

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

### D1.4 Molecular QTL analyses in relevant healthy and diseased cells and tissues

Nature:	Report
Dissemination Level:	Public
Contractual Date of Delivery to the EC:	31/12/22
Actual Date of Delivery to the EC:	25/07/23
WP number and Title:	WP1 Cloud infrastructure, semantic interlinking of data and construction of a functional framework of molecular interactions
Lead Beneficiary:	UMCU
Version 0.1:	Authored by Dr. Sander W. van der Laan and Lennart Landsmeer on behalf of the UMCU team in collaboration with TAUH on 16/12/2022 – Initial draft
Version 0.2:	Reviewed by Dr. Sander W. van der Laan, UMCU on 12/01/2023
Version 0.3:	Reviewed by Dr. Sander W. van der Laan, UMCU on 27/01/2023
Version 0.4:	Reviewed by the Work Package members and the coordinator on 04/02/2023
Version 1.0:	Submitted on 01/02/2023
Version 1.1:	Revised after comments from the reviewers by Dr. Sander W. van der Laan, UMCU and the coordinator on 23/07/2023
Version 2.0:	Resubmitted on 25/07/23
Version 2.1:	Revised after comments from reviewers by Dr. Sander W. van der Laan, UMCU and the coordinator and resubmitted on 04/11/2024.

# 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
<b>11</b>	<b>UNIVERSITAIR MEDISCH CENTRUM UTRECHT</b>	<b>UMCU</b>	<b>NETHERLANDS</b>
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](https://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<sup>5</sup> 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

---

<b>1.</b>	<b><i>Introduction</i></b> .....	<b>6</b>
<b>2.</b>	<b><i>Objectives</i></b> .....	<b>6</b>
<b>3.</b>	<b><i>Methods</i></b> .....	<b>6</b>
<b>3.1.</b>	<b>Study participants Tampere Vascular and Athero-Express Biobank Study</b> .....	<b>6</b>
<b>3.2.</b>	<b>Tampere Vascular Study methods</b> .....	<b>7</b>
3.2.1.	RNA isolation, transcriptional profiling and preprocessing .....	7
3.2.2.	DNA extraction, genotyping and imputation.....	7
3.2.3.	Biostatistical analysis.....	8
<b>3.3.</b>	<b>Athero-Express Biobank Study</b> .....	<b>8</b>
3.3.1.	DNA extraction, genotyping and imputation.....	8
3.3.2.	RNA isolation, transcriptional profiling and preprocessing .....	9
3.3.3.	DNA extraction and methylation experiment .....	11
3.3.4.	Quality control of methylation data .....	12
<b>4.</b>	<b><i>Results</i></b> .....	<b>13</b>
<b>4.1.</b>	<b>Identification of eQTLs in carotid plaques from TVS</b> .....	<b>13</b>
<b>4.2.</b>	<b>Identification of cis-acting eQTLs in the Athero-Express Biobank Study</b> .....	<b>14</b>
<b>4.3.</b>	<b>Trans-acting mQTL analyses in carotid plaques</b> .....	<b>17</b>
<b>5.</b>	<b><i>Conclusions and Future perspectives</i></b> .....	<b>17</b>
<b>5.1.</b>	<b>Conclusions: with respect to deliverable.</b> .....	<b>17</b>
<b>5.2.</b>	<b>Future perspectives</b> .....	<b>17</b>
<b>6.</b>	<b><i>Data security, availability and sharing</i></b> .....	<b>19</b>
<b>7.</b>	<b><i>References</i></b> .....	<b>20</b>
<b>8.</b>	<b><i>Annex 1</i></b> .....	<b>21</b>

# Table of Figures

---

Figure 1: Overlap of AEGS1, AEGS2 and AEGS3 with the whole Athero-Express Biobank Study. .... 9

Figure 2: Flow diagram of sample and gene quality assessment. .... 10

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 ( $p < 0.05/n\text{genes}$ ) are shown in black and the left axis. .... 11

Figure 4: Flowchart of samples used in the analysis after quality control. .... 12

Figure 5: Distribution of the genes in the 247 significant eQTL pairs in TVS across the human genome. .... 14

Figure 6: Biological processes enriched in the genes in the 247 eQTL pairs in TVS ( $FDR < 0.25$ ). .... 14

Figure 7: Genome-wide cis-acting eQTL results at  $p_{\text{empirical}} < 0.05$ . .... 15

Figure 8: Genome-wide cis-acting eQTL results from smoking-interaction analyses. .... 15

Figure 9: Genome-wide cis-acting eQTL results from sex-interaction analyses. .... 16

Figure 10: Sex-interaction eQTL-effect at AOPEP. .... 16

Figure 11: Trans-acting mQTL results in the Athero-Express Biobank Study. .... 17

# List of Tables

---

Table 1: Final gene counts given different missingness thresholds after gene quality control. .... 10

# List of Abbreviations

---

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

# Report

---

## 1. Introduction

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), and use public data from healthy donors as available through GTEx<sup>1-3</sup> and Blueprint<sup>4</sup>.

The QTLToolKit pipeline ([github.com/swvanderlaan/QTLToolKit](https://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<sup>5</sup> 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.

