*[438,439] and *PSEN1* genes were identified as drug targets and *GATAD1*[282] and *YY1AP1*[286,285] have monogenic mutations. *ERBB2* is an anti-cancer drug target and *PSEN1* is a dementia and

Alzheimer's disease drug target. For stroke, none of the 5 genes were identified as drug targets or with monogenic mutations.

For AF, genes with such evidence include: *GATAD1* and *PSEN1*, linked to dilated cardiomyopathy[284,282,277,280] and implicated in biological process linked to cardiac development and in the regulation of intracellular trafficking of angiotensin receptor and subsequent heart function[440,441]; *YYA1P1,* associated with early onset hypertension and/or haemorrhagic stroke with Grange syndrome[289]; *ERBB2,* likely to be involved in cardiac development (e.g., trabeculation), regeneration and its electrical activity[442,443,444,274]; *CHI3L1*, linked to atherosclerotic plaques buildup in arteries and carotid artery disease[445], and with altered protein expression levels found in blood plasma of patients with CAD[446,447,448]; *KRIT1,* likely to be involved in cardiac morphology or morphologic features of the heart; *CAMK2G,* likely to be involved in heart's intracellular calcium (Ca^{2+}) handling[449,450,451]; *IGFBP4,* which is involved in the process of cardiomyogenesis[452] and angiogenesis[453]; and *JMJD1C,* which may regulate cellular metabolism[454] is involved in angiogenesis[303] and development of megakaryocyte and platelet production[455]. *JMJD1C* interacts with AR (androgen receptor) gene which is targeted by an approved drug (as an agonist) for heart failure (Open Targets annotations).

Interestingly, Folkerson et al also prioritized the CH13L1 gene in a protein QTL GWAS[456]. Consistent with findings in this thesis, Folkersen et al conducted colocalization of Olink CVD-I pQTLs using PrediXcan[214] and MR analysis and not only identified association of CHI3L1 protein levels with its corresponding mRNA expression in GTEx tissues but also showed causal evidence for contribution of CHI3L1 protein in AF with no evidence of inverse direction of causality[456], confirming CHI3L1 as a promising target for drug discovery. In addition, differentially expressed proteins levels of CHI3L1 have been found in patients diagnosed with advanced atherosclerotic plaques or CHD compared with healthy individuals[446] and its elevated levels as inflammatory biomarker have been revealed to be associated with CAD[457]. Interestingly, population-based study showed association between increased levels of plasma CHI3L (YKL-40) and high risk of AF[458].

For stroke, genes with published evidence of a role in cardiovascular disease include: *SRF*, involved in angiogenesis, integrity of small vessels and cardiogenesis[335,459]; *CACFD1* (Open Targets annotations)*, HVCN1* and *TRAFD1* genes are involved in immune pathways and inflammation and have been linked to vascular phenotypes such as stroke or cerebral infarction*,* and also likely to have a role in homeostasis and metabolism (Open Targets, Mouse Phenotypes annotations)[339,341,342,345,346,349]; *SMARCA4* which is likely to be involved in structural development of the heart[354,353] and differentially expressed exons of *SMARCA4* gene have been reported in blood of patients with small vessel disease (cause of SVS) compared to controls using whole transcriptome microarrays[374].

ABO gene, which colocalized with LDL levels subclasses and stroke in this thesis, also emerged as a potentially interesting candidate due to its well-known role in lipid metabolism. *ABO* gene is located in the 9q34, encodes blood group, and plays a role in pathways related to lipid metabolism including globo sphingolipid metabolism and glycosphingolipid biosynthesis[460]. Consistent with my findings, Chong et al conducted MR between circulating proteome and three subtypes of IS and found biomarkers including histo-blood group ABO system transferase to have a causal mediatory role along the pathway to CES[461]. They also identified causal effects of apolipoprotein A on LAS subtype. Moreover, the mediatory role of total cholesterol has been previously shown along the pathway from the ABO blood group to CAD by mediation analysis adjusted for CVD risk factors[462].

7.4 Shared genetic pathway between the *HP* gene and AF and stroke

Colocalization can be used to inform us about shared genetic architectures[463]. This thesis found colocalization evidence connecting *HP* expression to both AF and stroke risk, thus a possible hypothesis is that AF might be an intermediate step before stroke and this pathway is partly mediated by the *HP* gene (*HP*=>AF=>CES). Haptoglobin is involved in preventing iron loss while allowing degradation of hemoglobin via binding to free hemoglobin in blood plasma. Elevated serum levels of haptoglobin were reported in patients with acute and chronic inflammatory diseases with potential anti-inflammatory function[464,465,466]. Haptoglobin protein has been suggested to be involved in cardiovascular disease development and arterial restructuring[467,468,386]. Moreover, prohaptoglobin (proHp) has also been shown to be involved in the angiogenic process in endothelial cells[469,470]. The causal role of haptoglobin in lipid metabolism and macroangiopathy has also been revealed through MR analysis in patients suffering from metabolic traits[471,472]. A previous study has reported that high cholesterol is linked to increased levels of inflammation-sensitive plasma proteins (ISP) (including haptoglobin) associated with higher risk of IS in men[417]. Circulating HP as an inflammatory biomarker has been linked to metabolic and cardiovascular traits including carotid arterial functions and stroke[473,474,475,476,415]. Indeed, HP is an established risk factor for stroke and its circulating concentrations have also been shown to be predictive for stroke and IS[415]. Studies in patients with AF revealed that inflammation linked to perpetuation of AF[414]. Taken together, these findings indicate the importance of *HP* gene as an inflammatory biomarker for metabolic-associated cardiovascular diseases such as AF and stroke.

In addition to the *HP* gene, the zinc finger homeobox 3, *ZFHX3* has also been investigated as a potential causative gene for AF at the 16q22 locus[87,431,41,67]. *ZFHX3* encodes a transcription factor (TF) which is expressed abundantly in atrial tissue and involved in development[395,396,477]. ZFHX3 was shown to be involved in inflammatory signaling pathways in a porcine model of AF induced by pacing[478]. The ZFHX3 TF has also been reported to contribute to regulation of nuclear protein SUMOylation in epithelial cells[479,480,481]. A study integrating epigenetic data, chromatin states, eQTL data along with GWAS of AF in left atrial tissue found a gene interaction network regulated by *ZFHX3* expression and chromatin elements underpinning AF[482]. However, as *ZFHX3* gene expression did not colocalize with AF risk at this locus, the findings in this thesis support *HP* as being the potential causal gene for AF susceptibility instead. However, it is possible that *ZFHX3* could have a causative role in pathways involved in AF progression which was not investigated in this thesis.

For this locus it was also evaluated whether the *HP* gene expression signal overlapped with HP protein abundance (pQTL). In the primary analysis, mQTL-cg03463523 and eQTL-*HP* gene colocalized with AF and stroke risk but secondary analysis showed little evidence for AF variant association with haptoglobin protein level (i.e., rs2359171-*HP cis*-eQTL is not a *cis*-pQTL), inconsistent with the simple model that gene expression influences protein levels. As it is relatively common for eQTLs and pQTLs to not overlap[410,-], this does not preclude *HP* as the candidate causal gene, but does suggest that complex transcriptional control could be happening at this locus, which would need further studies to resolve. Furthermore, the effect on protein levels (and AF) might only occur in older people as AF begins to develop.

Interestingly, in a phenome-wide MR study of the proteome, Zheng et al[240] found a shared effect between plasma haptoglobin pQTLs and LDL cholesterol. However, in the moloc analyses conducted in this thesis, no colocalization was detected between *HP* expression, AF and circulating LDL cholesterol. Therefore, a possible explanation for the different effects observed between gene expression and protein abundance, could be that *HP* expression is not mediated through lipid metabolism pathway, whereas the HP protein abundance is, which again would need further studies for clarification.

In addition to the *HP/ZFHX3* genes in the 16q22, the other gene which has been postulated in the literature[97,87] as being involved in the AF to stroke pathway is *PITX2* which is located in the 4q25 genomic region[97,107]. The *PITX2* region has the strongest GWAS association signal with AF[41]. However, no evidence of colocalization between *PITX2* expression and AF or stroke was detected in this thesis either.

MR was used to evaluate whether AF is a causative risk factor for stroke. Although MR did confirm a positive relationship between genetically predicted AF and stroke liability, the direction of effect was ambiguous, therefore it remains uncertain whether AF increases the liability to stroke or vice versa. Therefore, it is reasonable to conclude that *HP* could be involved in a shared molecular pathway for AF and stroke, but there is a lack of clear evidence to be able to attribute a causal direction to this (i.e., that AF is an intermediate phenotype whereby *HP* influences AF risk which consequently influences stroke which was the original hypothesis).

7.5 Context-specificity of molQTLs

Tissue and cell type specific effects of QTLs are important to consider. In this thesis, I used molQTL profiled in whole blood tissue as this had the largest sample size and best statistical power to detect effects, but this tissue is not necessarily the most disease relevant one (such as heart and atrial tissues) for some of the underlying biological pathways due to the fact that specific CpG sites and genes might be differentially regulated between whole blood and different tissues or cell types[255,237].

Overall, for 88 out of the 111 AF loci and 11 out of the 22 stroke loci we did not find evidence of colocalization with either methylation or metabolite and therefore could not map to a gene. These loci were not taken forward for further moloc analysis in this thesis (i.e., I excluded genes which did not have mQTL or mbQTL evidence). In a previous study, Malik et al[67] performed lookups of stroke-associated variants in public and nonpublic databases of genetic associations in different stroke-relevant tissues (including blood and immune cells) and found overlap of stroke risk variants with QTLs associated with expression of different genes (or protein levels or methylation) either nearest to the lead SNP or more distant from the lead SNP, supporting the plausibility of detecting different pleiotropic mechanisms across cell types and tissues. This suggests that the lack of colocalization in this study may be due to tissue-specific effects, for example colocalization analysis of QTLs in heart, atrial, brain, aortic endothelial cells and immune cells may be more informative in identifying the potential link between different molecular phenotypes and in prioritizing candidate CpG sites and causal genes. Furthermore, blood has a variety of cell types and cell type specific effects are missed due to the measurement of bulk tissue.

7.6 Strengths

This research work has several strengths. One of the key strengths is the sample size and statistical power from combining some of the largest molecular QTL datasets published along with large-scale summary statistics of AF and stroke. The scale of these data ensured that the moloc analyses were well powered to identify evidence of colocalization between combinations of traits at AF and stroke GWAS loci. Another strength of this research is the integration of multiple molecular traits in a multi-trait colocalization framework to establish shared aetiology, which provides more mechanistic insights than a pairwise colocalization with a single molecular trait.

7.7 Limitations and future directions

The current analysis pipeline could be extended by investigating the role of different types of QTLs as well as looking at a broader range of phenotypes. For example, a potential future work would be to additionally integrate pQTL effects across the proteome, which could be done via incorporating the UK Biobank O-link proteomics data, which are currently in progress but not yet available. Other QTLs which would be useful to consider when datasets of a sufficient sample size become available, are splicing QTLs (sQTLs) and cell type-specific QTLs. Extending the colocalization to look at more traits (for example by using HyperColoc[230]) would allow for a fuller understanding of the phenotypic landscape and potential shared pathways involved with other cardiovascular related risk factors and disease endpoints. This could be extended to investigate association with a variety of traits phenome-wide, which would be valuable in an early drug discovery context, enabling both drug repurposing opportunities and adverse side effects to be assessed for nominated causal genes. To complement this analysis, a potential future avenue for methodological improvement would be to develop a colocalization method which can automatically evaluate both multiple independent variants and multiple traits (for example fine-mapping combined with moloc) to streamline the analysis required for genes in more LD complex regions. Finally, tissue and cell type specific effects of QTLs are important to consider. In this thesis, I used molQTL profiled in whole blood tissue as this had the largest sample size and best statistical power to detect effects, but as described in section **7.5** molQTLs are context specific. Therefore, the availability of QTL catalogs from a wider range of cell types and tissues as well as functional features will be useful. Furthermore, the 450K methylation array covers only 2% of the methylome and newer arrays or methylome sequencing may improve coverage of the methylome. Statistical follow-up with MR and functional follow-up analyses e.g., gene editing might be needed to confirm a true causal gene(s).

7.8 Main conclusions

In the work presented here, multi-omics approaches were systematically applied using large datasets that represent intermediate molecular layers to explore the evidence of molecular mechanisms underlying AF and stroke associations. My findings provide informative evidence of potential biological links and a shared genetic architecture between different intermediate molecular phenotypes and pathogenesis of AF and stroke. My multi-trait colocalization results suggest complex relationships between gene expression, DNA methylation and metabolite levels.

Whilst this approach does not distinguish horizontal from vertical pleiotropy, I showed that integrating genetic studies of molecular phenotypes with results from GWAS of AF and stroke can provide more insights into the pathways which are influenced by genetic liability to AF and stroke and facilitate the prioritization of promising genes. Importantly, my findings could be of value in drug target prioritization or in identifying biomarkers for early detection of AF or stroke. For example, this study was able to prioritize the *HP* gene as a potential causal factor for AF and CES subtype. However, the 16q22 locus also illustrated the complexity of integrated analysis on molQTLs as *HP* colocalized with gene expression, but not protein expression so regulation through this gene may not follow a simple transcriptional model.

The approach taken here could be applied to a range of other disease areas and is likely to yield more informative results as new large-scale molecular datasets become available in relevant tissues.

References

1. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
2. Durbin, R. M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
3. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
4. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
5. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
6. Kathiresan, S. & Srivastava, D. Genetics of Human Cardiovascular Disease. *Cell* **148**, 1242–1257 (2012).
7. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **33**, 228–237 (2003).
8. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
9. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
10. Elsworth, B. *et al.* The MRC IEU OpenGWAS data infrastructure. *bioRxiv* 2020.08.10.244293 (2020) doi:10.1101/2020.08.10.244293.
11. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

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

13. Pritchard, J. K. & Przeworski, M. Linkage Disequilibrium in Humans: Models and Data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).

14. Gallagher, M. D. & Chen-Plotkin, A. S. The Post-GWAS Era: From Association to Function. *Am. J. Hum. Genet.* **102**, 717–730 (2018).

