Horizon 2020
WORK PROGRAMME 2018 – 2020
H2020-SC1-2019-Two-Stage-RTD

Project Acronym: **TO_AITION**

Grant Agreement N°: 848146
Project Full Title: **A high-dimensional approach for unwinding immune-metabolic causes of cardiovascular disease-depression multimorbidities**

Starting Date: 01/01/2020

Duration in months: 60

D3.1 Tissue pleiotropic effects of depression and cardiovascular disease

Nature:	Report
Dissemination Level:	Public
Contractual Date of Delivery to the EC:	30 June 2023
Actual Date of Delivery to the EC:	30 July 2023
WP number and Title:	WP3 Identification of common mechanisms, pathways and molecules driving co/multimorbidity through causal inference modelling
Lead Beneficiary:	UMCU
Version 0.9:	Authored by dr. Sander W. van der Laan and Lennart Landsmeer on behalf of the UMCU team on [2023-05-23] – Initial draft
Version 1.0:	Reviewed by dr. Sander W. van der Laan, UMCU on 28/07/2023
Version 2.0:	Reviewed by the coordinator on 29/07/23
Final Version:	Final version submitted

List of Beneficiaries

In bold the lead beneficiary.

Ben. No	Beneficiary name	Short name	Country
1	IDRYMA IATROVIOLOGIKON EREUNON AKADEMIAS ATHINON	BRFAA	GREECE
2	UNIVERSITEIT VAN AMSTERDAM	UVA	NETHERLANDS
3	PIRKANMAA HOSPITAL DISTRICT	TAUH	FINLAND
4	UNIVERSITATSKLINIKUM BONN	UKB	GERMANY
5	PANEPISTIMIO IOANNINON	UOI	GREECE
6	MICRONIT HOLDING BV	MICRONIT	NETHERLANDS
7	GENOWAY	GENOWAY	FRANCE
8	RUPRECHT-KARLS-UNIVERSITAET HEIDELBERG	UHEI	GERMANY
9	STICHTING VUMC	VUMC	NETHERLANDS
10	UNIVERSYTET MEDYCZNY W LODZI	LODZ	POLAND
11	**UNIVERSITAIR MEDISCH CENTRUM UTRECHT**	**UMCU**	**NETHERLANDS**
12	UNIVERSITE DE GENEVE	UNIGE	SWITZERLAND
13	EXELIXIS RESEARCH MANAGEMENT & COMMUNICATION	EXELIXIS	GREECE
14	SOCIETE EUROPEENNE DE CARDIOLOGIE	ESC	FRANCE

Executive Summary

Identification of key driving variants and genes in a large range of cell types and tissues from GTEx and other similar expression quantitative trait loci (eQTL) datasets. While we hypothesize that an inflammatory component connects depression and cardiovascular disease, the exact biological mechanism and site is unknown. In this deliverable we executed molecular (expression and epigenomic) quantitative trait loci (molecular QTL, molQTL) analyses in cardiovascular disease patients in Athero-Express Biobank Study (AE) and Tampere Vascular Study (TVS).
The QTLToolKit pipeline (github.com/swvanderlaan/QTLToolKit) was adapted and employed for cis-acting and trans-acting molQTL analysis; this state-of-the-art pipeline is based on QTLTools[5] and TensorQTL which enables rapid parallelized analyses of thousands of samples. Here we present the development of the methods – which needed adaptation to work with plaque-derived data – and results from the AE and TVS. We studied the balance between missing gene counts per sample and the number of identified eQTLs in the AE. We discovered thousands of nominally associated eGenes and confirmed 951 eGenes after permutation testing. Sex-interaction analyses identified *AOPEP* where the same allele (G) has different effects between the sexes. Future research will focus on integrating plaque-derived molQTL results with summary statistics from depression, risk factor and cardiovascular disease (CVD) genome-wide association studies, as well as causal network inference. These analyses are geared at identifying genetic loci and driver genes overlapping CVD and depression and may point to prospective druggable targets and biomarkers of disease.

Contents

List of Figures

List of Tables

List of Abbreviations

AE	Athero- Express Biobank Study
eQTL	expression quantitative trait loci
mQTL	methylation quantitative trait loci
molQTL	molecular quantitative trait loci, (expression, methylation, protein, etc)
TVS	Tampere Vascular Study

Report

1. Introduction

Key driving variants, which are derived from specific tissues and cells, identified in **Task 1.3** and **Task 1.5,** are used individually and in aggregate as instruments for Mendelian randomization. We aimed to investigate the causal relation between cardiovascular disease (CVD) and major depressive disorder (MDD) using summary statistics from genome-wide association studies (GWAS) of depression phenotypes, cardiovascular outcomes, relevant risk factors, and molecular and physiological biomarkers. This enables the examination of pleiotropic effects on gene regulation and transcription in CVD and MDD. Combining tissue-derived driving variants in this framework with diseases and intermediate traits, we aim to identify the causal path to disease. While we hypothesize that an inflammatory component connects depression and cardiovascular disease, the exact biological mechanism and site is unknown. Here we specifically tested whether this relation (between MDD and CVD) is mediated through atherosclerotic plaques. This will be informative for the selection of the proper cell and gene for follow-up experiments by TO_AITION partners or others. We utilize the plaque specific expression quantitative trait loci (eQTL) datasets we generated in the previous tasks (**1.4/1.5**).

2. Objectives

The main objective of these studies is to explore 1) the causal relation between major depressive disorder (MDD) and cardiovascular diseases (CVD), 2) to test for mediation through the plaque, and 3) to assess the role of inflammatory components in the causal relation between MDD and CVD.

3. Methods

3.1. Study participants Athero-Express Biobank Study

The Athero-Express Biobank Study (AE, approved and registered under number TME/C-01.18 and biobanknumber 22/088 entitled “Utrechts Cardiovasculair Cohort - The Second Manifestations of ARTerial disease Study (UCC-SMART/Athero-Express Biobank)” with study protocol 13-597) is an ongoing cohort study started in 2002[6] and includes patients undergoing arterial endarterectomy surgery in the University Medical Center Utrecht (Utrecht, The Netherlands) and the St. Antonius Hospital Nieuwegein (Nieuwegein, The Netherlands). The study design was described before[6]. Briefly, blood and plaque samples are obtained during surgery, and routinely stored at -80°C and plaque material is used for standardized (immuno)histochemical analysis[8]. Extensive data on clinical outcome up to 3 years after surgery, baseline clinical characteristics, medication use, and (prior) medical and family history are recorded. For this study we only included carotid endarterectomy (CEA) patients. The AE was approved by the respective hospitals’ Ethics Committees and follows the European and national guidelines regarding data security and GDPR. Only patients providing written informed consent are included and the study conforms to the Declaration of Helsinki.