## 2. Objectives

The main objective of these studies was to identify genetic variants that affect DNA methylation, and gene expression in human carotid artery plaques from the Tampere Vascular Study (TVS) and Athero-Express participants.

## 3. Methods

### 3.1. Study participants Tampere Vascular and Athero-Express Biobank Study

Vascular sample series from Tampere Vascular Study (TVS)<sup>6</sup> including femoral arteries, carotid arteries, abdominal aortas and ascending aorta were obtained during open vascular procedures during 2005–2009 from patients fulfilling the following inclusion criteria: 1) carotid endarterectomy due to asymptomatic or symptomatic and hemodynamically significant (>70%) carotid stenosis, or 2) femoral or 3) aortic endarterectomy with aortoiliac or aortobifemoral bypass due to symptomatic peripheral arterial disease. An exclusion criterion was a patient's denial to participate in the study. Gene expression was analyzed from carotid (n = 26), femoral (n = 16), abdominal aortic (n = 8) and ascending aortic (n=10) plaques. The samples were taken from patients subjected to open vascular surgical procedures in the Division of Vascular Surgery and Heart Centre, Tampere University Hospital.

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<sup>7</sup> 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<sup>7</sup>. 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<sup>8</sup>. 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 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.

## 3.2. Tampere Vascular Study methods

### 3.2.1. RNA isolation, transcriptional profiling and preprocessing

The fresh arterial tissue samples were soaked in RNALater solution (Ambion Inc., Austin, TX, USA) and isolated with Trizol reagent (Invitrogen, Carlsbad, CA, USA) and the RNeasy Kit with DNase Set (Qiagen, Valencia, CA, USA). Total-RNA was then extracted using RNeasy® Mini Kit (Qiagen). Manufacturers' instructions were followed in all isolation protocols. The quality of the RNA samples was evaluated spectrophotometrically, and the samples were stored in  $-80^{\circ}\text{C}$ . The expression levels of the arterial samples were analyzed with an Illumina HumanHT-12 v3 Expression BeadChip (Illumina, San Diego, CA, USA) analyzing 46,435 transcripts of all known genes, gene candidates and splice variants. The array was run according to given instructions by the manufacturer and scanned with the Illumina iScan system. More detailed descriptions of the methodology and qRT-PCR validation of the microarrays have been published previously<sup>9</sup>.

After background subtraction, raw intensity data was exported using the Illumina GenomeStudio software. Raw expression data were imported into R ([www.r-project.org/](http://www.r-project.org/)),  $\log_2$ -transformed and normalized by the LOESS normalization method implemented in the R/Bioconductor package lumi ([www.bioconductor.org](http://www.bioconductor.org)). LOESS normalization was selected for the data from all three tissues because it gave the best accuracy in comparison to qRT-PCR data for artery samples<sup>9</sup>. Data quality control criteria included detection of outlier arrays based on the low number of robustly expressed genes and hierarchical clustering. The artery samples ( $n=60$ ) fulfilled all data quality control criteria and had both the genotype and expression data available.

We used the Re-Annotator pipeline<sup>10</sup> to map the probe sequences provided by the array manufacturer onto the *in-silico* mRNA reference database formed from the NCBI Reference Sequence Database (hg19, build 37) gene annotation data (obtained from the UCSC Genome Browser) and the sequence of the exons. Array probe sequences that did not align to the mRNA reference database with a maximum of four mismatches were mapped to the reference genome (hg19) as described in <sup>1</sup>. Of all 48,803 probe sequences, 94% were aligned to either the mRNA reference database ( $n=30,657$ ) or if no hit was found to the reference genome ( $n=15,370$ ). Of all aligned probes, 28,415 were mapped to a distinct region (defined as a maximum of 25 bp distance between multiple hits for the same sequence) in the genome with a known autosomal gene and were included. The majority (95.3%) of these probes were aligned without any mismatches. Probes ( $n=2,246$ ) that resided in regions with common (MAF  $>1\%$ ) SNPs or indels in the European population ( $n=379$ ) of the 1000 Genomes Project (March 2012) were further excluded, leaving 26,169 reliable probes free of common polymorphisms. Moreover, only probes with an Illumina detection  $p < 0.01$  in more than 5% of the samples were included in further analysis. Therefore, the final expression data set included 13,910 probes representing 11,077 genes in the artery data.

### 3.2.2. DNA extraction, genotyping and imputation

Genomic DNA was extracted from peripheral blood leukocytes using QIAamp DNA Blood Minikit and automated biorobot M48 extraction (Qiagen, Hilden, Germany). Genotyping was done using the Illumina HumanHap660W-Quad BeadChip (Illumina, Inc, San Diego, CA) according to the manufacturer's recommendation at Helmholtz Zentrum, München, Germany. The following quality control filters were applied: GenCall score  $< 0.2$ , sample and SNP call rate  $< 0.95$ , HWE  $p$ -value  $< 10^{-6}$ , excess heterozygosity, cryptic relatedness ( $\pi$ -hat  $> 0.2$ ) and gender check. After quality control there were 538,851 SNPs available. Haplotype phasing was performed using SHAPEIT v2.r644 and imputation using SNPTEST v2.3.0 and 1000 Genomes

March 2012 haplotypes as reference. We used the resulting 6,521,532 SNPs and indels with good post imputation quality (info >0.4) and minor allele frequency (MAF) of at least 5%.