15. Bhartiya, D. & Scaria, V. Genomic variations in non-coding RNAs: Structure, function and regulation. *Genomics* **107**, 59–68 (2016).

16. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).

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

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

19. Ward, L. D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.* **30**, 1095–1106 (2012).

20. Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507**, 371–375 (2014).

21. Claussnitzer, M. *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N. Engl. J. Med.* **373**, 895–907 (2015).

22. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).

23. Hannon, E., Weedon, M., Bray, N., O'Donovan, M. & Mill, J. Pleiotropic Effects of Trait-Associated Genetic Variation on DNA Methylation: Utility for Refining GWAS Loci.

Am. J. Hum. Genet. **100**, 954–959 (2017).

24. Roger, V. L. *et al.* Executive summary: heart disease and stroke statistics--2012 update: a report from the American Heart Association. *Circulation* **125**, 188–197 (2012).

25. Mozaffarian, D. *et al.* Executive Summary: Heart Disease and Stroke Statistics—2016 Update. *Circulation* **133**, 447–454 (2016).

26. Chugh, S. S. *et al.* Worldwide epidemiology of atrial fibrillation: a Global Burden of Disease 2010 Study. *Circulation* **129**, 837–847 (2014).

27. Benjamin, E. J. *et al.* Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association. *Circulation* **139**, e56–e528 (2019).

28. Benjamin, E. J. *et al.* Impact of atrial fibrillation on the risk of death: the Framingham Heart Study. *Circulation* **98**, 946–952 (1998).

29. Vermond, R. A. *et al.* Incidence of Atrial Fibrillation and Relationship With Cardiovascular Events, Heart Failure, and Mortality: A Community-Based Study From the Netherlands. *J. Am. Coll. Cardiol.* **66**, 1000–1007 (2015).

30. Soliman, E. Z. *et al.* Atrial fibrillation and the risk of myocardial infarction. *JAMA Intern. Med.* **174**, 107–114 (2014).

31. Kokubo, Y. & Matsumoto, C. Traditional Cardiovascular Risk Factors for Incident Atrial Fibrillation. *Circ. J. Off. J. Jpn. Circ. Soc.* **80**, 2415–2422 (2016).

32. Janssens, A. C. J. W. Validity of polygenic risk scores: are we measuring what we think we are? *Hum. Mol. Genet.* **28**, R143–R150 (2019).

33. Lin, H. J. *et al.* Stroke severity in atrial fibrillation. The Framingham Study. *Stroke* **27**, 1760–1764 (1996).

34. Bushnell, C. *et al.* Guidelines for the prevention of stroke in women: a statement for healthcare professionals from the American Heart Association/American Stroke

Association. *Stroke* **45**, 1545–1588 (2014).

35. Weng, L.-C. *et al.* Heritability of Atrial Fibrillation. *Circ. Cardiovasc. Genet.* **10**, e001838 (2017).

36. Kotecha, D. *et al.* Efficacy of β blockers in patients with heart failure plus atrial fibrillation: an individual-patient data meta-analysis. *Lancet Lond. Engl.* **384**, 2235–2243 (2014).

37. Lubitz, S. A. *et al.* Independent susceptibility markers for atrial fibrillation on chromosome 4q25. *Circulation* **122**, 976–984 (2010).

38. Oyen, N. *et al.* Familial aggregation of lone atrial fibrillation in young persons. *J. Am. Coll. Cardiol.* **60**, 917–921 (2012).

39. Christophersen, I. E. *et al.* Familial aggregation of Atrial Fibrillation – a study in Danish Twins. *Circ. Arrhythm. Electrophysiol.* **2**, 378–383 (2009).

40. Lubitz, S. A. *et al.* Association between familial atrial fibrillation and risk of new-onset atrial fibrillation. *JAMA* **304**, 2263–2269 (2010).

41. Nielsen, J. B. *et al.* Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat. Genet.* **50**, 1234–1239 (2018).

42. Feghaly, J., Zakka, P., London, B., MacRae, C. A. & Refaat, M. M. Genetics of Atrial Fibrillation. *J. Am. Heart Assoc.* **7**, e009884 (2018).

43. Hayashi, K., Tada, H. & Yamagishi, M. The genetics of atrial fibrillation. *Curr. Opin. Cardiol.* **32**, 10–16 (2017).

44. Fatkin, D., Santiago, C. F., Huttner, I. G., Lubitz, S. A. & Ellinor, P. T. Genetics of Atrial Fibrillation: State of the Art in 2017. *Heart Lung Circ.* **26**, 894–901 (2017).

45. Wang, J. *et al.* Pitx2 prevents susceptibility to atrial arrhythmias by inhibiting left-sided pacemaker specification. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 9753–9758 (2010).

46. Wang, J. *et al.* Pitx2-microRNA pathway that delimits sinoatrial node development and inhibits predisposition to atrial fibrillation. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 9181–9186 (2014).

47. Franco, D., Chinchilla, A., Daimi, H., Dominguez, J. N. & Aránega, A. Modulation of conductive elements by Pitx2 and their impact on atrial arrhythmogenesis. *Cardiovasc. Res.* **91**, 223–231 (2011).

48. Kirchhof, P. *et al.* PITX2c is expressed in the adult left atrium, and reducing Pitx2c expression promotes atrial fibrillation inducibility and complex changes in gene expression. *Circ. Cardiovasc. Genet.* **4**, 123–133 (2011).

49. Olesen, M. S., Nielsen, M. W., Haunsø, S. & Svendsen, J. H. Atrial fibrillation: the role of common and rare genetic variants. *Eur. J. Hum. Genet. EJHG* **22**, 297–306 (2014).

50. Kalstø, S. M., Siland, J. E., Rienstra, M. & Christophersen, I. E. Atrial Fibrillation Genetics Update: Toward Clinical Implementation. *Front. Cardiovasc. Med.* **6**, 127 (2019).

51. Campbell, H. M. & Wehrens, X. H. T. Genetics of atrial fibrillation: an update. *Curr. Opin. Cardiol.* **33**, 304–310 (2018).

52. Roselli, C. *et al.* Multi-ethnic genome-wide association study for atrial fibrillation. *Nat. Genet.* **50**, 1225–1233 (2018).

53. Nielsen, J. B. *et al.* Genome-wide Study of Atrial Fibrillation Identifies Seven Risk Loci and Highlights Biological Pathways and Regulatory Elements Involved in Cardiac Development. *Am. J. Hum. Genet.* **102**, 103–115 (2018).

54. Lubitz, S. A. *et al.* Novel genetic markers associate with atrial fibrillation risk in Europeans and Japanese. *J. Am. Coll. Cardiol.* **63**, 1200–1210 (2014).

55. Chinchilla, A. *et al.* PITX2 insufficiency leads to atrial electrical and structural remodeling linked to arrhythmogenesis. *Circ. Cardiovasc. Genet.* **4**, 269–279 (2011).

56. Zhang, M. *et al.* Long-range Pitx2c enhancer-promoter interactions prevent predisposition to atrial fibrillation. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 22692–22698 (2019).

57. Hodgson-Zingman, D. M. *et al.* Atrial Natriuretic Peptide Frameshift Mutation in Familial Atrial Fibrillation. *N. Engl. J. Med.* **359**, 158–165 (2008).

58. Donkor, E. S. Stroke in the 21st Century: A Snapshot of the Burden, Epidemiology, and Quality of Life. *Stroke Res. Treat.* **2018**, 3238165 (2018).

59. GBD 2015 DALYs and HALE Collaborators. Global, regional, and national disability-adjusted life-years (DALYs) for 315 diseases and injuries and healthy life expectancy (HALE), 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet Lond. Engl.* **388**, 1603–1658 (2016).

60. GBD 2015 Mortality and Causes of Death Collaborators. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet Lond. Engl.* **388**, 1459–1544 (2016).

61. Roth, G. A. *et al.* Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. *J. Am. Coll. Cardiol.* **70**, 1–25 (2017).

62. Sacco, R. L. *et al.* An updated definition of stroke for the 21st century: a statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* **44**, 2064–2089 (2013).

63. Markus, H. Stroke: causes and clinical features. *Medicine (Baltimore)* **44**, 515–520 (2016).

64. GBD 2016 Stroke Collaborators. Global, regional, and national burden of stroke, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* **18**, 439–458 (2019).

65. Murphy, S. J. & Werring, D. J. Stroke: causes and clinical features. *Med. Abingdon Engl. UK Ed* **48**, 561–566 (2020).

66. Boehme, A. K., Esenwa, C. & Elkind, M. S. V. Stroke Risk Factors, Genetics, and Prevention. *Circ. Res.* **120**, 472–495 (2017).

67. 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).

68. Tirschwell, D. L. *et al.* Association of cholesterol with stroke risk varies in stroke subtypes and patient subgroups. *Neurology* **63**, 1868–1875 (2004).

69. Meschia, J. F. *et al.* Guidelines for the Primary Prevention of Stroke. *Stroke* **45**, 3754–3832 (2014).

70. Kishore, A. *et al.* Detection of atrial fibrillation after ischemic stroke or transient ischemic attack: a systematic review and meta-analysis. *Stroke* **45**, 520–526 (2014).

71. Grond, M. *et al.* Improved detection of silent atrial fibrillation using 72-hour Holter ECG in patients with ischemic stroke: a prospective multicenter cohort study. *Stroke* **44**, 3357–3364 (2013).

72. Wang, Q. *et al. A phenome-wide multi-directional Mendelian randomization analysis of atrial fibrillation*. http://medrxiv.org/lookup/doi/10.1101/2020.10.15.20212654 (2020) doi:10.1101/2020.10.15.20212654.

73. Menezes Cordeiro, I. *et al.* Shifting the CARASIL Paradigm. *Stroke* **46**, 1110–1112 (2015).

74. Coucke, P. J. *et al.* Mutations in the facilitative glucose transporter GLUT10 alter angiogenesis and cause arterial tortuosity syndrome. *Nat. Genet.* **38**, 452–457 (2006).

75. Peters, N. *et al.* Spectrum of mutations in biopsy-proven CADASIL: implications for

diagnostic strategies. *Arch. Neurol.* **62**, 1091–1094 (2005).

76. Joutel, A. *et al.* The ectodomain of the Notch3 receptor accumulates within the cerebrovasculature of CADASIL patients. *J. Clin. Invest.* **105**, 597–605 (2000).

77. Flossmann, E., Schulz, U. G. R. & Rothwell, P. M. Systematic review of methods and results of studies of the genetic epidemiology of ischemic stroke. *Stroke* **35**, 212–227 (2004).

78. Seshadri, S. *et al.* Parental occurrence of stroke and risk of stroke in their children: the Framingham study. *Circulation* **121**, 1304–1312 (2010).

79. Schulz, U. G. R., Flossmann, E. & Rothwell, P. M. Heritability of ischemic stroke in relation to age, vascular risk factors, and subtypes of incident stroke in population-based studies. *Stroke* **35**, 819–824 (2004).

80. Zöller, B. & Dahlbäck, B. Linkage between inherited resistance to activated protein C and factor V gene mutation in venous thrombosis. *Lancet Lond. Engl.* **343**, 1536–1538 (1994).

81. Bertina, R. M. *et al.* Mutation in blood coagulation factor V associated with resistance to activated protein C. *Nature* **369**, 64–67 (1994).

82. de Maat, M. P., Kluft, C., Jespersen, J. & Gram, J. World distribution of factor V Leiden mutation. *Lancet Lond. Engl.* **347**, 58 (1996).

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

84. Devan, W. J. *et al.* Heritability estimates identify a substantial genetic contribution to risk and outcome of intracerebral hemorrhage. *Stroke* **44**, 1578–1583 (2013).

85. Chauhan, G. & Debette, S. Genetic Risk Factors for Ischemic and Hemorrhagic Stroke. *Curr. Cardiol. Rep.* **18**, (2016).

86. Gudbjartsson, D. F. *et al.* Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature* **448**, 353–357 (2007).

87. Gudbjartsson, D. F. *et al.* A sequence variant in ZFHX3 on 16q22 associates with atrial fibrillation and ischemic stroke. *Nat. Genet.* **41**, 876–878 (2009).

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

89. Traylor, M. *et al.* A novel MMP12 locus is associated with large artery atherosclerotic stroke using a genome-wide age-at-onset informed approach. *PLoS Genet.* **10**, e1004469 (2014).

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

91. Neurology Working Group of the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium, the Stroke Genetics Network (SiGN), and the International Stroke Genetics Consortium (ISGC). Identification of additional risk loci for stroke and small vessel disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* **15**, 695–707 (2016).

92. Traylor, M. *et al.* Genetic variation at 16q24.2 is associated with small vessel stroke. *Ann. Neurol.* **81**, 383–394 (2017).

93. Woo, D. *et al.* Meta-analysis of genome-wide association studies identifies 1q22 as a susceptibility locus for intracerebral hemorrhage. *Am. J. Hum. Genet.* **94**, 511–521 (2014).

94. Kilarski, L. L. *et al.* Meta-analysis in more than 17,900 cases of ischemic stroke reveals a novel association at 12q24.12. *Neurology* **83**, 678–685 (2014).

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

96. Williams, F. M. K. *et al.* Ischemic stroke is associated with the ABO locus: the EuroCLOT study. *Ann. Neurol.* **73**, 16–31 (2013).

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

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

99. Matarin, M. *et al.* Whole genome analyses suggest ischemic stroke and heart disease share an association with polymorphisms on chromosome 9p21. *Stroke* **39**, 1586–1589 (2008).

100. von Eisenhart Rothe, A. *et al.* Depressed mood amplifies heart-related symptoms in persistent and paroxysmal atrial fibrillation patients: a longitudinal analysis--data from the German Competence Network on Atrial Fibrillation. *Eur. Eur. Pacing Arrhythm. Card. Electrophysiol. J. Work. Groups Card. Pacing Arrhythm. Card. Cell. Electrophysiol. Eur. Soc. Cardiol.* **17**, 1354–1362 (2015).

101. Marzona, I. *et al.* Increased risk of cognitive and functional decline in patients with atrial fibrillation: results of the ONTARGET and TRANSCEND studies. *CMAJ Can. Med. Assoc. J. J. Assoc. Medicale Can.* **184**, E329-336 (2012).