3.2. DNA extraction, genotyping and imputation

We genotyped the AE in three separate, but consecutive experiments. In short, DNA was extracted from EDTA blood or (when no blood was available) plaque samples (regardless of arterial source) of 1,858 consecutive patients from the Athero-Express Biobank Study and genotyped in 3 batches. For the Athero-Express Genomics Study 1 (AEGS1) 891 patients (602 males, 262 females, 27 unknown sex), included between 2002 and 2007, were genotyped (440,763 markers) using the Affymetrix Genome-Wide Human SNP Array 5.0 (SNP5) chip (Affymetrix Inc., Santa Clara, CA, USA) at Eurofins Genomics(www.eurofinsgenomics.eu/, formerly known as AROS). For the Athero-Express Genomics Study 2 (AEGS2) 954 patients (640 makes, 313 females, 1 unknown sex), included between 2002 and 2013, were genotyped (587,351 markers) using the Affymetrix Axiom® GW CEU 1 Array (AxM) at the Genome Analysis Center (www.helmholtz-muenchen.de). The two first batches, AEGS1 and AEGS2, were described before[8,9]. For the Athero-Express Genomics Study 3 (AEGS3) 658 patients (448 males, 203 females, 5 unknown sex), included between 2002 and 2016, were genotyped (693,931 markers) using the Illumina GSA MD v1 BeadArray (GSA) at Human Genomics Facility, HUGE-F (glimdna.org/index.html). All experiments were carried out according to OECD standards. We used the genotyping calling algorithms as advised by Affymetrix (AEGS1 and AEGS2) and Illumina (AEGS3): BRLMM-P, AxiomGT1, and Illumina GenomeStudio respectively.

After genotype calling, we adhered to community standard quality control and assurance (QCA) procedures of the genotype data from AEGS1, AEGS2, and AEGS3[10]. Samples with low average genotype calling and sex discrepancies (compared to the clinical data available) were excluded. The data was further filtered per sample set on 1) individual (sample) call rate > 97%, 2) SNP call rate > 97%, 3) minor allele frequencies (MAF) > 3%, 4) average heterozygosity rate ± 3.0 s.d., 5) relatedness (pi-hat > 0.20), 6) Hardy–Weinberg Equilibrium (HWE $p < 1.0\times10-3$), and 7) Monomorphic SNPs ($< 1.0\times10-6$). After QCA 2,493 samples remained, 108 of non-European descent/ancestry, and 156 related pairs. These comprise 890 samples and 407,712 SNPs in AEGS1, 954 samples and 534,508 SNPs in AEGS2, and 649 samples and 534,508 SNPs in AEGS3 remained.

Before phasing using SHAPEIT2, data was lifted to genome build b37 using the liftOver tool from UCSC (genome.ucsc.edu/cgi-bin/hgLiftOver). Finally, data was imputed with 1000G phase 3, version 5 and HRC release 1.1 as a reference using the Michigan Imputation Server (imputationserver.sph.umich.edu/)[11]. These results were further integrated using QCTOOL v2, where HRC imputed variants are given precedence over 1000G phase 3 imputed variants. After imputation we merge dataset and re-evaluated the quality and relatedness of samples. This resulted in a final list of 2,124 samples of good quality (Figure 1), including family relations of which we randomly chose 1 for downstream analyses leaving 2,060 unique samples. We also re-evaluated the ancestral background and determined that 33 are from non-European ancestry applying PCA and using data from the 1000G phase 3.

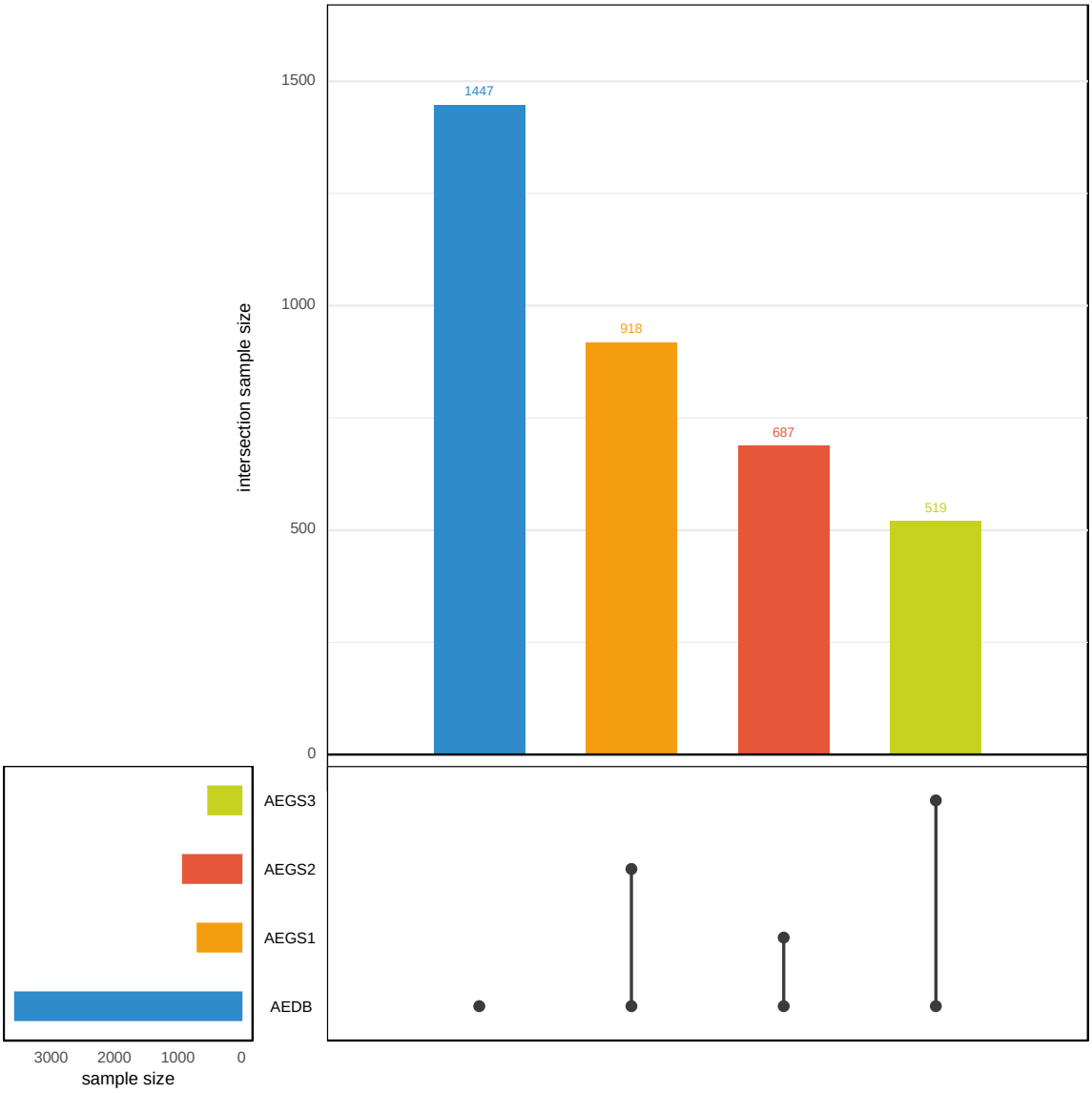

Figure 1: Overlap of AEGS1, AEGS2 and AEGS3 with the whole Athero-Express Biobank Study.