### 3.2.3. Biostatistical analysis

Expression quantitative trait loci (eQTL) analysis of the artery plaques was performed using QTLTools<sup>3</sup>. The analysis was adjusted for eight first principal components (PCs) from genome-wide expression data, two first PCs from genotype data and the four tissue types. Biological process enrichment analysis of the genes from identified eQTL-gene pairs was done using ShinyGO v0.741<sup>11</sup>.

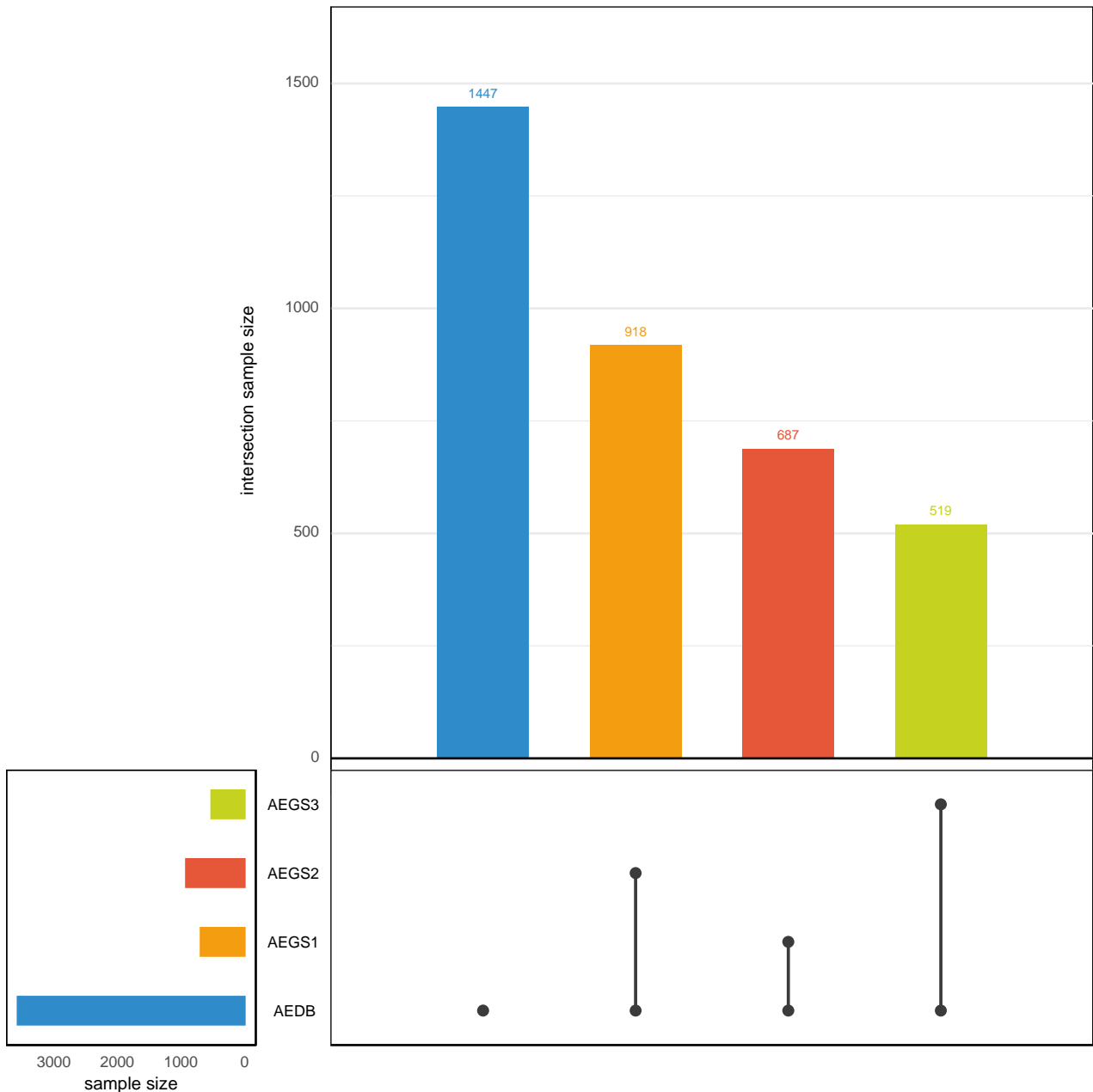
## 3.3. Athero-Express Biobank Study

### 3.3.1. 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/](http://www.eurofinsgenomics.eu/), formerly known as AROS). For the Athero-Express Genomics Study 2 (AEGS2) 954 patients (640 males, 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](http://www.helmholtz-muenchen.de)). The two first batches, AEGS1 and AEGS2, were described before<sup>12</sup>. 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](http://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<sup>12,13</sup>. 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  $\pm$  3.0 s.d., 5) relatedness ( $\pi$ -hat > 0.20), 6) Hardy–Weinberg Equilibrium (HWE  $p < 1.0 \times 10^{-3}$ ), and 7) Monomorphic SNPs ( $< 1.0 \times 10^{-6}$ ). 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](http://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/](http://imputationserver.sph.umich.edu/))<sup>14</sup>. 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.