102. Thrall, G., Lane, D., Carroll, D. & Lip, G. Y. H. Quality of life in patients with atrial fibrillation: a systematic review. *Am. J. Med.* **119**, 448.e1–19 (2006).

103. Knecht, S. *et al.* Atrial fibrillation in stroke-free patients is associated with memory impairment and hippocampal atrophy. *Eur. Heart J.* **29**, 2125–2132 (2008).

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

105. Wolf, P. A., Abbott, R. D. & Kannel, W. B. Atrial fibrillation as an independent risk factor for stroke: the Framingham Study. *Stroke* **22**, 983–988 (1991).

106. Sinner, M. F. *et al.* Integrating genetic, transcriptional, and functional analyses to identify 5 novel genes for atrial fibrillation. *Circulation* **130**, 1225–1235 (2014).

107. Lemmens, R. *et al.* The association of the 4q25 susceptibility variant for atrial fibrillation with stroke is limited to stroke of cardioembolic etiology. *Stroke* **41**, 1850–1857 (2010).

108. Tada, H. *et al.* Twelve-single nucleotide polymorphism genetic risk score identifies individuals at increased risk for future atrial fibrillation and stroke. *Stroke* **45**, 2856–2862 (2014).

109. Tao, Y. *et al.* Pitx2, an atrial fibrillation predisposition gene, directly regulates ion transport and intercalated disc genes. *Circ. Cardiovasc. Genet.* **7**, 23–32 (2014).

110. Cheng, W.-L. *et al.* MicroRNA-133 suppresses ZFHX3-dependent atrial remodelling and arrhythmia. *Acta Physiol. Oxf. Engl.* **227**, e13322 (2019).

111. Pulit, S. L. *et al.* Atrial fibrillation genetic risk differentiates cardioembolic stroke from other stroke subtypes. *Neurol. Genet.* **4**, e293 (2018).

112. Christophersen, I. E. *et al.* Erratum: Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation. *Nat. Genet.* **49**, 1286 (2017).

113. 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**, 940–947 (2011).

114. Verhaaren, B. F. J. *et al.* Multiethnic genome-wide association study of cerebral white

matter hyperintensities on MRI. *Circ. Cardiovasc. Genet.* **8**, 398–409 (2015).

115. Biffi, A. *et al.* Variants at APOE influence risk of deep and lobar intracerebral hemorrhage. *Ann. Neurol.* **68**, 934–943 (2010).

116. Brouwers, H. B. *et al.* Apolipoprotein E genotype predicts hematoma expansion in lobar intracerebral hemorrhage. *Stroke* **43**, 1490–1495 (2012).

117. Lorenzen, J. M., Martino, F. & Thum, T. Epigenetic modifications in cardiovascular disease. *Basic Res. Cardiol.* **107**, 245 (2012).

118. Kohli, R. M. & Zhang, Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* **502**, 472–479 (2013).

119. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14**, 204–220 (2013).

120. Relton, C. L. & Davey Smith, G. Epigenetic epidemiology of common complex disease: prospects for prediction, prevention, and treatment. *PLoS Med.* **7**, e1000356 (2010).

121. Bell, J. T. *et al.* DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.* **12**, R10 (2011).

122. Drong, A. W. *et al.* The presence of methylation quantitative trait loci indicates a direct genetic influence on the level of DNA methylation in adipose tissue. *PloS One* **8**, e55923 (2013).

123. Gaunt, T. R. *et al.* Systematic identification of genetic influences on methylation across the human life course. *Genome Biol.* **17**, 61 (2016).

124. Min, J. L. *et al.* Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nat. Genet.* **53**, 1311–1321 (2021).

125. Zhang, D. *et al.* Genetic control of individual differences in gene-specific methylation in human brain. *Am. J. Hum. Genet.* **86**, 411–419 (2010).

126. Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat. Genet.* **44**, 1084–1089 (2012).

127. Gutierrez-Arcelus, M. *et al.* Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *eLife* **2**, e00523 (2013).

128. Bonder, M. J. *et al.* Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.* **49**, 131–138 (2017).

129. Crujeiras, A. B. *et al.* Genome-wide DNA methylation pattern in visceral adipose tissue differentiates insulin-resistant from insulin-sensitive obese subjects. *Transl. Res. J. Lab. Clin. Med.* **178**, 13-24.e5 (2016).

130. Miao, L. *et al.* Integrated DNA methylation and gene expression analysis in the pathogenesis of coronary artery disease. *Aging* **11**, 1486–1500 (2019).

131. Deng, G.-X. *et al.* Association between promoter DNA methylation and gene expression in the pathogenesis of ischemic stroke. *Aging* **11**, 7663–7677 (2019).

132. Gibbs, J. R. *et al.* Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.* **6**, e1000952 (2010).

133. Gutierrez-Arcelus, M. *et al.* Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *eLife* **2**, e00523 (2013).

134. van Eijk, K. R. *et al.* Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. *BMC Genomics* **13**, 636 (2012).

135. Battram, T. *et al.* The EWAS Catalog: a database of epigenome-wide association studies. (2021) doi:10.31219/osf.io/837wn.

136. Lin, H. *et al.* Methylome-wide Association Study of Atrial Fibrillation in Framingham Heart Study. *Sci. Rep.* **7**, 40377 (2017).

137. Soriano-Tárraga, C. *et al.* Identification of 20 novel loci associated with ischaemic

stroke. Epigenome-wide association study. *Epigenetics* **15**, 988–997 (2020).

138. Portela, A. & Esteller, M. Epigenetic modifications and human disease. *Nat. Biotechnol.* **28**, 1057–1068 (2010).

139. Zeng, M. *et al.* The Role of DNA Methylation in Ischemic Stroke: A Systematic Review. *Front. Neurol.* **11**, 566124 (2020).

140. Bergman, Y. & Cedar, H. DNA methylation dynamics in health and disease. *Nat. Struct. Mol. Biol.* **20**, 274–281 (2013).

141. Wang, X. *et al.* Genome-wide DNA methylation patterns in coronary heart disease. *Herz* **43**, 656–662 (2018).

142. Agha, G. *et al.* Blood Leukocyte DNA Methylation Predicts Risk of Future Myocardial Infarction and Coronary Heart Disease. *Circulation* **140**, 645–657 (2019).

143. Bain, C. R. *et al.* DNA methylation patterns from peripheral blood separate coronary artery disease patients with and without heart failure. *ESC Heart Fail.* **7**, 2468–2478 (2020).

144. Nazarenko, M. S. *et al.* A comparison of genome-wide DNA methylation patterns between different vascular tissues from patients with coronary heart disease. *PloS One* **10**, e0122601 (2015).

145. Westra, H.-J. *et al.* Cell Specific eQTL Analysis without Sorting Cells. *PLOS Genet.* **11**, e1005223 (2015).

146. Võsa, U. *et al.* Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv* 447367 (2018) doi:10.1101/447367.

147. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).

148. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

149. Consortium, T. Gte. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

150. Liu, B., Gloudemans, M. J., Rao, A. S., Ingelsson, E. & Montgomery, S. B. Abundant associations with gene expression complicate GWAS follow-up. *Nat. Genet.* **51**, 768–769 (2019).

151. Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: present and future. *Philos. Trans. R. Soc. B Biol. Sci.* **368**, 20120362 (2013).

152. Nicolae, D. L. *et al.* Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLOS Genet.* **6**, e1000888 (2010).

153. Kasowski, M. *et al.* Variation in Transcription Factor Binding Among Humans. *Science* **328**, 232–235 (2010).

154. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res.* **22**, 1748–1759 (2012).

155. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).

156. Grubert, F. *et al.* Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell* **162**, 1051–1065 (2015).

157. Waszak, S. M. *et al.* Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. *Cell* **162**, 1039–1050 (2015).

158. Cotney, J. *et al.* Chromatin state signatures associated with tissue-specific gene expression and enhancer activity in the embryonic limb. *Genome Res.* **22**, 1069–1080 (2012).

159. Kilpinen, H. *et al.* Coordinated effects of sequence variation on DNA binding,

chromatin structure, and transcription. *Science* **342**, 744–747 (2013).

160. Xiao, S. *et al.* Comparative epigenomic annotation of regulatory DNA. *Cell* **149**, 1381–1392 (2012).

161. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

162. Matsuda, R. *et al.* STUDIES OF METABOLITE-PROTEIN INTERACTIONS: A REVIEW. *J. Chromatogr. B Analyt. Technol. Biomed. Life. Sci.* **966**, 48–58 (2014).

163. Kuehnbaum, N. L. & Britz-McKibbin, P. New advances in separation science for metabolomics: resolving chemical diversity in a post-genomic era. *Chem. Rev.* **113**, 2437–2468 (2013).

164. Patti, G. J., Yanes, O. & Siuzdak, G. Metabolomics: the apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.* **13**, 263–269 (2012).

165. Kaddurah-Daouk, R., Kristal, B. S. & Weinshilboum, R. M. Metabolomics: a global biochemical approach to drug response and disease. *Annu. Rev. Pharmacol. Toxicol.* **48**, 653–683 (2008).

166. Lee, D. Y., Bowen, B. P. & Northen, T. R. Mass spectrometry-based metabolomics, analysis of metabolite-protein interactions, and imaging. *BioTechniques* **49**, 557–565 (2010).

167. Soininen, P., Kangas, A. J., Würtz, P., Suna, T. & Ala-Korpela, M. Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Cardiovascular Epidemiology and Genetics. *Circ. Cardiovasc. Genet.* **8**, 192–206 (2015).

168. Castelli, W. P. Cholesterol and lipids in the risk of coronary artery disease--the Framingham Heart Study. *Can. J. Cardiol.* **4 Suppl A**, 5A-10A (1988).

169. Flora, G. D. & Nayak, M. K. A Brief Review of Cardiovascular Diseases, Associated

Risk Factors and Current Treatment Regimes. *Curr. Pharm. Des.* **25**, 4063–4084 (2019).

170. Würtz, P. *et al.* Metabolite profiling and cardiovascular event risk: a prospective study of 3 population-based cohorts. *Circulation* **131**, 774–785 (2015).

171. Hansson, G. K. Inflammation, atherosclerosis, and coronary artery disease. *N. Engl. J. Med.* **352**, 1685–1695 (2005).

172. Taleb, S. Inflammation in atherosclerosis. *Arch. Cardiovasc. Dis.* **109**, 708–715 (2016).

173. Roederer, M. *et al.* The Genetic Architecture of the Human Immune System: A Bioresource for Autoimmunity and Disease Pathogenesis. *Cell* **161**, 387–403 (2015).

174. 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).

175. Shin, S.-Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46**, 543–550 (2014).

176. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).

177. Rusk, N. The UK Biobank. *Nat. Methods* **15**, 1001–1001 (2018).

178. Nordestgaard, B. G. & Langsted, A. Lipoprotein (a) as a cause of cardiovascular disease: insights from epidemiology, genetics, and biology. *J. Lipid Res.* **57**, 1953–1975 (2016).

179. Kronenberg, F. & Utermann, G. Lipoprotein(a): resurrected by genetics. *J. Intern. Med.* **273**, 6–30 (2013).

180. Langsted, A., Nordestgaard, B. G. & Kamstrup, P. R. Elevated Lipoprotein(a) and Risk of Ischemic Stroke. *J. Am. Coll. Cardiol.* **74**, 54–66 (2019).

181. Burgess, S. *et al.* Association of LPA Variants With Risk of Coronary Disease and the Implications for Lipoprotein(a)-Lowering Therapies: A Mendelian Randomization

Analysis. *JAMA Cardiol.* **3**, 619–627 (2018).

182. Kasowski, M. *et al.* Extensive variation in chromatin states across humans. *Science* **342**, 750–752 (2013).

183. Stadler, M. B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490–495 (2011).

184. Lienert, F. *et al.* Identification of genetic elements that autonomously determine DNA methylation states. *Nat. Genet.* **43**, 1091–1097 (2011).

185. Hannon, E. *et al.* Leveraging DNA-Methylation Quantitative-Trait Loci to Characterize the Relationship between Methylomic Variation, Gene Expression, and Complex Traits. *Am. J. Hum. Genet.* **103**, 654–665 (2018).

186. Mancuso, N. *et al.* Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. *Am. J. Hum. Genet.* **100**, 473–487 (2017).

187. Burkhardt, R. *et al.* Integration of Genome-Wide SNP Data and Gene-Expression Profiles Reveals Six Novel Loci and Regulatory Mechanisms for Amino Acids and Acylcarnitines in Whole Blood. *PLoS Genet.* **11**, e1005510 (2015).

188. Richardson, T. G. *et al.* Systematic Mendelian randomization framework elucidates hundreds of CpG sites which may mediate the influence of genetic variants on disease. *Hum. Mol. Genet.* **27**, 3293–3304 (2018).

189. Nath, A. P. *et al.* An interaction map of circulating metabolites, immune gene networks, and their genetic regulation. *Genome Biol.* **18**, 146 (2017).

190. Mihalik, S. J. *et al.* Increased levels of plasma acylcarnitines in obesity and type 2 diabetes and identification of a marker of glucolipotoxicity. *Obes. Silver Spring Md* **18**, 1695–1700 (2010).

191. Adams, S. H. *et al.* Plasma acylcarnitine profiles suggest incomplete long-chain fatty acid beta-oxidation and altered tricarboxylic acid cycle activity in type 2 diabetic African-American women. *J. Nutr.* **139**, 1073–1081 (2009).

192. 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).

193. Ndungu, A., Payne, A., Torres, J. M., van de Bunt, M. & McCarthy, M. I. A Multi-tissue Transcriptome Analysis of Human Metabolites Guides Interpretability of Associations Based on Multi-SNP Models for Gene Expression. *Am. J. Hum. Genet.* **106**, 188–201 (2020).

194. Irvin, M. R. *et al.* Epigenome-wide association study of fasting blood lipids in the Genetics of Lipid-lowering Drugs and Diet Network study. *Circulation* **130**, 565–572 (2014).

195. Pfeiffer, L. *et al.* DNA methylation of lipid-related genes affects blood lipid levels. *Circ. Cardiovasc. Genet.* **8**, 334–342 (2015).

196. Sayols-Baixeras, S. *et al.* Identification and validation of seven new loci showing differential DNA methylation related to serum lipid profile: an epigenome-wide approach. The REGICOR study. *Hum. Mol. Genet.* **25**, 4556–4565 (2016).