3.3. RNA isolation, transcriptional profiling and preprocessing

A total of 700 segments were selected from patients who were included in the study between 2002 and 2016. The RNA isolated from the archived advanced atherosclerotic lesion is fragmented. We have ultimately employed the CEL-seq2 method[7]. CEL-seq2 yielded the highest mappability reads to the annotated genes compared to other library preparation protocols. The methodology captures 3'-end of polyadenylated RNA species and includes unique molecular identifiers (UMIs), which allow direct counting of unique RNA molecules in each sample.

Libraries were sequenced on the Illumina Nextseq500 platform; a high output paired-end run of 2 × 75 bp was performed (Utrecht Sequencing Facility). The reads were demultiplexed and aligned to human cDNA reference (Ensembl 84) using the BWA (0.7.13). Multiple reads mapping to the same gene with the same unique molecular identifier (UMI, 6bp long) were counted as a single read. The raw read counts were corrected for UMI sampling (corrected_count=-4096*(ln(1-(raw_count/4096)))), normalized for sequencing depth and quantile normalized (core scripts can be found in github.com/mmokry/bulkCEL-seq2 and github.com/mmokry/seurat_meets_bulk_AE). We

have detected a median of 19.501 (SD = 5.874) genes per sample with at least one unique read and discarded samples (n=46) with less than 9,000 detected genes from further analysis.

3.3.1. Gene quality control

Gene exclusion was performed in the Python package *pandas*. The UMI corrected RNAseq count and corresponding hg19 biomart gene information were loaded into a dataframe. Non-protein coding genes were excluded as well as those lying on non-standard (alternative) chromosome. UMI-corrected counts reported as infinite float values were replaced by the largest observed finite count value. Next, a sweep over a missingness threshold from 10% to 100% in steps of 10% was conducted, and a separate gene dataset was prepared for each threshold. The filter is applied by removing genes with zero counts for more than the threshold-portion of samples. Next, TMM normalization (Robinson 2010) as provided by the *conorm* package and inverse normal transform (INT) normalization using the *scipy.stats* package. A comparison of counts per missingness threshold is reported in Table 1. A flow diagram of sample and gene quality assessment is given in Figure 2.

Table 1: Final gene counts given different missingness thresholds after gene quality control.

Threshold	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Final count	10,006	11,651	12,710	13,620	14,358	15,057	15,682	16,258	17,052	18,462

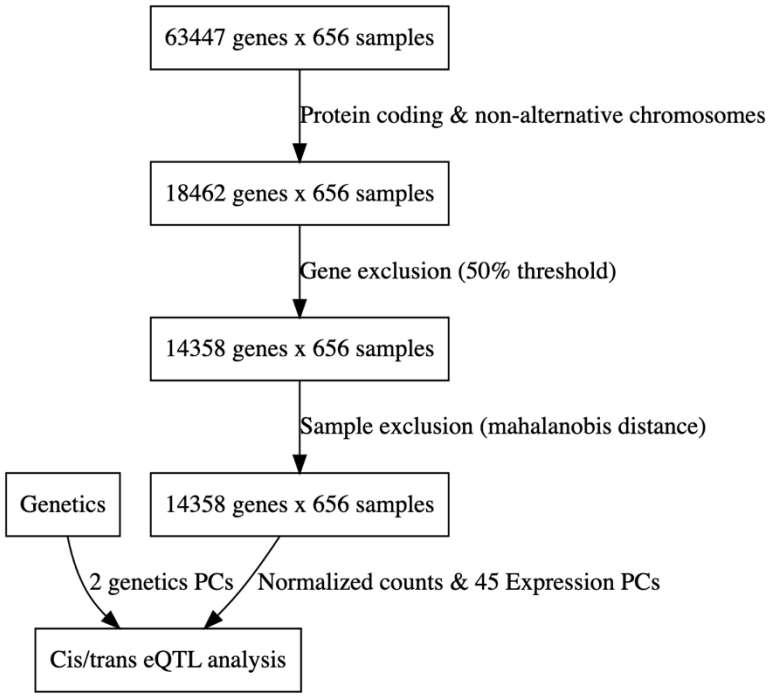

Figure 2: Flow diagram of sample and gene quality assessment.

3.3.2. Sample quality control

For sample covariates and subsequent sample exclusion, 2 genetics PC's and 100 PC's expression were calculated. For expression, randomized truncated PCA estimation[8] was used due to the large dataset. For the first 2 expression covariation, the sample Mahalanobis distance was calculated, and samples below the Chi^2 (n=3, alpha=0.95) threshold were selected. Then a sweep over the amount

of included expression PCs was performed from n=0 to 100 components and each total (2+n) covariates was saved to a different file.

3.3.3. Missingness threshold and covariate count selection and cis/trans expression-QTL mapping

`QTLtools`[8] was used to select the missingness threshold (10-100%) and expression covariate counts (0-100) from all generated combinations for subsequent analysis. For this, SNPs were filtered on MAF larger than 0.03 and INFO score larger than 0.4 and stored as VCF file as required for 80% power. `QTLtools` was run in *cis* permutation mode with a window of 1Mb and the amount of gene-level and genome-wide level results from the permutation test adjusted p-values was determined. Here it was found that the amount of genome-wide significant results flattens at a missingness value of 50%, where then a peak is found for 45 expression covariates. Plots of significant hits are shown in Figure 3.

For the final *cis*- and *trans*-molQTL mapping we employed the `TensorQTL`[9] package on an Nvidia RTX6000 GPU using the previously determined missingness threshold and number of expression covariates and 10 genetic PCs for the eQTL analyses, and only genetic PCs for the molQTL analyses. For final QTL mapping, VCF files were converted to `PLINK`[10] BED file format as required by `TensorQTL`.

