Genetic investigation of coronary artery disease by Mendelian randomization and transcriptome-wide association studies

Ling Li

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung einer Doktorin der Naturwissenschaften (Dr. rer. nat.) genehmigten Dissertation.

Vorsitz: Prof. Dr. Burkhard Rost
Prüfer*innen der Dissertation: 1. Prof. Dr. Julien Gagneur
2. Prof. Dr. Heribert Schunkert
3. apl. Prof. Dr. Bertram Müller-Myhsok

Die Dissertation wurde am 08.08.2022 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 30.07.2023 angenommen.
Acknowledgments

Foremost, I would like to thank the members of my dissertation committee including: Prof. Dr. Burkhard Rost, Prof. Dr. Julien Gagneur, Prof. Dr. Bertram Muller-Myhsok, and Prof. Dr. Heribert Schunkert, not only for their time and extreme patience, but for their intellectual contributions to my development as a scientist.

I want to express my sincere gratitude to my advisor Prof. Heribert Schunkert. It’s my honor to join his group. He is an inspiring mentor, very open-minded to new ideas, and willing to give support to young researchers. During my PhD study, he has closely supervised my work and given full support and guidance at all levels of my PhD research. He gave me great opportunities to train in research presentation and scientific writing, which are critical for my academic career. I appreciate all his knowledgeable input, kindness, enthusiasm, and encouragement throughout my PhD training.

To my supervisor Prof. Julien Gagneur, I thank him for providing the opportunity to pursue my PhD research at the Deutsches Herzzentrum München and Technische Universität München, where I can apply advanced bioinformatic techniques in translational research. He is a leading expert in bioinformatics and I have enhanced my technical skills under his guidance. I acknowledge the collaboration between the labs of Prof. Gagneur and Prof. Schunkert which inspired me to perform the first CAD TWAS allowing the great integration of ‘wet’ and ‘dry’ lab-based research. I thank his valuable advice for my PhD work.

I would like to express my special thanks to Dr. Lingyao Zeng and Dr. Zhifen Chen. I thank Lingyao for introducing me to Prof. Schunkert’s lab and giving me lots of help at beginning of my PhD study when I faced challenges as an international student. Zhifen is a passionate researcher who always has lots of new ideas. It’s very enjoyable to discuss science
topics with her. I appreciate her valuable advice and experiment validation for my TWAS findings. I also want to thank Zhifen for helping revise my dissertation.

To my lab mates, I thank them for the enjoyable lab life and support to my research work. I’d like to give special thanks to Dr. Moritz von Scheidt and Dr. Thorsten Keßler. Moritz has been continuously helping me to obtain data from external collaborators. Thorsten shared creative ideas for my PhD work. I also want to thank Shiyue Li who helped conduct validation experiments.

Last but not least, I want to thank my family and friends for their constant support through the ups and downs during my PhD study. This dissertation is dedicated to the memory of my father.
Abstract

Coronary artery disease (CAD) is the leading cause of death globally, which is in part caused by genetic variants. The majority of risk loci of CAD identified by genome-wide association study (GWAS) are located in non-coding regions that hamper their function interpretation. So, follow-up studies are needed.

Polygenic risk score (PRS) is a genetic estimate of an individual’s liability to a trait or disease, calculated from a set of independent risk variants based on large-scale GWASs. The PRS has been widely applied to study the genetic association between complex traits. Mendelian randomization (MR) uses genetic variants as instrumental variables to infer whether risk factors (exposures) causally affect a health outcome, which is broadly used in observational epidemiology. The combination of PRS and MR (‘PRS+MR’ strategy) could improve detection rates for causal relationships which can be particularly useful when evaluating associations between genetic liability for a given trait and hundreds of diverse health outcomes. Transcriptome-wide association studies (TWAS) have been recently proposed as an invaluable tool for annotating GWAS risk loci by investigating the potential gene expression regulatory mechanisms underlying variant-trait associations. The signals identified by TWAS reflect associations between genetically regulated expression (GReX) and complex traits. In this dissertation, we applied these two bioinformatic approaches to study the genetic association between intelligence and CAD and to filter risk genes for CAD in a tissue-specific fashion.

The first work is to study the genetic association between intelligence and CAD using the ‘PRS+MR’ strategy. In this study, we first estimated a genetic intelligence score (gIQ) in samples from ten CAD case-control studies (n=34,083) from CARDIoGRAMplusC4D cohorts and UK Biobank (UKB) (n=427,306) based on 242 variants independently associated with
intelligence. Meta-analysis using a fixed-effect size model indicated the increase of gIQ by one standard deviation (SD) led to a relative decrease of CAD risk by 5% (95% confidence interval [CI], 0.93 to 0.98; \( P=4.93\times10^{-5} \)). From UKB, we observed significant inverse correlations between gIQ and lifestyle factors of CAD including smoking, body mass index (BMI), type 2 diabetes (T2D), hypertension, and a positive correlation with high-density lipoprotein cholesterol (HDL-C). We also observed positive correlations for gIQ with measured intelligence and educational attainment. The association between gIQ and CAD risk was largely attenuated after the adjustment of lifestyle factors, suggesting mediatory roles of these lifestyle factors in the pathway of linking high gIQ and low risk of CAD. The effects of gIQ on lifestyle factors were also largely attenuated after the adjustment of measured intelligence and educational attainment, suggesting their mediatory roles in the pathways between gIQ and lifestyle factors. Finally, the associations between intelligence and CAD as well as its lifestyle factors were confirmed through the two-sample MR method. In conclusion, using genetic approaches, we depicted a pathway from gIQ to CAD risk. The higher gIQ is associated with the higher measured intelligence and longer educational attainment, both of which appear to reduce the prevalence of unfavourable factors of CAD including BMI, smoking, T2D, and hypertension, and increase HDL-C, which subsequently reduce the incidence of CAD.

The second work applied the TWAS methodology for filtering risk genes of CAD as well as for identification of the tissues in which differential expression affect rise. In this study, we first trained expression prediction models for nine risk tissues of CAD from the two largest reference panels, the Stockholm-Tartu Atherosclerosis Reverse Network Engineering Task (STARNET) and the Genotype-Tissue Expression (GTEx). We next applied these prediction models to imputing individual-level GReX from genotype data of 11 cohorts. By performing association analysis between GReX and CAD risk, we identified 114 transcriptome-wide significant genes (\( P<3.85\times10^{-6} \)). Of these, 96 resided within GWAS loci of CAD (known), and
18 independent of GWAS loci (novel). Stepwise analyses bridged the known genes with a series of pathophysiological pathways related to CAD, including lipid metabolism, inflammation, angiogenesis, high blood pressure, etc. The in-silico analyses showed that the novel genes were associated with lipid metabolism in both genotype data of human samples and expression data of an atherosclerosis mouse model. Of these novel genes, KPTN and RGS19, which were rarely studied before, gave significant signals in liver tissue by TWAS analysis. The CRISPR/Cas9-based gene knockdown of the two genes in the human hepatocyte cell line resulted in reduced secretion of APOB100 and lipids in the cell culture medium. In conclusion, our CAD TWAS work i) prioritized candidate genes in a tissue-specific fashion, and ii) identified 18 novel genes to be associated with CAD by their connection to lipid metabolism.

In summary, the two studies of my dissertation using different genetic approaches, on the one hand depicted the genetic pathway from high intelligence to low risk of CAD, and on the other hand pinpointed risk genes for CAD in a tissue-specific fashion. These studies expand our knowledge scope of the genetic etiology of CAD from different perspectives.
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAD</td>
<td>Coronary artery disease</td>
</tr>
<tr>
<td>GWAS</td>
<td>Genome-wide association study</td>
</tr>
<tr>
<td>PheWAS</td>
<td>Phenome-wide association study</td>
</tr>
<tr>
<td>TWAS</td>
<td>Transcriptome-wide association study</td>
</tr>
<tr>
<td>MR</td>
<td>Mendelian randomization</td>
</tr>
<tr>
<td>PRS</td>
<td>Polygenic risk score</td>
</tr>
<tr>
<td>GReX</td>
<td>Genetically regulated expression</td>
</tr>
<tr>
<td>NGS</td>
<td>Next-generation sequencing</td>
</tr>
<tr>
<td>1KG</td>
<td>The 1000 Genome project</td>
</tr>
<tr>
<td>HRC</td>
<td>The Haplotype Reference Consortium</td>
</tr>
<tr>
<td>WGS</td>
<td>Whole-genome sequencing</td>
</tr>
<tr>
<td>WES</td>
<td>Whole-exome sequencing</td>
</tr>
<tr>
<td>RNAseq</td>
<td>RNA sequencing</td>
</tr>
<tr>
<td>eQTL</td>
<td>Expression quantitative trait locus</td>
</tr>
<tr>
<td>GTEx</td>
<td>The Genotype-Tissue Expression project</td>
</tr>
<tr>
<td>STARNET</td>
<td>The Stockholm-Tartu Atherosclerosis Reverse Network Engineering Task</td>
</tr>
<tr>
<td>UKB</td>
<td>UK Biobank</td>
</tr>
<tr>
<td>LD</td>
<td>Linkage disequilibrium</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>gIQ</td>
<td>Genetic intelligence score</td>
</tr>
<tr>
<td>AOR</td>
<td>Atherosclerotic aortic root</td>
</tr>
<tr>
<td>Term</td>
<td>Description</td>
</tr>
<tr>
<td>-------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>COR</td>
<td>Coronary</td>
</tr>
<tr>
<td>MAM</td>
<td>Internal mammary artery</td>
</tr>
<tr>
<td>TIB</td>
<td>Tibial artery</td>
</tr>
<tr>
<td>LIV</td>
<td>Liver</td>
</tr>
<tr>
<td>SF</td>
<td>Subcutaneous fat</td>
</tr>
<tr>
<td>VAF</td>
<td>Visceral abdominal fat</td>
</tr>
<tr>
<td>SKLM</td>
<td>Skeletal muscle</td>
</tr>
<tr>
<td>BLD</td>
<td>Blood</td>
</tr>
<tr>
<td>LDL-C</td>
<td>Low-density lipoprotein cholesterol</td>
</tr>
<tr>
<td>HDL-C</td>
<td>High-density lipoprotein cholesterol</td>
</tr>
<tr>
<td>TG</td>
<td>Triglycerides</td>
</tr>
<tr>
<td>TC</td>
<td>Total cholesterol</td>
</tr>
<tr>
<td>APOA</td>
<td>Apolipoprotein A</td>
</tr>
<tr>
<td>APOB</td>
<td>Apolipoprotein B</td>
</tr>
</tbody>
</table>
# Table of Contents

Acknowledgments............................................................................................................. II

Abstract............................................................................................................................. IV

Abbreviations.................................................................................................................... VII

Table of Contents............................................................................................................... IX

I Introduction ..................................................................................................................... 1

1.1 Coronary artery disease (CAD) ................................................................................. 1

1.1.1 Epidemiology of CAD ....................................................................................... 1

1.1.2 Biological mechanisms underlying CAD ......................................................... 1

1.1.3 Risk factors for CAD ......................................................................................... 2

1.2 Genome-wide association study (GWAS) ............................................................... 2

1.3 Post-GWAS era ......................................................................................................... 5

1.3.1 Fine-mapping ..................................................................................................... 5

1.3.2 Phenome-wide association study (PheWAS) .................................................... 9

1.3.3 Mendelian randomization (MR) ........................................................................ 10

1.4 Genetics-of-gene-expression panels ....................................................................... 14

1.5 UK Biobank (UKB) ................................................................................................. 16

1.6 Rare mutation association study in complex disease .......................................... 17

1.7 Transcriptome-wide association study .................................................................. 20

1.7.1 Individual-level predictors ............................................................................... 21

1.7.2 Summary-based predictors ............................................................................... 22
1.7.3 Opportunities and challenges of TWAS .......................................................... 24

1.8 Polygenic risk score (PRS) .............................................................................. 25

1.8.1 Definition and analysis pipeline of PRS .................................................... 25

1.8.2 Application of PRS .................................................................................. 26

1.8.3 Further discussion about PRS ................................................................. 28

1.9 Genetics of CAD ............................................................................................ 30

1.9.1 GWAS of CAD ....................................................................................... 30

1.9.2 Prioritization of CAD causal genes by in-silico methods ......................... 31

II Methods .......................................................................................................... 32

2.1 Preprocessing of 11 GWAS cohorts .............................................................. 32

2.1.1 Ten CARDIoGRAMplusC4D cohorts .................................................... 35

2.1.2 Genotype data of UKB ............................................................................ 35

2.2 Methods used in the Intelligence-CAD project ............................................ 36

2.3 Methods used in the TWAS project ............................................................. 38

III Discussion ..................................................................................................... 40

3.1 Genetic association between Intelligence and CAD ..................................... 40

Author contributions ......................................................................................... 42

3.2 TWAS for CAD ............................................................................................ 42

Author contributions ......................................................................................... 46

IV Conclusion and outlook ................................................................................. 46

References ......................................................................................................... 48
Publications during PhD .................................................................102
Approval from each publisher ..........................................................103
Appendix I Genetically determined intelligence and coronary artery disease risk........104
Appendix II Transcriptome-wide association study of coronary artery disease identifies novel susceptibility genes ..................................................................................................................114
I Introduction

1.1 Coronary artery disease (CAD)

1.1.1 Epidemiology of CAD

Epidemiologic and experimental cardiovascular research has improved the understanding of coronary artery disease (CAD) pathophysiology. Preventive and therapeutic strategies based on this knowledge decreased age-adjusted CAD mortality by over 50% in the past 30 years[1]. However, CAD remains the leading cause of death worldwide. Globally, CAD affects ~126 million individuals globally (1,655 persons per 100,000) representing 1.72% of the world population, and caused ~9 million deaths in 2017[2]. With the aging of the population, prevalence is expected to exceed 1845 persons per 100,000 by 2030. According to the European Commission and European Society of Cardiology (ESC), CAD causes 12.5% of all death and costs over €100 billion yearly in Europe. The enormous health care and economic burdens urge better preventive and therapeutic approaches to tackle the disease.

1.1.2 Biological mechanisms underlying CAD

CAD is clinically presented as the accumulation of atherosclerotic plaques within the wall of coronary arteries that provide nutrients and oxygen to the heart. CAD can be caused by dyslipidemia, dysfunction of endothelial cells (EC), vascular smooth muscle cells (VSMC) or fibroblasts, immune-inflammatory reactions, hyperinsulinism, and abnormal glucose metabolism[3]. Although the underlying mechanisms of CAD are intensively investigated the pathophysiology of CAD has not been fully elucidated yet. Generally, EC damage is thought to be the initial step of CAD, which may start early in life. These cells show dysfunction in regulating vascular tone via nitric oxide signaling. The increased amount of lipoprotein particles enhances activation and adhesion of monocytes and promotes the cholesterol-loading
of macrophages, or ‘foam cells’. VSMC and fibroblasts underlying the EC layer proliferate, migrate and remodel the vessels. In addition, inflammation and immune response trigger the vicious cycle of EC damage and deposition of fat and calcium inside the artery wall, all of which accelerate atherosclerotic plaque formation, narrow the vessels, and eventually obstruct the blood flow[4].

1.1.3 Risk factors for CAD

Various risk factors affect the multifactorial etiology of CAD. Risk factors of CAD could be classified into non- and modifiable types[5–7]. Non-modifiable risk factors include increased age, male gender, ethnicity, and family history of CAD. Modifiable risk factors include dyslipidemia, smoking, body-mass index (BMI), hypertension, diabetes mellitus, obesity, poor diet, lack of physical activity, and stressful lifestyle. Dyslipidemia could be caused by dysregulated levels of lipoprotein (a) (LPA), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), triglycerides (TG), and other lipoproteins.

1.2 Genome-wide association study (GWAS)

Genome-wide association study (GWAS) is a methodology for studying associations between genetic variants and phenotypes across the entire genome by testing allele frequency differences of genetic variants among individuals with and without the trait. Importantly, ancestry background needs to be similar between affected and control individuals[8]. Genetic variants in GWAS include mainly single-nucleotide polymorphisms (SNPs) and a small portion of structured variants, e.g., insertion/deletion polymorphism (InDel). The classic workflow of conducting a GWAS test is shown in Figure 1[9]. First, researchers need to recruit appropriate participants, from whom they collect phenotype data and tissue samples for DNA extraction and genotyping. Second, genetic data will be obtained using microarray genotyping technology, or next-generation sequencing (NGS) technology like whole-genome sequencing
Introduction

(WGS) and whole-exome sequencing (WES). In addition, currently, targeted sequencing is becoming a rapid and cost-effective way of detecting known and novel variants in selected genomic regions. Third, researchers have to conduct quality control (QC) on both variant and sample levels after calling genotype from the array or NGS platforms. The QC is an essential step to remove factors that might cause bias on association test, such as removing low-quality variants, removing variants and individuals with low calling rate, detecting the population strata, and removing samples of genetically related. Fourth, the imputation step predicts undetected variants from detected variants based on genetics reference panels. The 1000 Genome (1KG) from the International Genome Sample Resource[10] and the Haplotype Reference Consortium (HRC)[11] are the two most popular genetic reference panels. Fifth, the imputed genotype data will be used for testing associations between genetic variants and phenotypes. According to the International HapMap project and other studies, the human genome has approximately one million independent common genetic variants on average[12], which makes the Bonferroni correcting threshold of $P < 5e^{-8}$ (false discovery rate 0.05/1e6) the most popular significance threshold of GWAS. PLINK[13] and CGTA[14] (Genome-wide Complex Trait Analysis) are the two most popular tools of GWAS analysis.
GWAS tests associations of genetic variants with traits on the basis of the single-SNP model at a genomic scale. A group of significant variants in the same locus are frequently observed in association tests because of the linkage disequilibrium (LD) relationship among common variants, which challenges the identification of the causal variants. Several attempts could help to prioritize the candidate causal variants at a locus. The simplest method is to pick the lead SNP (the one with the lowest association $P$-value) in a genomic region (e.g., a 1-Mb window centered on the locus). The method is based on the hypothesis of a single causal variant at a genomic region, and the top SNP captures the maximum amount of variation by its LD relations with other significant SNPs in the vicinity. Yet, the hypothesis may have several limitations. First, even if the hypothesis is true, the genotyping and imputation may not capture
all the variants at a locus. Second, the hypothesis might not be true given that multiple causal variants could exist at a locus. In this case, a single SNP does not fully represent potential causal variants, especially the ones not LD-linked with the lead[15]. To overcome these limitations, a tool, conditional and joint multiple-SNP analysis of GWAS summary statistics (COJO), was developed to identify additional variants associated with complex diseases in a genomic locus[15]. COJO, embedded in CGTA, requires two inputs, the GWAS summary statistics, and the reference genome of the same ethnic background, to estimate LD-correlations between SNPs, which largely facilitate the studies without individual-level genotype data[14, 15]. COJO starts with conditional analysis of the lead-associated SNPs, followed by a stepwise selection of regional SNPs by conditioning the effects of lead SNPs. COJO performs a stepwise model to select independently associated SNPs (e.g., \( P < 5 \times 10^{-8} \)) based on conditional \( P \) values and estimates the joint effects of all selected SNPs after the model has been optimized. Compared with the method of regional lead SNP picking, COJO may identify multiple independent candidate causal variants in the same regions and can capture larger phenotype variations in the genomic regions.

1.3 Post-GWAS era

1.3.1 Fine-mapping

A decade of GWASs have uncovered thousands of genomic variants associated with complex human traits and diseases. In the post-GWAS era, efforts were directed to delineate the biological mechanisms underlying the associations between risk variants and traits/diseases. The biggest challenge originates from the non-coding feature of the majority of the GWAS loci, which hampers their functional interpretation. To address this, many different approaches have been developed including expression quantitative trait loci (eQTL) analysis, colocalization analysis, and tissue of action (TOA) analysis. Many collective resources, like
NIH Roadmap consortium, ENCODE, FANTOM, and the IHEC consortium provide functional annotation to genetic variants and regions. These indispensable context-specific resources can be integrated into fine-mapping methods to pinpoint regulatory mechanisms of GWAS variants on disease risk.