197. Braun, K. V. E. *et al.* Epigenome-wide association study (EWAS) on lipids: the Rotterdam Study. *Clin. Epigenetics* **9**, 15 (2017).

198. Hedman, Å. K. *et al.* Epigenetic Patterns in Blood Associated With Lipid Traits Predict Incident Coronary Heart Disease Events and Are Enriched for Results From Genome-Wide Association Studies. *Circ. Cardiovasc. Genet.* **10**, e001487 (2017).

199. Mittelstraß, K. & Waldenberger, M. DNA methylation in human lipid metabolism and related diseases. *Curr. Opin. Lipidol.* **29**, 116–124 (2018).

200. Chambers, J. C. *et al.* Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: a nested case-control study. *Lancet Diabetes Endocrinol.* **3**, 526–534 (2015).

201. Kriebel, J. *et al.* Association between DNA Methylation in Whole Blood and Measures of Glucose Metabolism: KORA F4 Study. *PloS One* **11**, e0152314 (2016).

202. Dekkers, K. F. *et al.* Blood lipids influence DNA methylation in circulating cells. *Genome Biol.* **17**, 138 (2016).

203. 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 Med.* **14**, e1002215 (2017).

204. Kulkarni, H. *et al.* Novel epigenetic determinants of type 2 diabetes in Mexican-American families. *Hum. Mol. Genet.* **24**, 5330–5344 (2015).

205. Richard, M. A. *et al.* Abstract 32: Novel Genetic Loci for Blood Pressure Regulation Identified by the Analysis of DNA Methylation. *Circulation* **133**, A32–A32 (2016).

206. Dekkers, K. F., Slagboom, P. E., Jukema, J. W. & Heijmans, B. T. The multifaceted interplay between lipids and epigenetics. *Curr. Opin. Lipidol.* **27**, 288–294 (2016).

207. Wahl, S. *et al.* Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* **541**, 81–86 (2017).

208. Zaghlool, S. B. *et al.* Deep molecular phenotypes link complex disorders and physiological insult to CpG methylation. *Hum. Mol. Genet.* **27**, 1106–1121 (2018).

209. Reed, Z. E., Suderman, M. J., Relton, C. L., Davis, O. S. P. & Hemani, G. The association of DNA methylation with body mass index: distinguishing between predictors and biomarkers. *Clin. Epigenetics* **12**, 50 (2020).

210. Sayols-Baixeras, S., Tiwari, H. K. & Aslibekyan, S. W. Disentangling associations

between DNA methylation and blood lipids: a Mendelian randomization approach. *BMC Proc.* **12**, 23 (2018).

211. Frazier-Wood, A. C. *et al.* Methylation at CPT1A locus is associated with lipoprotein subfraction profiles. *J. Lipid Res.* **55**, 1324–1330 (2014).

212. Gomez-Alonso, M. del C. *et al.* DNA methylation and lipid metabolism: an EWAS of 226 metabolic measures. *Clin. Epigenetics* **13**, 7 (2021).

213. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).

214. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).

215. Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1825 (2018).

216. Giambartolomei, C. *et al.* A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinforma. Oxf. Engl.* **34**, 2538–2545 (2018).

217. Smith, G. D. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32**, 1–22 (2003).

218. Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N. & Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat. Med.* **27**, 1133–1163 (2008).

219. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **23**, R89-98 (2014).

220. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic

association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).

221. Wallace, C. *et al.* Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. *Hum. Mol. Genet.* **21**, 2815–2824 (2012).

222. Nica, A. C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* **6**, e1000895 (2010).

223. He, X. *et al.* Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. *Am. J. Hum. Genet.* **92**, 667–680 (2013).

224. Plagnol, V., Smyth, D. J., Todd, J. A. & Clayton, D. G. Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13. *Biostat. Oxf. Engl.* **10**, 327–334 (2009).

225. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).

226. Wen, X., Pique-Regi, R. & Luca, F. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS Genet.* **13**, e1006646 (2017).

227. 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).

228. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375, S1-3 (2012).

229. Wallace, C. A more accurate method for colocalisation analysis allowing for multiple causal variants. *bioRxiv* 2021.02.23.432421 (2021) doi:10.1101/2021.02.23.432421.

230. Foley, C. N. *et al.* A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nat. Commun.* **12**, 764 (2021).

231. Bowden, J. *et al.* A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Stat. Med.* **36**, 1783–1802 (2017).

232. Neumeyer, S., Hemani, G. & Zeggini, E. Strengthening Causal Inference for Complex Disease Using Molecular Quantitative Trait Loci. *Trends Mol. Med.* **26**, 232–241 (2020).

233. Burgess, S., Davies, N. M. & Thompson, S. G. Bias due to participant overlap in two-sample Mendelian randomization. *Genet. Epidemiol.* **40**, 597–608 (2016).

234. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7**, e34408 (2018).

235. Hauberg, M. E. *et al.* Large-Scale Identification of Common Trait and Disease Variants Affecting Gene Expression. *Am. J. Hum. Genet.* **100**, 885–894 (2017).

236. Pavlides, J. M. W. *et al.* Predicting gene targets from integrative analyses of summary data from GWAS and eQTL studies for 28 human complex traits. *Genome Med.* **8**, 84 (2016).

237. Baird, D. A. *et al.* Identifying drug targets for neurological and psychiatric disease via genetics and the brain transcriptome. *PLOS Genet.* **17**, e1009224 (2021).

238. Wu, Y. *et al.* Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nat. Commun.* **9**, 918 (2018).

239. Qi, T. *et al.* Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat. Commun.* **9**, 2282 (2018).

240. Zheng, J. *et al.* Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat. Genet.* **52**, 1122–1131 (2020).

241. Võsa, U. *et al.* Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv* 447367 (2018) doi:10.1101/447367.

242. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from

1,092 human genomes. *Nature* **491**, 56–65 (2012).

243. Drenos, F. *et al.* Metabolic Characterization of a Rare Genetic Variation Within APOC3 and Its Lipoprotein Lipase–Independent Effects. *Circ. Cardiovasc. Genet.* **9**, 231–239 (2016).

244. Boyd, A. *et al.* Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. *Int. J. Epidemiol.* **42**, 111–127 (2013).

245. 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**, 955–959 (2012).

246. Shah, T. *et al.* Population genomics of cardiometabolic traits: design of the University College London-London School of Hygiene and Tropical Medicine-Edinburgh-Bristol (UCLEB) Consortium. *PloS One* **8**, e71345 (2013).

247. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

248. Boef, A. G. C., Dekkers, O. M. & le Cessie, S. Mendelian randomization studies: a review of the approaches used and the quality of reporting. *Int. J. Epidemiol.* **44**, 496–511 (2015).

249. Pierce, B. L. & Burgess, S. Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *Am. J. Epidemiol.* **178**, 1177–1184 (2013).

250. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* **37**, 658–665 (2013).

251. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid

instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).

252. Bowden, J. *et al.* Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. *Int. J. Epidemiol.* **45**, 1961–1974 (2016).

253. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.* **40**, 304–314 (2016).

254. Hartwig, F. P., Davey Smith, G. & Bowden, J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int. J. Epidemiol.* **46**, 1985–1998 (2017).

255. Taylor, K., Davey Smith, G., Relton, C. L., Gaunt, T. R. & Richardson, T. G. Prioritizing putative influential genes in cardiovascular disease susceptibility by applying tissue-specific Mendelian randomization. *Genome Med.* **11**, 6 (2019).

256. Wang, B. *et al.* Integrative Omics Approach to Identifying Genes Associated With Atrial Fibrillation. *Circ. Res.* **126**, 350–360 (2020).

257. Assum, I. *et al. Tissue-specific multiOMICs analysis of atrial fibrillation*. 2020.04.06.021527 https://www.biorxiv.org/content/10.1101/2020.04.06.021527v2 (2020) doi:10.1101/2020.04.06.021527.

258. Opacic, D., van Bragt, K. A., Nasrallah, H. M., Schotten, U. & Verheule, S. Atrial metabolism and tissue perfusion as determinants of electrical and structural remodelling in atrial fibrillation. *Cardiovasc. Res.* **109**, 527–541 (2016).

259. Iwasaki, Y., Nishida, K., Kato, T. & Nattel, S. Atrial fibrillation pathophysiology: implications for management. *Circulation* **124**, 2264–2274 (2011).

260. Ghezelbash, S., Molina, C. E. & Dobrev, D. Altered atrial metabolism: an underappreciated contributor to the initiation and progression of atrial fibrillation. *J. Am. Heart Assoc.* **4**, e001808 (2015).

261. Christophersen, I. E. *et al.* Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation. *Nat. Genet.* **49**, 946–952 (2017).

262. Lin, H. *et al.* Whole blood gene expression and atrial fibrillation: the Framingham Heart Study. *PloS One* **9**, e96794 (2014).

263. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).

264. Min, J. L., Hemani, G., Davey Smith, G., Relton, C. & Suderman, M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinforma. Oxf. Engl.* **34**, 3983–3989 (2018).

265. Koscielny, G. *et al.* Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res.* **45**, D985–D994 (2017).

266. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514-517 (2005).

267. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

268. Liu, Y. *et al.* GeMes, clusters of DNA methylation under genetic control, can inform genetic and epigenetic analysis of disease. *Am. J. Hum. Genet.* **94**, 485–495 (2014).

269. Slamon, D. J. *et al.* Use of Chemotherapy plus a Monoclonal Antibody against HER2 for Metastatic Breast Cancer That Overexpresses HER2. *http://dx.doi.org/10.1056/NEJM200103153441101*

https://www.nejm.org/doi/10.1056/NEJM200103153441101 (2009) doi:10.1056/NEJM200103153441101.

270. Ewer, M. S. & Ewer, S. M. Trastuzumab cardiotoxiciy: the age-old balance of risk and benefit. *Br. J. Cancer* **115**, 1441–1442 (2016).

271. Crone, S. A. *et al.* ErbB2 is essential in the prevention of dilated cardiomyopathy. *Nat. Med.* **8**, 459–465 (2002).

272. Ozcelik, C. *et al.* Conditional mutation of the ErbB2 (HER2) receptor in cardiomyocytes leads to dilated cardiomyopathy. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 8880–8885 (2002).

273. Negro, A. *et al.* erbB2 is required for G protein-coupled receptor signaling in the heart. *Proc. Natl. Acad. Sci.* **103**, 15889–15893 (2006).

274. Tenin, G. *et al.* Erbb2 is required for cardiac atrial electrical activity during development. *PloS One* **9**, e107041 (2014).

275. Ma, H., Yin, C., Zhang, Y., Qian, L. & Liu, J. ErbB2 is required for cardiomyocyte proliferation in murine neonatal hearts. *Gene* **592**, 325–330 (2016).

276. Honkoop, H. *et al.* Single-cell analysis uncovers that metabolic reprogramming by ErbB2 signaling is essential for cardiomyocyte proliferation in the regenerating heart. *eLife* **8**, e50163 (2019).

277. Li, D. *et al.* Mutations of presenilin genes in dilated cardiomyopathy and heart failure. *Am. J. Hum. Genet.* **79**, 1030–1039 (2006).

278. Nakajima, M., Moriizumi, E., Koseki, H. & Shirasawa, T. Presenilin 1 is essential for cardiac morphogenesis. *Dev. Dyn. Off. Publ. Am. Assoc. Anat.* **230**, 795–799 (2004).

279. Chen, W.-T. *et al.* G206D Mutation of Presenilin-1 Reduces Pen2 Interaction, Increases Aβ42/Aβ40 Ratio and Elevates ER Ca(2+) Accumulation. *Mol. Neurobiol.* **52**, 1835–1849

(2015).

280. Song, X.-W. *et al.* Conditionally targeted deletion of PSEN1 leads to diastolic heart dysfunction. *J. Cell. Physiol.* **233**, 1548–1557 (2018).

281. Nakajima, M. *et al.* Abnormal blood vessel development in mice lacking presenilin-1. *Mech. Dev.* **120**, 657–667 (2003).

282. Theis, J. L. *et al.* Homozygosity mapping and exome sequencing reveal GATAD1 mutation in autosomal recessive dilated cardiomyopathy. *Circ. Cardiovasc. Genet.* **4**, 585–594 (2011).

283. Sasaki, T. *et al.* Elevated intraocular pressure, optic nerve atrophy, and impaired retinal development in ODAG transgenic mice. *Invest. Ophthalmol. Vis. Sci.* **50**, 242–248 (2009).

284. Yang, J., Shah, S., Olson, T. M. & Xu, X. Modeling GATAD1-Associated Dilated Cardiomyopathy in Adult Zebrafish. *J. Cardiovasc. Dev. Dis.* **3**, 6 (2016).

285. Grange, D. K., Balfour, I. C., Chen, S. & Wood, E. G. Familial syndrome of progressive arterial occlusive disease consistent with fibromuscular dysplasia, hypertension, congenital cardiac defects, bone fragility, brachysyndactyly, and learning disabilities. *Am. J. Med. Genet.* **75**, 469–480 (1998).

286. Guo, D. *et al.* Loss-of-Function Mutations in YY1AP1 Lead to Grange Syndrome and a Fibromuscular Dysplasia-Like Vascular Disease. *Am. J. Hum. Genet.* **100**, 21–30 (2017).

287. Rath, M. *et al.* Identification of pathogenic YY1AP1 splice variants in siblings with Grange syndrome by whole exome sequencing. *Am. J. Med. Genet. A.* **179**, 295–299 (2019).

288. Ciuffetelli Alamo, I. V. *et al.* Grange syndrome due to homozygous YY1AP1 missense rare variants. *Am. J. Med. Genet. A.* **179**, 2500–2505 (2019).

289. Saida, K. *et al.* Hemorrhagic stroke and renovascular hypertension with Grange syndrome arising from a novel pathogenic variant in YY1AP1. *J. Hum. Genet.* **64**, 885–890

(2019).

290. Deng, Y. *et al.* Upregulated microRNA-381-5p strengthens the effect of dexmedetomidine preconditioning to protect against myocardial ischemia-reperfusion injury in mouse models by inhibiting CHI3L1. *Int. Immunopharmacol.* **92**, 107326 (2021).

291. Maegdefessel, L. *et al.* miR-24 limits aortic vascular inflammation and murine abdominal aneurysm development. *Nat. Commun.* **5**, 5214 (2014).