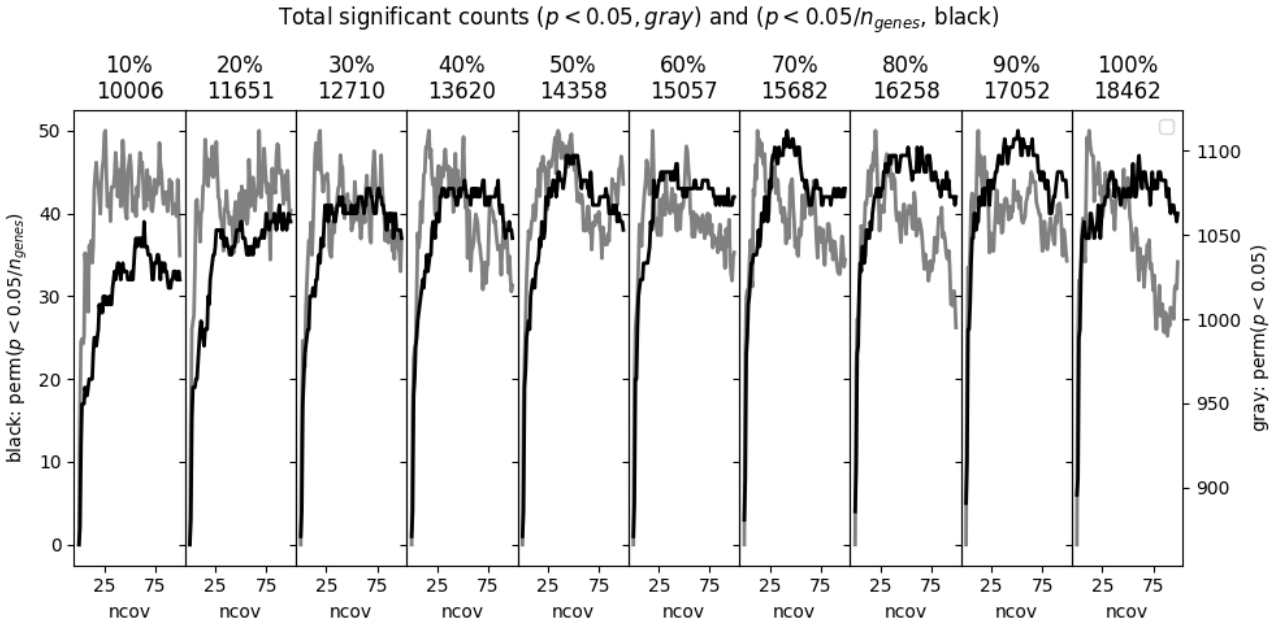

Figure 3: Missingness threshold and covariate counts selection.
Total significant counts (p < 0.05) are shown in gray and the right axis and genome wide significant hits (p < 0.05/ngenes) are shown in black and the left axis.

3.4. Genome-wide association studies summary statistics

Genome-wide association studies summary statistics were leveraged from publicly available datasets. Mining these datasets uncovered a plethora of challenges. Not all GWAS summary statistics are available for download, can be requested by the authors, and/or only provide the lead single nucleotide polymorphisms (SNPs). Other issues included ambiguous reports of sample size, genome build, or missing information such as standard errors and effect sizes. We were able to partly resolve these issues by using a custom pipeline (`gwas2cojo`, available here

https://github.com/CirculatoryHealth/gwas2cojo), which automates tasks such as genome build liftover and SNP alignment and provides a uniform file format. This significantly decreased workload and improved downstream applications.

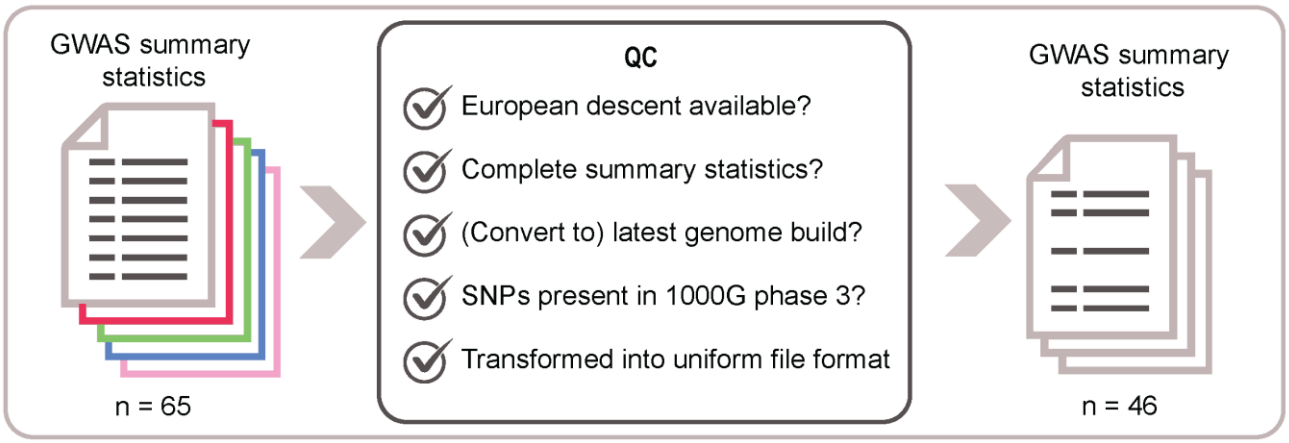

Figure 4: Parsing and harmonizing GWAS summary statistics.
We collected 46 GWAS datasets focusing on data from individuals of European descent. When needed we converted data to genome build 37, and checked that variants were present in 1000G phase 3[11]. Lastly, all data were transformed into a uniform file format.

Genome-wide association studies summary statistics were selected by means of relevance, 'latest and the largest' and publicly available, easily accessible and/or open-source data was given priority. Incomplete summary statistics were omitted during this phase. A custom Python[17]pipeline (see *Data availability*) was used to pre-process GWAS summary statistics and primary quality control. This primary control step was used to uniformly transform GWAS summary statistics because of a lack of a standardized format. In short, our custom pipeline was built to automate effect/other allele alignment with 1000 Genomes phase 3[18] release with a maximum frequency distance of 25%, maximum minor allele frequency (MAF) of 45% for ambivalent variants and lifting genomic positions to hg19 where needed. In preparation for GWAS alignment, dbSNP153 (GCF1405) was referenced to translate MEGASTROKE rs-ids to genetic positions and to augment AF, ASD, NICM, and TAGC (GWAS abbreviations are listed in Table 2) with allele frequency information. Standard errors for HF and NICM were recovered as follows:

```
Z=sign(beta)*abs(qnorm(p/2))
SE=beta/Z
```

Where beta is the effect size, *P* is the *P*-value of association which is quantile normalized using the qnorm-function in R. SE is the standard error calculated by dividing the beta by Z. Effect sizes were calculated from odds ratios as beta = log(OR) for BIP, insomnia, IBD, and MDD before entering our custom pipeline.
The final selection was grouped into one of three categories: atherosclerotic disease and other cardiovascular disease (12), risk factors (14), and other (20). Data were further processed with MAGMA[19] (version 1.07) for genome-wide analysis, annotation and characterization of significant hits via the FUMA[20] web platform (version 1.3.6, accessed June 2020, https://fuma.ctglab.nl/). We applied genome-wide significance ($P < 5 \times 10^{-8}$) and linkage disequilibrium ($r^2 > 0.05$) filtering for clumping of independent loci within 1000 kb of the lead variant and included only variants with MAF >1%. For SVS, logOnset, EvrSmk and SVD no SNP associations below the genome-wide significance threshold we found. Therefore, lead variant discovery was performed with a relaxed threshold of 5×10^{-6}.

Table 2: GWAS summary statistics used.

We applied GWAS2COJO for the Slenders L et al paper (EHJ open 2022). GWAS2COJO is a software to GWAS summary statistics into one common format (for COJO-analyses) using either rsIDs or chr:bp as identifiers and aligning results with a common reference (in this 1000G phase 3).