A direct way of understanding the effects of variants showing association by GWAS is to test the effects of these variants on gene expression of cells or tissues, namely eQTL analysis. An eQTL explains a fraction of gene expression regulated by genetic variants at this locus[16]. The genetics-of-gene-expression panels, such as the Stockholm-Tartu Atherosclerosis Reverse Network Engineering Task (STARNET)[17] and the Genotype-Tissue Expression (GTEx)[18], have largely facilitated eQTL studies. These panels usually collected genotype and tissue or cell-type transcriptomic data from hundred participants. Standard eQTL analysis firstly maps GWAS variants to a gene region and then tests the association between individual respective genotypes and expression levels in a tissue- or cell-type-specific manner. Based on the distance between variants and genes in genome architecture, eQTLs have been classified into cis and trans. Conventionally, variants residing within ±1Mb of a gene’s transcription starting site (TSS) or gene-body regions are classified as cis variants; otherwise, variants are classified as trans variants. Although most studies to date focus on cis-eQTLs, trans-eQTL studies are catching up[17, 19]. Besides eQTL, splicing QTL (sQTL), protein QTL (pQTL), and methylation QTL (mQTL) are increasingly investigated driven by the need of detangling the complexity of variants’ function and availability of corresponding omics datasets[20–22].

A QTL variant is sometimes associated with multiple genes and a gene often has multiple QTL variants, which challenges the identification of causal variants and genes. Based on QTL analysis, scientists have introduced another strategy, colocalization, to prioritize variants and genes by integrating GWAS and QTL signals[23–25]. Colocalization seeks to
identify shared causal variants between a molecular profile (e.g., gene expression or protein level) and a disease trait in a genomic locus, therefore prioritizing candidate causal genes from GWAS loci. COLOC is a popular tool for colocalization between pairs of genetic association studies using the Bayesian model (Figure 2)[26]. COLOC takes summary data of eQTL and GWAS as input to compute the probability of five hypotheses: H0 corresponding to no eQTL and no GWAS association in the region; H1 corresponding to association with eQTL but no GWAS association, or vice-versa (H2); H3 corresponding to eQTL and GWAS association but independent signals; H4 corresponding to shared eQTL and GWAS associations[23, 27]. MOLOC, a multi-trait extension of COLOC, was designed to compare association signals for multiple traits which could be used to pinpoint the regulatory mechanism of GWAS variants[26]. In the case of colocalization among eQTL, mQTL, and GWAS signals, the eQTL signal helps pinpoint the responsible genes in GWAS loci, while the mQTL signal helps identify the epigenetic mechanisms that impact gene expression, and which in turn affect disease risk[28].
Introduction

Figure 2 Example of one configuration under different hypotheses. A configuration is represented by a binary vector for each trait of (0,1) at a length of n=9, the number of shared variants in a region. The two traits shown in the figure are eQTL and GWAS signal. The value of 1 means the variant of the position is causally related to the trait, 0 vice versa. The top plot, corresponding to hypotheses 1 and 2, means that only one trait has a causal variant in the locus. The middle plot, corresponding to hypothesis 3, means that two traits have different causal variants in the locus. The bottom plot, corresponding to hypothesis 4, means that both traits share the same causal variant in the locus. The figure was adapted from Giambartolomei[26].

The tissue of action (TOA) score exerts a new strategy to partition the effects of variants on disease in the context of tissues or cell types[29]. For each GWAS signal, the TOA method systematically partitions posterior probabilities from the Bayesian fine-mapping method by
integrating tissue-specific functional annotation and expression data. The TOA score reflects the probability of genetic variant functioning in a specific tissue or cell type. As complex diseases are usually multiple tissues or cell types involved, the TOA score provides valuable guidance for in vivo/vitro experiment validation in the post-GWAS era.

1.3.2 Phenome-wide association study (PheWAS)

Phenome-wide association study (PheWAS), an alternative/complementary approach to the GWAS methodology, is to estimate associations of single genetic variants with a wide range of phenotypes. A fundamental difference between GWAS and PheWAS designs is the direction of inference: PheWAS studies pleiotropy of SNPs, a genotype-to-phenotype strategy; GWAS studies causality of SNPs, a phenotype-to-genotype strategy[30].

The PheWAS approach was originally developed due to the widespread availability of both anonymized human clinical electronic health record (EHR) data and matched genotype data. One of the main components of EHR is the International Classifications of Diseases version 9 (ICD9) codes, which includes information of 17,000 phenotypes binned into different hierarchy codes. In most cases, PheWASs divide cohorts into “cases” if participants have ICD9 codes related to a specific trait/disease, or “controls” if participants do not. Usually, GWAS SNPs are logical starting points for PheWAS because of the availability of association data. Staring with a known disease-associated SNP, a PheWAS study performs association tests for the specific SNP with a series of related traits. The significance level of PheWAS analysis is estimated via a Bonferroni correction as \( \alpha=0.05/N \) (N the number of feasible models tested).

One of the major advantages of PheWAS is its potential to identify genomic variants with pleiotropic properties. Investigating cross-phenotype associations of SNPs makes PheWAS an important tool for understanding genetic associations among diseases/traits and gene-disease associations, as well as elucidating mechanisms of GWAS risk loci[31–34].
instance, a PheWAS work revealed that \textit{CCDC92}, a novel GWAS risk locus of CAD, likely affected CAD through insulin resistance pathways[31]. PheWAS has been proposed as a possible aid in drug development through elucidating mechanisms of action, identifying alternative indications, or predicting adverse drug events[35].

PheWAS design has challenges. PheWAS is a hypothesis-generating approach, which is challenged by multiple comparison testing. The $N$ used for calculating the significance threshold can be varied greatly in different PheWAS designs. Suppose a SNP is tested for associations with all traits listed in the ICD9, and the Bonferroni significance is $2.94 \times 10^{-6} = 0.05/17,000$. If we had multiple SNPs tested at the same time, the threshold would be even higher. If we only tested associations of SNPs or SNP sets with several phenotypes, the threshold would be lower. This phenomenon could lead to false positive or false negative associations. Since PheWAS heavily relies on the annotation of traits/diseases, the depth and breadth of annotation data influence the PheWAS results. Like GWAS, association regression in PheWAS faces the problem of covering all possible covariates, which makes further validation of PheWAS findings necessary.

1.3.3 Mendelian randomization (MR)

As mentioned in section 1.1.3, many risk factors are associated with CAD, such as LDL-C, HDL-C, TG, inflammatory cytokines, larger BMI, Diabetes, and blood pressure. Finding the causal nature of risk factors on the disease process is critically important in epidemiology because these modifiable risk factors represent promising targets for primary intervention and drug development[36]. Mendelian randomization (MR) uses genetic variants as instrumental variables (IVs) to infer whether risk factors (exposures) causally affect a health outcome which is CAD in our case (Figure 3)[37]. In addition, the associations between exposures and
outcomes reflected by MR analyses are genetic associations that are independent of confounding factors.

In MR studies variants must fulfill three assumptions if they shall be used as instrumental variables: i) the genetic variants are associated with the exposure (i.e. the risk factor-relevance assumption); ii) the genetic variants are associated with the outcome through the studied exposure only (exclusion restriction assumption); iii) the variants are independent of other confounders which could affect the outcome (independence assumption)[37, 38]. Mendel’s Law of Independent Assortment that the alleles segregate randomly when passed from parents to offspring, forms the foundation of MR. In addition, the germline genotypes are determined at conception, they precede the investigated outcomes, and therefore observed associations cannot be explained by reverse causation[39]. However, the wide existing pleiotropy of genetic variants and the LD association among genetic variants raise scientists’ caution when performing MR analysis.

There are two types of MR designs: one-sample MR and two-sample MR (Figure 3)[39]. For one-sample MR analysis, the SNP-exposure associations and SNP-outcome associations are estimated from the same sample, set. In two-sample MR, the SNP-exposure associations are estimated in one sample set and the SNP-outcome association is estimated in a second sample set. Compared with the one-sample MR study, the two-sample MR study doesn’t require exposure and outcome to be measured in all data sets, which allows researchers to use GWAS summary statistics from large consortia. This also improves transparency and reproducibility.

The classic steps of performing two-sample MR analysis are: i) identify genetic variants from a well-conducted GWAS; ii) obtain SNP-exposure associations from data set 1; iii) obtain SNP-outcome associations from data set 2; iv) harmonize SNP effects on exposure and
outcome; v) generate MR estimates; vi) perform sensitivity analyses. The inverse variance weighted (IVW) method is the most efficient estimate of the causal effect when all genetic variants are valid instruments[40]. IVV causal estimates could be biased when genetic variants exhibit horizontal pleiotropy because it violates the assumption of exclusion restriction. As mentioned in section 1.3.2, pleiotropy of genetic variants is inevitable in most MR studies with the increased knowledge of the genetic basis of complex diseases. For sensitivity analysis, scientists developed other robust methods, like MR Egger[41] and weighted median[42], that can provide valid causal inferences under weaker assumptions than the standard IVW mythology. Egger regression can identify the presence of ‘directional’ pleiotropy and provide a less biased causal estimate. The weighted median estimator provides a consistent estimate of the causal effect even when up to 50% of the information contributing to the analysis comes from genetic variants that are invalid IVs[42].

Figure 3 One-sample and two-sample Mendelian randomization designs. (A) One-sample Mendelian randomization (MR) measures both exposure and outcome in the same population.
In two-sample MR, associations of genetic variants with exposures and outcomes are measured in non-overlapping populations. The Figure was adapted from Zheng[39].

The MR studies have identified many risk biomarkers, traits, and diseases relevant to the pathogenesis of CAD (Figure 4)[43]. Risk factors displaying confirmation by MR may be considered causal and include LDL-C, TG, LPA, IL6, non-fasting glucose, diabetes, obesity, adiponectin, blood pressure, and telomere length. Vice versa, MR studies challenged the roles of many biomarkers including HDL-C, CRP, lipoprotein-associated phospholipase A2 (LP-PLA2), homocysteine, fibrinogen, bilirubin, and uric acid. Drugs targeting causal risk factors defined by MR are attractive treatment targets for CAD. For example, statins are well-known to reduce cardiovascular events and mortality in CAD patients because of their significant lipid-lowering functions[44]. We have shown in our MR studies that genetically modulated educational attainment may have implications for a series of unfavorable risk factor profiles, such as BMI and hypertension, and, subsequently, affects the prevalence of CAD[45]. These findings strengthen the importance of campaigns for enabling adequate schooling for preventing CAD.
Figure 4. Risk factors of coronary artery diseases identified by MR analyses. The figure was adapted from Jansen[43].

1.4 Genetics-of-gene-expression panels

In the post-GWAS era, studying GWAS variants’ effects on gene expression is a critical step to elucidating the genetic basis of common diseases or traits, which makes large-scale examination of genotype and transcript data in the context of disease-relevant tissues or cell types indispensable. The fast development of high-throughput technologies, such as the genotyping array, whole-genome sequencing (WGS), whole-exome sequencing (WES), RNA sequencing (RNA-seq), and single-cell RNA sequencing (scRNA-seq) accelerates the efficiency and affordability of generating such massive genotype and gene expression data.

The Genotype-Tissue Expression project (GTEx) is one of the most popular comprehensive public resources to study tissue-specific gene expression and regulation[18]. Samples of the GTEx project were collected from 54 non-diseased tissue sites of up to 1000 individuals for molecular assays including genotyping, WGS, WES, and RNA-seq.
GTEx donors, 84.6% are white, 67.1% are male, 32.1% are between 50-59 years old, and 36% are between 60-70 years old. Furthermore, the GTEx consortium has also profiled scRNA data for eight tissues including breast, esophagus mucosa, esophagus muscularis, heart, lung, skeletal muscle, prostate, and skin. Scientists can download the processed data, such as expression and eQTL data, from the project portal, or apply for the secured data, such as genotyping data and phenotype data, from the dbGAP platform. The GTEx portal also provides a user-friendly interface and search engine at https://gtexportal.org/home/.

Different from GTEx which collected samples from healthy tissues, the Stockholm-Tartu Atherosclerosis Reverse Network Engineering Task (STARNET) represents the unique genetics-of-gene-expression panel for CAD study[17]. Tissue samples of STARNET were obtained from ~600 CAD patients during open thorax surgery, including atherosclerotic aortic root (AOR), atherosclerotic-lesion-free internal mammary artery (MAM), blood (BLD), subcutaneous fat (SF), visceral abdominal fat (VAF), skeletal muscle (SKLM), and liver (LIV). The STARNET patients are mainly Caucasians (31% females), 32% had diabetes, 75% had hypertension, 67% had hyperlipidemia, and 33% had an MI before age 60. By New York Heart Association criteria, 45% were class I, 42% class II, 9% class III, and 1% class IV CAD. DNA was genotyped with the OmniExpress Exome array (Illumina, ~900k SNPs), and RNA was sequenced using the HighSeq2000 platform. Researchers could apply the data from the dbGAP database. A study comparing tissue-specific eQTLs between STARNET and GTEs showed that STARNET had more eQTLs coinciding with CAD-associated risk SNPs suggesting that risk SNPs were more active in the focus of infection[17]. STARNET also provides a browser for exploring co-expression models inside and between tissues, which is at http://starnet.mssm.edu/[46].
1.5 UK Biobank (UKB)

UK Biobank (UKB) is the largest publicly available resource with deep phenotyping and genomic data of around 500,000 UK residents aged 40-69 years old initially recruited from 2006 to 2010[47]. The study continuously collects extensive phenotypic and genotypic data from all participants, including data from questionnaires, physical measures, sample assays, accelerometry, multimodal imaging, genome-wide genotyping, and longitudinal follow-up for a wide range of health-related outcomes.

The genome-wide genotyping data of ~500,000 participants released in 2016 was performed using the UK Biobank Axiom Array. Approximately 850,000 variants were directly measured, with >90 million variants imputed using the HRC and UK10K + 1KG reference panels. UKB provides different levels of genotyping data, such as before or after imputation, to meet the needs of most applicants. Starting from 2016, UKB gradually released WES data of participants. By the fourth season of 2021, the WES data of ~500,000 participants were available for all participants[48]. The WES data measures the regions of the genome (about 2%) that encode proteins and is particularly suitable for identifying disease-causing rare variants (see section 1.6). By 2021, the database also released telomere data for the ~500,000 participants and metabolic data for ~127,000 participants. UKB’s imaging study includes measures such as white matter hyper intensities derived from the brain scans, visceral fat derived from the abdominal scans, and left ventricular ejection fraction derived from the cardiac scans. The imaging data of ~50,000 participants were released in 2020, extra 50,000 participants’ data could be expected in the near future. All data release information can be found at https://www.ukbiobank.ac.uk/enable-your-research/about-our-data. UKB also developed a Research Analysis Platform (RAP) powered by the DNAnexus (https://www.dnanexus.com/), providing a secure, scalable, and cloud-based environment for researchers to use UKB biomedical resources.
The greatest sample size and widest scope of phenotype data make UKB the most valuable resource for researchers of multiple fields. According to their statistics, the number of annual publications using data of UKB increased from only one in 2012 to 690 in 2021. The UKB resource was applied to multiple investigations of CAD revealing novel risk loci for CAD[49], identifying risk loci associated with aortic valve area from genotyping and imaging data[50], and studying the interaction between polygenic risk scores and monogenic mutations[51], etc.

1.6 Rare mutation association study in complex disease

There are two controversial hypotheses about the contribution of genetic variants to individual susceptibility to common complex diseases: the common diseases common variant (CDCV) hypothesis, and the common disease rare variant (CDRV) hypothesis. The CDCV assumes that common variants (e.g., MAF>0.01) are major contributors to complex diseases. The CDRV supposes that rare genetic variants (e.g. MAF<0.01) play a more important role compared to common variants[52]. Many studies have shown that rare variants usually have a larger effect size on disease development than common variants which empowers the unique role of rare variants in investing genetic pathogenesis of complex diseases[53]. Same to GWAS, the association test for rare variants is a kind of hypothesis-free approach in which researchers do not start with a certain functional hypothesis. The enrichment of rare variants provides hypothesis-free evidence for gene causality. The rare missense, loss-of-function (LoF), and gain-of-function (GoF) variants mimic the impact of gene knocking out or enhancing experiments, which are ideal for functional analysis to elucidate new disease mechanisms. Rare variants showing association for a disease could provide favorable targets for drug development as their alleles mimic the effect of a modulated drug target. In addition, rare variants with larger effect size on disease development could be biomarkers for personalized medicine.
The workflow of the rare variant association study is shown in Figure 5[54]. The first step is to set up a work target that determines the variant calling platforms, genotyping, or sequencing. Compared to the genotyping array, the NGS technology has the advantages of massive scaling and high sensitivity in identifying rare variants. The next step is to call genotype data and perform quality filtering on the raw data. For data called from genotyping array, the quality control (QC) step is similar to GWAS, except that rare variants would be kept. When data is called from NGS platforms, extra QC parameters are required, such as sequencing depth, coverage, heterozygosity/contamination rate, etc. The third step is to assay function of rare variants using different in-silico methods, such as LRT score[55], MutationTaster[56], PolyPhen-2 HumDiv, PolyPhen-2 HumVar[57], and SIFT[58]. There is a variety of annotation sources to predict the consequence of genetic variants. sdNSFP (v4.1) is the largest database designed to facilitate the functional annotation step by providing deleteriousness prediction and functional annotation for all potential non-synonymous and splice-site single nucleotide variants (a total of 84,013,093) in the human genome[59]. ClinVar is a freely accessible, public archive of reports of the associations of genetic variations with human phenotypes, with supporting evidence[60]. Tools like Variant Effect Predictor (VEP)[61], snpEff[62], and ANNOVAR[63] assemble these tools and resources into a set providing a single-line command to annotate the effect of genetic variants.

The next step is to perform an association test between functional variants and diseases/traits. The models designed for association analysis can be generally classified into two types: single-variant level and gene- or region-based aggression tests. In single-variant tests, the association is typically evaluated by linear regression for continuous traits and by logistic regression for binary traits. Single variant tests for rare variants are only suitable when the sample size is large enough; otherwise, they are less powerful than the tests for common variants. Different from single-variant tests, gene- or region-based aggression tests increase the
statistical power by evaluating the cumulative effects of multiple genetic variants in a gene or region. Based on varying assumptions about the underlying genetic models, methods for aggression test can be broadly categorized into five classes: burden test, adaptation burden test, variance-component test, combined burden and variance-component (VC) test, and exponential-combination (EC) test[54]. Fisher’s exact test for categorical phenotypes is the simplest combined test. The Sequence Kernel Association Test (SKAT), a widely used score-based VC test, is robust to groupings including variants with both positive and negative effects[64].

Figure 5 Workflow of rare variant association study. The figure was adapted from Lee[54].

The association test of rare variants informs the posterior probability of the disease relevance of a gene, which could guide the decision-making for biological validation. By performing a meta-analysis of exome-chip studies of European descent involving 42,335 patients and 78,240 controls, the CARDIoGRAMplusC4D Consortium identified associations of variants in ANGPTL4, LPL, and SVEP1 with CAD[65]. Following the genetic study, an in-
vivo investigation was carried out to elucidate the mechanism of the novel gene SVEP1 in CAD development[66].

1.7 Transcriptome-wide association study

Transcriptome-wide association studies (TWAS) have been recently proposed as an invaluable tool for annotating GWAS risk loci by investigating the potential gene regulatory mechanisms underlying variant-trait associations. The expression level of a gene can be decomposed into three components: a genetically determined component, a component altered by the trait itself, and a component determined by the remaining factors (including environment)[67]. Based on the assumption, TWAS reflects associations between genetically regulated gene expression (GReX) and diseases or traits.