### 3.3.2. 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<sup>15</sup>. 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 ( $\text{corrected\_count} = -4096 * (\ln(1 - (\text{raw\_count} / 4096)))$ ), normalized for sequencing depth and quantile normalized (core scripts can be found in [github.com/mmokry/bulkCEL-seq2](https://github.com/mmokry/bulkCEL-seq2) and

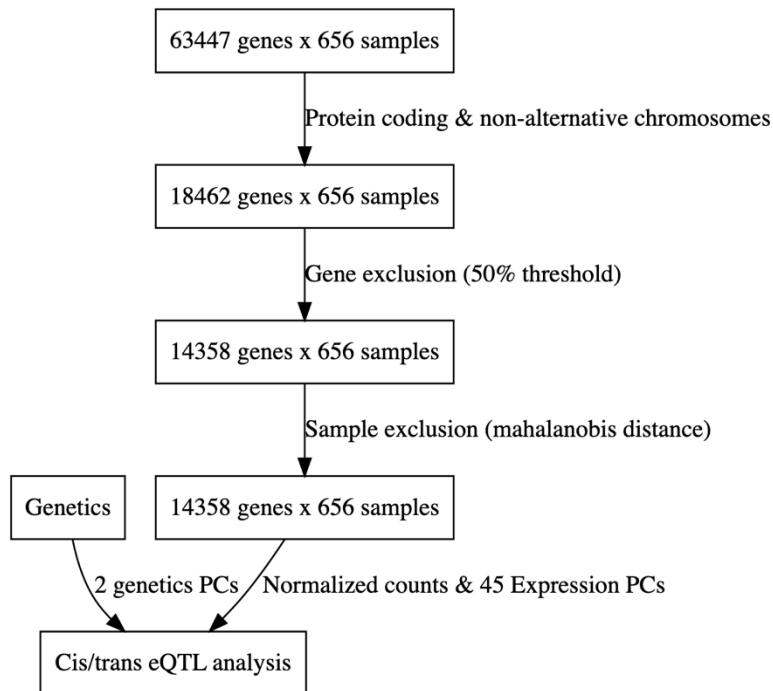
[github.com/mmokry/seurat\\_meets\\_bulk\\_AE](https://github.com/mmokry/seurat_meets_bulk_AE)). We have detected a median of 19.501 (SD = 5.874) genes per sample with at least one unique read and discarded samples (n=46) with less than 9,000 detected genes from further analysis.

### 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.

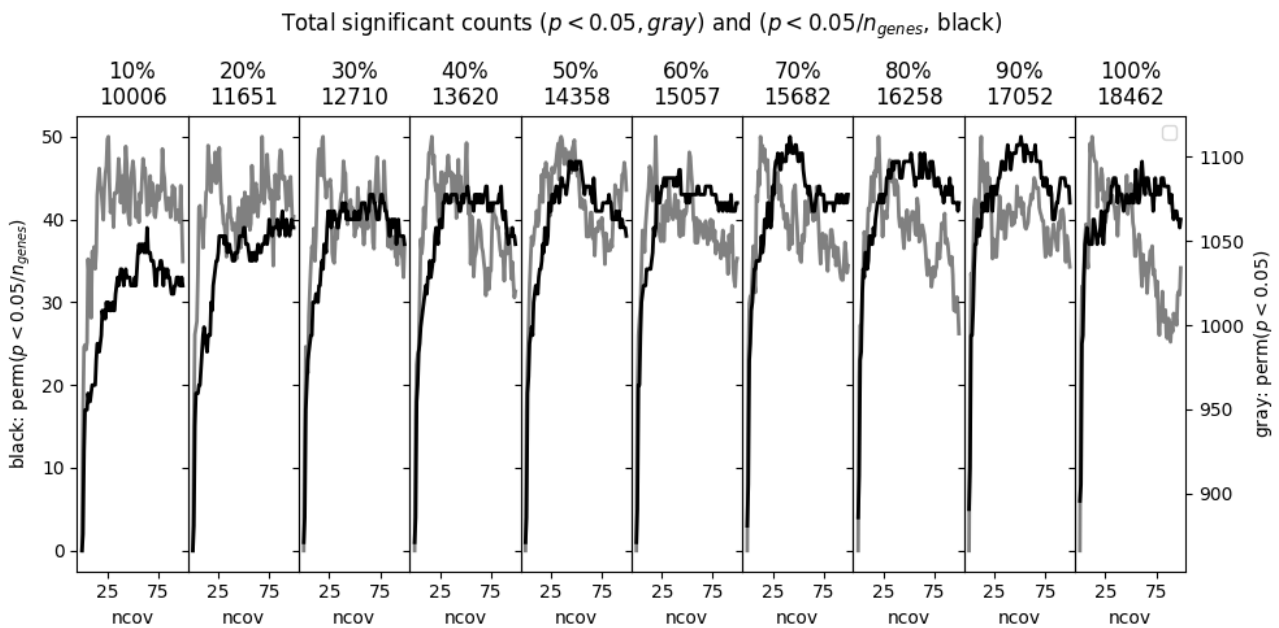
### Sample quality control

For sample covariates and subsequent sample exclusion, 2 genetics PC's and 100 PC's expression were calculated. For expression, randomized truncated PCA estimation<sup>16</sup> 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.