292. Mleynek, T. M. *et al.* Lack of CCM1 induces hypersprouting and impairs response to flow. *Hum. Mol. Genet.* **23**, 6223–6234 (2014).

293. Goitre, L. *et al.* KRIT1 regulates the homeostasis of intracellular reactive oxygen species. *PloS One* **5**, e11786 (2010).

294. Whitehead, K. J., Plummer, N. W., Adams, J. A., Marchuk, D. A. & Li, D. Y. Ccm1 is required for arterial morphogenesis: implications for the etiology of human cavernous malformations. *Dev. Camb. Engl.* **131**, 1437–1448 (2004).

295. Zhou, Z. *et al.* The cerebral cavernous malformation pathway controls cardiac development via regulation of endocardial MEKK3 signaling and KLF expression. *Dev. Cell* **32**, 168–180 (2015).

296. Vieceli Dalla Sega, F. *et al.* KRIT1 Deficiency Promotes Aortic Endothelial Dysfunction. *Int. J. Mol. Sci.* **20**, E4930 (2019).

297. Daniels, L. J. *et al.* Inhibition of calcium/calmodulin-dependent kinase II restores contraction and relaxation in isolated cardiac muscle from type 2 diabetic rats. *Cardiovasc. Diabetol.* **17**, 89 (2018).

298. Chelu, M. G. *et al.* Calmodulin kinase II-mediated sarcoplasmic reticulum Ca2+ leak promotes atrial fibrillation in mice. *J. Clin. Invest.* **119**, 1940–1951 (2009).

299. Neef, S. *et al.* CaMKII-dependent diastolic SR Ca2+ leak and elevated diastolic Ca2+

levels in right atrial myocardium of patients with atrial fibrillation. *Circ. Res.* **106**, 1134–1144 (2010).

300. Purohit, A. *et al.* Oxidized Ca(2+)/calmodulin-dependent protein kinase II triggers atrial fibrillation. *Circulation* **128**, 1748–1757 (2013).

301. Mesubi, O. O. *et al.* Oxidized CaMKII and O-GlcNAcylation cause increased atrial fibrillation in diabetic mice by distinct mechanisms. *J. Clin. Invest.* **131**, e95747.

302. Zhu, W. *et al.* IGFBP-4 is an inhibitor of canonical Wnt signalling required for cardiogenesis. *Nature* **454**, 345–349 (2008).

303. Yu, S., Li, Y., Zhao, H., Wang, Q. & Chen, P. The Histone Demethylase JMJD1C Regulates CAMKK2-AMPK Signaling to Participate in Cardiac Hypertrophy. *Front. Physiol.* **11**, 539 (2020).

304. Pierce, B. L. *et al.* Co-occurring expression and methylation QTLs allow detection of common causal variants and shared biological mechanisms. *Nat. Commun.* **9**, 804 (2018).

305. Guan, W. *et al.* Genome-wide association study of plasma N6 polyunsaturated fatty acids within the cohorts for heart and aging research in genomic epidemiology consortium. *Circ. Cardiovasc. Genet.* **7**, 321–331 (2014).

306. Yasumo, H. *et al.* Nuclear receptor binding factor-2 (NRBF-2), a possible gene activator protein interacting with nuclear hormone receptors. *Biochim. Biophys. Acta* **1490**, 189–197 (2000).

307. Chiazza, F. & Collino, M. Chapter 9 - Peroxisome Proliferator-Activated Receptors (PPARs) in Glucose Control. in *Molecular Nutrition and Diabetes* (ed. Mauricio, D.) 105–114 (Academic Press, 2016). doi:10.1016/B978-0-12-801585-8.00009-9.

308. Guo, T. *et al.* Histone Modifier Genes Alter Conotruncal Heart Phenotypes in 22q11.2 Deletion Syndrome. *Am. J. Hum. Genet.* **97**, 869–877 (2015).

309. Zhang, S., Lu, Y. & Jiang, C. Inhibition of histone demethylase JMJD1C attenuates cardiac hypertrophy and fibrosis induced by angiotensin II. *J. Recept. Signal Transduct. Res.* **40**, 339–347 (2020).

310. Eicher, J. D., Xue, L., Ben-Shlomo, Y., Beswick, A. D. & Johnson, A. D. Replication and hematological characterization of human platelet reactivity genetic associations in men from the Caerphilly Prospective Study (CaPS). *J. Thromb. Thrombolysis* **41**, 343–350 (2016).

311. Schooling, C. M. *et al.* Genetic predictors of testosterone and their associations with cardiovascular disease and risk factors: A Mendelian randomization investigation. *Int. J. Cardiol.* **267**, 171–176 (2018).

312. Zhang, T., Zhao, J. V. & Schooling, C. M. The associations of plasma phospholipid arachidonic acid with cardiovascular diseases: A Mendelian randomization study. *EBioMedicine* **63**, 103189 (2021).

313. Bell, J. T. *et al.* Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genet.* **8**, e1002629 (2012).

314. Acharya, C. R., Owzar, K. & Allen, A. S. Mapping eQTL by leveraging multiple tissues and DNA methylation. *BMC Bioinformatics* **18**, 455 (2017).

315. Hannon, E. *et al.* Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nat. Neurosci.* **19**, 48–54 (2016).

316. Sanderson, E., Davey Smith, G., Windmeijer, F. & Bowden, J. An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. *Int. J. Epidemiol.* **48**, 713–727 (2019).

317. Sanderson, E. Multivariable Mendelian Randomization and Mediation. *Cold Spring*

Harb. Perspect. Med. **11**, a038984 (2021).

318. Carter, A. R. *et al.* Mendelian randomisation for mediation analysis: current methods and challenges for implementation. *Eur. J. Epidemiol.* **36**, 465–478 (2021).

319. Zhao, S., Jiang, H., Liang, Z.-H. & Ju, H. Integrating Multi-Omics Data to Identify Novel Disease Genes and Single-Neucleotide Polymorphisms. *Front. Genet.* **10**, 1336 (2019).

320. Holmes, M. V. *et al.* Lipids, Lipoproteins, and Metabolites and Risk of Myocardial Infarction and Stroke. *J. Am. Coll. Cardiol.* **71**, 620–632 (2018).

321. Freiberg, J. J., Tybjaerg-Hansen, A., Jensen, J. S. & Nordestgaard, B. G. Nonfasting triglycerides and risk of ischemic stroke in the general population. *JAMA* **300**, 2142–2152 (2008).

322. Tirosh, A. *et al.* Changes in triglyceride levels and risk for coronary heart disease in young men. *Ann. Intern. Med.* **147**, 377–385 (2007).

323. Kasai, T. *et al.* Mortality risk of triglyceride levels in patients with coronary artery disease. *Heart Br. Card. Soc.* **99**, 22–29 (2013).

324. Hindy, G. *et al.* Role of Blood Lipids in the Development of Ischemic Stroke and its Subtypes: A Mendelian Randomization Study. *Stroke* **49**, 820–827 (2018).

325. Sun, L. *et al.* Causal associations of blood lipids with risk of ischemic stroke and intracerebral hemorrhage in Chinese adults. *Nat. Med.* **25**, 569–574 (2019).

326. Yuan, S., Tang, B., Zheng, J. & Larsson, S. C. Circulating Lipoprotein Lipids, Apolipoproteins and Ischemic Stroke. *Ann. Neurol.* **88**, 1229–1236 (2020).

327. Richardson, T. G. *et al.* Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis. *PLOS Med.* **17**, e1003062 (2020).

328. Kurth, T. *et al.* Lipid levels and the risk of ischemic stroke in women. *Neurology* **68**, 556–562 (2007).

329. Allara, E. *et al.* Genetic Determinants of Lipids and Cardiovascular Disease Outcomes: A Wide-Angled Mendelian Randomization Investigation. *Circ. Genomic Precis. Med.* **12**, e002711 (2019).

330. Yao, C. *et al.* Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat. Commun.* **9**, 3268 (2018).

331. Dichgans, M. *et al.* Shared genetic susceptibility to ischemic stroke and coronary artery disease: a genome-wide analysis of common variants. *Stroke* **45**, 24–36 (2014).

332. Chow, N. *et al.* Serum response factor and myocardin mediate arterial hypercontractility and cerebral blood flow dysregulation in Alzheimer's phenotype. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 823–828 (2007).

333. Norman, C., Runswick, M., Pollock, R. & Treisman, R. Isolation and properties of cDNA clones encoding SRF, a transcription factor that binds to the c-fos serum response element. *Cell* **55**, 989–1003 (1988).

334. Chai, J., Jones, M. K. & Tarnawski, A. S. Serum response factor is a critical requirement for VEGF signaling in endothelial cells and VEGF-induced angiogenesis. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* **18**, 1264–1266 (2004).

335. Franco, C. A. *et al.* Serum response factor is required for sprouting angiogenesis and vascular integrity. *Dev. Cell* **15**, 448–461 (2008).

336. Barron, M. R. *et al.* Serum response factor, an enriched cardiac mesoderm obligatory factor, is a downstream gene target for Tbx genes. *J. Biol. Chem.* **280**, 11816–11828 (2005).

337. Niu, Z. *et al.* Conditional mutagenesis of the murine serum response factor gene blocks cardiogenesis and the transcription of downstream gene targets. *J. Biol. Chem.* **280**, 32531–

32538 (2005).

338. Yao, C.-K. *et al.* A synaptic vesicle-associated Ca2+ channel promotes endocytosis and couples exocytosis to endocytosis. *Cell* **138**, 947–960 (2009).

339. Sasaki, M., Takagi, M. & Okamura, Y. A voltage sensor-domain protein is a voltage-gated proton channel. *Science* **312**, 589–592 (2006).

340. Musset, B. *et al.* Aspartate 112 is the selectivity filter of the human voltage-gated proton channel. *Nature* **480**, 273–277 (2011).

341. Ramsey, I. S., Ruchti, E., Kaczmarek, J. S. & Clapham, D. E. Hv1 proton channels are required for high-level NADPH oxidase-dependent superoxide production during the phagocyte respiratory burst. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 7642–7647 (2009).

342. Wu, L. *et al.* The Voltage–gated Proton Channel, Hv1, Enhances Brain Damage from Ischemic Stroke. *Nat. Neurosci.* **15**, 565–573 (2012).

343. Sieber, M. W. *et al.* Age-specific transcriptional response to stroke. *Neurobiol. Aging* **35**, 1744–1754 (2014).

344. Tian, D.-S. *et al.* Deficiency in the voltage-gated proton channel Hv1 increases M2 polarization of microglia and attenuates brain damage from photothrombotic ischemic stroke. *J. Neurochem.* **139**, 96–105 (2016).

345. Mashima, R. *et al.* FLN29, a novel interferon- and LPS-inducible gene acting as a negative regulator of toll-like receptor signaling. *J. Biol. Chem.* **280**, 41289–41297 (2005).

346. Sanada, T. *et al.* FLN29 deficiency reveals its negative regulatory role in the Toll-like receptor (TLR) and retinoic acid-inducible gene I (RIG-I)-like helicase signaling pathway. *J. Biol. Chem.* **283**, 33858–33864 (2008).

347. Caso, J. R. *et al.* Toll-like receptor 4 is involved in brain damage and inflammation after experimental stroke. *Circulation* **115**, 1599–1608 (2007).

348. Cao, C.-X. *et al.* Reduced cerebral ischemia-reperfusion injury in Toll-like receptor 4 deficient mice. *Biochem. Biophys. Res. Commun.* **353**, 509–514 (2007).

349. Hua, F. *et al.* Activation of Toll-like Receptor 4 Signaling Contributes to Hippocampal Neuronal Death Following Global Cerebral Ischemia/Reperfusion. *J. Neuroimmunol.* **190**, 101–111 (2007).

350. Ziegler, G. *et al.* TLR2 has a detrimental role in mouse transient focal cerebral ischemia. *Biochem. Biophys. Res. Commun.* **359**, 574–579 (2007).

351. Lehnardt, S. *et al.* Toll-like receptor 2 mediates CNS injury in focal cerebral ischemia. *J. Neuroimmunol.* **190**, 28–33 (2007).

352. Park, J.-I. *et al.* Telomerase modulates Wnt signalling by association with target gene chromatin. *Nature* **460**, 66–72 (2009).

353. Zhang, M. *et al.* SWI/SNF complexes containing Brahma or Brahma-related gene 1 play distinct roles in smooth muscle development. *Mol. Cell. Biol.* **31**, 2618–2631 (2011).

354. Akerberg, B. N., Sarangam, M. L. & Stankunas, K. Endocardial Brg1 disruption illustrates the developmental origins of semilunar valve disease. *Dev. Biol.* **407**, 158–172 (2015).

355. He, M. *et al.* ABO blood group and risk of coronary heart disease in two prospective cohort studies. *Arterioscler. Thromb. Vasc. Biol.* **32**, 2314–2320 (2012).

356. Dentali, F., Sironi, A. P., Ageno, W., Crestani, S. & Franchini, M. ABO blood group and vascular disease: an update. *Semin. Thromb. Hemost.* **40**, 49–59 (2014).

357. Vasan, S. K. *et al.* ABO Blood Group and Risk of Thromboembolic and Arterial Disease: A Study of 1.5 Million Blood Donors. *Circulation* **133**, 1449–1457; discussion 1457 (2016).

358. Fry, A. E. *et al.* Common variation in the ABO glycosyltransferase is associated with

susceptibility to severe Plasmodium falciparum malaria. *Hum. Mol. Genet.* **17**, 567–576 (2008).

359. Meade, T. W. *et al.* Factor VIII, ABO blood group and the incidence of ischaemic heart disease. *Br. J. Haematol.* **88**, 601–607 (1994).

360. Sabino, A. de P. *et al.* ABO blood group polymorphisms and risk for ischemic stroke and peripheral arterial disease. *Mol. Biol. Rep.* **41**, 1771–1777 (2014).

361. Garrison, R. J. *et al.* ABO blood group and cardiovacular disease: the Framingham study. *Atherosclerosis* **25**, 311–318 (1976).

362. Song, J. *et al.* Quantitative Influence of ABO Blood Groups on Factor VIII and Its Ratio to von Willebrand Factor, Novel Observations from an ARIC Study of 11,673 Subjects. *PLoS ONE* **10**, (2015).