The TWAS methodologies are broadly separated into two categories: the individual-level predictor and the GWAS summary-based predictor. The first step of both methodologies is to train prediction models from reference genetics-of-gene-expression panels which scale genotype and expression data simultaneously (Figure 6). Then the individual-level predictors would apply prediction models to impute expression profiles from individual genotypes of GWAS cohorts and perform association test between predicted expression and traits (section 1.7.1). However, the GWAS summary-based predictors take GWAS summary data as input to examine associations between intermediate gene expression levels and phenotypes (section 1.7.2).
Figure 6 An overview of the TWAS. Briefly, TWAS involves i) training tissue-specific prediction models from references; ii) predicting genetically regulated expression (GReX) from genotype data using prediction models; iii) associating GReX with phenotypes. The figure was adapted from Wainberg[68].

1.7.1 Individual-level predictors

PrediXcan was the first individual-level predictor which aggregates impacts of variant set on gene expression[67]. Studies have suggested that aggregating variants by integrating gene expression or other omics may better explain underlying biological mechanisms and increase the power of association studies beyond GWAS[7]. Based on PrediXcan, Gao developed EpiXcan[69] which outperformed PrediXcan in prediction performance by integrating epigenome annotation from Roadmap Epigenomics Mapping Consortium (REMC)[70]. It firstly calculates SNP priors by using a hierarchical Bayesian model (qtlBHM) that jointly analyzes REMC epigenome annotations and eQTL statistics. Then, priors are transformed with an adaptive mapping function to penalty factors, which are then utilized by the weighted elastic net (WENet), a model selection technique that combines LASSO and Ridge regression and
seeks to identify which predictor variables to include in a regression model. The WENet model analyzes SNP priors, genotypes, and gene expression traits to estimate genetically regulated expression across different. The simulated expression value for each gene is

\[ y = X \times \beta + \varepsilon \]

Here, \( X \) denotes the genotype matrix of cis-SNPs included in the prediction model, \( \beta \) denotes the coefficient vector of the cis-SNPs, ‘\( \times \)' denotes matrix-vector product, and \( \varepsilon \) denotes residual value. The criterion of the WENet model can be written as:

\[ C_{WENet}(\theta, \lambda, \alpha) = \sum_{i=1}^{n} [y_i - X_i\beta]^2 + \lambda \alpha |\beta|_w + \lambda (1 - \alpha)\beta^T W \beta, \]

\[ |\beta|_w = \sum_{j=1}^{m} w_j |\beta_j| \]

In the above equations, \( n \) is the number of samples used for training model; \( X_i, 1 \leq i \leq n \), is the \( i \)-th row-vector of matrix \( X \) containing genotypes with dosages from 0 to 2; \( m \) is the number of cis-SNPs included in the model; \( w \) is the weight matrix that stores the penalty factors for SNPs. The \( \alpha \) parameter is set to 0.5 and \( \lambda \) is estimated via cross-validation (CV).

We could apply the EpiXcan pipeline to train prediction models in a tissue-specific fashion from genetics-of-gene-expression panels, such as GTEx and STARNET (see section 1.4), and use these models to predict gene expression from individual-level genotype data. Finally, we can perform association analyses on predicted expression using regular models, such as logistic regression for categorical traits, and linear regression for quantitative traits.

### 1.7.2 Summary-based predictors

In many cases, the individual-level genotype data are not available, but the GWAS summary data are accessible via general collective resources, such as the GWAS catalog, or trait-specific
consortia, such as CARDIoGRAMplusC4D for CAD and GLGG for lipid traits[71]. In this case, the GWAS summary-based predictors become valuable. The tool Summary PrediXcan (S-PrediXcan) was derived from PrediXcan by the same research group[27]. S-PrediXcan uses prediction models trained by PrediXcan, takes summary data as input, and estimates the Z-score (Wald statistic) of the association between predicted gene expression and a phenotype. The Z-score for gene g is estimated as:

$$Z_g \approx \sum_{i}^{m} w_{ij} \frac{\hat{\sigma}_i}{\hat{\sigma}_g} \frac{\hat{\beta}_i}{se(\hat{\beta}_i)}$$

In this formula, $m$ is the number of cis-SNPs included in a gene’s prediction model, $w_{ij}$ is the weight of $i$-th SNP in the expression prediction model, $\hat{\sigma}_i$ is the estimated variance of $i$-th SNP, $\hat{\sigma}_g$ is the estimated variance of gene g, $\hat{\beta}_i$ is the estimated effect size of $i$-th SNP from GWAS summary data, and $se(\hat{\beta}_i)$ is the standard error of $i$-th variant in GWAS summary data. The weights of cis-SNPs on gene expression derived from EpiXcan can also be applied.

FUSION is another summary-based TWAS tool[72]. Different from the elastic-net(ENet) model used by PrediXcan or EpiXcan, FUSION uses the Bayesian sparse linear mixed model (BSLMM)[73] in estimating weights for cis-SNPs. The FUSION package also provides other weights-estimating models including BLUP, ENet, and top SNPs. A comparison study has shown that FUSION and S-PrediXcan are consistent and complementary to each other[74]. Different from most TWAS tools focusing on cis-SNPs, Bayesian genome-wide TWAS (BGW-TWAS) leverages both cis- and trans-eQTL information for a TWAS[75]. With the accumulation of knowledge about functional genomics, the strategy of integrating multi-omics is en vogue. MOSTWAS is the recently published multi-omics strategy for TWAS analyses[76].
1.7.3 Opportunities and challenges of TWAS

TWAS, on the one hand, is a promising approach to prioritizing causal genes at GWAS loci in a tissue-specific fashion, on the other hand, the method can identify novel risk genes residing outside of GWAS risk loci. For instance, Wu prioritized 48 risk genes for breast cancer by performing TWAS analyses on 229,000 participants, of which 14 genes were independent of any reported GWAS loci[77]. Gene silence experiments on novel genes identified 11 novels that had effects on cell proliferation and/or colony-forming efficiency. In another work, the gene expression imputations across multiple brain regions in over one million participants identified 67 non-MHC (major histocompatibility complex) risk genes for schizophrenia, of which 14 did not fall within previous GWAS loci[78].

Despite the great contribution of providing insights into complex diseases, TWAS has certain limitations. TWAS combines eQTL reference panels with large-scale genotype data to test associations between genes and diseases. The pleiotropy features of eQTLs, and LD associations among eQTLs often result in multiple gene hits at the same locus[68]. In some cases, multiple genes significant at the same locus are due to co-expression with causal genes or co-regulated by the same set of eQTLs or those in high LD. Therefore, further validation is needed to check the causality. Another big challenge of TWAS is how to select appropriate tissues for testing association with diseases. Most candidate causal genes drop out after switching to tissue with has no clear mechanistic relationship to the trait due to the lack of sufficient expression or sufficiently heritable expression[68]. For TWAS analysis, the eQTL reference panels are critical in building prediction models. The differences in sample collections, expression scaling techniques, disease status of panel samples, etc., could create consistent outcomes of TWAS but might also be complementary to each other. Therefore,
combing TWAS results from different reference panels could provide richer insights into diseases[17].

1.8 Polygenic risk score (PRS)

1.8.1 Definition and analysis pipeline of PRS

A polygenic risk score (PRS) is a genetic estimate of an individual’s liability to a trait or disease, calculated from a set of independent variants usually based on large-scale GWAS data[79]. For each individual, PRS is a sum of the number of risk alleles at each variant (0, 1, 2), which could be unweighted, or weighted by its effect size estimated derived from GWAS[80].

The PRS analysis process is shown in Figure 7[79]. PRS can be characterized by the use of base and target data. The risk loci and betas or odds ratio (OR) weights are retrieved from base data and then applied to calculate PRS for target data. The QC for both base and target data sets are similar to standards of classic GWAS. In addition, the “QC checklist” specific to PRS analysis has to be outlined: i) using base GWAS data with heritability $h^2 > 0.05$; ii) specifying effect and non-effect alleles from base GWAS data; iii) using target data with sample size of $\geq 100$ individuals (or effective sample sizes $> 100$ for case/control data) if PRS would be applied for association test[79]. There are two main options for approximating PRS. The classic method is called the clumping + thresholding (‘C+T’) method, which clumps SNPs passing the GWAS significance threshold (see section 1.2) so that SNPs retained are largely independent of each other. Another method is beta shrinkage which all SNPs are included, accounting for the LD between them. Many tools have been developed for calculating PRS, like PLINK[81], PRSice-2[82], LDpred[83], lassosum[84], etc.
1.8.2 Application of PRS

In basic science, PRS has been used to evaluate associations between phenotypes and to elucidate risk factors that may play a mediating role along the causal pathway to disease[85].
The MR method has the advantage of featuring causal relationships among traits (see section 1.3.3). Therefore, the combination of PRS and MR analysis can improve detection rates for causal relationships which can be particularly useful when evaluating associations between genetic liability for a given trait and hundreds of diverse health outcomes[85]. We named this ‘PRS+MR’ strategy. Zeng et al. detected a negative association between PRS of education attainment and CAD risk, then they used the MR method to deduct the causal pathway of how genetic education attainment reduces CAD risk through its impacts on a series of risk factors[45]. This strategy was also applied to one of my PhD works which studied the genetic association between intelligence score and CAD risk (see section 3.1)[86].

In the clinical setting, PRS can be implemented in three key stages (Figure 8)[87]. First, PRS contributes to risk stratification in an apparently healthy population by screening for high-risk groups followed by intensified strategies for disease prevention and early intervention. Second, PRS could be used in clinical diagnosis when people are in the early phase of diseases without significant clinical diagnostic signs. Third, it is possible that in the future, PRS could contribute to personalized medicine and outcome prediction.

There is a long way to go before PRS could be massively applied in clinical settings. PRS itself is not strong enough in the prediction of disease incidence. In CAD, PRS didn’t outperform in predicting subsequent CAD events as compared to other clinical risk predictors. But when both were combined, it was more accurate than either PRS or clinical risk predictors alone[87]. Studies also showed that the combination of PRS in subjects affected by monogenic variants with low to moderate penetrance could increase the accuracy of risk prediction. The average probability of CAD by the age of 75 years old subjects was found to increase from 13% in noncarriers of familial hypercholesterolemia (FH) variants to 41% in carriers of FH variants[88]. Likewise, in carriers of FH variants, a substantial gradient of risk was observed.
depending on the PRS. Similar trends were also observed in breast cancer[88] and prostate cancer[89].

Figure 8 An overview of the population cohorts where polygenic risk scores could be applied. The figure was adapted from Wray[87].

1.8.3 Further discussion about PRS

**Ethnicity**

By far, most GWAS studies were conducted on individuals of European decedent, and genomic research of the non-European population is significantly underrepresented. More and more scientists put their attention to trans-ethnic GWASs because of genetic diversity among ethnicities[49, 90]. Along with increased identification of ethnicity-specific and trans-ethnic loci, a question about PRS study was raised: which is better, ethnicity-specific or trans-ethnic PRS? The latest work about CAD GWAS identified 8 novel Japanese-specific risk loci as well as 35 novel trans-ethnic loci[90]. This work also suggested that PRS derived from trans-ethnic
loci outperformed PRS derived either from Japanese or European risk loci. However, further work is needed to confirm this phenomenon.

**Multi-PRS strategy**

In classic studies, the PRS is estimated based on risk loci associated with the target trait. Inspired by the pleiotropy feature of genetic variants and genetic correlations between complex traits, the multiple polygenic risk score (multiPRS) approach improved prediction performance by using the joint predictive power of multiple polygenic scores in one regression model[91, 92]. For instance, the formula of multiPRS for CAD could be represented as

$$multiPRS_{CAD} = \beta_1PRS_{CAD} + \beta_2PRS_{Diabetes} + \beta_3PRS_{Hypertension} + \beta_3PRS_{LDL-C} + \cdots covrs$$

, in which $multiPRS_{CAD}$ is a weighted combination of CAD and other risk traits’ PRS. The multiPRS approach may be useful in investigating developmental, multivariate and gene-environment interplay issues, stratifying individuals according to the risk of conditions, and eventually, improving performance in personalized medicine.

**Polygenic resilience score**

Polygenic resilience is a reverse concept of PRS. It studies genetic variants that promote resistance to disease by reducing the penetrance of risk loci, wherein resilience and risk loci operate orthogonally to each other[93]. Currently, there is only one study about polygenic resilience score conducted by Hess et al[93]. Based on schizophrenia PRS, they stratified samples into different risk groups. The controls from high PRS groups were defined as high resistance controls. From the high-risk group, they retrieved genetic resilience variants by performing GWAS analysis. The works suggested that positive correlation between polygenic resilience score and PRS in the whole cohort, but no correlation in case samples. The
correlation coefficient reached the highest in controls of the high PRS group which verified the hypothesis of the orthogonal relation between genetic risk and resilience. Studying polygenic resilience factors could give us extra insights into why participants in the high PRS group are resilient to developing disease and might promote a more sophisticated PRS model with better performance than the current one.

1.9 Genetics of CAD

1.9.1 GWAS of CAD

Decades of GWASs have generated wealthy knowledge of genetic factors contributing to CAD etiology. Due to the endeavor of large CAD consortia, as well as national and international collaboration, 321 genome-wide significant loci have been associated with CAD (Figure 8)[94]. The sample size and diversity of participants are still increasing in multiple biobanks, such as CARDIoGRAMplusC4D, UKB, Japanese biobank, Million Hearts GWAS, Million Veteran Program, and All of Us Research Program. A combination of these biobanks would increase the possibility of identifying novel risk loci and studying ethnicity-specific risk lock. Current knowledge of CAD GWAS brings us to the post-GWAS era to i) elucidate the disease-associated mechanisms underlying CAD loci, ii) prioritize potential causal genes and novel drug targets for the disease, and iii) harness CAD genetic variations as the age-independent biomarkers for risk stratification, disease prevention, and personalized medicine.
1 Introduction

Figure 9 Genes mapped to 321 CAD risk loci and related pathophysiological pathways of atherosclerosis. The figure was adapted from Chen[94].

1.9.2 Prioritization of CAD causal genes by in-silico methods

The in-silico methods of prioritizing risk genes have been discussed in section 1.3. Strategies of pinpointing causal genes for CAD include but are not limited to i) vicinity of genes to risk loci; ii) eQTL mapping in CAD-relevant tissues according to reference panels like GTEx and STARNET; iii) monogenic mutations on candidate genes associated with CAD or its risk factors; iv) fine-mapping with functional annotation; v) colocalization between GWAS signal and genes’ eQTL or other quantitative signals; vi) phenome-wide screening for pleiotropy of risk loci; vii) cross-mapping with consortia of mouse models like Hybrid Mouse Diversity Panel (HMDP)[95], the International Mouse Phenotyping Consortium (IMPC)[96] or Mouse
Genome Informatics database (MGI)[97]; viii) integrative genomic analysis (IGA) to combine results from multiple in-silico methods, which has emerged as a powerful strategy for identifying causal genes[98]. Moreover, the fast accumulation of single-cell data and metabolomics data would greatly benefit causal gene selections and pathogenic mechanism study[99, 100]. However, TWAS of CAD, which could systematically identify tissue-specific risk genes for CAD has not been performed yet. This forms the rationale of my PhD research on CAD TWAS.

II Methods

2.1 Preprocessing of 11 GWAS cohorts

Genotype data of ten CARDIoGRAMplusC4D[86, 101–109] cohorts and UKB[47] were used in both projects building the PhD thesis (Table 1).
Table 1 Overview of eleven individual-level genotype cohorts.

<table>
<thead>
<tr>
<th>Study</th>
<th>Array</th>
<th>CAD case</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>N_case</td>
<td>N_control</td>
</tr>
<tr>
<td></td>
<td></td>
<td>female( %N_cases)</td>
<td>female( %N Controls)</td>
</tr>
<tr>
<td>GerMIFS</td>
<td>Affymetrix Mapping 500K Array Set</td>
<td>207(33.3)</td>
<td>795(51.3)</td>
</tr>
<tr>
<td>GerMIFSII</td>
<td>Affymetrix Genome-Wide Human SNP Array 6.0</td>
<td>244(20.5)</td>
<td>604(48.1)</td>
</tr>
<tr>
<td>GerMIFSIII</td>
<td>Affymetrix Genome-Wide Human SNP Array 5.0/6.0</td>
<td>212(20.1)</td>
<td>696(48.3)</td>
</tr>
<tr>
<td>GerMIFSIV</td>
<td>Affymetrix Genome-Wide Human SNP Array 6.0</td>
<td>336(35.2)</td>
<td>697(61.4)</td>
</tr>
<tr>
<td>GerMIFSV</td>
<td>Illumina HumanOmniExpress/Omniuni_2.5</td>
<td>593(24.3)</td>
<td>827(52.5)</td>
</tr>
<tr>
<td>GerMIFSVI</td>
<td>Illumina PsychChip_v1-1</td>
<td>492(30)</td>
<td>609(51.3)</td>
</tr>
<tr>
<td>GerMIFSVII</td>
<td>Infinium Global Screening Array-24</td>
<td>1031(33.7)</td>
<td>1886(54.5)</td>
</tr>
<tr>
<td>WTCCC</td>
<td>Affymetrix Genome-Wide Human SNP Array 6.0</td>
<td>395(20.8)</td>
<td>1481(50.9)</td>
</tr>
<tr>
<td>Cardiogenics</td>
<td>Illumina Human660W-Quad</td>
<td>49(12.8)</td>
<td>382</td>
</tr>
<tr>
<td>-------------</td>
<td>-------------------------</td>
<td>----------</td>
<td>-----</td>
</tr>
<tr>
<td>MIGen</td>
<td>Affymetrix Mapping 500K Array Set</td>
<td>646(22.3)</td>
<td>2901</td>
</tr>
<tr>
<td>UKB</td>
<td>UK BiLEVE Axiom array n~=50,000</td>
<td>4465 (22.0)</td>
<td>20310</td>
</tr>
<tr>
<td></td>
<td>UK Biobank Axiom array n~=450,000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2.1.1 Ten CARDIoGRAMplusC4D cohorts

Individua-level genotyping data of 17,687 cases and 17,854 controls were collected from ten CARDIoGRAMplusC4D cohorts including the German Myocardial Infarction Family Studies (GerMIFS) I-VII[86, 101–106], Wellcome Trust Case Control Consortium (WTCCC)[107], LURIC[108] study and Myocardial Infarction Genetics Consortium (MIGen)[109]. All samples were of European descent, mostly from Germany or UK. The subjects building the MIGen cohort were from several European countries and the United States.

All studies went through the following quality filtering steps (as mentioned in section 2) using plink (v1.9)[13] before imputation: individual-level calling rate≥0.98, SNP-level calling rate≥0.95, MAF>0.01, sex consistency between the reported and the genotype-derived, deviation from HWE $P>$1e-5, Identity By Descent (IBD)<0.125 (individuals were distant to each other more than the third generation), heterozygosity rate within mean ± 3*SD. After quality filtering of genotype level, all studies except GerMIFS VII were imputed according to the reference of 1000 Genomes Phase I integrated variant (v3)[10] using SHAPEIT2[110] and IMPUTE2[111]. For GerMIFS VII, the reference data was HRC[11]. All SNPs were mapped to NCBI GRCh37/hg19. After genotype imputation, we conducted quality filtering again using the following parameters: SNP-level calling rate>0.98, MAF>0.01, and deviation from HWE $P>$1e-5.

2.1.2 Genotype data of UKB

The UKB is globally the largest biobank with deep phenotypic and genomic data (see section 1.5)[47]. We obtained imputed genotype data from 3rd release of UKB. The data were imputed using a combination of two reference panels. The first panel is the HRC reference panel[11]. The second is a merged reference panel of the UK10K haplotype panel and the 1000 Genomes references. If variants were imputed from both reference panels, the HRC imputation result
was retained. A series of QC were done on the imputed data: MAF>0.00001, imputation info>0.4, SNP-level calling rate>0.95, individual-level calling rate≥0.98, sex consistency, kinship coefficient<0.088, and deviation from HWE $P>1e^{-5}$.