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

QTLtools<sup>5</sup> 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.

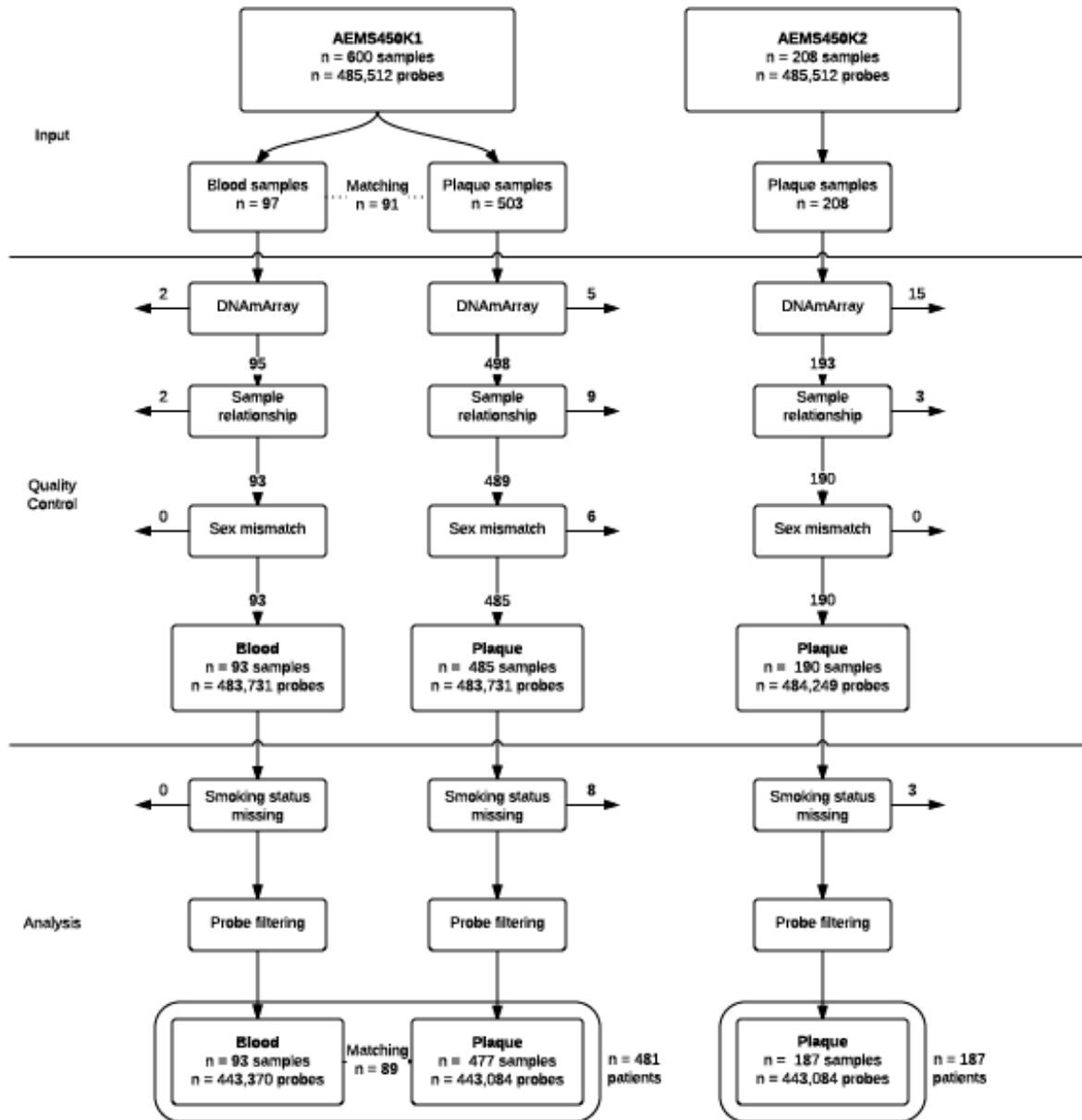
For the final cis- and trans-moQTL mapping we employed the TensorQTL<sup>17</sup> package on an Nvidia RTX6000 GPU using the previously determined missingness threshold and number of expression covariates and 10 genetic PCs for the eQTL analyses, and only genetic PCs for the moQTL analyses. For final QTL mapping, VCF files were converted to PLINK<sup>18</sup> 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 ( $p < 0.05/n_{genes}$ ) are shown in black and the left axis.

### 3.3.3. DNA extraction and methylation experiment

For the purpose of executing the DNA methylation experiments, DNA was extracted from stored plaque segments and stored blood samples of patients using standardized in-house protocols as described before in Van der Laan et al<sup>19</sup>. DNA purity and concentration were assessed using the Nanodrop 1000 system (Thermo Scientific, Massachusetts, USA). DNA concentrations were equalized at 600 ng, randomized over 96-well plates and bisulfite converted using a cycling protocol, and the EZ-96 DNA methylation kit (Zymo Research, Orange County, USA). Subsequently, DNA methylation was measured on the Infinium HumanMethylation450 Beadchip Array (HM450k, Illumina, San Diego, USA), which was performed at the Erasmus Medical Center Human Genotyping Facility in Rotterdam, the Netherlands. Processing of the sample and array was performed according to the manufacturer's protocol. Following these protocols, we isolated DNA of 509 patients across 503 plaque samples and 97 blood samples in the discovery study, called Athero-Express Methylation Study 1 (AEMS450K1). The replication study, called Athero-Express Methylation Study 2 (AEMS450K2), included 208 plaque samples (Figure 4) but was not used in this study due to insufficient overlap with the genetic data.