363. Larson, N. B. *et al.* ABO blood group associations with markers of endothelial dysfunction in the Multi-Ethnic Study of Atherosclerosis. *Atherosclerosis* **251**, 422–429 (2016).

364. Heit, J. A. *et al.* A genome-wide association study of venous thromboembolism identifies risk variants in chromosomes 1q24.2 and 9q. *J. Thromb. Haemost. JTH* **10**, 1521–1531 (2012).

365. Klarin, D. *et al.* Genome-wide association analysis of venous thromboembolism identifies new risk loci and genetic overlap with arterial vascular disease. *Nat. Genet.* **51**, 1574–1579 (2019).

366. Ripatti, P. *et al.* Polygenic Hyperlipidemias and Coronary Artery Disease Risk. *Circ. Genomic Precis. Med.* **13**, e002725 (2020).

367. Nikpay, M. *et al.* A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47**, 1121–1130 (2015).

368. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).

369. Arvanitis, M. *et al.* Genome-wide association and multi-omic analyses reveal ACTN2 as a gene linked to heart failure. *Nat. Commun.* **11**, 1122 (2020).

370. van der Harst, P. & Verweij, N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circ. Res.* **122**, 433–443 (2018).

371. DeCoursey, T. E. Voltage-Gated Proton Channels: Molecular Biology, Physiology, and Pathophysiology of the HV Family. *Physiol. Rev.* **93**, 599–652 (2013).

372. Kulminski, A. M., Loika, Y., Nazarian, A. & Culminskaya, I. Quantitative and Qualitative Role of Antagonistic Heterogeneity in Genetics of Blood Lipids. *J. Gerontol. A. Biol. Sci. Med. Sci.* **75**, 1811–1819 (2020).

373. Xiong, X. *et al.* BRG1 variant rs1122608 on chromosome 19p13.2 confers protection against stroke and regulates expression of pre-mRNA-splicing factor SFRS3. *Hum. Genet.* **133**, 499–508 (2014).

374. Dykstra-Aiello, C. *et al.* Alternative Splicing of Putative Stroke/Vascular Risk Factor Genes Expressed in Blood Following Ischemic Stroke Is Sexually Dimorphic and Cause-Specific. *Front. Neurol.* **11**, 584695 (2020).

375. Wang, Z. *et al.* Identification of pleiotropic genes between risk factors of stroke by multivariate metaCCA analysis. *Mol. Genet. Genomics MGG* **295**, 1173–1185 (2020).

376. Schaer, C. A. *et al.* Mechanisms of haptoglobin protection against hemoglobin peroxidation triggered endothelial damage. *Cell Death Differ.* **20**, 1569–1579 (2013).

377. Andersen, C. B. F. *et al.* Haptoglobin. *Antioxid. Redox Signal.* **26**, 814–831 (2017).

378. Allison, A. C. & Rees, W. A. The binding of haemoglobin by plasma proteins

(haptoglobins); its bearing on the renal threshold for haemoglobin and the aetiology of haemoglobinuria. *Br. Med. J.* **2**, 1137–1143 (1957).

379. Langlois, M. R. & Delanghe, J. R. Biological and clinical significance of haptoglobin polymorphism in humans. *Clin. Chem.* **42**, 1589–1600 (1996).

380. Salvatore, A., Cigliano, L., Carlucci, A., Bucci, E. M. & Abrescia, P. Haptoglobin binds apolipoprotein E and influences cholesterol esterification in the cerebrospinal fluid. *J. Neurochem.* **110**, 255–263 (2009).

381. Spagnuolo, M. S. *et al.* Analysis of the haptoglobin binding region on the apolipoprotein A-I-derived P2a peptide. *J. Pept. Sci. Off. Publ. Eur. Pept. Soc.* **19**, 220–226 (2013).

382. Cigliano, L., Pugliese, C. R., Spagnuolo, M. S., Palumbo, R. & Abrescia, P. Haptoglobin binds the antiatherogenic protein apolipoprotein E - impairment of apolipoprotein E stimulation of both lecithin:cholesterol acyltransferase activity and cholesterol uptake by hepatocytes. *FEBS J.* **276**, 6158–6171 (2009).

383. Carter, K. & Worwood, M. Haptoglobin: a review of the major allele frequencies worldwide and their association with diseases. *Int. J. Lab. Hematol.* **29**, 92–110 (2007).

384. Kristiansen, M. *et al.* Identification of the haemoglobin scavenger receptor. *Nature* **409**, 198–201 (2001).

385. Bulters, D. *et al.* Haemoglobin scavenging in intracranial bleeding: biology and clinical implications. *Nat. Rev. Neurol.* **14**, 416–432 (2018).

386. Levy, A. P. *et al.* Haptoglobin phenotype is an independent risk factor for cardiovascular disease in individuals with diabetes: The Strong Heart Study. *J. Am. Coll. Cardiol.* **40**, 1984–1990 (2002).

387. Chapelle, J. P., Albert, A., Smeets, J. P., Heusghem, C. & Kulbertus, H. E. Effect of the

haptoglobin phenotype on the size of a myocardial infarct. *N. Engl. J. Med.* **307**, 457–463 (1982).

388. Ubaĭdullaev, A. M., Kazakov, K. S. & Khakimov, M. A. [Genetic variants of haptoglobin in patients with nephrotuberculosis]. *Probl. Tuberk.* 42–43 (2002).

389. Teye, K. *et al.* A novel I247T missense mutation in the haptoglobin 2 β-chain decreases the expression of the protein and is associated with ahaptoglobinemia. *Hum. Genet.* **114**, 499–502 (2004).

390. Bjornsson, E. *et al.* A rare splice donor mutation in the haptoglobin gene associates with blood lipid levels and coronary artery disease. *Hum. Mol. Genet.* **26**, 2364–2376 (2017).

391. Cahill, L. E. *et al.* Currently available versions of genome-wide association studies cannot be used to query the common haptoglobin copy number variant. *J. Am. Coll. Cardiol.* **62**, 860–861 (2013).

392. Boettger, L. M. *et al.* Recurring exon deletions in the HP (haptoglobin) gene contribute to lower blood cholesterol levels. *Nat. Genet.* **48**, 359–366 (2016).

393. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).

394. Kazmi, N. *et al.* Genetic determinants of circulating haptoglobin concentration. *Clin. Chim. Acta Int. J. Clin. Chem.* **494**, 138–142 (2019).

395. Berry, F. B. *et al.* Positive and negative regulation of myogenic differentiation of C2C12 cells by isoforms of the multiple homeodomain zinc finger transcription factor ATBF1. *J. Biol. Chem.* **276**, 25057–25065 (2001).

396. Parsons, M. J. *et al.* The Regulatory Factor ZFHX3 Modifies Circadian Function in SCN via an AT Motif-Driven Axis. *Cell* **162**, 607–621 (2015).

397. Kao, Y.-H. *et al.* ZFHX3 knockdown increases arrhythmogenesis and dysregulates

calcium homeostasis in HL-1 atrial myocytes. *Int. J. Cardiol.* **210**, 85–92 (2016).

398. van Ouwerkerk, A. F. *et al.* Identification of atrial fibrillation associated genes and functional non-coding variants. *Nat. Commun.* **10**, 4755 (2019).

399. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).

400. Brown, B. C. & Knowles, D. A. Phenome-scale causal network discovery with bidirectional mediated Mendelian randomization. *bioRxiv* 2020.06.18.160176 (2020) doi:10.1101/2020.06.18.160176.

401. Lim, W.-H. *et al.* Impact of Hemoglobin Levels and Their Dynamic Changes on the Risk of Atrial Fibrillation: A Nationwide Population-Based Study. *Sci. Rep.* **10**, 6762 (2020).

402. Pritchard, J. K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).

403. Inouye, M. *et al.* Metabonomic, transcriptomic, and genomic variation of a population cohort. *Mol. Syst. Biol.* **6**, 441 (2010).

404. Kettunen, J. *et al.* Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat. Genet.* **44**, 269–276 (2012).

405. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

406. Lyon, M. S. *et al.* The variant call format provides efficient and robust storage of GWAS summary statistics. *Genome Biol.* **22**, 32 (2021).

407. Arnold, M., Raffler, J., Pfeufer, A., Suhre, K. & Kastenmüller, G. SNiPA: an interactive, genetic variant-centered annotation browser. *Bioinforma. Oxf. Engl.* **31**, 1334–1336 (2015).

408. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19**, 807–812 (2011).

409. Fraser, A. *et al.* Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int. J. Epidemiol.* **42**, 97–110 (2013).

410. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).

411. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535–550 (2016).

412. Eraslan, B. *et al.* Quantification and discovery of sequence determinants of protein-per-mRNA amount in 29 human tissues. *Mol. Syst. Biol.* **15**, e8513 (2019).

413. Martin, R. I. R. *et al.* Genetic variants associated with risk of atrial fibrillation regulate expression of PITX2, CAV1, MYOZ1, C9orf3 and FANCC. *J. Mol. Cell. Cardiol.* **85**, 207–214 (2015).

414. Issac, T. T., Dokainish, H. & Lakkis, N. M. Role of inflammation in initiation and perpetuation of atrial fibrillation: a systematic review of the published data. *J. Am. Coll. Cardiol.* **50**, 2021–2028 (2007).

415. Holme, I., Aastveit, A. H., Hammar, N., Jungner, I. & Walldius, G. Haptoglobin and risk of myocardial infarction, stroke, and congestive heart failure in 342,125 men and women in the Apolipoprotein MOrtality RISk study (AMORIS). *Ann. Med.* **41**, 522–532 (2009).

416. Engström, G. *et al.* Inflammation-sensitive plasma proteins, diabetes, and mortality and incidence of myocardial infarction and stroke: a population-based study. *Diabetes* **52**, 442–447 (2003).

417. Engström, G. *et al.* Effects of cholesterol and inflammation-sensitive plasma proteins

on incidence of myocardial infarction and stroke in men. *Circulation* **105**, 2632–2637 (2002).

418. January, C. T. *et al.* 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on practice guidelines and the Heart Rhythm Society. *Circulation* **130**, e199-267 (2014).

419. Kirchhof, P. *et al.* 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS. *Eur. Heart J.* **37**, 2893–2962 (2016).

420. Haycock, P. C. *et al.* Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. *Am. J. Clin. Nutr.* **103**, 965–978 (2016).

421. Fill, J., Fokina, A. & Klappacher, G. Genetically determined atrial fibrillation and risk of stroke: a Mendelian randomization study. *Eur. Heart J.* **41**, (2020).

422. Hou, L. *et al.* Exploring the causal pathway from ischemic stroke to atrial fibrillation: a network Mendelian randomization study. *Mol. Med.* **26**, 7 (2020).

423. Timpson, N. J. *et al.* C-reactive protein levels and body mass index: elucidating direction of causation through reciprocal Mendelian randomization. *Int. J. Obes. 2005* **35**, 300–308 (2011).

424. Hemani, G., Tilling, K. & Davey Smith, G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet.* **13**, e1007081 (2017).

425. Hemani, G., Tilling, K. & Davey Smith, G. Correction: Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet.* **13**, e1007149 (2017).

426. Davies, N. M. *et al.* The many weak instruments problem and Mendelian randomization. *Stat. Med.* **34**, 454–468 (2015).

427. Davies, N. M., Holmes, M. V. & Davey Smith, G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ* **362**, k601 (2018).

428. Zeng, P., Wang, T., Zheng, J. & Zhou, X. Causal association of type 2 diabetes with amyotrophic lateral sclerosis: new evidence from Mendelian randomization using GWAS summary statistics. *BMC Med.* **17**, 225 (2019).

429. Cragg, J. G. & Donald, S. G. Testing Identifiability and Specification in Instrumental Variable Models. *Econom. Theory* **9**, 222–240 (1993).

430. Cochran, W. G. The comparison of percentages in matched samples. *Biometrika* **37**, 256–266 (1950).

431. Kiliszek, M. *et al.* Association between Variants on Chromosome 4q25, 16q22 and 1q21 and Atrial Fibrillation in the Polish Population. *PLOS ONE* **6**, e21790 (2011).

432. Ntalla, I. *et al.* Multi-ancestry GWAS of the electrocardiographic PR interval identifies 202 loci underlying cardiac conduction. *Nat. Commun.* **11**, 2542 (2020).

433. Verweij, N. *et al.* The Genetic Makeup of the Electrocardiogram. *Cell Syst.* **11**, 229-238.e5 (2020).

434. Burgess, S., Thompson, S. G., & CRP CHD Genetics Collaboration. Avoiding bias from weak instruments in Mendelian randomization studies. *Int. J. Epidemiol.* **40**, 755–764 (2011).

435. Pietzner, M. *et al.* Mapping the proteo-genomic convergence of human diseases. *Science* **374**, eabj1541 (2021).

436. Franceschini, N. *et al.* GWAS and colocalization analyses implicate carotid intima-media thickness and carotid plaque loci in cardiovascular outcomes. *Nat. Commun.* **9**, 5141 (2018).

437. Liu, B. *et al.* Genetic Regulatory Mechanisms of Smooth Muscle Cells Map to

Coronary Artery Disease Risk Loci. *Am. J. Hum. Genet.* **103**, 377–388 (2018).

438. Maximiano, S., Magalhães, P., Guerreiro, M. P. & Morgado, M. Trastuzumab in the Treatment of Breast Cancer. *BioDrugs Clin. Immunother. Biopharm. Gene Ther.* **30**, 75–86 (2016).

439. Derakhshani, A. *et al.* Overcoming trastuzumab resistance in HER2-positive breast cancer using combination therapy. *J. Cell. Physiol.* **235**, 3142–3156 (2020).

440. Dong, C. & Wu, G. Regulation of anterograde transport of adrenergic and angiotensin II receptors by Rab2 and Rab6 GTPases. *Cell. Signal.* **19**, 2388–2399 (2007).

441. Heidecker, B. *et al.* The gene expression profile of patients with new-onset heart failure reveals important gender-specific differences. *Eur. Heart J.* **31**, 1188–1196 (2010).

442. Fukuda, R. *et al.* Metabolic modulation regulates cardiac wall morphogenesis in zebrafish. *eLife* **8**, e50161.

443. D'Uva, G. *et al.* ERBB2 triggers mammalian heart regeneration by promoting cardiomyocyte dedifferentiation and proliferation. *Nat. Cell Biol.* **17**, 627–638 (2015).