All phenotypes were defined by either self-reported, hospital episode, and/or death registry data. The definition of CAD encompassed individuals with fatal or normal myocardial infarction (MI), percutaneous transluminal coronary angioplasty (PTCA), or coronary artery bypass grafting (GABAD). MI was defined as hospital admission or cause of death due to ICD10 I21-I24, I25.2, ICD9 410-412, and self-reported 1075. PTCA, CABG, and triple heart bypass were defined as hospital admission or cause of death due to OPCS-4 K40-K46, K49, K50.1, K75, and self-reported 1070, 1095, 1523. Finally, we got 20,310 hard CAD cases and randomly selected 25,000 non-CAD samples as controls.

2.2 Methods used in the Intelligence-CAD project

In this work, we firstly constructed a genetic intelligence score (gIQ) based on 242 SNPs independently associated with intelligence[112] for participants from ten CARDIoGRAMplusC4D[86, 101–109] cohorts and UKB[47]. The ‘C+T’ method (see section 1.8.1) of PRS estimating was applied to estimate gIQ. In short, gIQ is a sum of the weighted dosage of effect alleles of 242 independent intelligence variants. As to missing variants in the genotype data, we replaced them with reference allele frequencies.

We tested the association between gIQ and CAD risk in eleven cohorts using logistic regression models. For all cohorts except UKB, two principal components were added to the regression model to adjust the bias of population stratification. As the genotype data of UKB was scaled from two array platforms and has a relatively more complicated ethnic background, the top five principal components and array platform were added to the regression model.
In UKB, we also studied associations between gIQ and serial risk factors of CAD including smoking, hypertension, BMI, T2D, HDL-C, and LDL-C. Logistic regression models were applied for binary traits, and linear regression models were applied for quantitative traits. All regression models were adjusted by the top five principal components and array platforms. Next, the significant risk factors were applied as adjustments of regression models between gIQ and CAD to study their mediatory roles.

Because of genetic overlaps between intelligence and educational attainment, the effects of intelligence on CAD and its risk factors might be false positive. We defined the direct effect of intelligence as the effect of intelligence that was not mediated by educational attainment. Lee et al. reported 1271 independent SNPs associated with educational attainment through a meta-analysis in 1.1 million persons[113]. Seven SNPs associated with both intelligence and educational attainment were excluded during recalculating gIQ. Then same regression analysis was performed to study the direct effects of intelligence.

To further study causal pathways from intelligence/education to CAD risk, a multivariable two-sample MR analysis was carried out. The GWAS summary data of CAD and its risk factors, educational attainment were acquired from CARDIoGRAMplusC4D (CAD)[104], GIANT (BMI)[114], TAG (smoking)[115], GLGC (HDL-C, LDL-C)[71], SSGAC (educational attainment)[113], and DIAGRAM (T2D)[116]. Three MR methods were used including IVW, MR-egger, and weighted median. Lastly, MR sensitivity analyses were performed for intelligence and educational attainment respectively. SNPs moderately associated with CAD and risk factors ($P<0.001$) were removed from intelligence SNP and education SNPs respectively.
2.3 Methods used in the TWAS project

As introduced in section 1.7.1, the first step was to train prediction models for risk tissues of CAD. We adopted the existing expression prediction models trained using the EpiXcan pipeline by Zhang et al., including models of AOR, MAM, LIV, SF, VAF, BLD, and SKLM based on the STARNET panel[17], and models of AOR, LIV, BLD, SF, VAF, and SKLM based on the GTEx panel[18]. We trained prediction models for another two risk tissues of CAD, arterial wall coronary (COR) and tibial artery (TIB) datasets which are only available in the GTEx panel (V7), using the EpiXcan pipeline. All individuals used for prediction models were restricted to European descent. First, genetic variants were filtered out if they matched one of the QC parameters: calling rate<0.95, MAF<0.01, and HWE<1e-6. Second, for expression data, we did sample-level quantile normalization and gene-level inverse quantile normalization using preprocess codes of the PredicDB pipeline. Third, we calculated SNP priors using qtlBHM that jointly analyzed epigenome annotations of aorta derived from REMC[70]. Lastly, the SNP priors, genotype data, and expression data were jointly applied to 10-fold cross-validated WENet to train predicting models by deploying the EpiXcan pipeline[69]. The predictors were filtered by cross-validated prediction $R^2$>0.01.

Next, we applied prediction models of nine risk tissues to impute GReX from individual genotype data of ten CARDIoGRAMplusC4D cohorts[86, 101–109] and UKB[47], which, in total, having of 37,997 cases and 42,854 controls. For each tissue, we performed association tests between GReX and CAD using the logistic regression model in 11 genotype cohorts and performed meta-analysis for all cohorts to get summarized TWAS statistics. In total, we identified 114 genes representing 193 gene-tissue pairs thresholding Bonferroni-corrected significance.
Then, we did a series of *in-silico* analyses to check the plausibility of the TWAS gene list. i) We compared the genomic position between TWAS genes and GWAS loci using MAGMA[117]. ii) We compared TWAS genes with gene lists resulting from colocalization analyses. In this part, the tool coloc[23, 27] integrated eQTL data from either GTEx or STARNET and GWAS summary data from CARDIoGRAMplusC4D[65], providing significant gene-tissue pairs which posterior probability of hypothesis 4, PP4 ≥ 0.55 (section 1.3.1). iii) To check the biological function and pathogenicity of TWAS genes, we did pathway enrichment analysis using ClueGO[118] and disease enrichment analysis based on DisGeNET[119]. iv) We also performed gene-based rare variant association analysis using WES data from UKB on 200,632 participants. The damaging rare variants residing in TWAS genes were defined by one of five *in-silico* methods (LRT score, MutationTaster, PolyPhen-2 HumDiv, PolyPhen-2 HumVar, and SIFT) by the annotation resource dbNSFP 4.1a[59] using the VEP tool[61]. For gene-based tests, we used Fisher’s exact test for binary traits and the linear regression model for quantitative traits.

Finally, our attention was focused on 18 genes that resided outside known GWAS loci, which we called “novel”. To study the susceptibility of novel genes, we did *in-silico* analyses and *in-vitro* validation. We studied the genetic association between novel genes and a series of lipids traits in human genotype data from UKB. We also studied their association with lipid traits in atherosclerosis mouse models from the Hybrid Mouse Diversity Panel (HMDP). Both human genotype data and mouse expression data suggested connections between novel genes and lipid traits. Of these novel genes, *KPTN* and *RGS19*, which were rarely studied before, gave significant signals in liver tissue by TWAS analysis. So, we finally carried out knockdown experiments in human hepatocytes using dual CRISPR strategy for *KPTN* and *RGS19*. Cells for measurement of secretion of triglycerides, cholesterol, and APOL100 were cultured for 16
hours in serum-free medium. Medium triglycerides and cholesterol were enriched five times by vacuum centrifuge and measured with colorimetric kits, triglycerides (cobas) and CHOL2 (cobas), respectively.

III Discussion

3.1 Genetic association between Intelligence and CAD

This work applied ‘PRS+MR’ strategy which firstly use PRS to filter genetic association between traits, then apply MR method to study casual effects on associated traits (see section 1.8.2).

We firstly tested association of gIQ with CAD and its risk factors. In this part, we observed that one standard deviation (SD) increase of gIQ was related to a 5% decrease of CAD risk (odds ratio [OR] of 0.95; 95% confidence interval [CI] 0.93 to 0.98; $P=4.93e^{-5}$), which was validated in UKB (OR=0.97; 95% CI 0.96 to 0.99; $P=6.4e^{-4}$). In UKB data, we also observed significant inverse correlations between gIQ and risk factors of CAD including BMI (OR=0.899; 95% CI 0.886 to 0.911; $P=5.4e^{-49}$), smoking (OR=0.981; 95% CI 0.975 to 0.987; $P=8.3e^{-10}$), T2D (OR=0.966; 95% CI 0.951 to 0.980; $P=4.1e^{-6}$), hypertension (OR=0.987; 95% CI 0.981 to 0.993; $P=3.8e^{-5}$), and a positive correlation with HDL-C (OR=1.007; 95% CI 1.006 to 1.008; $P=1.3e^{-29}$). The associations of gIQ with CAD and its risk factors were largely attenuated after the adjustment of measured intelligence and educational attainment. Same phenomena happened between gIQ and CAD risk with the adjustment of risk factors significantly associated with gIQ. These findings suggested intermediate role of risk factors, measured intelligence, and educational attainment between gIQ and CAD.

Next, we applied two-sample MR analyses to depict casual pathway from genetic intelligence to CAD risk. One SD increase of intelligence resulted in decrease of CAD risk by
25% (OR=0.75; 95% CI 0.69 to 0.81; \( P<1\times10^{-10} \)), decrease of BMI by 0.1 kg/m² (95% CI -0.16 to -0.14; \( P=1.02\times10^{-3} \)), decrease of T2D risk by 15% (OR=85; 95% CI 0.77 to 0.95). One SD increase in the education years resulted in decrease of risk of CAD by 38% (OR=0.62; 95% CI 0.58 to 0.66; \( P<1\times10^{-10} \)), decrease of BMI by 0.32 kg/m² (95% CI -0.37 to -0.27; \( P<1\times10^{-10} \)), increase of HDL-C by 0.19 mmol/L (95% CI 0.14 to 0.25; \( P<1\times10^{-10} \)), decrease of the risk of smoking by 43% (OR=0.57; 95% CI 0.501 to 0.642; \( P<1\times10^{-10} \)), and decrease of T2D risk by 47% (OR=0.53; 95% CI 0.49 to 0.57; \( P<1\times10^{-10} \)). The effects of educational attainment on CAD and its risk factors displayed the same direction as intelligence but were stronger in magnitude.

Figure 10 Pathways from higher gIQ to lower risk of CAD.

In conclusion, using genetic approaches, we depicted a pathway from gIQ to CAD risk. The higher gIQ is associated with the higher measured intelligence and longer educational attainment, both of which appear to reduce the prevalence of risk factors of CAD including BMI, smoking, T2D, and hypertension, and increase HDL-C, which in concert subsequently reduce the prevalence of CAD (Figure 10). Moreover, the effects of educational attainment on risk factors and CAD appear to be stronger than the effects of intelligence. Thus, repetitive campaigns throughout schooling may be worthwhile for preventive reasons as they may ameliorate the association between gIQ and unhealthy lifestyle.
III Discussion

Author contributions

Heribert Schunkert and Ling Li designed the study and wrote the manuscript. Shichao Pang, Lingyao Zeng, and Ulrich Güldener provided technical support and gave conceptual advice.

3.2 TWAS for CAD

Our work is the first to systematically identify susceptibility genes of CAD in tissue-specific fashion using TWAS method. The main findings of this work were summarized in Figure 11. We first trained gene expression prediction models for nine CAD-relevant tissues using EpiXcan based on two largest genetics-of-gene-expression panels, STARNET and GTEx. We next explored these prediction models on individual level genotype data of 37,997 CAD cases and 42,854 controls. Our TWAS identified 114 CAD risk genes to be differentially expressed by genetic means. Of these, 96 genes were resided within ±1Mb region of previously identified by GWAS loci (a sort of positive control) and 18 were novel. For 114 TWAS genes, we performed stepwise analyses to prove their plausibility, biological function, and pathogenicity to CAD, including analyses for colocalization, damaging mutation, pathway enrichment, phenome-wide associations with human data and expression-traits correlations using mouse data. Finally, we focused on two novel genes, RSG19 and KPTN, and conducted CRISPR/Cas9-based knockdown experiments for them in human hepatocytes. We observed reduced secretion of APOB100 and lipids in the cell culture medium, i.e., a functional explanation for the association findings.
Discussion

Figure 11 Schematic illustration of CAD TWAS.

This work discovered 18 novel genes to be associated with CAD, and functionally evaluated 96 genes within CAD GWAS loci, for example by indicating their tissue(s) of action. The downstream analyses of these genes revealed their intermediate cardiometabolic phenotypes bridging gene variants with their effects on CAD. Our result provides a substantial step towards prioritization of genes at respective GWAS loci as well as their tissues of actions. In this respect, 46 genes identified by this TWAS are known for effects in pathophysiological pathways related to CAD, including lipid metabolism, inflammation, angiogenesis,
transcriptional regulation, cell proliferation, NO signaling, and high blood pressure, to name a
few giving credibility to the association findings.

Our *in-silico* analyses on novel genes suggested they were associated with lipid traits in both human genotype data and mouse expression data. The hypothesis was furtherly confirmed by the latest lipid GWAS paper conducted by GLGLC consortium (Table 2)[120]. Two novel genes, *KPTN* and *RGS19*, were firstly confirmed to be associated with lipid metabolism from both *in-silico* and *in-vitro* data. So, we believe that our study on novel genes may provide novel insights into molecular etiology of CAD.

Table 2 8 novel genes were within 500kb of lipid risk loci

<table>
<thead>
<tr>
<th>Gene</th>
<th>Lipid risk loci</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOMER3</td>
<td>EUR_TC_rs10423802;EUR_HDL_rs60570301;EUR_logTG_rs1966500;EUR_nonHDL_HOMER3</td>
</tr>
<tr>
<td>KPTN</td>
<td>EUR_HDL_rs12609461;EUR_logTG_rs10408163;EUR_nonHDL_rs12461923;EUR_HDL_KPTN</td>
</tr>
<tr>
<td>NLRC4</td>
<td>EUR_TC_rs62142080</td>
</tr>
<tr>
<td>RGS19</td>
<td>EUR_TC_rs35201382;EUR_LDL_rs6090040;EUR_nonHDL_20_62692060_C_A;EUR_HDL_RGS19</td>
</tr>
<tr>
<td>SDCCAG3</td>
<td>EUR_nonHDL_rs3780190;EUR_TC_rs3780190;EUR_LDL_rs13301660</td>
</tr>
<tr>
<td>STX4</td>
<td>EUR_TC_rs1870293;EUR_LDL_rs35468353;EUR_nonHDL_rs73530203;EUR_logTG_STX4</td>
</tr>
<tr>
<td>TXNRD3</td>
<td>EUR_LDL_rs9862203;EUR_TC_rs9862203</td>
</tr>
</tbody>
</table>
We must admit that our study has limitation embedded in TWAS mythology. Since TWAS are strongly dependent on the reference panel linking genetic signatures with gene expression, it had to be expected that STARNET- and GTEx-based predictive models display differences in gene-CAD associations. The difference may be due to different sample sizes used for training predictions models, different disease states (subjects with and without CAD), intravital or postmortem sample collection, as well as different transcript abundance and genotype coverage leading to differences in expression associated SNPs in our reference panels. Even so, a fair consistency of TWAS results between STARNET- and GTEx-based models gave us rationality of combing results derived from both panels to increase the power of capturing risk genes. Second, although TWAS facilitates candidate risk gene prioritization, LD-link between SNPs, co-regulation or co-expression in cis at a given locus limits the precise determination of the culprit gene. Indeed, at 12 loci we observed signals for three or more TWAS genes. This strengthens the importance of integration of other genetic analyses to improve risk gene prioritization. In our work, a series of stepwise analyses were performed on TWAS gene list to furtherly study their disease-causing mechanism, like damaging variant association, pathway enrichments, genetic association with other phenotypes and expression-trait association statistics. Lastly, all findings by in-silico methods have to be furtherly validated and functionally explained by in-vitro or in-vivo methods. Our prove-of-concept experiment on KPTN and RGS19 in human hepatocytes validated our hypothesis about their associations with lipid traits. But extra efforts are necessary to clearly depict the molecular mechanisms.
In summary, our TWAS study based on two genetics-of-gene-expression panels created a set of gene-centered and tissue-annotated associations for CAD, providing insightful guidance for further biological investigation and therapeutic development.

**Author contributions**

Heribert Schunkert, Ling Li, and Zhifen Chen designed the study and wrote the manuscript. Ling Li ran the bioinformatic analyses. Zhifen Chen, Shuangyue Li, and Andrea Steiner performed the wet lab experiments. Julien Gagneur, Moritz von Scheidt, Ulrich Güldene1, Simon Koplev, Angela Ma, Ke Hao, Calvin Pan, Aldons J. Lusis, Shichao Pang, Thorsten Kessler, Raili Ermel, Katyayani Sukhavasi, Arno Ruusalepp, Jeanette Erdman, Jason C. Kovacic, Johan L.M. Björkegren provided research data, technical support and gave conceptual advice.

**IV Conclusion and outlook**

I conducted two original research on post-GWAS studies during my graduate training, including a PRS study and the CAD TWAS. In the PRS study, I used genetic methods to firstly verify the epidemiology phenomenon of reverse association between intelligence and CAD risk and depicted a pathway between them. In the TWAS study, we pinpointed risk genes of CAD as well as their action tissues which empowered our understanding of the molecular mechanisms of CAD. The positive experimental validations of two novel genes might lay the foundation of the therapeutic development of CAD. The two projects inspired me to explore potential transcriptome changes due to the polygenic risk in a tissue-specific fashion.

Emerging wave of TWAS studies focuses on investigating the role of spicing variants and transcripts in CAD. The data sets of tissue splicing variants (sQTL) and transcripts are
available in several public resources, such as STARNet and GTEx. However, the lack of cellular level omics data is a disadvantage of the current TWAS analysis. The CAD risk linked to the cell-type specific function of a gene might be obscured due to the mix cellular profile in a tissue. The booming single-cell technology will soon tackle the current disadvantage and TWAS analysis will be performed in a cell type- or subtype-specific manner. Moreover, the concept of TWAS analysis will be applied to several other omics datasets, such as epigenomics, proteomics, and metabolomics, when tissue- or cell-specific data will be available in scale. The different layers of biological data could be integrated for a multi-omics TWAS analysis to uncover the interaction among the layers of cellular profiles. The increasing complexity in the data integration will undoubtedly challenge the statistic algorithms. In this scenario, machine learning (ML) will embark to unlock the myth of complex common diseases.

Finally, I’d like close my thesis using poetry from a Chinese famous poet Qu Yuan: Long, long had been my road and far, far was the journey; I would go up and down to seek my heart's desire (路漫漫其修远兮，我将上下而求索).
References

References

References

Sawhney M, Saylan M, Schmidt MI, Schutte AE, Senthilkumaran S, Sepanlou SG, Sha
F, Shahabi S, Shahid I, Shaikh MA, Shamali M, Shamsizadeh M, Shawon MSR, Sheikh
Silva DAS, Singh JA, Singh A, Skryabin VY, Skryabina AA, Soheili A, Spurlock EE,
Stockfelt L, Stortecky S, Stranges S, Abdulkader RS, Tadbiri H, Tadesse EG, Tadesse
DB, Tajdini M, Tariqujjaman M, Teklehaimanot BF, Temsah MH, Tesema AK, Thakur
B, Thankappan KR, Thapar R, Thrift AG, Timalsina B, Tonelli M, Touvier M, Tovani-
Palone MR, Tripathi A, Tripathy JP, Truelsen TC, Tsegay GM, Tsegaye GW,
Tsilimparis N, Tusa BS, Tyrovolas S, Umapathi KK, Unim B, Unnikrishnan B, Usman
MS, Vaduganathan M, Valdez PR, Vasankari TJ, Velazquez DZ, Venkatasubramanian
BL, Xu G, Yadollahpour A, Yamagishi K, Yan LL, Yandrapalli S, Yano Y, Yatsuya H,
Yeheyis TY, Yeshaw Y, Yilgwan CS, Yonemoto N, Yu C, Yusefzadeh H, Zachariah G,
Zaman S Bin, Zaman MS, Zamanian M, Zand R, Zandifar A, Zarghi A, Zastrozhin MS,
CJL, Fuster V (2020) Global Burden of Cardiovascular Diseases and Risk Factors, 1990-
2019: Update From the GBD 2019 Study. J Am Coll Cardiol 76:2982–3021. doi:
10.1016/J.JACC.2020.11.010

2. Khan MA, Hashim MJ, Mustafa H, Baniyas MY, Suwaidi SKBM Al, AlKatheeri R,
Alblooshi FMK, Almatrooshi MEAH, Alzaabi MEH, Darmaki RS Al, Lootah SNAH
(2020) Global Epidemiology of Ischemic Heart Disease: Results from the Global
Burden of Disease Study. Cureus 12. doi: 10.7759/CUREUS.9349

References


References


13. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, De Bakker PIW, Daly MJ, Sham PC (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575. doi: 10.1086/519795


PLoS Genet 16. doi: 10.1371/JOURNAL.PGEN.1008538


References


36. Østergaard SD, Mukherjee S, Sharp SJ, Proitsi P, Lotta LA, Day F, Perry JRB, Boehme
References

References


62


References

doi: 10.1161/CIRCGEN.120.003014


65. Investigators MIG and CardiEC (2016) Coding Variation in ANGPTL4, LPL, and


References


References

References

45:1274–1283. doi: 10.1038/ng.2797


Transcriptome-Wide Association Studies. PLOS Genet 17:e1009398. doi: 10.1371/JOURNAL.PGEN.1009398

References


References

References


References


97. Shaw DR (2016) Searching the mouse genome informatics (MGI) resources for information on mouse biology from genotype to phenotype. Curr Protoc Bioinforma 2016:1.7.1-1.7.16. doi: 10.1002/cpbi.18

doi: 10.1161/CIRCGEN.121.003365


References

References

References


References


References


References


References


Publications during PhD


Approval from each publisher

Publication 1

Genetically determined intelligence and coronary artery disease risk
Author: Ling J, et al.
Publication: Clinical Research in Cardiology
Publisher: Springer Nature
Date: Aug 1, 2013
Copyright © 2013, The Author(s)

Creative Commons
This is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
To request permission for a type of use not listed, please contact Springer Nature.