**Figure 4:** Flowchart of samples used in the analysis after quality control.

Flow-chart depicting the number of *input* samples, and *quality control* and *analysis* sample removal. Technical outliers were identified using DNAmArray<sup>19</sup> which includes MethylAid<sup>20</sup>. *Sample relationships* were identified through correlation of methylation data derived genotypes based on work by Chen *et al.*<sup>21</sup> and Zhou *et al.*<sup>22</sup>; where available we also compared the raw data of the 65 SNPs included on the HM450k array with those of SNP-chip derived data using the `--genome` function in PLINK<sup>18</sup>, and samples with poor correlation ( $\hat{\rho} \leq 0.8$ , indicative of possible mix-up) across these 65 SNPs were excluded. In addition, sex mismatches were identified by comparing sex-chromosomes (X and Y) beta-value distribution with the sex status derived from the medical records. *Matching* shows number of patients with both plaque and blood data in AEMS450K1 (n = 89).

### 3.3.4. Quality control of methylation data

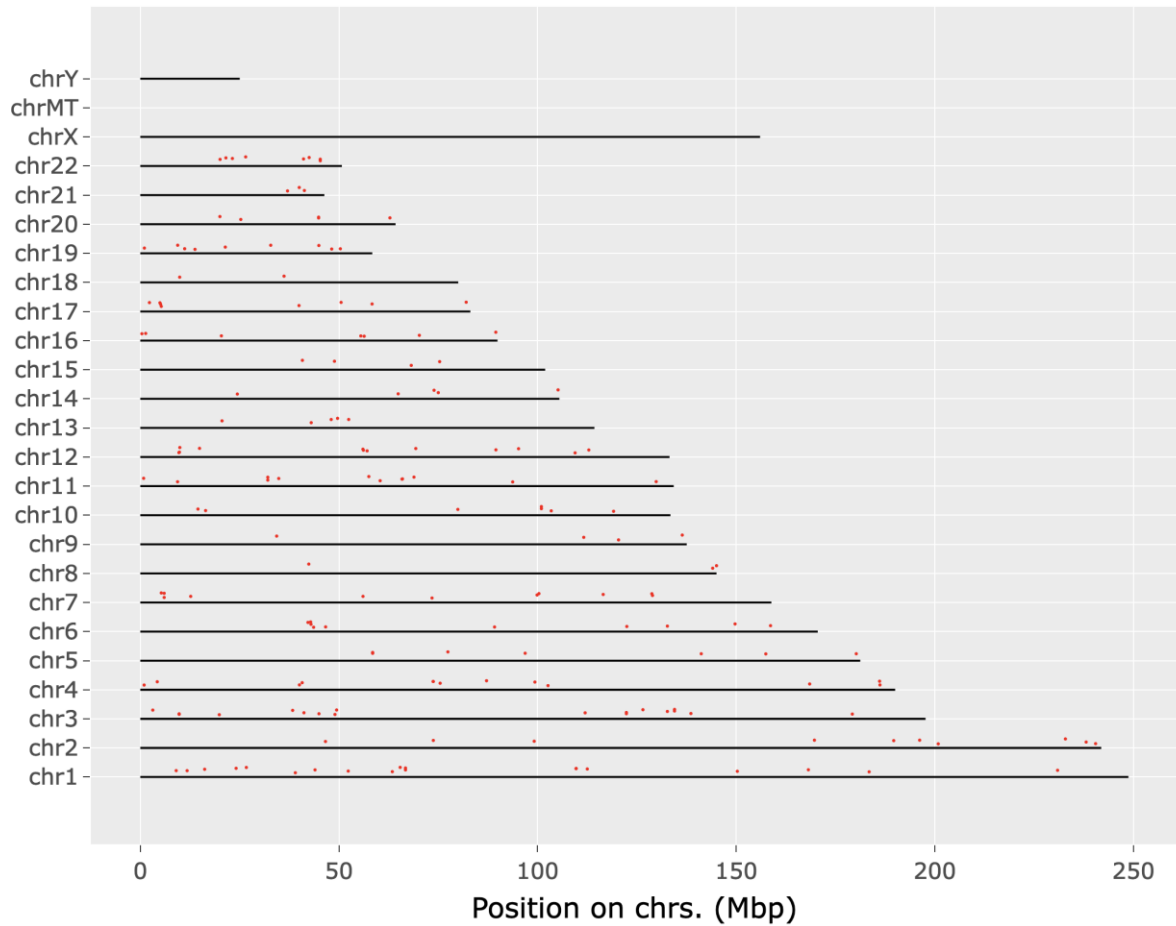
Quality control (QC) of the HM450k array data was performed following the workflow from the DNAmArray R-package<sup>23</sup> ([github.com/molepi/DNAmArray](https://github.com/molepi/DNAmArray)) using default settings, controlling for sample-dependent and probe-dependent parameters. Bisulfate conversion efficiency was determined using dedicated probes on the HM450k. We performed a principal component (PC) analysis for exploratory data analysis using the irlba R-package<sup>24</sup> ([github.com/bwlewis/irlba](https://github.com/bwlewis/irlba)) and to determine the number of PCs to use for normalization.