TRAIT NUMBER	TRAIT OR DISEASE	ABBREVIATION	CATEGORY	SAMPLE SIZE	CONSORTIUM	PMID	REFERENCE	JOURNAL	ANCESTRY
1	Coronary Artery Disease	CAD	AD	154,654	CARDIoGRAMplusC4D +UKBB	28714975	Nelson et. al, 2017[3]	nature genetics	European
2	Coronary Artery Calcification	CAC	AD	15,523		23561647	Van Setten et al. 2013[5]	Atherosclerosis	European
3	Carotid Intima-Media Thickness	cIMT	AD	71,128	CHARGE	30510157	Franceschini et al. 2018[6]	nature communication	Multi ancestry
4	Plaque Presence	Plaque	AD	48,434	CHARGE	30510157	Franceschini et al. 2018[6]	nature communication	Multi ancestry
5	Cardio- Embolic stroke	CES	CM	521,612	MEGASTROKE	29531354	Malik et al., 2019[4]	nature genetics	European
6	Any stroke	AS	CM	446,696	MEGASTROKE	29531354	Malik et al., 2019[4]	nature genetics	European
7	Any Ischemic stroke	IS	CM	446,696	MEGASTROKE	29531354	Malik et al., 2019[4]	nature genetics	European
8	Small Vessel Disease	SVD	CM	446,696	MEGASTROKE	29531354	Malik et al., 2019[4]	nature genetics	European
9	Large Artery Stroke	LAS	CM	446,696	MEGASTROKE	29531354	Malik et al., 2019[4]	nature genetics	European
10	Atrial Fibrillation	AF	CM	588,190		29892015	Roselli et al., 2018[44]	nature genetics	Multi ancestry
11	Heart Failure	HF	CM	488,010		30586722	Aragam et al. 2018[45]	Circulation	European
12	Nonischemic Cardiomyopa thy	NICM	CM	488,010		30586722	Aragam et al. 2018[45]	Circulation	European
13	High-Density Lipoprotein	HDL	CM	1,888,577	GLGC	24097068	Global lipids Genetiscs Consortium, 2013[30]	nature genetics	European
14	Low-Density Lipoprotein	LDL	CM	1,888,577	GLGC	24097068	Global lipids Genetiscs Consortium, 2013[30]	nature genetics	European
15	Total Cholesterol	TC	CM	1,888,577	GLGC	24097068	Global lipids Genetiscs Consortium, 2013[30]	nature genetics	European
16	Triglycerides	TG	CM	1,888,577	GLGC	24097068	Global lipids Genetiscs Consortium, 2013[30]	nature genetics	European
17	Pulse Pressure	PP	CM	1,050,906		30224653	Evangelou et al., 2018[46]	nature genetics	European
18	Diastolic Blood Pressure	DBP	CM	1,050,906		30224653	Evangelou et al., 2018[46]	nature genetics	European
19	Systolic Blood Pressure	SBP	CM	1,050,906		30224653	Evangelou et al,. 2018[46]	nature genetics	European
20	Cigarettes per Day	CpD	CM	74,035	TAG	20418890	The Tobacco and Genetics Consortium, 2010[47]	nature genetics	European
21	Ever smoked	EvrSmk	CM	74,035	TAG	20418890	The Tobacco and Genetics Consortium, 2010[47]	nature genetics	European

22	Onset Smoking	logOnset	CM	74,035	TAG	20418890	The Tobacco and Genetics Consortium, 2010[47]	nature genetics	European
23	Former Smoker	FrmrSmk	CM	74,035	TAG	20418890	The Tobacco and Genetics Consortium, 2010[47]	nature genetics	European
24	Type 2 Diabetes	T2D	CM	898,130	DIAGRAM	30297969	Mahajan et al., 2016[48]	nature genetics	Multi ancestry
25	Type 2 Diabetes adjusted for BMI	T2DadjBMI	CM	898,130	DIAGRAM	30297969	Mahajan et al., 2016[48]	nature genetics	European
26	Body Mass Index	BMI	CM	693,529	GIANT	30124842	Yengo et al., 2018[49]	Human Molecular Genetics	European
27	Alzheimer's Disease	AD	Other	455,258		30617256	Jansen et al., 2019[50]	nature genetics	European
28	Autism Spectrum Disorder	ASD	Other	46,350	PGC	30804558	Grove et al., 2019[51]	nature genetics	European
29	Bipolar Disorder	BIP	Other	198,882	PGC	31043756	Stahl et al., 2019[52]	nature genetics	European
30	Depressive symptoms	DS	Other	298,420		27089181	Okbay et al., 2016[53]	nature genetics	European
31	Educational Attainment	EA	Other	293,723		27225129	Okbay et al., 2016[54]	nature genetics	European
32	Insomnia	Insomnia	Other	386,533		30804565	Jansen et al., 2018[50]	nature genetics	European
33	Intelligence Quotient	IQ	Other	269,867		29942086	Savage et al., 2018[55]	nature genetics	European
34	Major Depression Disorder	MDD	Other	480,359	PGC	29700475	Wray et al., 2018[56]	nature genetics	European
35	Neuroticism	Neuroticism	Other	298,420		27089181	Okbay et al., 2016[53]	nature genetics	European
36	Parkinson's Disease	PD	Other	1,456,306		31701892	Nalls et al.,2019[57]	The Lancet. Neurology	European
37	Asthma	Asthma	Other	142,486		29273806	Demenais et al., 2018[58]	nature genetics	Multi ancestry
38	Inflammatory Bowel Disease	IBD	Other	86,640		26192919	Liu et al., 2015[59]	nature genetics	Multi ancestry
39	Breast Cancer	BC	Other	256,123		29059683	Michailidou et al., 2017[60]	nature	Multi ancestry
40	Prostate Cancer	PrCa	Other	140,306		29892016	Schumacher et al., 2018[61]	nature genetics	European
41	Amyotrophic Lateral Sclerosis	ALS	Other	43,259		27455348	van Rheenen et al., 2016[62]	nature genetics	European
42	Femoral Neck Bone Mass Density	FNBMD	Other	53,236	GEFOS	26367794	Zheng et al., 2015[63]	nature	European
43	Forarm Bone Mass Density	FABMD	Other	53,236	GEFOS	26367794	Zheng et al., 2015[63]	nature	European
44	Lumbar Spine Bone Mass Density	LSBMD	Other	53,236	GEFOS	26367794	Zheng et al., 2015[63]	nature	European
45	Height	Height	Other	693,529	GIANT	30124842	Yengo et al., 2018[49]	Human Molecular Genetics	European

46	Subjective Well-Being	SWB	Other	298,420	27089181	Okbay et al., 2016[53]	nature genetics	European
47	Coronary Artery Calcification	CHARGE_CAC_EA_AA	AD	35,776		Kavousi et al., 2022[64]	medRxiv (accept. nature genetics)	Multi ancestry
48	Type 2 Diabetes	T2D_EU	CM	442,817	30297969	Mahajan et al., 2016[48]	nature genetics	European