Publication 2

Transcriptome-wide association study of coronary artery disease identifies novel susceptibility genes
Author: Ling J, et al.
Publication: Basic Research in Cardiology
Publisher: Springer Nature
Date: Feb 17, 2012
Copyright © 2012, The Author(s)

Creative Commons
This is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
To request permission for a type of use not listed, please contact Springer Nature.
Appendix I Genetically determined intelligence and coronary artery disease risk
Genetically determined intelligence and coronary artery disease risk

Ling Li1 · Shichao Pang1 · Lingyao Zeng1 · Ulrich Gündener1 · Heribert Schunkert1,2

Received: 7 May 2020 / Accepted: 23 July 2020 © The Author(s) 2020

Abstract

Background Epidemiological studies have shown inverse association between intelligence and coronary artery disease (CAD) risk, but the underlying mechanisms remain unclear.

Methods Based on 242 SNPs independently associated with intelligence, we calculated the genetic intelligence score (gIQ) for participants from 10 CAD case-control studies (n = 34,083) and UK Biobank (n = 427,306). From UK Biobank, we extracted phenotypes including body mass index (BMI), type 2 diabetes (T2D), smoking, hypertension, HDL cholesterol, LDL cholesterol, measured intelligence score, and education attainment. To estimate the effects of gIQ on CAD and its related risk factors, regression analyses were applied. Next, we studied the mediatory roles of measured intelligence and educational attainment. Lastly, Mendelian randomization was performed to validate the findings.

Results In CAD case-control studies, one standard deviation (SD) increase of gIQ was related to a 5% decrease of CAD risk (odds ratio [OR] of 0.95; 95% confidence interval [CI] 0.93 to 0.98; \(P = 4.93 \times 10^{-5}\)), which was validated in UK Biobank (OR = 0.97; 95% CI 0.96 to 0.99; \(P = 6.4 \times 10^{-4}\)). In UK Biobank, we also found significant inverse correlations between gIQ and risk factors of CAD including smoking, BMI, T2D, hypertension, and a positive correlation with HDL cholesterol. The association signals between gIQ and CAD as well as its risk factors got largely attenuated after the adjustment of measured intelligence and educational attainment. The causal role of intelligence in mediating CAD risk was confirmed by Mendelian randomization analyses.

Conclusion Genetic components of intelligence affect measured intelligence and educational attainment, which subsequently affect the prevalence of CAD via a series of unfavorable risk factor profiles.

Keywords Coronary artery disease · Intelligence · Educational attainment · Genetic association · Genetic risk score · Smoking · Obesity

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00392-020-01727-x) contains supplementary material, which is available to authorized users.

Extended author information available on the last page of the article

Published online: 02 August 2020 © Springer
Appendix I Genetically determined intelligence and coronary artery disease risk

Introduction

Epidemiological studies have shown an inverse association between intelligence score and risk of coronary artery disease (CAD) [1, 2]. Higher intelligence is also inversely associated with risk factors of CAD, like smoking and obesity [3–5]. Moreover, there is evidence for association between higher intelligence and longer educational attainment [6, 7] which might be an important mediator in reducing CAD risk [8]. However, the mechanisms linking higher intelligence with a decreased risk of CAD remain unclear.

Genome wide association studies (GWAS) have identified large numbers of genetic variants, typically single nucleotide polymorphisms (SNPs), associated with a wide range of complex traits providing opportunities of exploring the relationships between traits. Polygenic risk scores defined as sum of trait-associated SNPs weighted by effect size derived from large-scale GWAS measure the liability of individuals developing such traits [9, 10]. Thereby polygenic risk scores become an important genetic tool for studying association between traits [8, 11]. Two-sample Mendelian randomization (MR) is another genetic method of accessing causal relationships among traits which requires summary statistics of GWAS instead of full individual level genotype data and phenotypic measurements [12].

Savage et al., performed genome-wide association meta-analysis in 269,967 individuals and identified 242 SNPs independently associated with intelligence [13]. We used the statistics of these intelligence SNPs to perform both regression analysis of the individual-level polygenic score and two-sample MR analysis to study the association between intelligence and CAD risk, and to explore potential pathways from a higher genetic intelligence score to lower CAD risk.

Methods

Cohorts description of individual-level genotype data

Individual level genotype data were collected from ten case–control studies of CAD as discovery set [14–21]. All participants were of European descent, mostly from the Germany and UK. The replication set was from UK Biobank [22] which includes genotypes of 487,409 individuals derived from two different genotyping array platforms.

The data of UK Biobank were also applied to characterize interplay between intelligence and risk factors of CAD including body mass index (BMI), type 2 diabetes (T2D), HDL cholesterol, LDL cholesterol, hypertension, and smoking behavior. These traits were either self-reported or extracted from hospital episodes or death registries as reported by UK Biobank [22]. Intelligence scores were measured in UK Biobank through a 13-item verbal-numeric reasoning test designed to assess the ability of solving problems that require logic and reasoning ability, independent of acquired knowledge (field ID 20016). The total range of intelligence as measured by this score was from 0 to 13 arbitrary unit. Details of corresponding studies, data preprocessing and traits definition of data from UK Biobank are shown at Supplementary Notes and Table S1.

Intelligence associated variants

Savage et al., performed GWAS meta-analysis of 14 independent epidemiological cohorts of European descent and reported 242 independent SNPs with genome-wide significant association (\(P<5\times10^{-8}\)) to intelligence scores [13]. We estimated effect size for each SNP from GWAS summary statistic table using method by Zhu et al. [23]. Details are shown at Supplementary Notes and Table S2.

Statistics

The summary statistics of 242 independent SNPs of intelligence were applied to calculate the individual-level weighted genetic score of intelligence for each study. Firstly, each variant was given a value from 0 to 2 according to the presence of the intelligence allele in the imputed genotype data of each participant, which was then multiplied with the effect size of the variant on intelligence. For variants with missing genotypes in the imputed data, the reference allele frequency was applied. Then we summed these values of 242 variants for each participant as the polygenic score of intelligence, namely the genetic intelligence score (gIQ). Afterwards, the continuous gIQ was standardized into z-scores with mean of 0 and standard deviation (SD) of 1. By logistic regression analyses, we estimated effects of gIQ on CAD risk for each study separately. To control the bias due to population stratification or different genotyping platforms, the first two principle components for 10 CAD studies were added as adjustments of the regression model. In UK Biobank, because of more complex population structure, we employed top five principle components and array platforms for this data set. Lastly, the fixed-effect size meta-analysis was performed to estimate the combined effects across all CAD studies. Based on gIQ, all individuals were evenly separated into low, medium and high groups to study the distribution of cases and controls along with increasing gIQ.
Appendix I Genetically determined intelligence and coronary artery disease risk

Clinical Research in Cardiology

Although the gIQ reflects intelligence at first place, the SNPs utilized in this score may be pleiotropic and thus affects other traits [24-26]. Seven of 242 intelligence SNPs were reported to be associated with educational attainment through a large scale GWAS cohort which detected 1271 education-associated SNPs [27]. We thus re-evaluated the association between gIQ and CAD risks after exclusion of 7 SNPs overlapping with educational attainment to estimate the direct effects of intelligence.

In UK Biobank, we estimated effects of gIQ on measured intelligence, educational attainment, risk factors of CAD including BMI, T2D, smoking, HDL cholesterol, LDL cholesterol and hypertension. Definitions of these traits are shown at Supplementary Notes and Table S3. Logistic regression was applied to binary traits like T2D, smoking, hypertension; and linear regression was for continuous traits like measured intelligence, educational attainment, BMI, HDL cholesterol, and LDL cholesterol. Top five principle components and array platforms were used as adjustment of regression models. We also studied phenotypic association of measured intelligence with educational attainment and CAD incidence in UK Biobank. Additionally, to avoid the genetic influence of education derived from genetic overlaps between intelligence and education, we re-estimated the effects of intelligence on CAD and its risk factors by eliminating seven overlapping SNPs [27].

Two-sample Mendelian randomization analysis

Mendelian randomization (MR) is a method using genetic variants as instruments to study causal relationships between exposures and outcomes [28]. We introduced the multivariable two-sample MR analysis to investigate the direct causal effects of intelligence and educational attainment on CAD and its risk factors. This approach taking GWAS summary statistics as input measures effects of one standard deviation (SD) change in intelligence or educational attainment. As bias can be introduced in two-sample MR when using genetic consortia that have partially overlapping sets of participants, we selected consortia without overlaps. The GWAS summary statistics of CAD and its risk factors, educational attainment were acquired from CARDioGRAMplus45CAD (CAD) [17], GIANT (BMI) [29], TAG (smoking) [30], GLGC (HDL cholesterol, LDL cholesterol) [31], SSSAGAC (educational attainment) [27], and DIAGRAM (T2D) [32]. Elaborate description of these five studies can be found at Supplementary Notes and Table S4.

To address the influence of genetic overlaps between education and intelligence, we eliminated seven SNPs that are both associated with intelligence and educational attainment in MR analysis. Three MR methods including inverse-variance-weighted average (IVW) [33], MR-egger [34] and weighted median [35] were applied. Relationships significant ($p < 0.05$) in at least two of three methods were identified to be reliable and shown by IVW results. Lastly, sensitivity analysis of effects of intelligence and educational attainment on CAD were performed by excluding SNPs that were moderately associated with risk factors of CAD ($p < 0.001$) from intelligence SNPs and education SNPs respectively. Details are shown at Supplementary Notes.

Results

Effect of gIQ on the risk of CAD

Ten case–control studies of CAD with 16,144 CAD cases and 17,939 controls were included in this study. Majority of participants were from the Germany and UK. Individual-level genotype data and elaborate phenotype data from UK Biobank were used as validating set containing 20,310 CAD cases which were defined by either self-reported, or hospital episode and death registry data, and 406,996 controls. (Supplementary Notes and Table S1). For each cohort, we generated gIQ based on 242 SNPs reported to be genome-wide significantly associated with intelligence [13].

The score in participants of the 10 CAD studies was normally distributed (Fig. S1). Meta-analysis using fixed-effect size model indicated relative decrease of CAD risk by 5% (95% confidence interval [CI], 0.93 to 0.98; $p = 4.93e-5$) along with per 1-SD increase in gIQ (Fig. 1). When individuals were equally grouped into a low, medium and high group of gIQ, risk of CAD steadily decreased with an odds ratio (OR) of the high group vs low group being 0.89 (95% CI 0.84 to 0.93; $p = 6.2e-6$, Fig. 2).

Data from the UK Biobank confirmed the inverse association between gIQ and CAD risk with an OR = 0.97 (95% CI 0.96 to 0.99; $p = 6.4e-4$, Fig. 1). The risk of high gIQ group was 7% lower than the low gIQ group ($p = 0.0005$) in UK Biobank. As expected, the association between gIQ and CAD risk was abolished after adjustment for measured intelligence defining measured intelligence as an intermediary trait between gIQ and CAD risk (Fig. 3).

Bidirectional association between intelligence and education

In UK Biobank, we found that 1-SD increase of gIQ increased measured intelligence score by 0.29 unit ($p < 1e-10$) and prolonged years spent in school by 0.45 year. In addition, one more year spent in school increased the measured intelligence score by 0.16 unit ($p < 1e-10$). Vice versa, one unit increase in measured intelligence prolonged years spent in school by 0.98 year ($p < 1e-10$). Both the measured intelligence and educational attainment had inverse effects on CAD risk. See results in Table S5.
Appendix I Genetically determined intelligence and coronary artery disease risk

<table>
<thead>
<tr>
<th>Study</th>
<th>N CAD cases</th>
<th>N controls</th>
<th>Odds ratio [95% CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>GerMFSc</td>
<td>622</td>
<td>1551</td>
<td>1.000 [0.90, 1.12]</td>
</tr>
<tr>
<td>GerMFScv</td>
<td>1192</td>
<td>1256</td>
<td>0.900 [0.83, 0.96]</td>
</tr>
<tr>
<td>GerMFScV</td>
<td>1065</td>
<td>1441</td>
<td>0.970 [0.89, 1.06]</td>
</tr>
<tr>
<td>GerMFScvV</td>
<td>954</td>
<td>1136</td>
<td>0.900 [0.73, 1.00]</td>
</tr>
<tr>
<td>GerMFSc</td>
<td>2407</td>
<td>1574</td>
<td>1.000 [0.99, 1.12]</td>
</tr>
<tr>
<td>GerMFScv</td>
<td>1639</td>
<td>1136</td>
<td>0.970 [0.90, 1.05]</td>
</tr>
<tr>
<td>GerMFScV</td>
<td>2062</td>
<td>3462</td>
<td>1.020 [0.95, 1.09]</td>
</tr>
<tr>
<td>CARDIOPEN</td>
<td>382</td>
<td>404</td>
<td>0.790 [0.67, 0.90]</td>
</tr>
<tr>
<td>WTCCC</td>
<td>1900</td>
<td>2911</td>
<td>1.010 [0.95, 1.07]</td>
</tr>
<tr>
<td>MGan</td>
<td>2961</td>
<td>3018</td>
<td>0.970 [0.83, 1.10]</td>
</tr>
<tr>
<td>FE meta-analysis</td>
<td>$p = 4.93 \times 10^{-5}$</td>
<td>0.900 [0.93, 0.98]</td>
<td></td>
</tr>
<tr>
<td>UK Biobank</td>
<td>20310</td>
<td>406966</td>
<td>0.970 [0.96, 0.98]</td>
</tr>
<tr>
<td>FE meta-analysis</td>
<td>$p = 7.55 \times 10^{-7}$</td>
<td>0.970 [0.95, 0.98]</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 1 Association of gIQ and CAD risk. The genetic intelligence score was calculated in 10 case-controls studies of CAD and UK Biobank respectively. Logistic regression was performed to evaluate the association between gIQ and CAD risks in each study. Fixed-effect size meta-analysis was performed to combine all studies. Forest plot shows regression result in each study and the overall effect size. The gIQ was inversely associated with CAD risk.

Fig. 2 Distribution of cases and controls according to gIQ. Individuals in 10 CAD studies were evenly grouped into a low (score = 1), medium (score = 2) and high (score = 3) group according to their gIQ. The OR for incidence of CAD relative to low group. Risk of CAD decreases along the increase of gIQ.

Effects of gIQ on risk factors of CAD

We next asked, in UK Biobank data, whether the association between gIQ and CAD risk was mediated by traditional risk factors of CAD, and whether such effects were dependent on measured intelligence and educational attainment. We found strong associations of gIQ with BMI, smoking, T2D, HDL cholesterol, and hypertension (Fig. 3). The effects of gIQ on CAD risk factors were largely attenuated after adjustment for measured intelligence or educational attainment (Fig. 3 and Table S6), suggesting that measured intelligence and educational attainment mainly mediated associations between gIQ and these risk factors. The analyses after removal of seven SNPs overlapping between intelligence and educational attainment obtained quantitatively and qualitatively similar effects of gIQ on CAD and its risk factors (Fig. S2).

We also studied the mediatory roles of these risk factors on the association between gIQ and CAD risk by applying them as adjustments to the regression model. Adjusting for individual risk factor or risk factors combined markedly attenuated association signals between gIQ and CAD risk (Fig. S3), indicating these risk factors were involved in mediating the association between gIQ and CAD risk.

Mendelian randomization validation

To substantiate our observations, we performed multi-variable two-sample MR analysis taking intelligence or
Appendix I Genetically determined intelligence and coronary artery disease risk

### Clinical Research in Cardiology

<table>
<thead>
<tr>
<th>Traits</th>
<th>OR[95%CI]</th>
<th>Pvalue</th>
<th>Adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAD</td>
<td>0.976[0.962,0.990]</td>
<td>6.4e-04</td>
<td>No</td>
</tr>
<tr>
<td>CAD</td>
<td>1.010[0.986,1.036]</td>
<td>0.412</td>
<td>IQ</td>
</tr>
<tr>
<td>CAD</td>
<td>1.006[0.991,1.022]</td>
<td>0.420</td>
<td>Edu</td>
</tr>
<tr>
<td>BMI</td>
<td>0.899[0.886,0.911]</td>
<td>5.4e-49</td>
<td>No</td>
</tr>
<tr>
<td>BMI</td>
<td>0.933[0.912,0.955]</td>
<td>5.3e-09</td>
<td>IQ</td>
</tr>
<tr>
<td>BMI</td>
<td>0.846[0.832,0.861]</td>
<td>1.0e-12</td>
<td>Edu</td>
</tr>
<tr>
<td>Smoking</td>
<td>0.987[0.975,0.987]</td>
<td>6.3e-10</td>
<td>No</td>
</tr>
<tr>
<td>Smoking</td>
<td>0.984[0.974,0.994]</td>
<td>0.002</td>
<td>IQ</td>
</tr>
<tr>
<td>Smoking</td>
<td>1.000[0.984,1.007]</td>
<td>0.950</td>
<td>Edu</td>
</tr>
<tr>
<td>T2D</td>
<td>0.966[0.951,0.980]</td>
<td>4.1e-06</td>
<td>No</td>
</tr>
<tr>
<td>T2D</td>
<td>0.966[0.961,1.011]</td>
<td>0.278</td>
<td>IQ</td>
</tr>
<tr>
<td>T2D</td>
<td>0.997[0.981,1.014]</td>
<td>0.742</td>
<td>Edu</td>
</tr>
<tr>
<td>Hypertension</td>
<td>0.897[0.891,0.903]</td>
<td>3.9e-05</td>
<td>No</td>
</tr>
<tr>
<td>Hypertension</td>
<td>1.000[0.991,1.010]</td>
<td>0.923</td>
<td>IQ</td>
</tr>
<tr>
<td>Hypertension</td>
<td>1.000[0.994,1.007]</td>
<td>0.883</td>
<td>Edu</td>
</tr>
<tr>
<td>LDL cho.</td>
<td>0.999[0.996,1.001]</td>
<td>0.341</td>
<td>No</td>
</tr>
<tr>
<td>LDL cho.</td>
<td>0.998[0.993,1.002]</td>
<td>0.274</td>
<td>IQ</td>
</tr>
<tr>
<td>LDL cho.</td>
<td>0.996[0.996,1.002]</td>
<td>0.547</td>
<td>Edu</td>
</tr>
<tr>
<td>HDL cho.</td>
<td>1.007[1.0016,1.008]</td>
<td>1.3e-29</td>
<td>No</td>
</tr>
<tr>
<td>HDL cho.</td>
<td>1.006[1.004,1.008]</td>
<td>1.0e-09</td>
<td>IQ</td>
</tr>
<tr>
<td>HDL cho.</td>
<td>1.005[1.003,1.007]</td>
<td>5.9e-06</td>
<td>Edu</td>
</tr>
</tbody>
</table>

![Graph](image)

**Fig. 3** Associations of gIQ with CAD and its risk factors including BMI, smoking, T2D, HDL cholesterol, LDL cholesterol, and hypertension in UK Biobank. The OR for BMI is shown as logarithm of the linear regression coefficient. ‘Adjusted’ indicates the regression model between gIQ and trait after adjustment for measured intelligence (IQ), or length of school years completed (Edu), or neither of the two (No). The gIQ had inverse effects on BMI, T2D, smoking, and hypertension and a positive effect on HDL cholesterol. The association signals were largely attenuated by measured intelligence and educational attainment.