'Functional Normalization'<sup>25</sup> with 4 control-probe principal components was used for normalization and correction of batch effects. We computed sex based on sex-chromosome beta-value distribution and compared this to the known sex-status in order to determine possible sample mix-ups. We further assessed sample relations using beta-value extracted genotypes as calculated by the `omicsPrint` R-package ([github.com/molepi/omicsPrint](https://github.com/molepi/omicsPrint) and [bioconductor.org/packages/release/bioc/html/omicsPrint.html](https://bioconductor.org/packages/release/bioc/html/omicsPrint.html))<sup>26</sup>. Where available we also compared genotype data to the raw data of the 65 SNPs included on the HM450k array, to determine possible mix-up (as indicated by  $R \leq 0.8$  across these 65 SNPs). All samples for which sample mix-up could not be confidently ruled out were excluded from further analysis. A total of 42,428 probes were excluded based on above QC steps and the intersection of AEMS450K1 and AEMS450K2, with 443,084 probes (91.3 %) of good quality remaining. After QC, imputation of missing data (average 0.14% and 0.07% missing in AEMS450K1 and AEMS450K2, respectively) was performed using the `knn` algorithm in the `impute` R package ([bioconductor.org/packages/release/bioc/html/impute.html](https://bioconductor.org/packages/release/bioc/html/impute.html)). For analyses we also excluded probes containing SNPs or which mapped to multiple locations<sup>9</sup>. Samples with missing covariates (*i.e.* age, sex, hospital of inclusion) were excluded. After quality control, 485 plaque samples and 93 blood samples obtained from 485 unique patients were remaining in AEMS450K1. A flow-chart summarizing quality control of samples is presented in Figure 4.

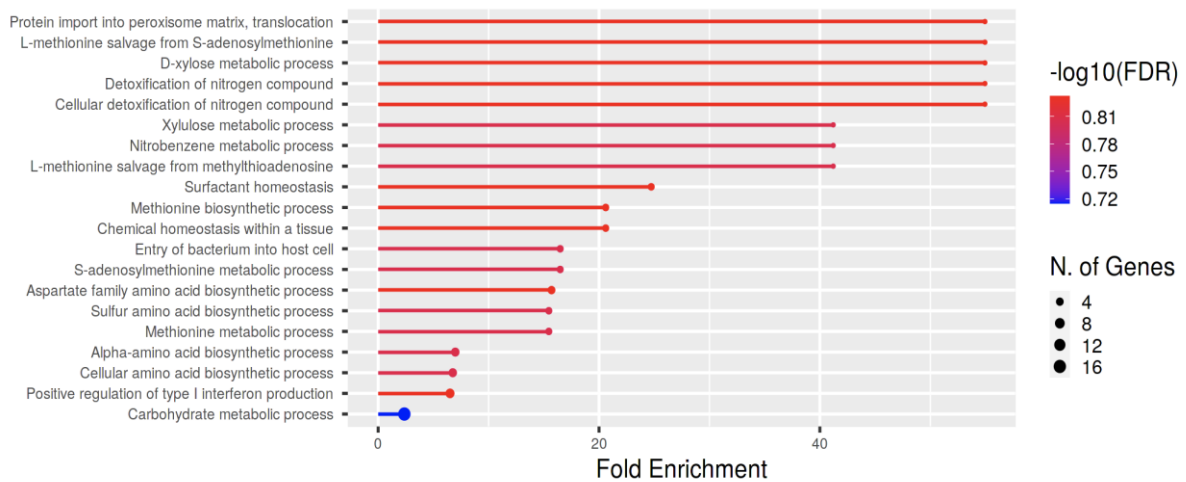
## 4. Results

### 4.1. Identification of eQTLs in carotid plaques from TVS

We identified 247 eQTL-gene pairs in TVS. Distribution of the genes in the 247 significant eQTL pairs across the human genome is shown in Figure 5. Biological processes enriched in the genes in the 247 eQTL pairs with  $FDR < 0.25$  are shown in Figure 6.