References used: 3. Nelson, C. P. et al. Association analyses based on false discovery rate implicate new loci for coronary artery disease. Nat. Genet. 49, 1385–1391 (2017).; 4. Malik, R. et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. Nat. Genet. 50, 524–537 (2018).; 5. van Setten, J. et al. Genome-wide association study of coronary and aortic calcification implicates risk loci for coronary artery disease and myocardial infarction. Atherosclerosis 228, 400–405 (2013).; 6. Franceschini, N. et al. GWAS and colocalization analyses implicate carotid intima-media thickness and carotid plaque loci in cardiovascular outcomes. Nat. Commun. 9, 5141 (2018).; 30. Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. Nat. Genet. 45, 1274–1283 (2013).; 44. Roselli, C. et al. Multi-ethnic genome-wide association study for atrial fibrillation. Nat. Genet. 50, 1225–1233 (2018).; 45. Aragam, K. G. et al. Phenotypic Refinement of Heart Failure in a National Biobank Facilitates Genetic Discovery. Circulation (2018) doi:10.1161/CIRCULATIONAHA.118.035774.; 46. Evangelou, E. et al. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. Nat. Genet. 50, 1412–1425 (2018).; 47. Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat. Genet. 42, 441–447 (2010).; 48. Mahajan, A. et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat. Genet. 50, 1505–1513 (2018).; 49. Yengo, L. et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. Hum. Mol. Genet. 27, 3641–3649 (2018).; 50. Jansen, I. E. et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nat. Genet. 51, 404– 413 (2019).; 51. Grove, J. et al. Identification of common genetic risk variants for autism spectrum disorder. Nat. Genet. 51, 431–444 (2019).; 52. Stahl, E. A. et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat. Genet. 42, 508–514 (2010).; 53. Okbay, A. et al. Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. Nat. Genet. 48, 624–633 (2016).; 54. Okbay, A. et al. Genome-wide association study identifies 74 loci associated with educational attainment. Nature 533, 539–542 (2016).; 55. Savage, J. E. et al. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. Nat. Genet. 50, 912–919 (2018).; 56. Wray, N. R. et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat. Genet. 50, 668–681 (2018).; 57. Nalls, M. A. et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet Neurol. 18, 1091–1102 (2019).; 58. Demenais, F. et al. Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. Nat. Genet. 50, 42–53 (2018).; 59. Liu, J. Z. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat. Genet. 47, 979–986 (2015).; 60. Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. Nature 551, 92–94 (2017).; 61. Schumacher, F. R. et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. Nat. Genet. 50, 928–936 (2018).; 62. van Rheenen, W. et al. Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. Nat. Genet. 48, 1043–1048 (2016).; 63. Zheng, H.-F. et al. Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. Nature 526, 112–117 (2015).; 64. Kavousi, M., et al. Multi-ancestry genome-wide analysis identifies effector genes and druggable pathways for coronary artery calcification. medRxiv 2022. medRxiv 2022.05.02.22273844; doi: https://doi.org/10.1101/2022.05.02.22273844.

4. Results

4.1. Identification of cis-acting eQTLs in the Athero-Express Biobank Study

After extensive quality control we included 624 samples with overlapping carotid plaque gene expression and genetic data for *cis*-acting eQTL analyses. A nominal analysis identified 14,284 unique eQTL-eGene pairs at p < 0.05. Next, we performed permutation testing (1000x) and identified 951 *cis*-acting eQTLs across all 22 chromosomes at $p_{emperical}$ < 0.05 (Figure 5).

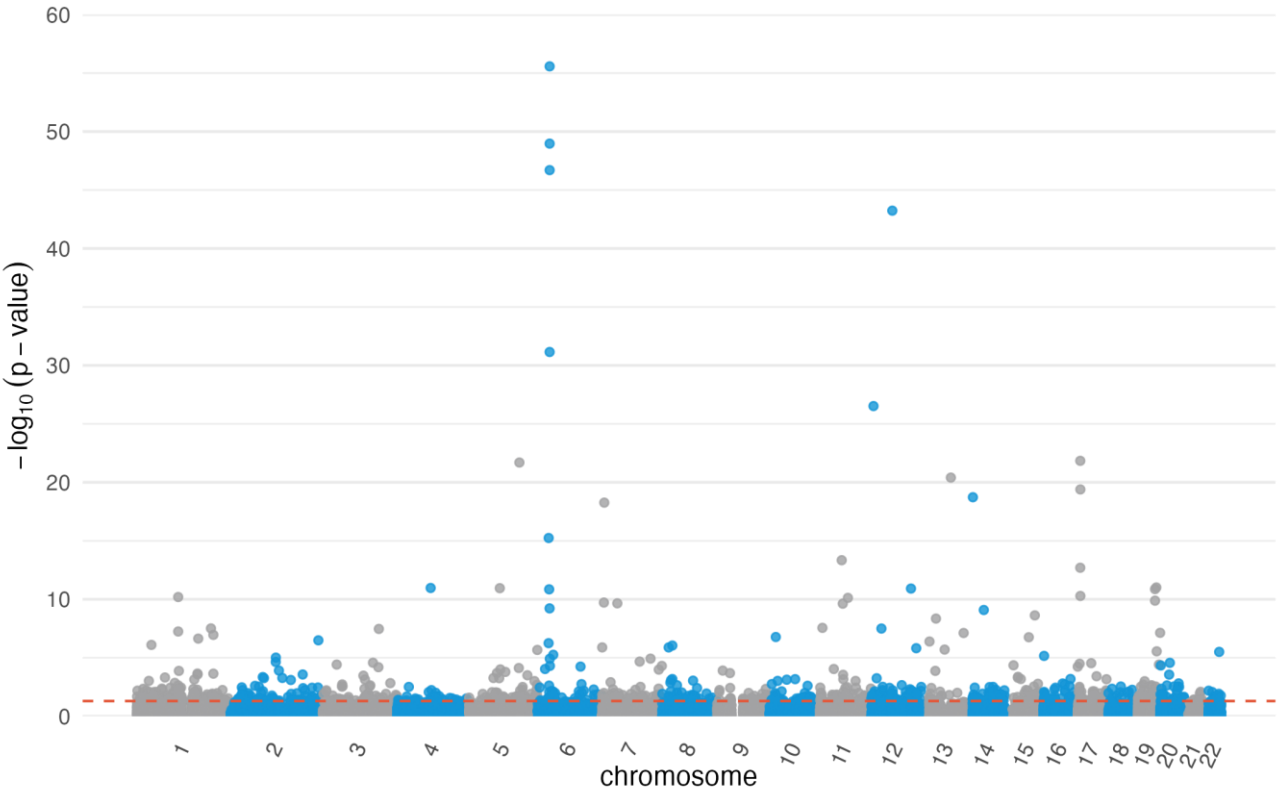

Figure 5: Genome-wide *cis*-acting eQTL results at $p_{emperical}$ < 0.05.

4.2. Effects of MDD associated loci on plaque derived gene expression

We tested the association of 6 known MDD loci with gene expression for over 14,000 plaque derived genes. Out of 14356 genes tested, the 6 MDD loci associated with 610 genes at $p \leq 0.05$ – which is less than expected by chance. The most significant was ENSG00000186918 (β = 0.587 ± 0.146 s.e., p = 5.56x10^{-5}).