Educational attainment as exposures, CAD and its risk factors as outcomes. The estimates of the direct effects on outcomes for intelligence and education were generally in a consistent direction (Fig. 4). A 1 SD increase of intelligence resulted in decrease of CAD risk by 25% (OR = 0.73; 95% CI 0.69 to 0.81; P < 1e-10), decrease of BMI by 0.1 kg/m^2 (95% CI − 0.16 to − 0.14; P = 1.0e-3), decrease of T2D risk by 15% (OR = 0.85; 95% CI 0.77 to 0.95). A SD increase in the education years resulted in decrease of risk of CAD by 38% (OR = 0.62; 95% CI 0.58 to 0.66; P < 1e-10), decrease of BMI by 0.32 kg/m^2 (95% CI − 0.37 to − 0.27; P < 1e-10), increase of HDL cholesterol by 0.19 mmol/L (95% CI 0.14 to 0.25; P < 1e-10), decrease of the risk of smoking by 43% (OR = 0.57; 95% CI 0.501 to 0.642; P < 1e-10), and decrease of T2D risk by 47% (OR = 0.53; 95% CI 0.49 to 0.57; P < 1e-10). The effects of educational attainment on CAD and its risk factors displayed the same direction as intelligence but were stronger in magnitude. See details at supplementary notes and Table S7.

Lastly, MR sensitivity analysis were performed for intelligence and educational attainment respectively. For intelligence, SNPs moderately associated (P < 0.001) with CAD (n = 5), BMI (n = 45), and HDL cholesterol (n = 5) were removed from intelligence SNPs. The sensitivity analysis showed 1 SD increase of intelligence to decrease the risk of CAD by 22% (OR for IVW method of 0.78; 95% CI 0.72 to 0.84; P = 5.6e-10). Same as intelligence, SNPs moderately associated (P < 0.001) with CAD (n = 13), BMI (n = 155), HDL cholesterol (n = 6), LDL cholesterol (n = 5), and smoking (n = 2) were removed from education SNPs. The sensitivity analysis showed 1 SD increase in education years to decrease the risk of CAD by 34% (OR for IVW method of 0.66; 95% CI 0.62 to 0.70; P < 1e-10). Results are shown at Table S8.
Appendix I Genetically determined intelligence and coronary artery disease risk

Clinical Research in Cardiology

Discussion

Epidemiological studies have revealed that increased intelligence correlates with reduced CAD risk [1, 2]. Consistently, our study shows that 1-SD increase of gIQ based on accumulated effects of genetic variants associated with intelligence, results in 5% decrease in the risk of CAD. The CAD risk in the high group of gIQ is relatively lower by 11% than in the low group. The observation was replicated in the UK Biobank. Interestingly, the inverse association got largely attenuated after adjustment for measured intelligence and educational attainment supporting the hypothesis that these traits play a role in modulating CAD risk.

Our study also shows the inverse effects of gIQ on health-related outcomes including BMI, smoking, T2D, hypertension, and a positive effect on HDL cholesterol, which are well-known for their influences on CAD risk [36-40]. Same as for CAD, these association signals appear to be largely mediated by measured intelligence and educational attainment. It can be concluded that these risk factors mediate the association between gIQ and CAD risk individually and collectively.

Our study confirms that intelligence and educational attainment are genetically and phenotypically associated with each other [6, 7]. Like in the present study, a recent study by our group states that educational attainment is inversely associated with CAD risk which appears to be mediated by risk factors such as BMI and smoking [8, 26]. Interestingly, our current study indicates that the effects of educational attainment on CAD and its risk factors are quantitatively stronger than respective effects of intelligence. All these findings indicate that improving educational attainment can have potential benefits in improving decision-making regarding health-relevant lifestyle factors and reducing risk of CAD and other health-related outcomes.

Polygenic risk score and two-sample MR are two genetic approaches of investigating association between traits. Compared with the traditional epidemiologic approach, the genetic approach is unlikely to be confounded by lifestyle or environmental factors as genotypes are stable over lifetime [11]. The utilization of genetic methods is limited, however, by false discovery because of horizontal pleiotropy, a phenomenon explained by the fact that variants may affect multiple traits through different pathways [9, 10]. The complex interplay of intelligence and educational attainment caused by their genetic roots limits a precise causal relationship between intelligence and CAD as well as its risk factors. In our study, we aimed to exclude genetic overlaps between intelligence and education to highlight putative causal effects of intelligence on CAD and its risk factors. Indeed, this notion was furtherly confirmed by MR analysis and the MR sensitivity analysis after excluding SNPs marginally associated with risk factors of CAD from intelligence (or education).

There are some limitations in our study. First, the intelligence SNPs utilized in this study were identified from a large GWAS meta-analysis based 14 independent epidemiological cohorts of European ancestry [13]. To avoid bias due to difference in population genetics, we restricted our analysis to cohorts from Germany, UK, and others of European ancestry. Second, there might be other health-related or socioeconomic factors that interplay with intelligence and CAD risk [26]. Specially, environmental exposures can be important confounders of association between intelligence and CAD risk. Third, the measured intelligence obtained in UK Biobank through a 13-item verbal-numeric reasoning test does not equal to real intelligence whose full scopes are unspecifiable. Moreover, educational attainment defined as years spent in schools in this study has a wide spectrum in various countries. Last, the two-sample MR analyses are likely to be biased if two studies contains overlapping participants or cohorts which are quite common in large-scale GWAS meta-analysis [41]. We tried best to choose studies that are of European ancestry and have minimal overlaps to avoid such bias in two-sample MR analysis.

In conclusion, using genetic approaches, we depicted a pathway from gIQ to CAD risk (Fig. 5). The higher gIQ is associated with the higher measured intelligence and longer educational attainment, both of which appear to reduce the prevalence of risk factors of CAD including BMI, smoking,
Appendix I Genetically determined intelligence and coronary artery disease risk

Clinical Research in Cardiology

Fig 5 Pathway from higher gIQ to lower risk of CAD. Our study shows inverse effects of genetic determinants of intelligence on CAD and its risk factors including BMI, smoking, hypertension, and T2D and positive effects on HDL cholesterol. These association signals are mediated by measured intelligence and educational attainment, which are bidirectionally associated with each other.

T2D and hypertension, and increase HDL cholesterol, which in concert subsequently reduce the prevalence of CAD. Moreover, the effects of educational attainment on risk factors and CAD appear to be stronger than the effects of intelligence. Thus, repetitive campaigns throughout schooling may be worthwhile for preventive reasons as they may ameliorate the association between gIQ and unhealthy lifestyle.

Acknowledgements Open Access funding provided by Projekt DEAL.

Funding We gratefully acknowledge the support of the German Federal Ministry of Education and Research (BMBF) within the framework of ERA-NET on Cardiovascular Disease (Drugable-Mi- genes: 01KL1802), within the scheme of target validation (BLOCK- CAD: 16GW0168K), the German Centre of Cardiovascular Research (DZHK) Munich Heart Alliance, within the framework of the eMed research and funding concept (ANCD-Net: 10ZX1706C). As a Co-applicant for the British Heart Foundation (BHF)/German Centre of Cardiovascular Research (DZHK)-collaboration, we kindly acknowledge their support and funding. Further support has been received from the German Research Foundation (DFG) as part of the Sonderforschungsbereich SFB 1123 (B02) and the Sonderforschungsbereich SFB TRO 267 (B05). Furthermore, Bavarian State Ministry of Health and Care funded this work within its framework of DigiMed Bayern (Grant No: DMB-1805-0001). The work has also been supported by the German Federal Ministry of Economics and Energy in its scheme of ModulMax (Grant No ZF45902018A). We got authorization for applying data from UK Biobank (project ID: 25214).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Appendix I Genetically determined intelligence and coronary artery disease risk

Clinical Research in Cardiology
Appendix I Genetically determined intelligence and coronary artery disease risk

Clinical Research in Cardiology

Affiliations

Ling Li1 · Shichao Pang1 · Lingyao Zeng1 · Ulrich Gülde1er1 · Heribert Schunkert1,2

1 Heribert Schunkert
schunkert@dhm.mhn.de

2 Deutsches Zentrum für Herz- und Kreislaufforschung
(DZHK), Munich Heart Alliance, Munich, Germany

1 Department of Cardiology, Deutsches Herzzentrum
München, Technische Universität München, Lazarettstr. 36,
80636 Munich, Germany
Appendix II Transcriptome-wide association study of coronary artery disease identifies novel susceptibility genes

Appendix II Transcriptome-wide association study of coronary artery disease identifies novel susceptibility genes
Appendix II Transcriptome-wide association study of coronary artery disease identifies novel susceptibility genes

Transcriptome-wide association study of coronary artery disease identifies novel susceptibility genes

Ling Li1,2,3, Zhifen Chen1,3,2, Moritz von Scheidt1,3,2, Shuangyue Li1,3, Andrea Steiner1,3,2, Ulrich Güldener1,2,2, Simon Koplev1,2, Angela Ma1,2, Ke Hao1, Calvin Fan1,2, Aldons J. Luis1,3,2, Shichao Fang1,2, Thorsten Kessler1,3,2, Ralf Irmfelt1,2, Katayani Sukhavastu1,2, Arno Ruusalepp1,2,1,1, Julien Gagneur1,2, Jeanette Erdmann1,3,2, Jason C. Kovacic1,2,3,4,4, Johan L. M. Björkegren1,2,4,5,6,7,8,9,10,11, and Heribert Schunkert1,2,3

Received: 2 November 2021 / Revised: 18 January 2022 / Accepted: 1 February 2022 / Published online: 17 February 2022
© The Author(s) 2022, corrected publication 2022

Abstract

The majority of risk loci identified by genome-wide association studies (GWAS) are in non-coding regions, hampering their functional interpretation. Instead, transcriptome-wide association studies (TWAS) identify gene-trait associations, which can be used to prioritize candidate genes in disease-relevant tissue(s). Here, we aimed to systematically identify susceptibility genes for coronary artery disease (CAD) by TWAS. We trained prediction models of nine CAD-relevant tissues using EpiXcm based on two genetics-of-gene-expression panels, the Stockholm-Tartu Atherosclerosis Reverse Network Engineering Task (STARNET) and the Genotype-Tissue Expression (GTeX). Based on these prediction models, we imputed gene expression of respective tissues from individual-level genotype data on 37,997 CAD cases and 42,854 controls for the subsequent gene-trait association analysis. Transcriptome-wide significant association (i.e., P < 3.85e−6) was observed for 114 genes. Of these, 96 resided within previously identified GWAS risk loci and 18 were novel. Stepwise analyses were performed to study their plausibility, biological function, and pathogenicity in CAD, including analyses for colocalization, damaging mutations, pathway enrichment, phenotype-wide associations with human data and expression-trait correlations using mouse data. Finally, CRISPR/Cas9-based gene knockdown of two newly identified TWAS genes, RGS19 and KPTN, in a human hepatocyte cell line resulted in reduced secretion of APOB100 and lipids in the cell culture medium. Our CAD TWAS work (i) prioritized candidate causal genes at known GWAS loci, (ii) identified 18 novel genes to be associated with CAD, and (iii) suggested potential tissues and pathways of action for these TWAS CAD genes.

Keywords Coronary artery disease · Transcriptome-wide association study · Genome-wide association study · Genetically regulated expression

Introduction

Coronary artery disease (CAD), a leading cause of premature death worldwide, is influenced by interactions of lifestyle, environmental, and genetic risk factors [43]. Genomewide association studies (GWAS) have identified over 200 risk loci for CAD [11, 17, 35]. Most of them are located in non-coding regions which hampers their functional interpretation. Expression quantitative traits loci (eQTLs) to some extent explain the genomic effects of GWAS signals [19, 61, 64]. By leveraging effects of multiple cis-eQTL variants on gene expression, transcriptome-wide association studies (TWAS) search primarily for gene-trait associations. The approach builds on prediction models of gene expression from reference panels that correlate genotype patterns with transcript levels in tissues which are relevant for the phenotype. Prediction models are then used to impute tissue-specific gene expression based on genotypes with a given trait in individuals of GWAS cohorts [21]. Since TWAS signals reflect association between transcriptome-wide genetically regulated expression (GREx) and traits or diseases, the approach can be used to prioritize candidate genes across disease-relevant tissues. Thereby, TWAS may point to causal genes at risk loci identified by GWAS and

---

Ling Li and Zhifen Chen have contributed equally to this work.

Heribert Schunkert
schunkert@dhm.mhn.de

Extended author information available on the last page of the article.

---

115
Appendix II Transcriptome-wide association study of coronary artery disease identifies novel susceptibility genes

thus provide further insights into biological mechanisms [62, 70]. Moreover, TWAS increase the sensitivity to identify susceptibility genes missed by traditional GWAS analyses. Here we performed a TWAS to identify novel susceptibility genes for CAD comprising more than 80,000 individuals with genotype data along with validation and exploratory analyses for the associated genes.

Materials and methods

Prediction models of nine tissues based on two reference panels

The starting point of this investigation was two large human biobanks with individual-level data on genome-wide genotypes as well as mRNA expression levels in multiple tissues with relevance for CAD. These include atherosclerotic aortic wall (AOR), atherosclerotic-lesion-free internal mammary artery (MAM), liver (LIV), blood (BLD), subcutaneous fat (SF), visceral abdominal fat (VAF), and skeletal muscle (SKLM) in the Stockholm-Tartu Atherosclerosis Reverse Network Engineering Task (STARNET) [20], and AOR, LIV, BLD, SF, VAF, and SKLM in the Genotype-Tissue Expression (GTEx) [1] (Supplementary Table 1). Arterial wall coronary (COR) and tibial artery (TIB) datasets were only available in the GTEx. The pipeline used for training prediction models was EpixXcan which was built on the basis of PrediXcan but with improved prediction performance by integrating epigenomic annotation data into model-training process [21, 70]. The samples used for training models were restricted to European ancestry. We adopted the existing expression prediction models established by Zhang except COR and TIB tissues which were not covered yet [70].

We established predictive models for COR and TIB tissues using the same parameters as other tissues [70]. In brief, we first filtered the genotype and expression data of COR and TIB from GTEx v7 [1]. For genotype data, variants with call rate <0.95, minor allele frequency (MAF) <0.01, and Hardy Weinberg equilibrium (HWE) <1e-6 were removed. For expression, we used quality-controlled data and performed sample-level quantile normalization and gene-level inverse quantile normalization using preprocess codes of PredicDB pipeline [21]. We then calculated SNP priors using hierarchical Bayesian model (qtlBHM) [40] that jointly analyzed epigenome annotations of aorta derived from Roadmap Epigenomics Mapping Consortium (REMC) [5], and eQTL statistics. The SNP priors (Supplementary Table 2), genotype data and expression data were jointly applied to tenfold cross-validated weighted elasticnet to train prediction models [70].

Both STARNET- and GTEx-based models were filtered by cross-validated prediction $R^2 > 0.01$ [28, 68]. The summary statistics of sample sizes used for training models and the transcript numbers of genes covered by each predicting model are shown in Supplementary Table 1.

GWAS cohorts

For the discovery cohort, individual-level genotyping data were collected from ten CAD GWAS, a subset of CARDIoGRAMplusCAD, including the German Myocardial Infarction Family Studies (GerMIFS) I-VIII [16, 18, 38, 47, 48, 52, 56], Wellcome Trust Case Control Consortium (WTCCC) [7], LURIC [65], and Myocardial Infarction Genetics Consortium (MiGen) [2]. We used a part of individual-level data from UK Biobank (UKB) as the replication cohort [8], by extracting 20,310 CAD cases according to hospital episodes or death registries as reported and randomly selecting 25,000 non-CAD participants as controls. The detailed information about selection criteria of case and control were described at elsewhere [38]. In total, genotypes of 37,997 cases and 42,854 controls were included in our transcriptome-wide association studies (TWAS) of CAD (Supplementary Table 3). The preprocessing steps of genotyping data are as previously described [38].

TWAS analysis

The transcriptome-wide association analysis was performed using prediction models of nine tissues for imputing individual-level GReX from CAD cases and controls of 11 GWAS cohorts and by association of these tissue-specific GReX with CAD risk in each cohort. To test the replicability of TWAS results, we used ten GWAS cohorts as discovery set and UKB as the replication set to test replicability within STARNET- and GTEx-based models, respectively. We compared the consistency of TWAS results between STARNET- and GTEx-based models of the six overlapping tissues using all genotype data. Then, we merged TWAS genes resulted from two reference-based panels as the final list. Finally, we annotated the TWAS genes list by over 200 CAD loci identified by GWASs [17, 35] using MAGMA [37]. Gene resided in the ± 1 Mb regions around known GWAS loci were marked as the known, otherwise genes were marked as the novel.

Colocalization of the eQTL and GWAS signals

Colocalization analysis was performed using COLOC, a Bayesian statistical methodology that takes GWAS and eQTL data as inputs, and tests the posterior probabilities of hypothesis #4 (PP4) that there are shared casual variants for each locus [23]. The summary statistics of GWAS meta-analysis were obtained from CARDioGRAMplusCAD Consortium [47], and the eQTL data of nine tissues from
Appendix II Transcriptome-wide association study of coronary artery disease identifies novel susceptibility genes

STARNET [20] and GTEx [1]. The significance threshold is PP4 > 0.55.

Co-expression network for protein coding and IncRNA genes

We used RNA-seq data of STARNET [20] to calculate expression correlations between long non-coding RNA (IncRNA) genes and protein-coding genes in seven tissues. Co-expression pairs with absolute Pearson correlation coefficient larger than 0.4 were considered to be significant. The co-expression network was displayed by cytoscape [34].

Gene set enrichment analyses

Pathway enrichment analysis was carried out using ClueGO (<2.5.2) [6], a plugin of cytoscape [34], based on collated gene sets from public databases including Gene Ontology (GO) [26], KEGG [30], Reactome [12], and WikiPathways [35]. Gene sets with false discovery rate (FDR) by right-sided hypergeometric test less than 0.05 were considered to be significant.

Furthermore, we also studied the diseases or traits associated with risk genes by performing disease enrichment analysis based on DisGeNET [50], the publicly available datasets of genes and variants association of human diseases. FDR < 0.05 was used for thresholding.

Rare damaging variants association analysis

To investigate association of damaging variants in TWAS genes with CAD, we used whole-exome sequencing (WES) data of 200,632 participants from UKB [27]. The WES data were processed following the Functional Equivalence (FE) protocol. We performed quality control on the WES data by filtering variants with calling rate < 0.9 and variants with HWE < 1e−6. For the relevant traits, besides CAD, we considered i) three lifestyle factors including body mass index (BMI), diabetes, hypertension; ii) seven categories of blood lipids including low-density lipoproteins cholesterol (LDL-C), high-density lipoproteins cholesterol (HDL-C), apolipoprotein A (APOA), apolipoprotein B (APOB), Lipoprotein(a) (LPA), total cholesterol (TC) and triglycerides (TG); iii) four inflammation related factors including C-reactive protein (CRP), lymphocyte count (Lymphocyte), monocyte count (Monocyte) and neutrophil count (Neutrophil). In total, 15 traits were studied.