444. Aharonov, A. *et al. ERBB2 drives YAP activation and EMT-like processes during cardiac regeneration.* 2020.01.07.897199 https://www.biorxiv.org/content/10.1101/2020.01.07.897199v1 (2020) doi:10.1101/2020.01.07.897199.

445. Tsantilas, P. *et al.* Chitinase 3 like 1 (CHI3L1) is a regulator of smooth muscle cell physiology and atherosclerotic lesion stability. *Cardiovasc. Res.* cvab014 (2021) doi:10.1093/cvr/cvab014.

446. Bai, H.-L. *et al.* Microarray profiling analysis and validation of novel long noncoding RNAs and mRNAs as potential biomarkers and their functions in atherosclerosis. *Physiol. Genomics* **51**, 644–656 (2019).

447. Rani, R. & Singh, V. Overexpression of YKL-40 (CHI3L1 gene) in patient fluids may be a potential predictive marker for early detection of comorbidity in non-communicable disease. *Med. Hypotheses* **143**, 110076 (2020).

448. Wallentin, L. *et al.* Plasma proteins associated with cardiovascular death in patients with chronic coronary heart disease: A retrospective study. *PLoS Med.* **18**, e1003513 (2021).

449. Kepenek, E. S. *et al.* Differential expression of genes participating in cardiomyocyte electrophysiological remodeling via membrane ionic mechanisms and Ca2+-handling in human heart failure. *Mol. Cell. Biochem.* **463**, 33–44 (2020).

450. Sag, C. M. *et al.* Calcium/calmodulin-dependent protein kinase II contributes to cardiac arrhythmogenesis in heart failure. *Circ. Heart Fail.* **2**, 664–675 (2009).

451. Lu, Y.-M. *et al.* DY-9760e inhibits endothelin-1-induced cardiomyocyte hypertrophy through inhibition of CaMKII and ERK activities. *Cardiovasc. Ther.* **27**, 17–27 (2009).

452. Xue, Y. *et al.* Insulin-like growth factor binding protein 4 enhances cardiomyocytes induction in murine-induced pluripotent stem cells. *J. Cell. Biochem.* **115**, 1495–1504 (2014).

453. Wo, D. *et al.* IGFBP-4 enhances VEGF-induced angiogenesis in a mouse model of myocardial infarction. *J. Cell. Mol. Med.* **24**, 9466–9471 (2020).

454. Lynch, J. R. *et al.* JMJD1C-mediated metabolic dysregulation contributes to HOXA9-dependent leukemogenesis. *Leukemia* **33**, 1400–1410 (2019).

455. Björn, N. *et al.* Genes and variants in hematopoiesis-related pathways are associated with gemcitabine/carboplatin-induced thrombocytopenia. *Pharmacogenomics J.* **20**, 179–191 (2020).

456. Folkersen, L. *et al.* Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat. Metab.* **2**, 1135–1148 (2020).

457. Ściborski, K. *et al.* Plasma YKL-40 levels correlate with the severity of coronary atherosclerosis assessed with the SYNTAX score. *Pol. Arch. Intern. Med.* **128**, 644–648 (2018).

458. Marott, S. C. W. *et al.* YKL-40 levels and atrial fibrillation in the general population. *Int. J. Cardiol.* **167**, 1354–1359 (2013).

459. Li, S. *et al.* Requirement for serum response factor for skeletal muscle growth and maturation revealed by tissue-specific gene deletion in mice. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 1082–1087 (2005).

460. Gaudet, P., Livstone, M. S., Lewis, S. E. & Thomas, P. D. Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Brief. Bioinform.* **12**, 449–462 (2011).

461. Chong, M. *et al.* Novel Drug Targets for Ischemic Stroke Identified Through Mendelian Randomization Analysis of the Blood Proteome. *Circulation* **140**, 819–830 (2019).

462. Gong, P. *et al.* [Total cholesterol mediates the effect of ABO blood group on coronary heart disease]. *Zhonghua Xin Xue Guan Bing Za Zhi* **43**, 404–407 (2015).

463. Pickrell, J. K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709–717 (2016).

464. Yang, F. *et al.* Haptoglobin reduces lung injury associated with exposure to blood. *Am. J. Physiol.-Lung Cell. Mol. Physiol.* (2003) doi:10.1152/ajplung.00115.2002.

465. Arredouani, M. S. *et al.* Haptoglobin dampens endotoxin-induced inflammatory effects both in vitro and in vivo. *Immunology* **114**, 263–271 (2005).

466. Lee, P.-L. *et al.* Relationships of Haptoglobin Phenotypes with Systemic Inflammation and the Severity of Chronic Obstructive Pulmonary Disease. *Sci. Rep.* **9**, 189 (2019).

467. de Kleijn, D. P. V. *et al.* Acute-phase protein haptoglobin is a cell migration factor

involved in arterial restructuring. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* **16**, 1123–1125 (2002).

468. Lohr, N. L., Warltier, D. C., Chilian, W. M. & Weihrauch, D. Haptoglobin expression and activity during coronary collateralization. *Am. J. Physiol. Heart Circ. Physiol.* **288**, H1389-1395 (2005).

469. Oh, M.-K., Park, H.-J., Lee, J.-H., Bae, H.-M. & Kim, I.-S. Single chain precursor prohaptoglobin promotes angiogenesis by upregulating expression of vascular endothelial growth factor (VEGF) and VEGF receptor2. *FEBS Lett.* **589**, 1009–1017 (2015).

470. Oh, M.-K. & Kim, I.-S. Involvement of placental growth factor upregulated via TGF-β1-ALK1-Smad1/5 signaling in prohaptoglobin-induced angiogenesis. *PLOS ONE* **14**, e0216289 (2019).

471. Levy, A. P. *et al.* Haptoglobin genotype is a determinant of iron, lipid peroxidation, and macrophage accumulation in the atherosclerotic plaque. *Arterioscler. Thromb. Vasc. Biol.* **27**, 134–140 (2007).

472. Wang, S. *et al.* Mendelian randomization analysis to assess a causal effect of haptoglobin on macroangiopathy in Chinese type 2 diabetes patients. *Cardiovasc. Diabetol.* **17**, 14 (2018).

473. Can, U. *et al.* Investigation of the inflammatory biomarkers of metabolic syndrome in adolescents. *J. Pediatr. Endocrinol. Metab. JPEM* **29**, 1277–1283 (2016).

474. Awadallah, S. *et al.* Plasma levels of Apolipoprotein A1 and Lecithin:Cholesterol Acyltransferase in type 2 diabetes mellitus: Correlations with haptoglobin phenotypes. *Diabetes Metab. Syndr.* **11 Suppl 2**, S543–S546 (2017).

475. Melander, O. *et al.* New circulating biomarkers for predicting cardiovascular death in healthy population. *J. Cell. Mol. Med.* **19**, 2489–2499 (2015).

476. Wang, S. *et al.* Association between serum haptoglobin and carotid arterial functions: usefulness of a targeted metabolomics approach. *Cardiovasc. Diabetol.* **18**, 8 (2019).

477. Wilcox, A. G., Vizor, L., Parsons, M. J., Banks, G. & Nolan, P. M. Inducible Knockout of Mouse Zfhx3 Emphasizes Its Key Role in Setting the Pace and Amplitude of the Adult Circadian Clock. *J. Biol. Rhythms* **32**, 433–443 (2017).

478. Tsai, C.-T., Lin, J.-L., Lai, L.-P., Lin, C.-S. & Huang, S. K. S. Membrane translocation of small GTPase Rac1 and activation of STAT1 and STAT3 in pacing-induced sustained atrial fibrillation. *Heart Rhythm* **5**, 1285–1293 (2008).

479. Sun, X., Li, J., Dong, F. N. & Dong, J.-T. Characterization of nuclear localization and SUMOylation of the ATBF1 transcription factor in epithelial cells. *PloS One* **9**, e92746 (2014).

480. Maejima, Y. & Sadoshima, J. SUMOylation: a novel protein quality control modifier in the heart. *Circ. Res.* **115**, 686–689 (2014).

481. Zhou, M., Liao, Y. & Tu, X. The role of transcription factors in atrial fibrillation. *J. Thorac. Dis.* **7**, (2015).

482. Hall, A. W. *et al.* Epigenetic Analyses of Human Left Atrial Tissue Identifies Gene Networks Underlying Atrial Fibrillation. *Circ. Genomic Precis. Med.* **13**, e003085 (2020).

Appendix

Appendix A. The CpG sites lookups

The CpG sites colocalized with AF in the association region in Phase I moloc were looked up for the annotated gene and physical position across the locus using the *meffil* R package. p-value (P)=0 represents P<10^{-325}

locus	colocalized CpG site	CpG pos (hg19, b37)	CpG gene annotation	Beta	P
1p36	cg16583536	22282684	NA	-0.437	2.64E-292
1q21	cg19233405	154988721	ZBTB7B	0.180	3.07E-76
1q24	cg22693806	170631847	PRRX1	0.208	2.83E-133
1q32	cg03900565	203031815	PPFIA4	-0.226	9.15E-153
	cg11656175	203040823	PPFIA4	-0.129	4.96E-51
	cg23098069	203055507	MYOG	0.162	1.74E-70
3p25	cg24848339	12840334	CAND2	0.325	3.64E-298
3p14	cg15724417	66433521	LRIG1	0.099	1.30E-25
4q34	cg24950233	174457944	NBLA00301	-0.520	0
	cg18575740	174457224	NBLA00301	-0.238	2.11E-139
	cg13935962	174451443	HAND2;NBLA00301	0.186	5.84E-86
5q35	cg13004182	172665707	NA	-0.214	3.44E-91
	cg12825773	172662463	NKX2-5	0.362	4.27E-267
	cg18839504	172665119	NA	0.254	2.93E-129
	cg06889108	173317342	CPEB4	0.322	1.68E-296
7q21	cg10481072	92459517	CDK6	-0.098	7.692e-25
7q32	cg18693656	128482739	FLNC	-0.498	3.59E-103
	cg13951589	128482287	FLNC	-0.679	1.69E-242
	cg10826733	128482561	FLNC	-0.751	1.15E-277
8q24	cg26291848	141608316	EIF2C2	0.404	0
	cg14396066	142010253	PTK2	0.316	6.15E-295
	cg10996527	141994633	PTK2	-0.324	9.88E-324
9q34	cg04455058	139085579	NA	0.433	4.03E-195
10q21	cg01631684	65280961	REEP3	-0.314	5.99E-304
10q22	cg16228286	75407372	SYNPO2L	-0.129	4.21E-25
	cg24637261	75410817	SYNPO2L	0.117	7.40E-21
10q24	cg17426192	105344747	NEURL	0.388	3.25E-228
12p12	cg22232504	26349469	SSPN	0.293	1.48E-216
	cg11332519	26348093	SSPN	0.199	5.55E-88
	cg07725355	26348333	SSPN	-0.252	0
	cg02593205	26348343	SSPN	0.219	2.96E-119
14q24	cg25949241	73264662	DPF3	-0.101	0

locus	colocalized CpG site	CpG pos (hg19, b37)	CpG gene annotation	Beta	P
15q24	cg10576051	73661652	HCN4	0.203	4.50E-64
	cg06071033	73662258	HCN4	-0.240	2.64E-88
	cg01796676	73680284	NA	-0.411	4.73E-265
15q25	cg12292492	80992262	FAM108C1	-0.055	8.01E-11
	cg13148921	80853140	ARNT2	-0.319	0
16q22	cg03463523	73099917	NA	0.143	1.87E-36
17p13	cg01557754	7342661	FGF11	-0.373	4.94e-324
17q12	cg22833065	38095691	NA	0.124	1.25E-44
17q25	cg23834688	76798392	USP36	-0.558	0

Appendix B. MR analyses results tables

MR estimates for five different methods at $5\text{x}10^{-8}$ instrument threshold in (a) the AF=>IS (IVs=111) and AF=>stroke (IVs=111) MR studies and in (b) IS=>AF (IVs=9) and stroke=>AF (IVs=8) MR studies.

a)

exposure	outcome	MR method	Beta	SE	P
AF	IS	IVW	0.213	0.019	$4.80\text{x}10^{-29}$
	stroke		0.200	0.018	$1.16\text{x}10^{-27}$
	IS	MR Egger	0.212	0.037	$8.62\text{x}10^{-8}$
	stroke		0.188	0.036	$6.81\text{x}10^{-7}$
	IS	Simple mode	0.272	0.062	$2.89\text{x}10^{-5}$
	stroke		0.237	0.056	$4.20\text{x}10^{-5}$
	IS	Weighted median	0.252	0.026	$4.78\text{x}10^{-22}$
	stroke		0.225	0.024	$4.59\text{x}10^{-21}$
	IS	Weighted mode	0.248	0.031	$1.61\text{x}10^{-12}$
	stroke		0.218	0.025	$2.26\text{x}10^{-14}$

b)

exposure	outcome	MR method	nSNP	Beta	SE	P
IS	AF	IVW	9	0.768	0.504	0.128
Stroke			8	0.854	0.586	0.145
IS		MR Egger	9	6.821	3.139	0.066
Stroke			8	6.425	3.812	0.143
IS		Simple mode	9	0.073	0.069	0.323
Stroke			8	0.120	0.089	0.218
IS		Weighted median	9	0.048	0.057	0.396
Stroke			8	0.101	0.064	0.115
IS		Weighted mode	9	0.046	0.061	0.470
Stroke			8	0.089	0.070	0.245

Appendix C. SNPs failed the Steiger directionality analysis.

SNPs failed the Steiger directionality analysis after performing the Steiger test for each SNP. Leave-one-out (LOO), SNP found as an outlier in the LOO analysis (TRUE), SNP was not found as an outlier in the LOO analysis (FALSE).