4.3. Effects of cis-acting eQTLs in plaque on the risk of coronary artery disease

Next, we tested the association of each variant associated with plaque derived genes expression on the risk of coronary artery disease using GWAS summary statistics. Out of the 14,356 genes tested, 718 were nominally associated to CAD ($p \leq 0.05$). The most significant association was for ENSG00000131095 (β = -0.078 ± 0.020 s.e., p = 9.29x10^{-5}).

4.4. Mediation analyses of MDD loci on CAD through plaque derived gene expression

Central to the TO_AITION project is the notion that MDD may be causing CAD, or vice versa. A key question is whether this causal effect may be mediated directly through atherogenesis, i.e. atherosclerotic plaques. Thus, we assessed whether the effects of MDD associated loci are mediated through atherosclerotic plaques on CAD. Out of 14,356 genes tested, 38 are nominally associated with both MDD and CAD. We observed a negative effect of plaque mediated MDD associated loci on CAD risk ($r^2 = 0.0289$), albeit nominally significant ($p = 0.05$, Figure 6). Our per plaque derived gene interaction analyses revealed no significant association (Table 3).

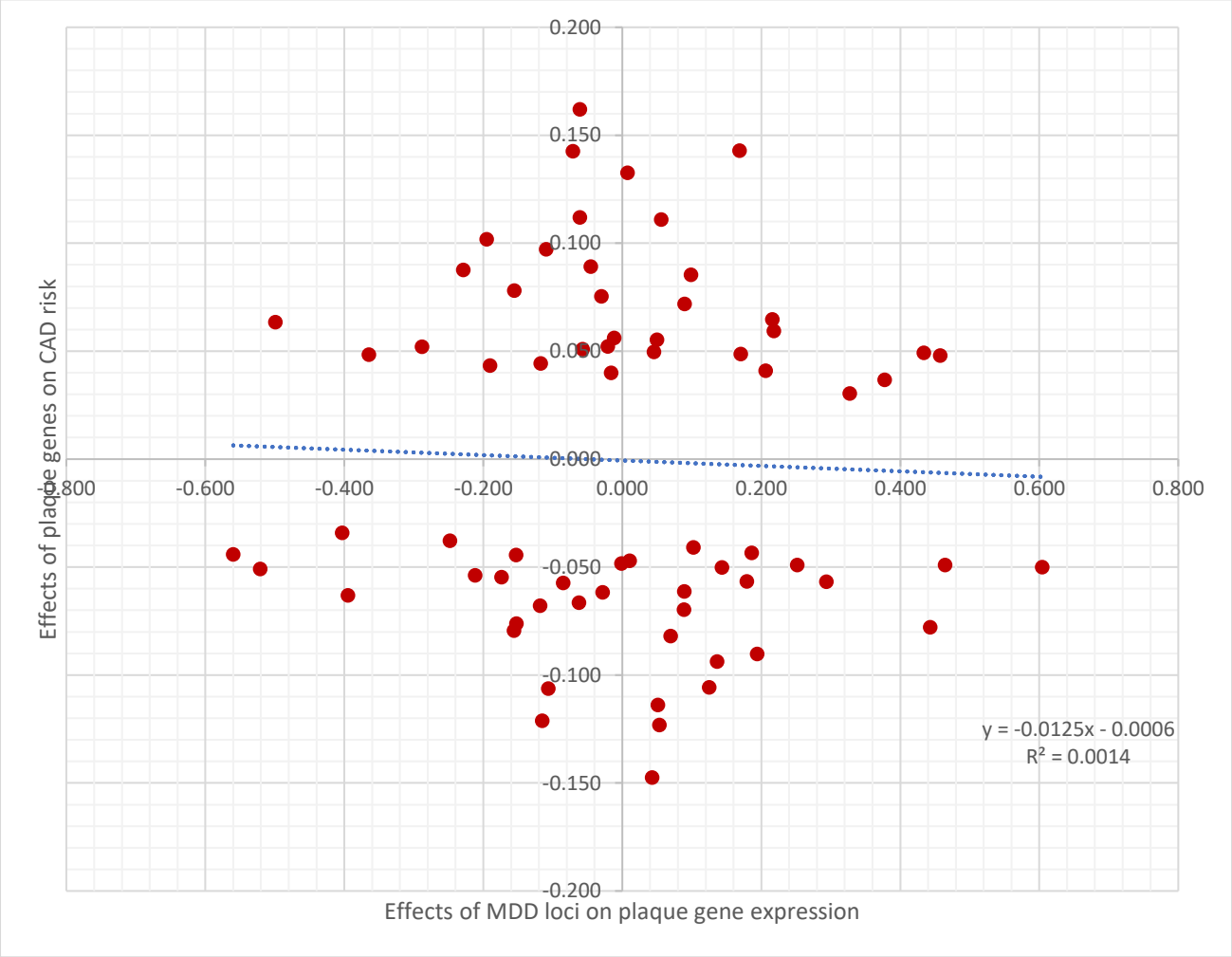

Figure 6: Plaque mediated association between MDD and CAD.

Table 3: Plaque derived gene expression mediation analysis of MDD (exposure) associate loci (6 in total) with CAD (outcome).

b_em: effect size exposure; se_em: standard error exposure; p_em: p-value exposure; nsnp_em: number of SNPs tested for exposure; b_mo: effect size outcome; se_mo: standard error outcome; p_mo: p-value outcome; nsnp_mo: number of SNPs tested for outcome; b_in: effect size interaction; se_in: standard error interaction; p_in: p-value interaction.