We defined damaging variants as (i) MAF < 0.01; (ii) annotated into following one of the three classes: loss-of-function (LoF) variants (stop-gained, splice site disrupting, or frameshift variants), pathogenic variants in ClinVar [36], or missense variants predicted to be damaging by one of five computer prediction algorithms (LRT score, MutationTaster, PolyPhen-2 HumDiv, PolyPhen-2 HumVar, and SIFT). The Ensembl Variant Effect Predictor (VEP) [45] and its plugin loftee [31], and annotation databases dbNSFP 4.1a [15] and ClinVar (GRCh38) [36] were used for annotating damaging mutations.

For each analysis, samples were classified into carriers or noncarriers of the gene’s damaging mutations. For binary traits, we used Fisher’s exact test to check if there was incidence difference of mutation carrying between case and controls. For the quantitative traits, we used linear regression model with adjustments of sex, first five principal components, and lipid medication status to investigate the associations between mutation carrying status and traits. We used nominal significance threshold (P < 0.05), given that coding variants are rather rare, and the case–control sample sizes were not balanced which might increase false negative rate.

Association of variants in novel genes with lipid traits

For 18 novel risk genes, we performed association analysis for variants located in respective loci (± 1 Mb) with lipid traits using genotyping data of UKB. The lipid traits include levels of LDL-C, HDL-C, APOA, APOB, LPA, TC, and TG. The variants were filtered by MAF > 0.01, and imputation info score > 0.4. The association test was performed using PLINK2 [10] with adjustment of sex, age, first five principal components, and lipid medication status. The numbers of independent SNPs were estimated using Genetic type 1 error calculator (GEC) tool [39].

Expression-traits association study using mouse data

The hybrid mouse diversity panel (HMDP) is a set of 105 well-characterized inbred mouse strains on a 50% C57BL/6J genetic background [42]. To specifically study atherosclerosis in the HMDP, transgene implementation of human APOE-Leiden and cholesteryl ester transfer protein was performed, promoting distinct atherosclerotic lesion formation [4]. A Western diet containing 1% cholesterol was fed for 16 weeks. Subsequently, gene expression was quantified in aorta and liver of these mice and lesion size was assessed in the proximal aorta using oil red O staining. Fourteen other related traits were measured too, including liver fibrosed area, body weight, TC, VLDL-C (very low-density lipoprotein cholesterol) + LDL-C, HDL-C, TGs, unesterified cholesterol, free fatty acid (FFA), IL-1b, IL-6, TNfa, Mcp-1, and M-csf. From HMDP, we extracted significant association pairs between TWAS genes and 15 risk traits by applying significance P < 0.05.
Appendix II Transcriptome-wide association study of coronary artery disease identifies novel susceptibility genes

Experimental validation of KPTN and RGS19 in human cells

To knock down KPTN and RGS19, two sgRNAs targeting shared exons of all transcript isoforms were delivered by lentivirus into a Cas9-expression huh7, a human hepatoma cell line. Exon 4 of KPTN and exon 5 of RGS19 were targeted by a dual CRISPR strategy to create a 40 bp and 130 bp frame shift deletion, respectively. sgRNAs were carried by Lent-Guide-Puro vector (addgene, #52963) and infected cells were treated with 10 μg/ml puromycin treatment for 3 days to eliminate the negative cell. Positive targetted cells were expanded in culture and passed for assays. Cells for measurement of secretive triglycerides, cholesterol, and APOB100 were cultured for 16 h in serum-free medium. Medium triglycerides and cholesterol were enriched for five times by vacuum centrifuge and measured with colorimetric kits, triglyceride (Kobas), and CHOL2 (Kobas), respectively. The amount of medium APOB100 was measured with an ELISA kit (MA-BTECH).

RNA isolation and sequencing

Total RNA from huh7 cells was isolated using RNAeasy Plus Mini Kit (Qiagen) (control cells, n = 3; knockout cells, n = 3). Quantity and quality of the isolated RNAs were measured by Fragment Analyzer (Agilent). RNA sequencing (RNA-seq) was performed by BGI TECH SOLUTIONS (HONGKONG) using strand specific library preparation with mRNA enrichment, paired-end sequencing with 100 bp read length on the DNBSEQ platform and 20 M clean read pairs per sample. Clean reads were mapped onto the GRCh38.p12. Expression quantifications, differential expression, and gene set enrichment were performed according to BGI RNA-seq pipeline.

Results

Transcriptome-wide significant genes for CAD

The study design is shown in Fig. 1. Expression prediction models of nine tissues were derived from two reference panels, STARNET [20] and GTEx [1], using EpiXcan pipeline [70] (Materials and methods). We applied these models to impute transcriptome-wide GReX of nine tissues from individual-level genotype data of 11 GWAS cohorts (Supplementary materials; Supplementary Figs. 1–2; Supplementary Tables 1–3) [2, 7, 8, 16, 18, 38, 47, 52, 56, 65]. We next associated the GReX with CAD risk in each cohort (Supplementary materials). The results revealed replicability of TWAS genes when taking ten CARDIoGRAMplusCAD cohorts as discovery and UKB as replication set within the STARNET- and GTEx-based prediction models, respectively (Supplementary Figs. 21–II; Supplementary Fig. 3). The results also showed consistency and complementarity of TWAS findings in six shared tissues between two reference-based prediction models (Supplementary Figs. 3–4; Supplementary Figs. 4–5). Therefore, we combined the results based on the two reference models for the final list of TWAS genes (Supplementary Fig. 2 V).

From STARNET-based models 129 gene-tissues pairs and from GTEx-based models 106 gene-tissue pairs were significantly associated with CAD (Bonferroni-corrected significance based on 12,995 genes, P < 3.85e−6). Since 42 pairs overlapped between the two panels (Supplementary Fig. 5), the total number of gene-tissue pairs was 193. Given that some genes displayed association in multiple tissues, the final list of significant TWAS genes for CAD was 114 genes (Fig. 2; Supplementary Fig. 6; Supplementary Table 4). Of these, 95 gene-tissue association pairs were confirmed using another commonly used fine-mapping tool (COLOC) [23] with posterior probabilities of shared causal variants in each locus larger than 0.55 (PP4 > 0.55; Materials and methods; Supplementary Table 5; Supplementary Fig. 7).

Of the 114 TWAS genes, 46 genes displayed genetically mediated differential expression in AOR, 28 in MAM, 25 in LIV, 23 in VAF, 22 in SKLM, 18 in SF, 16 in BLD, 10 in TIB, and 5 in COR (Supplementary Table 3; Supplementary Table 4). Most genes revealed significant associations in only a single tissue; 38 were significant in more than one, almost all having consistent directions of association between predicted expression levels and CAD across tissues (Supplementary Fig. 3b).

Among the 114 genes, 102 were protein-coding and 12 were IncRNA genes (Supplementary Table 4). The STAR- NET data showed that most lncRNAs were positively coexpressed with a surrounding gene in affected tissues (Supplementary Fig. 8). LINC00370 was the only exception, which displayed complex co-expression patterns with other genes.

Respective genes were found in 63 genomic regions, thus several regions represented multiple genes with significant associations. Six regions had multiple TWAS genes with shared GWAS and eQTL signals in respective tissues, like 1p13.3 and 2p33.2 (Supplementary Figs. 9–10; Supplementary Table 5). On the other hand, in 39 regions expression of only a single gene was found to be significantly associated, which makes these genes likely candidates for mediating causal effects, particularly, if these genes reside within GWAS risk loci for CAD (these genes are indicated in Supplementary Table 6).

Most TWAS genes (n = 96) could be positionally annotated to the 1 Mb region around one of the over 200 GWAS loci that are currently known to be genome-wide significantly associated with CAD [11, 17, 35]. Therefore, we marked these as known genes (Supplementary
Appendix II Transcriptome-wide association study of coronary artery disease identifies novel susceptibility genes

Step 1: Training expression prediction models
- STARNet: genotype and expression data of 7 tissues
- GTEx: genotype and expression data of 8 tissues

Prediction models of 9 tissues

Step 2: Imputing genetically regulated expression from 11 GWAS cohorts
- CARDiosGRAMplusCAD: 10 studies (17,687 CAD and 17,854 controls)
- UKB: (20,310 CAD and 25,000 controls)

Step 3: Transcriptome-wide association analysis
- 114 TWAS genes associated with CAD (96 known, 18 novel)

Step 4: Plausibility, biological function and pathogenicity of TWAS genes

Colocalization analysis
- GWAS: CARDiosGRAMplusCAD Consortium
- eQTL: STARNet and GTEx
- the posterior probabilities (PPH) > 0.55

Associations between damaging variants in TWAS genes and CAD and its risk traits (UKB)

Gene set enrichment: • Gene-disease associations (DisGeNET)
• Pathways

Step 5: Susceptibility of novel genes
- Variant-trait associations in human genotype data (UKB)
- Expression-trait associations in atherosclerosis mouse model (HMDP)
- Gene knockouts of KPTN and RGS19 in huc7 hepatocytes

Fig. 1 The study design. Step 1, we trained prediction models using EpiXcan from two eQTL panels, the Stockholm-Tartu Atherosclerosis Reverse Network Engineering Task (STARNet) and the Genotype-Tissue Expression (GTEx) for nine tissues. Step 2, the prediction models were applied to impute genetically regulated expression (GRcX) from individual-level genotype data of ten CARDiosGRAMplusCAD sets and UK Biobank (UKB). Step 3, we associated transcriptome-wide GRcX with risk of coronary artery disease (CAD) (Supplementary Results) and identified 114 transcriptome-wide significant genes (TWAS genes). Of these, 96 resided within genomewide significant (GWAS) loci and 18 outside of known GWAS loci (novel genes). Step 4, we tested the plausibility of novel TWAS genes by conducting colocalization analysis and studying effects of damaging mutations, as well as gene set enrichment analyses. Step 5, we explored potential mechanisms of novel genes by testing association with risk traits of CAD in human genotype data of UKB, and association between expressions and risk traits of CAD in atherosclerosis mouse models from the Hybrid Mouse Diversity Panel (HMDP). Lastly, we carried out CRISPR/Cas9-based knockdown experiment for two novel genes RGS19 and KPTN in human hepatocyte cell lines to experimentally validate related functions.

Table 6). On the other hand, 18 genes resided outside of these regions and were labeled as novel genes (Table 1). Most novel genes were tissue-specific, except RGS19, FAM114A1 and UFL1 which displayed evidence for differential expression in multiple tissues.

Pathways and diseases enriched by TWAS genes

We carried out two types of gene set enrichment tests to further study the biological relevance of genes giving signals in this TWAS. First, we studied disease-gene sets from the DisGeNET platform which is one of the largest publicly
Appendix II Transcriptome-wide association study of coronary artery disease identifies novel susceptibility genes

Fig. 2 Manhattan plot of CAD TWAS results. The association results from STARNET- and GTEs-based models were integrated by lowest $P$ values. The blue line marks $P=3.85e-6$, i.e. transcriptome-wide significance. Each point corresponds to an association test between gene-tissue pair. 18 novel TWAS genes were highlighted. Supplementary Fig. 6 identifies all genes identified by their genetically-modulated association signals. The color code identifies the tissue in which the genes were differentially expressed by genetic means: AOR aorta, COR coronary artery, MAM mammary artery, BLD blood, LIV liver, SF subcutaneous fat, VAF visceral abdominal fat, SKLM skeletal muscle

Fig. 3 Tissue distribution of 114 TWAS genes of CAD. a Number of transcriptome-wide significant genes across tissues. b Heatmap plot of 38 genes identified in more than one tissues. The color codes indicate direction of effects. Cells marked with * represent significant gene-tissue pairs ($P<3.85e-6$). AOR aorta, COR coronary artery, MAM mammary artery, BLD blood, LIV liver, SF subcutaneous fat, VAF visceral abdominal fat, SKLM skeletal muscle, TIB tibial artery

available collections of genes and variants associated with human diseases [50]. The results showed that genes discovered by TWAS were primarily enriched for CAD, coronary atherosclerosis, and hypercholesterolemia (Supplementary Table 7), adding to the plausibility of our TWAS findings.
In line with these results, gene set enrichment analysis based on GO [26], KEGG [30], Reactome [12], and WikiPathways [55] databases showed that the TWAS genes were highly enriched for pathways involved in cholesterol metabolism and regulation of lipoprotein levels. To a lesser extent, risk genes were enriched in regulation of blood pressure as well as development and morphogenesis of the heart and the aortic valve (Supplementary Table 8).

**Effects of damaging variants in TWAS genes**

We next searched in exome-sequencing data of 200,643 participants from UKB for rare damaging variants in TWAS genes (either loss-of-function mutations or mutations predicted to be adverse by one of five in-silico methods, allele frequency <0.01) (Materials and methods). In 97 genes we detected such variants. Expectedly these damaging mutations were very rare which limits the power of gene-based burden tests to observe association with risk of CAD or one of 14 CAD-related cardiometabolic traits we tested (15 traits in total). Nevertheless, associations of eight genes with risk traits reached Bonferroni-corrected significance ($P < 3.44e-5; 0.05/97 genes × 15 traits$) (Fig. 4; Supplementary Tables 9–10). Mutations of lipoprotein lipase (LPL), a critical regulator of lipid metabolism [29, 60], were evidently associated with lipid traits, including levels of HDL-C (beta = −0.106; $P = 4.54e−68$), APOA1 (beta = −0.062; $P = 6.25e−47$), APOB (beta = 0.025; $P = 1.38e−12$), and TG (beta = 0.241; $P = 1.47e−68$). ABCG5, encoding a sterol transfer protein [69], was associated with LDL-C (beta = 0.12; $P = 3.66e−10$), TC (beta = 0.16; $P = 8.63e−10$). PCSK9, a drug target for cholesterol lowering [13], was associated with LDL-C (beta = −0.01; $P = 4.29e−7$) and APOB (beta = −0.03; $P = 4.4e−10$). A mutation of SARS was associated with APOB (beta = −0.02; $P = 5.92e−7$), MAT2A with lymphocyte counts (beta = 1.34; $P = 3.41e−28$), and JCAD (odds ratio [OR] = 1.31; 95% confidence interval [CI] 1.18–1.46; $P = 5.77e−7$) as well as ARHGAP42 (OR = 2.08; 95% CI 1.65–2.59; $P = 2.22e−9$) were associated with risk of diabetes. We also observed nominally
Appendix II Transcriptome-wide association study of coronary artery disease identifies novel susceptibility genes

Fig 4 Effects of damaging variants in TWAS genes on CAD and its risk traits. Sign(beta)**−log10(P) displays direction and significance of gene-trait associations. When the Sign(beta)**−log10(P)>8, they were trimmed to 8. The gene-trait association pairs reached Bonferroni-significance P<3.44e−5 were highlighted in box. CAD coronary artery disease, LDL-C low-density lipoproteins cholesterol, VLDL-C very low-density lipoprotein cholesterol, HDL-C high-density lipoprotein cholesterol, APOA apolipoprotein A, APOB apolipoprotein B, TC total cholesterol, TG triglycerides, CRP C-reactive protein, BMI body mass index

Fig 5 Novel risk genes were associated with lipid traits. a Data from UK Biobank (UKB) indicated that lead variants inside the boundary of risk genes were associated with lipid traits with Bonferroni-corrected significance levels (**P<8.09e−6), or by genome-wide significance (**P<5e−8). b Expression levels of novel genes were likewise associated with lipid traits and aortic lesion area in an atherosclerosis mouse model from the hybrid mouse diversity panel (HMDP). *P<0.05; **FDR <0.05. LDL-C low-density lipoproteins cholesterol, VLDL-C very low-density lipoprotein cholesterol, HDL-C high-density lipoprotein cholesterol, APOA apolipoprotein A, APOB apolipoprotein B, TC total cholesterol, TG triglycerides, FFA free fatty acid

Novel genes associated with risk factors in human and mouse data

We next associated single nucleotide polymorphisms (SNPs) in the regions of ±1 Mb around the 18 novel TWAS genes to study their associations with a series of lipid traits including LDL-C, HDL-C, APOA, APOB, LPA, TC, and TG in UKB (Materials and methods). There were 883 independent SNPs estimated by GIEC. Bonferroni-corrected significance P<8.09e−6 (0.05/883 × 7 lipid traits) was observed for numerous respective lead variants, of which RGS19, SDC2, CAG3, EZR, HOMER3, and WPP2 reached genome-wide significant association (P<5e−8) with multiple lipid traits (Fig. 5a; Supplementary Table 11).

Next, we extracted expression-trait association statistics of TWAS genes from HMDP, which brings together genotypes and expression data from atherosclerosis mouse models [42]. Based on the expression data from mouse aorta and liver tissues, 55 TWAS genes were significantly associated with aortic lesion area and 14 further cardiovascular traits (P<0.05; Supplementary Table 12). Expression levels of seven novel genes, i.e. Rgs19, Ktn1, Ecr, Stre4a, Cnd1, Focal, and Wasf1, were associated with aortic lesion...
Appendix II Transcriptome-wide association study of coronary artery disease identifies novel susceptibility genes

Fig. 6 Targeting of KPTN and RGS19 reduced lipids and APOB secretion of human liver cells. a Two sgRNAs were used to target the exon4 of KPTN (shared exon among isoforms) in a Cas9-expressing huh7 liver cell line. The dual CRISPR strategy created a 40 bp frame shift deletion in the gene and profound reduction of KPTN at both mRNA and protein levels (Supplementary Fig. 11c, d). The primers (P-Fw and P-Rv) used for analyzing the CRISPR editing as indicated. b The same strategy was used for RGS19 targeting, which resulted in a 130 bp frame shift deletion in the gene, and reduction of mRNA and protein (Supplementary Fig. 11c, d). c Reduced triglyceride and cholesterol levels in knockout (KO) cell lines were detected by colorimetric method and APOB100 secretion was measured by human APOB100 Elisa (n = 6). Triglyceride, cholesterol, and APOB100 levels were normalized to total protein and compared between the KO and control (CTR) cell lines.

Area (Fig. 5b), a commonly used measure for atherosclerotic plaque formation in mice. Additionally, we found the novel genes were associated with at least one lipid trait in the mouse model.

Knockdown of RGS19 and KPTN in human liver cells

Potential functional implications of all novel genes, based on the literature, are summarized in Supplementary Table 13. We additionally aimed to validate two exemplary novel TWAS genes by in vitro studies. Based on the above in-silico annotations we focused these studies on novel genes identified in liver with potential effects on lipids, the top risk factor for CAD (Fig. 5). Among the five genes identified in liver including NR2F4, LZR, GASS, KPTN, and RGS19, the last two were, not only the least studied but also associated with nearly a full spectrum of lipid traits in human or mouse data (Fig. 5). In addition, both KPTN and RGS19 are indeed expressed in hepatocyte (Supplementary Fig. 11a, b). Finally, both KPTN and RGS19 are located within lipid loci identified in more than one million individuals [24]. Therefore, we decided to test the influence of KPTN and RGS19 on lipid metabolism of liver cells.

We generated gene knockout (KO) huh7 cell lines by a dual CRISPR strategy (Materials and methods), which substantially reduced expression of the respective genes (Supplementary Fig. 11c, d). We measured secretion levels of TG, cholesterol and APOB in gene-targeted versus control cells. Notably, under normal circumstances, human hepatocytes synthesize cholesterol, assemble TG and APOB100, and secrete these particles in form of VLDL-C [58]. Compared to control huh7 cells, we found reduced APOB and cholesterol levels in culture medium of KPTN-KO cells (Fig. 6a, c). In culture medium of RGS19-KO cells we also detected reduced levels of APOB100, cholesterol, and TG (Fig. 6b, c), in line with strong associations of this gene with an array of lipid traits in both human genotyping and mouse expression data sets (Fig. 5).