Study	instrument threshold	MR analysis	SNP	effect allele	other allele	r^2 exposure	r^2 outcome	Steiger P	LOO SNP
Main	5x10^{-8}	AF to stroke	rs284277	A	C	3.58E-05	5.37E-05	0.452	not tested
		stroke to AF (reverse)	rs2634074	A	T	1.26E-04	8.86E-04	3.89E-25	TRUE
	5x10^{-5}	AF to stroke	rs2149515	G	T	1.64E-05	2.04E-05	0.792	not tested
			rs284277	A	C	3.58E-05	5.37E-05	0.452	not tested
			rs4980386	A	C	2.22E-05	2.40E-05	0.917	not tested
		stroke to AF (reverse)	rs2634074	A	T	1.26E-04	8.86E-04	3.89E-25	TRUE
			rs4151702	C	G	4.94E-05	5.17E-05	0.927	FALSE
			rs6838973	T	C	4.55E-05	4.89E-04	9.41E-18	TRUE
Validation	5x10^{-8}	AF to stroke	rs880315	C	T	6.36E-05	6.53E-05	0.957	not tested
		stroke to AF (reverse)	rs2634074	A	T	1.26E-04	1.70E-03	1.08E-49	TRUE
	5x10^{-5}	AF to stroke	rs7333028	C	T	4.56E-05	6.50E-05	0.518	not tested
			rs880315	C	T	6.36E-05	6.53E-05	0.957	not tested
		stroke to AF (reverse)	rs2634074	A	T	1.26E-04	1.70E-03	1.08E-49	TRUE
			rs4151702	C	G	4.94E-05	7.47E-05	0.425	FALSE
			rs6838973	T	C	4.55E-05	1.20E-03	2.81E-43	TRUE

Appendix D. MR analyses results table

Results from MR analyses before and after removing instruments which failed the Steiger test to determine if the MR estimates were altered.

Study	exposure	outcome	instrument threshold	MR method	nSNP	Beta	SE	P	MR analysis
Main	AF	Stroke	5x10^{-8}	IVW	111	0.200	0.018	1.16E-27	all SNPs
				IVW	110	0.197	0.018	1.03E-28	steiger filtered
				MR Egger	111	0.188	0.036	6.81E-07	all SNPs
				MR Egger	110	0.195	0.034	1.21E-07	steiger filtered
				Simple mode	111	0.237	0.052	1.45E-05	all SNPs
				Simple mode	110	0.239	0.056	3.82E-05	steiger filtered
				Weighted median	111	0.225	0.024	2.04E-20	all SNPs
				Weighted median	110	0.225	0.023	7.62E-22	steiger filtered
				Weighted mode	111	0.218	0.024	4.12E-15	all SNPs
				Weighted mode	110	0.222	0.026	1.05E-13	steiger filtered
			5x10^{-5}	IVW	337	0.186	0.013	3.87E-44	all SNPs
				IVW	334	0.181	0.013	2.32E-44	steiger filtered
				MR Egger	337	0.201	0.026	1.34E-13	all SNPs
				MR Egger	334	0.207	0.025	5.41E-15	steiger filtered
				Simple mode	337	0.184	0.055	8.81E-04	all SNPs
				Simple mode	334	0.185	0.054	7.00E-04	steiger filtered
				Weighted median	337	0.206	0.023	3.44E-19	all SNPs
				Weighted median	334	0.205	0.021	3.51E-23	steiger filtered
				Weighted mode	337	0.211	0.025	3.13E-16	all SNPs
				Weighted mode	334	0.213	0.026	7.47E-15	steiger filtered
	Stroke	AF	5x10^{-8}	IVW	8	0.854	0.586	0.145	all SNPs
				IVW	7	0.166	0.067	0.013	steiger filtered
				MR Egger	8	6.425	3.812	0.143	all SNPs
				MR Egger	7	-0.621	0.475	0.248	steiger filtered
				Simple mode	8	0.120	0.087	0.210	all SNPs
				Simple mode	7	0.060	0.103	0.580	steiger filtered
				Weighted median	8	0.101	0.064	0.114	all SNPs
				Weighted median	7	0.100	0.063	0.113	steiger filtered
				Weighted mode	8	0.089	0.068	0.233	all SNPs
				Weighted mode	7	0.063	0.090	0.513	steiger filtered
			5x10^{-5}	IVW	162	0.212	0.049	1.83E-05	all SNPs
				IVW	159	0.119	0.018	1.62E-11	steiger filtered
				MR Egger	162	0.081	0.126	0.519	all SNPs
				MR Egger	159	-0.005	0.044	0.910	steiger filtered

Study	exposure	outcome	instrument threshold	MR method	nSNP	Beta	SE	P	MR analysis
				Simple mode	162	0.062	0.052	0.229	all SNPs
				Simple mode	159	0.065	0.048	0.185	steiger filtered
				Weighted median	162	0.060	0.019	0.002	all SNPs
				Weighted median	159	0.060	0.020	0.003	steiger filtered
				Weighted mode	162	0.053	0.047	0.270	all SNPs
				Weighted mode	159	0.056	0.047	0.238	steiger filtered
Validation	AF	Stroke	5x10^{-8}	IVW	90	0.191	0.020	6.88E-22	all SNPs
				IVW	89	0.187	0.019	1.22E-23	steiger filtered
				MR Egger	90	0.133	0.049	7.99E-03	all SNPs
				MR Egger	89	0.150	0.046	1.66E-03	steiger filtered
				Simple mode	90	0.161	0.051	2.07E-03	all SNPs
				Simple mode	89	0.162	0.052	2.62E-03	steiger filtered
				Weighted median	90	0.164	0.024	1.72E-11	all SNPs
				Weighted median	89	0.164	0.025	6.48E-11	steiger filtered
				Weighted mode	90	0.179	0.036	3.96E-06	all SNPs
				Weighted mode	89	0.177	0.036	3.11E-06	steiger filtered
			5x10^{-5}	IVW	250	0.187	0.014	2.47E-38	all SNPs
				IVW	248	0.182	0.014	3.26E-40	steiger filtered
				MR Egger	250	0.166	0.028	7.46E-09	all SNPs
				MR Egger	248	0.171	0.026	4.14E-10	steiger filtered
				Simple mode	250	0.156	0.049	1.53E-03	all SNPs
				Simple mode	248	0.154	0.047	1.21E-03	steiger filtered
				Weighted median	250	0.193	0.021	3.11E-19	all SNPs
				Weighted median	248	0.193	0.023	2.50E-16	steiger filtered
				Weighted mode	250	0.184	0.021	3.84E-16	all SNPs
				Weighted mode	248	0.185	0.023	1.09E-14	steiger filtered
	Stroke	AF	5x10^{-8}	IVW	7	0.175	0.070	0.012	all SNPs
				IVW	7	0.175	0.070	0.012	steiger filtered
				MR Egger	7	-0.192	0.585	0.756	all SNPs
				MR Egger	7	-0.192	0.585	0.756	steiger filtered
				Simple mode	7	0.142	0.079	0.124	all SNPs
				Simple mode	7	0.142	0.083	0.139	steiger filtered
				Weighted median	7	0.129	0.061	0.036	all SNPs
				Weighted median	7	0.129	0.061	0.033	steiger filtered
				Weighted mode	7	0.140	0.074	0.107	all SNPs
				Weighted mode	7	0.140	0.075	0.112	steiger filtered
			5x10^{-5}	IVW	135	0.182	0.036	3.70E-07	all SNPs
				IVW	134	0.153	0.019	2.44E-15	steiger filtered

Study	exposure	outcome	instrument threshold	MR method	nSNP	Beta	SE	P	MR analysis
				MR Egger	135	0.020	0.089	0.824	all SNPs
				MR Egger	134	0.087	0.048	0.075	steiger filtered
				Simple mode	135	0.065	0.063	0.305	all SNPs
				Simple mode	134	0.067	0.061	0.278	steiger filtered
				Weighted median	135	0.123	0.022	2.95E-08	all SNPs
				Weighted median	134	0.123	0.022	1.30E-08	steiger filtered
				Weighted mode	135	0.065	0.058	0.266	all SNPs
				Weighted mode	134	0.060	0.060	0.316	steiger filtered

Appendix E. Table of outlier exclusion analysis in the main MR analysis.

exposure	outcome	cut-off	Outlier exclusion analysis	MR method	nSNP	Beta	SE	P
Stroke	AF	$5\text{x}10^{-8}$	exclude rs2634074 outlier SNP	MR Egger	7	-0.621	0.475	0.248
				Weighted median		0.100	0.066	0.127
				IVW		0.166	0.067	0.013
				Simple mode		0.060	0.108	0.597
				Weighted mode		0.063	0.085	0.489
		$5\text{x}10^{-5}$	exclude rs2634074 outlier SNP	MR Egger	161	-0.054	0.070	0.439
				Weighted median		0.060	0.020	0.002
				IVW		0.146	0.028	$2.04\text{x}10^{-7}$
				Simple mode		0.062	0.052	0.231
				Weighted mode		0.054	0.046	0.244
			exclude rs2634074 and rs6838973 outlier SNPs	MR Egger	160	-0.007	0.046	0.885
				Weighted median		0.060	0.019	0.002
				IVW		0.126	0.019	$1.79\text{x}10^{-11}$
				Simple mode		0.063	0.050	0.210
				Weighted mode		0.056	0.047	0.232

Appendix F. Phenome-wide association (PheWAS) analyses results.

Results from PheWAS analyses conducted on rs2634074 (effete allele=A and other allele=T) and rs6838973 (effete allele=T and other allele=C) outliers against all GWAS traits available in the IEU OpenGWAS database[10].

SNP	SE	Beta	P	Trait
rs2634074	0.012	-0.327	2.69E-154	Arrhythmia
	0.015	-0.374	4.54E-145	Atrial fibrillation
	0.025	-0.593	8.81E-121	Atrial Fibrillation
	0.000	-0.007	1.80E-112	Diagnoses - secondary ICD10: I48 Atrial fibrillation and flutter
	0.000	-0.006	7.69E-111	Diagnoses - main ICD10: I48 Atrial fibrillation and flutter
	0.000	-0.005	1.30E-106	Non-cancer illness code, self-reported: atrial fibrillation
	0.000	-0.009	2.11E-84	Cardiac arrythmias, COPD co-morbidities
	0.000	-0.006	4.74E-78	Diagnoses - main ICD10: I48 Atrial fibrillation and flutter
	0.000	-0.003	2.90E-60	Operative procedures - main OPCS: X50.1 Direct current cardioversion
	0.000	-0.004	5.70E-50	Treatment/medication code: warfarin
	0.000	-0.004	2.00E-44	Diagnoses - secondary ICD10: Z92.1 Personal history of long-term (current) use of anticoagulants
	0.022	-0.298	6.82E-41	Ischemic stroke (cardioembolic)
	0.000	-0.004	1.09E-34	Treatment/medication code: warfarin
	0.010	-0.077	8.54E-16	Ischemic stroke
	0.001	-0.009	3.16E-15	Diseases of the circulatory system
	0.009	-0.072	3.35E-15	Stroke
	0.012	-0.094	5.90E-15	Ischemic stroke
	0.039	-0.301	1.28E-14	Cardioembolic stroke
	0.011	-0.084	6.56E-14	Stroke
	0.000	-0.001	8.13E-14	Treatment/medication code: flecainide
	0.001	-0.005	1.80E-13	Treatment speciality of consultant (recoded): Cardiology
	0.001	-0.004	2.00E-12	Main speciality of consultant (recoded): Cardiology
	0.000	-0.002	3.10E-09	Treatment/medication code: bisoprolol
	0.000	-0.002	3.66E-09	Treatment/medication code: bisoprolol
	0.000	-0.002	4.37E-08	STROKE
	0.000	-0.002	1.17E-07	Ischaemic Stroke, excluding all haemorrhages
	0.000	-0.001	1.50E-07	Operative procedures - main OPCS: H01.2 Emergency excision of abnormal appendix NEC
	0.000	-0.002	2.15E-07	Stroke, excluding SAH
	0.000	-0.001	2.96E-07	Diagnoses - main ICD10: I63 Cerebral infarction
	0.000	-0.001	4.08E-07	Diagnoses - main ICD10: K35 Acute appendicitis
	0.020	-0.103	4.20E-07	Ischemic stroke
	0.000	-0.001	5.10E-07	Non-cancer illness code, self-reported: heart arrhythmia
	0.001	-0.006	7.60E-07	Operative procedures - main OPCS: X99.8 No procedure performed
	0.000	-0.002	2.06E-06	Stroke, including SAH

SNP	SE	Beta	P	Trait
	0.016	-0.076	3.23E-06	Congestive heart failure
	0.001	-0.004	7.50E-06	Main speciality of consultant (recoded): General medicine
rs6838973	0.007	-0.184	1.35E-142	Atrial fibrillation
	0.007	-0.151	1.03E-111	Atrial fibrillation
	0.011	-0.199	7.25E-69	Arrhythmia
	0.021	-0.304	1.72E-45	Atrial Fibrillation
	0.000	-0.003	7.40E-34	Diagnoses - main ICD10: I48 Atrial fibrillation and flutter
	0.013	-0.156	1.69E-33	Atrial fibrillation
	0.000	-0.004	2.46E-30	Cardiac arrhythmias, COPD co-morbidities
	0.000	-0.003	4.30E-27	Diagnoses - secondary ICD10: I48 Atrial fibrillation and flutter
	0.000	-0.002	9.70E-27	Non-cancer illness code, self-reported: atrial fibrillation
	0.000	-0.003	7.21E-25	Diagnoses - main ICD10: I48 Atrial fibrillation and flutter
	0.000	-0.002	1.30E-14	Treatment/medication code: warfarin
	0.023	-0.171	3.71E-14	Atrial fibrillation and flutter
	0.000	-0.002	1.27E-12	Treatment/medication code: warfarin
	0.000	-0.001	3.80E-12	Operative procedures - main OPCS: X50.1 Direct current cardioversion
	0.001	-0.006	1.75E-11	Diseases of the circulatory system
	0.000	-0.001	1.30E-10	Diagnoses - secondary ICD10: Z92.1 Personal history of long-term (current) use of anticoagulants
	0.016	-0.104	2.62E-10	Cardiac arrythmias, COPD co-morbidities
	0.000	-0.001	1.13E-08	Treatment/medication code: flecainide
	0.000	-0.001	2.80E-08	Treatment/medication code: digoxin
	0.020	-0.108	3.58E-08	Ischemic stroke (cardioembolic)
	0.010	-0.050	8.31E-07	Ischemic stroke
	0.009	-0.042	6.48E-06	Stroke