m	b_em	se_em	p_em	nsnp_em	b_mo	se_mo	p_mo	nsnp_mo	b_in	se_in	p_in
ENSG00000106628	-0.814	0.279	3.49E-03	6	-0.040	0.016	0.013	13	0.033	0.018	0.068
ENSG00000119231	-0.209	0.082	0.010	6	0.104	0.038	5.63E-03	18	-0.022	0.012	0.069
ENSG00000186566	-0.499	0.202	0.013	5	0.063	0.022	4.09E-03	18	-0.032	0.017	0.070
ENSG00000131095	0.443	0.215	0.039	6	-0.078	0.020	9.29E-05	16	-0.035	0.019	0.075
ENSG00000157954	-0.428	0.153	5.21E-03	6	-0.058	0.024	0.016	19	0.025	0.014	0.079
ENSG00000183814	-0.755	0.306	0.014	6	-0.040	0.015	0.010	13	0.030	0.018	0.086
ENSG00000146909	-0.195	0.090	0.030	6	0.102	0.034	2.44E-03	26	-0.020	0.012	0.088
ENSG00000167526	0.169	0.080	0.035	6	0.143	0.046	1.92E-03	17	0.024	0.014	0.092
ENSG00000008869	-0.514	0.210	0.014	6	0.042	0.017	0.013	22	-0.021	0.013	0.093
ENSG00000166199	0.604	0.306	0.048	6	-0.050	0.015	6.17E-04	13	-0.030	0.018	0.097
ENSG00000176209	-0.377	0.158	0.017	6	0.052	0.021	0.013	23	-0.020	0.012	0.098
ENSG00000111229	-0.403	0.159	0.011	6	-0.050	0.022	0.026	21	0.020	0.013	0.108
ENSG00000154518	0.382	0.161	0.018	5	-0.101	0.043	0.020	7	-0.039	0.024	0.112
ENSG00000213337	-0.654	0.299	0.029	6	-0.032	0.013	0.012	20	0.021	0.013	0.113
ENSG00000119392	-0.708	0.275	0.010	6	0.058	0.027	0.033	6	-0.041	0.026	0.116
ENSG00000121075	0.565	0.237	0.017	6	-0.035	0.016	0.027	17	-0.020	0.013	0.121
ENSG00000151388	0.687	0.313	0.028	6	-0.033	0.014	0.017	15	-0.023	0.015	0.123
ENSG00000136247	0.379	0.178	0.033	6	-0.063	0.026	0.014	14	-0.024	0.016	0.123
ENSG00000158477	0.719	0.335	0.032	6	-0.030	0.013	0.018	16	-0.021	0.014	0.130
ENSG00000077157	0.285	0.118	0.015	6	0.077	0.037	0.039	13	0.022	0.015	0.134
ENSG00000089177	-0.503	0.252	0.046	6	-0.037	0.015	0.012	19	0.019	0.013	0.135
ENSG00000133216	0.315	0.148	0.033	6	-0.076	0.033	0.021	12	-0.024	0.016	0.136
ENSG00000132405	0.533	0.265	0.044	6	0.044	0.018	0.014	12	0.023	0.016	0.138
ENSG00000141505	0.362	0.147	0.014	6	-0.026	0.013	0.045	18	-0.009	0.006	0.138
ENSG00000187824	0.359	0.168	0.032	6	-0.061	0.027	0.023	14	-0.022	0.015	0.138
ENSG00000204540	0.724	0.355	0.041	6	0.034	0.014	0.017	11	0.025	0.017	0.139

ENSG00000127663	0.219	0.110	0.045	6	-0.069	0.029	0.016	23	-0.015	0.010	0.142
ENSG00000255398	0.656	0.286	0.022	6	-0.029	0.014	0.045	17	-0.019	0.013	0.152
ENSG00000105364	-0.316	0.140	0.024	6	0.051	0.025	0.043	22	-0.016	0.011	0.153
ENSG00000174306	-0.239	0.105	0.023	6	-0.107	0.053	0.045	9	0.026	0.018	0.153
ENSG00000125965	-0.687	0.307	0.025	6	0.025	0.013	0.047	17	-0.017	0.012	0.158
ENSG00000138069	0.310	0.149	0.038	6	-0.051	0.024	0.035	20	-0.016	0.011	0.160
ENSG00000198912	0.273	0.137	0.047	6	0.059	0.027	0.031	20	0.016	0.012	0.166
ENSG00000100095	-0.645	0.324	0.046	6	-0.030	0.014	0.034	13	0.019	0.014	0.169
ENSG00000143436	0.467	0.233	0.045	6	-0.031	0.016	0.045	22	-0.015	0.011	0.181
ENSG00000160856	-0.587	0.289	0.043	6	-0.026	0.013	0.050	18	0.015	0.011	0.184
ENSG00000180953	0.430	0.219	0.049	6	-0.032	0.016	0.047	19	-0.014	0.010	0.188

5. Future perspectives

We identified 610 plaque genes nominally effected by 6 MDD associated loci, 38 of which at $p < 0.005$. We also found 718 plaque genes nominally associated with CAD risk, 75 at $p < 0.005$. Mediation analyses revealed 38 genes were nominally associated with both 6 MDD loci, and CAD risk, but mediation testing revealed no significant effects.

We will further intergrate these results with the other plaque derived data, including our single-cell datasets to identify per genes specific cells involved. Likewise, we will execute network analyses to determine driver nodes associated with plaque-derived gene expression on CAD risk, and MDD loci on plaque gene regulatory networks. We will further expand these analyses by studying the mediating effects of blood-derived gene expression on MDD and CAD risk. Combining tissue-derived driving variants in this framework with diseases and intermediate traits, we will identify tissues on the causal path to disease, and as such be informative for the selection of the proper cell for follow-up experiments in WP4. These analyses are planned for Q3-Q4 of 2023 in collaboration with UVA.

Data security, availability and sharing

The input data for these analyses are available among the collaborating partners, either through the secure platform developed by UOI, or privately between the respective partners (see **List of Beneficiaries**). The Athero-Express Biobank Study data is publicly available, but upon request, through DataverseNL (https://doi.org/10.34894/4IKE3T). Likewise, the codes used for this project are available here: https://github.com/CirculatoryHealth/molqtl (privately, as the project is ongoing). This project falls under the Data Management Plan of the Athero-Express Biobank Study attached as a separate appendix and approved by the Information Security Officer and Data Management Officer of the UMC Utrecht.

As mentioned in **section 3.1** the Athero-Express Biobank Study (AE, approved and registered under number TME/C-01.18 and biobanknumber 22/088 entitled "Utrechts Cardiovasculair Cohort - The Second Manifestations of ARTerial disease Study (UCC-SMART/Athero-Express Biobank)" with study protocol 13-597) is an ongoing cohort study started in 2002[6] and includes patients undergoing arterial endarterectomy surgery in the University Medical Center Utrecht (Utrecht, The Netherlands) and the St. Antonius Hospital Nieuwegein (Nieuwegein, The Netherlands). The studies were approved by the respective hospitals' Ethics Committees and follow the European and national guidelines regarding data security and GDPR. Only patients providing written informed consent are included and the studies conform to the Declaration of Helsinki.

References

1. Delaneau, O. *et al.* A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* **8**, 15452 (2017).
2. Verhoeven, B. A. N. *et al.* Athero-express: differential atherosclerotic plaque expression of mRNA and protein in relation to cardiovascular events and patient characteristics. Rationale and design. *Eur. J. Epidemiol.* **19**, 1127–1133 (2004).
3. van Lammeren, G. W. *et al.* Time-Dependent Changes in Atherosclerotic Plaque Composition in Patients Undergoing Carotid Surgery. *Circulation* **129**, 2269–2276 (2014).
4. van der Laan, S. W. *et al.* Genetic Susceptibility Loci for Cardiovascular Disease and Their Impact on Atherosclerotic Plaques. *Circ Genom Precis Med* **11**, e002115 (2018).
5. Laurie, C. C. *et al.* Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet. Epidemiol.* **34**, 591–602 (2010).
6. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
7. Hashimshony, T. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).
8. Halko, N., Martinsson, P.-G. & Tropp, J. A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *arXiv [math.NA]* (2009).
9. Taylor-Weiner, A. *et al.* Scaling computational genomics to millions of individuals with GPUs. *Genome Biol.* **20**, 228 (2019).
10. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
11. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).