We further corroborated our experimental results by performing RNA sequencing (RNA-seq) on KPTN-KO and
Appendix II Transcriptome-wide association study of coronary artery disease identifies novel susceptibility genes

RGS19-KO hepatocytes. In comparison to control cells, dysregulated genes in KPTN-KO and RGS19-KO hepatocytes (P < 0.05; Supplementary Tables 14–15) were indeed enriched for lipid metabolism (Supplementary Fig. 12). For KPTN-KO hepatocytes, the top four significantly enriched pathways plausibly contribute to CAD risk. Pathways ranked 1 and 3, ‘regulation of cholesterol esterification’ and ‘LDL particle remodeling’, strongly suggested that KPTN can affect CAD risk via cholesterol metabolism (Supplementary Fig. 12a, b). For RGS19-KO hepatocytes, the dysregulated genes were enriched for both cholesterol and triglycerides metabolisms (Supplementary Fig. 12c, d) and eight of the top ten significantly enriched pathways were related to lipid metabolism, consistent with the reduced secretion of cholesterol and triglyceride of RGS19-KO cells (Fig. 6c).

Discussion

In a stepwise approach, we first generated models which allow to predict gene expression based on genotypes in nine tissues. Next, we applied these models to individual-level genotype data on more than 80,000 CAD cases and controls to perform a transcriptome-wide association analysis. We identified 114 genes with differential expression by genetic means in CAD patients. Many signals were highly plausible as they resided within loci displaying genome-wide significant association with CAD by traditional GWAS. By in-silico analyses, these genes were markedly enriched in established pathways for the disease. Moreover, damaging variants in these genes showed association with CAD risk or its underlying traits in whole exome sequencing data from UKB. Importantly, we also identified 16 genes without prior evidence for their involvement in CAD by GWAS, many of which were found to be associated with lipid metabolism in human and mouse data.

Only a minority of genes residing within published CAD GWAS loci have been validated experimentally for their underlying causal role in atherosclerosis. Our data provide a substantial step towards prioritization of genes at respective GWAS loci [17, 35], because the TWAS association finding is based on expression levels of specific genes in defined tissues. In this respect, 46 genes identified by this TWAS are known for effects in pathophysiological pathways related to CAD, including lipid metabolism, inflammation, angiogenesis, transcriptional regulation, cell proliferation, NO signaling, and high blood pressure, to name a few (Supplementary Table 6), giving credility to the association findings.

Interestingly, our TWAS recovered eight novel gene-CAD associations in fat tissue, including MGP and WASP1 in SF, and CAND1, FAM114A1, FOCA1, RGS59, TSPAN11, and TXNRD2 in VAT, representing half of the novel genes. All these genes also showed significant association with multiple lipid traits in a mouse atherosclerosis model (Fig. 5b). Given many CAD patients that are overweight or obese, it will be of great interest to identify how these genes modify cardiometabolic traits leading to cardiovascular disorders. In this respect our TWAS could provide a list of candidate genes and related targetable cardiometabolic traits. In addition, it is of surprise to unveil 22 genes linking SKLM to CAD risk, and eight were unique to this tissue, including HOMER3, SDCCAG3, MTAP, NME9, PSMA4, SLC2A12, UNCI198R, and VAMP5, the first two being novel. SDCCAG3 or ENTR1 encodes endosome associated trafficking regulator 1 and involves in recycling of GLUT1 (glucose transporter type 1), supplying the major energy source for muscle contraction. SKLM-based metabolism may have a protective role in CAD as suggested by the many cardioprotective effects of sports [44, 54]. Gene targets enhancing SKLM function in this respect might be effective in CAD prevention, a field relatively unexplored thus far. Here, for the first time, quantitative traits regulated genes in SKLM were associated with CAD by TWAS, providing novel evidence for genes that could modulate CAD risk by their functions in SKLM.

Many novel TWAS genes revealed association with lipid traits in both genotype-trait data of human biobank and expression-trait data of atherosclerosis mouse model. For example, KPTN and RGS19, both novel genes displaying significant TWAS results for CAD—based on their genetically-modulated expression profiles of liver tissue—also showed significant association with various lipid traits as well as aortic lesion area in the atherosclerosis mouse model. Moreover, both gene loci harbor SNPs which are significantly associated with several lipids including LDL-C, HDL-C, TC, and/or TG in human genotype data. Based on these observations, we functionally validated the roles of these two novel genes by studying lipid levels in human liver cells, i.e. the tissue that displayed evidence for differential expression by TWAS. Indeed, we observed that knockout of the two genes lowered secretion of APOB and lipids into culture medium. KPTN, kaprin (actin binding protein), a member of the KPTN, IFFG2, C12orf66, and SMT2 (KiSTOR) protein complex, is a lysosome-associated negative regulator of the mechanistic target of rapamycin complex 1 (mTORC1) signaling [67]. By investigating dysregulated genes of KPTN-KO hepatocytes, we found many genes of mTORC1 pathway to be upregulated (Supplementary Fig. 12b), including a subunit of mTORC1, namely, MLST8 (MTOR associated protein, LST8 Homolog). Interestingly, many lysosome genes were also significantly upregulated including LAMP1 (lysosomal associated membrane protein 1), ACP2 (acid phosphatase 2, lysosomal), AP1B1 (adapter related protein complex 1 subunit beta 1), ATP5H0C (ATPase H+ transporting V0 subunit c), CTNS (cystinosin, lysosomal cystine transporter), CTSA
Appendix II Transcriptome-wide association study of coronary artery disease identifies novel susceptibility genes

Basic Research in Cardiology (2022) 117:6

(cathespins A), and CLTB (clathrin light chain B). Lysosomes promote lipid catabolism and transport, and maintain cellular lipid homeostasis [57]. Activated mTORC1 located on lysosome membrane [67], acts as a sensor of lysosomal lipids [57], such as cholesterol and phosphatidic acid, which exert as building block for cellular and subcellular membrane system. In fact, cholesterol mediates mTORC1 activation at the lysosome [9]. The interaction of mTORC1 and lysosome may promote lipid-sensing and lipid-trafficking to support the function of other subcellular organelles [46, 57]. These results hint the enhanced cellular usage of cholesterol via mTOR lysosome axis in KPTN-KO hepatocytes. In addition, several lipoprotein genes were downregulated as well, including APOA1, APOA2, APOA4, and APOB. Both processes might contribute to the reduced cholesterol secretion and the association with CAD.

RGS19 belongs to the RGS (regulators of G-protein signaling) family, who are regulators for G-protein-coupled receptors (GPCRs) [49]. RGS19 inhibits GPCR signal transduction by increasing the GTPase activity of G-protein alpha subunits, thereby transforming them into an inactive GDP-bound form [53, 59]. The targeting GPCR of RGS19 has not been observed before, and how RGS19 regulates lipid metabolism remains unclear. The RGS19 locus was first reported to be associated with TC and TG in 2017 [33, 41]. We observed significant association of this gene with CAD and functionally validated its role in TC and cholesterol secretion. A potential mechanism could be related to PPARalpha pathway that regulates the expression of genes involving hepatic lipogenesis and lipid storage [63, 66]. PPARalpha also regulates cholesterol, bile acid homeostasis, and sphingolipid metabolism in the liver [22]. Many genes in PPARalpha pathway were significantly downregulated in RGS19-KO hepatocytes, including FABP1 (also known as liver fatty acid binding protein), PLTP (phospholipid transfer protein), APOA1, APOA4, and APOC3 (Supplementary Fig. 12d). RGS19 is a regulator for G-protein-coupled receptors (GPCR). Interestingly, we found six dysregulated GPCRs in RGS19-KO hepatocytes, including ADGR2, CELSR1, ADGRV1, OXER1, LGR5, and LGR4 (Supplementary Fig. 12d). Furthermore, one of them, OXER1, an activator PPARalpha [51], was also downregulated in RGS19-KO cells. All in all, one hypothesis could be that RGS19 associated GPCR signaling affects the PPARalpha pathway, and thereby lipid metabolism and CAD risk. Previous and current studies concordantly suggest from different angles that RGS19 has a role in lipid metabolism and our data further indicate that this function might mediate its effects on CAD risk.

There are certain limitations in our study. First, we observed that about 15% of gene-tissue pairs displayed some degree of heterogeneity in the association findings with CAD risk across the cohorts (Supplementary Table 4). While this number is relatively low and likely result from a play of chance when association findings are being compared across individuals with relatively small case-control samples, it might also indicate some degree of population specific effects within European ancestries from UK, Germany, France, and Italy. Second, since TWAS are strongly dependent on the reference panel linking genetic signatures with gene expression, it had to be expected that STARNET- and GTeX-based predictive models display some differences in gene-CAD associations. STARNET-based TWAS identified 129 gene-tissue pairs, whereas GTeX-based TWAS identified 106 gene-tissue pairs. Yet, 42 gene-tissue pairs were shared between the two analyses, and effect sizes for the shared genes were highly concordant (r = 0.97). An average of 62% overlapping genes was observed in the matched tissues of two reference-based models, and the resulting size of expression-CAD associations was generally consistent with an average r = 0.72. The relatively small differences may be due to different sample sizes used for training predictive models [70], different disease states (subjects with and without CAD), intravital (STARNET) or post mortem (GTeX) sample collection, as well as different transcript abundance and genotype coverage leading to differences in expression associated SNPs in our reference panels [20, 25]. Given a fair consistency between the two data sources, we combined results derived from both panels to increase the power for capturing risk genes. Third, although TWAS facilitates candidate risk gene prioritization, co-regulation or co-expression in cis at a given locus limits the precise determination of the culprit gene [62]. Indeed, at 12 loci we observed signals for three or more TWAS genes. For instance, in LIV tissue TWAS identified five genes at 1p13.3, ATXN7L2, CELSR2, PSMA5, PSRC1, SARS, and SORT1 which were co-regulated by same risk variant set, confusing prioritization of the causal gene. While CELSR2, PSRC1, and SORT1 were previously shown to act on lipid metabolism [3], we found that damaging mutations in SARS were also associated with serum levels of HDL-C and APOA. Thus, a combined effects of some or all genes at this locus may contribute to the association signal. In addition, all lncRNA genes identified by our study displayed co-expression with their neighboring coding genes, which makes it difficult to determine their causal effects. Nevertheless, in combining TWAS data with other genetic analyses, e.g. effects of damaging mutations, genetic association with other phenotypes and expression-traits association statistics, we aimed to improve risk gene prioritization, and to provide deeper insights of possible disease-causing mechanisms. For instance, LPL is well-known for its protective role against CAD by lowering lipids [29, 60], and our analyses showed that damaging LPL mutations were associated with higher lipid levels. Last, as with all statistical methods, there are certain limitations and assumptions associated with TWAS. Further evolution and improvement of these methods, as well as functional
validation experiments, will assuredly improve the accuracy of these studies.

In summary, our TWAS study based on two genetics-of-
gene-expression panels identified 114 gene expression-CAD
associations, of which 18 were novel. The extended analyses
with multiple datasets supported the reliability of the CAD
TWAS signals in prioritizing candidate risk genes and iden-
tifying novel associations in a tissue-specific manner. Func-
tional validation of two novel genes, RGS19 and KPTN, lend
support to our TWAS findings and provide strong evidence
for their role in lipid metabolism. Thus, our study created
a set of gene-centered and tissue-annotated associations for
CAD, providing insightful guidance for further biological
investigation and therapeutic development.

Supplementary Information The online version contains supplemen-
tary material available at https://doi.org/10.1007/s00395-022-00917-8.

Acknowledgements We’d like to thank Dr. Wen Zhang, Pro. Panos Rousou
for sharing the prediction models of seven tissues trained from
either STARNE or GTEx panels, as well as the EpiXcan pipeline.

Author contributions HS, LL, ZC designed the study and wrote the
manuscript. LL ran analyses. ZC, SL and AS performed experiment.
MS, GU, SK, AM, KH, CP, AEL, SF, TK, RE, KS, AR, JG, JG, JCK,
JLMB provided research data, technical support and gave conceptual
advice.

Funding Open Access funding enabled and organized by Projekt
DEAL. The work was funded by the German Federal Ministry of Edu-
ation and Research (BMBF) within the framework of ERA-NET on
Cardiovascular Disease (Drugable-Mi-genes: 01KI1802), within the
scheme of target validation (BlockCAD: 16GW0198K), COM-
MITMENT (01ZX1904A), and within the scope of eMed research
and funding concept (AcCDNet: 01ZX1706C). As a Co-applicant
of the British Heart Foundation (BHF)German Centre of Cardio-
vascular Research (DZHK)-collaboration (DZHK-BHF: 81X260552)
and the Ludwig Foundation for Cardiovascular Research (PiafOn-
tics: 18VC002), we gratefully acknowledge their funding. Additional
support has been received from the German Research Foundation
(DFG) as part of the Sonderforschungsbereich SFB 1123 (B02)
and the Sonderforschungsbereich SFB TRR 267 (B05). Further, we
kindly acknowledge the support of the Bavarian State Ministry of Health
and Care who funded this work with DigItalBayern (grant No: DMB-
1805-0011) within its Masterplan “Bayern Digital II” and of the
German Federal Ministry of Economics and Energy in its scheme of
ModiMax (grant No: ZFZ5902018A8). Finally, we acknowledge the
support of DZHK with Postdoc Start-up Grant (“Förderkernzeiten”).

 Declarations
Conflict of interest The authors have no competing interests to declare
that are relevant to the content of this article.

Tools and data EpiXcan pipeline: https://bitbucket.org/roussoslab/
epixcan/src/master/ and predictive models based on STARNE and
GTEx databases: http://predictdb.org/

PredXcan pipeline: https://github.com/hukayim/PredXcan
qIBTM: https://github.com/qianjiaqi/qIBTM

STARNE database: https://www.ncbi.nlm.nih.gov/geo/projects/g phase/cla-
GTEx database: https://www.ncbi.nlm.nih.gov/geo/projects/g phase/gt-
UK Biobank: https://www.ukbiobank.ac.uk/. Project ID: 25214.
MAGMA: https://ncgc.ecr4allsoftware/bmagma.

R package for colocalization analysis, color: https://cran.r-project.org/
web/packages/color/cogenerators/snpcharts.html

DiGenet: https://www.digenet.org/
CARDIoGRAMplusC4D Consortium: http://www.cardiogramplusc4d.org/

Open Access This article is licensed under a Creative Commons Attri-
bution 4.0 International License, which permits use, sharing, adapta-
tion, distribution and reproduction in any medium or format, as long
as you give appropriate credit to the original author(s) and the source,
provide a link to the Creative Commons Licence, and indicate if changes
were made. The images or other third party material in this article are
your article’s Creative Commons Licence, unless indicated otherwise in
a credit line to the material. If material is not included in the
article’s Creative Commons Licence and your intended use is not
permitted by statutory regulation or exceeds the permitted use, you
will need to obtain permission directly from the copyright holder. To view
a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References
1. AguE F, Brown AA, Castel SE, Davis JR, He Y, Jo B, Moham-
madi P, Park VS, Parsana P, Segré AV, Strober BJ, Zappala Z,
Cummings BB, Gelfand ET, Hudley K, Huang KH, Lok M, Li
X, Nedzel JL, Nguyen DY, Noble MS, Sullivan TJ, Tukiainen T,
AR, Lockhart NC, Moore HM, Rao A, Struwing JP, Volpi S,
Brigham LE, Haug R, Hunter M, Johns C, Johnson M, Kopen G,
Leineweber WF, Lomsdale JT, McDonald A, Mestelchil B, Mu-
ery K, Roe B, Salvatore M, Shah S, Thomas JA, Walten G, Wash-
ton M, Wheeler J, Bridge J, Foster IA, Gillard BM, Karasik E,
Kumar R, Mikhail M, Moser MT, Jewell SD, Montog KR, Rober-
ter DC, Valley D, Mars DC, Davis DA, Sobin L, Barus ME, Brant-
on PA, Abell NS, Ballin B, Dalaman O, Freard L, Gamazon ER,
Garzillo-Martin D, Gewirtz ADH, Gliner G, Glodanski MJ, Han
B, He AZ, Homaouldian F, Li X, Liu B, Kang EY, McDowell IC,
Onges H, Palosvich JJ, Peterson CR, Quan G, Ripsk S, Saha A,
Shabalin AA, Shimko T, Sal JH, Tenan TA, Tsang EK, Zhang H,
Zhou YH, Brossande C, Con NJ, Giuppi R, Kellin M, McCarthy
MF, Conrad DE, Eskin E, Li G, Nobel AB, Salanti C, Stronger BE,
Wen X, Wright FA, Arttde LG, Dermitzakis ET, Luppalainen T,
Bashir R, Bacha O, Boga OK, Chen LS, Chiang D, Damani FN,
Fereira PG, Hall IM, Housal C, Im HK, Kin Y, KinellMnth
S, Mungol S, Mondej J, Malte-Aguire M, Mabuto AW, Nicco-
lale DL, Oliva M, Panosx N, Papasaikas P, Payne A, Quan J,
S, Van De Bunt M, Wang G, Xi HS, Yegor-Lotem E, Zang B,
Akem JY, Bates D, Chan J, Claessens M, Demani K, Diegel
M, Doherty JA, Feinberg AP, Fernando MS, Halow J, Hansen
KD, Haugen E, Hickey PE, Hou L, Jaimine J, Jin R, Jiang L,
Johnson A, Kaul R, Kihibya MG, Lee K, Li JB, Li Q, Lin J, Lin
S, Linder S, Linke C, Liu Y, Mauvano MT, Molinie B, Nelson J,
Neri EJ, Park Y, Pierce BL, Rinsland NJ, Rizzard LE, Sandstrom

Springer
Appendix II Transcription-wide association study of coronary artery disease identifies novel susceptibility genes
Appendix II Transcriptome-wide association study of coronary artery disease identifies novel susceptibility genes
Appendix II Transcriptome-wide association study of coronary artery disease identifies novel susceptibility genes
Appendix II Transcriptome-wide association study of coronary artery disease identifies novel susceptibility genes

Appendix II Transcriptome-wide association study of coronary artery disease identifies novel susceptibility genes


Appendix II Transcriptome-wide association study of coronary artery disease identifies novel susceptibility genes


Appendix II Transcriptome-wide association study of coronary artery disease identifies novel susceptibility genes


Appendix II Transcriptome-wide association study of coronary artery disease identifies novel susceptibility genes


Authors and Affiliations

Ling Li1,2,3, Zhifen Chen1,3, Moritz von Scheidt1,3, Shuangyue Li1,3, Andrea Steiner1,3, Ulrich Gülßener1,3, Simon Koplev4, Angela Ma4, Ke Hao4, Calvin Pan5, Aldons J. Lušis5,6,7, Shichao Pang1,3, Thorsten Kessler1,3,7, Ralli Ermel8, Katayani Sukhavasi8, Arno Ruusalepp8,9, Julien Gagneur7, Jeanette Erdmann10,11, Jason C. Kovacic12,13,14, Johan L. M. Björkegren4,9,15, Heribert Schunkert1,3

1 Department of Cardiology, German Heart Center Munich, Technical University Munich, Lazarettstraße 36, 80636 Munich, Germany
2 Fakultät für Informatik, Technische Universität München, Munich, Germany
3 Deutsches Zentrum für Herz- und Kreislaufforschung (DZHK), Partner Site Munich Heart Alliance, Munich, Germany
4 Department of Genetics and Genomic Sciences, Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY 10029-6574, USA
5 Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA, USA
6 Department of Medicine, David Geffen School of Medicine, University of California, Los Angeles, CA, USA
7 Department of Microbiology, Immunology and Molecular Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA, USA
8 Department of Cardiac Surgery, The Heart Clinic, Tartu University Hospital, Tartu, Estonia
9 Clinical Gene Networks AB, Stockholm, Sweden
10 DZHK (German Research Centre for Cardiovascular Research), Partner Site Hamburg/Lübeck/Kiel, Lübeck, Germany
11 Institute for Cardiogenetics, University of Lübeck, Lübeck, Germany
12 Victor Chang Cardiac Research Institute, Darlinghurst, Australia
13 St Vincent's Clinical School, University of New South Wales, Sydney, Australia
14 Icahn School of Medicine at Mount Sinai, Cardiovascular Research Institute, New York, NY 10029-6574, USA
15 Department of Medicine, Huddinge, Karolinska Institutet, Karolinska Universitetssjukhuset, Stockholm, Sweden

Springer

134