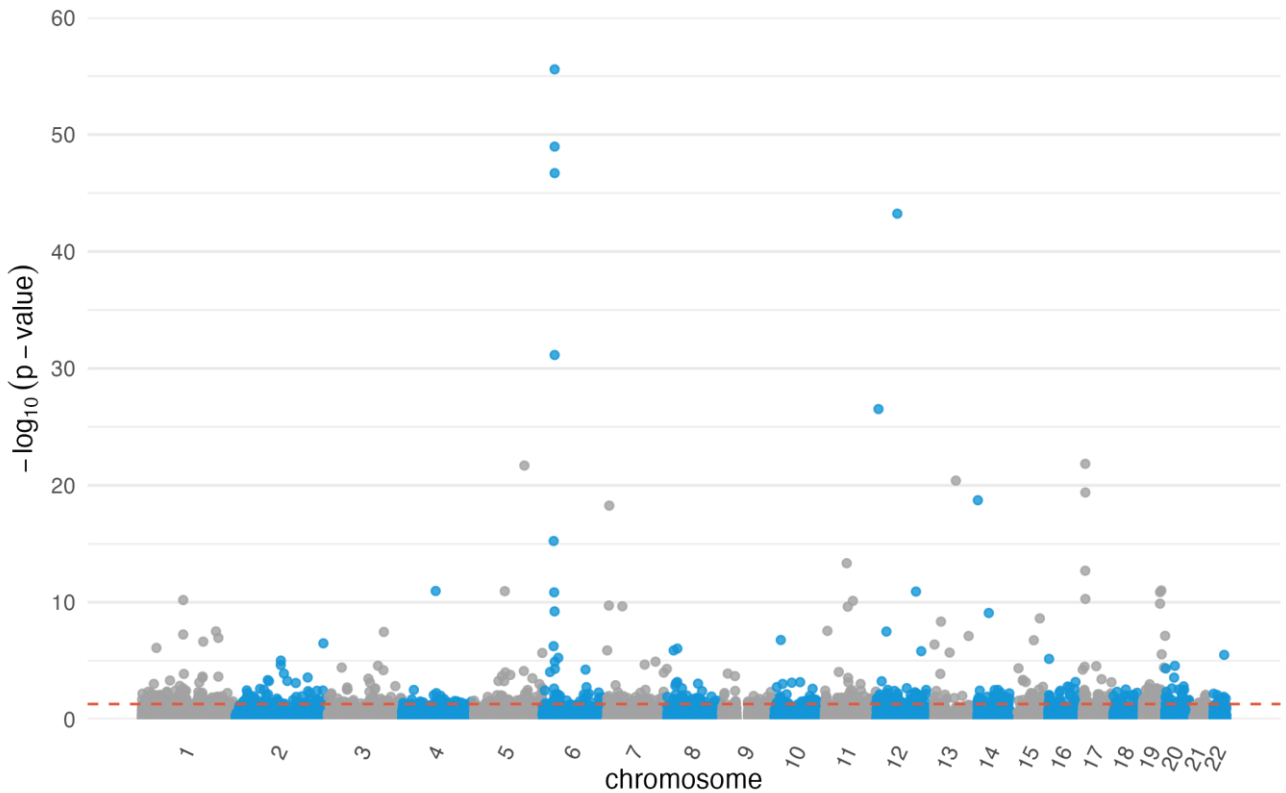
**Figure 5:** Distribution of the genes in the 247 significant eQTL pairs in TVS across the human genome.



**Figure 6:** Biological processes enriched in the genes in the 247 eQTL pairs in TVS (FDR < 0.25).

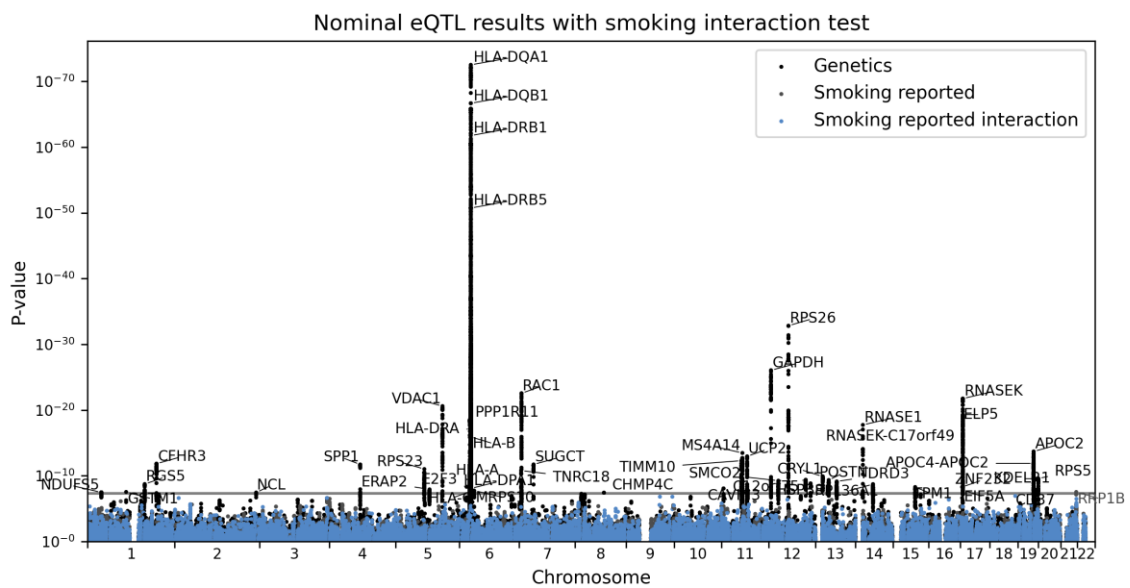
## 4.2. Identification of cis-acting eQTLs in the Athero-Express Biobank Study

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 (1,000x) and identified 951 *cis*-acting eQTLs across all 22 chromosomes at  $p_{\text{empirical}} < 0.05$  (Figure 7).

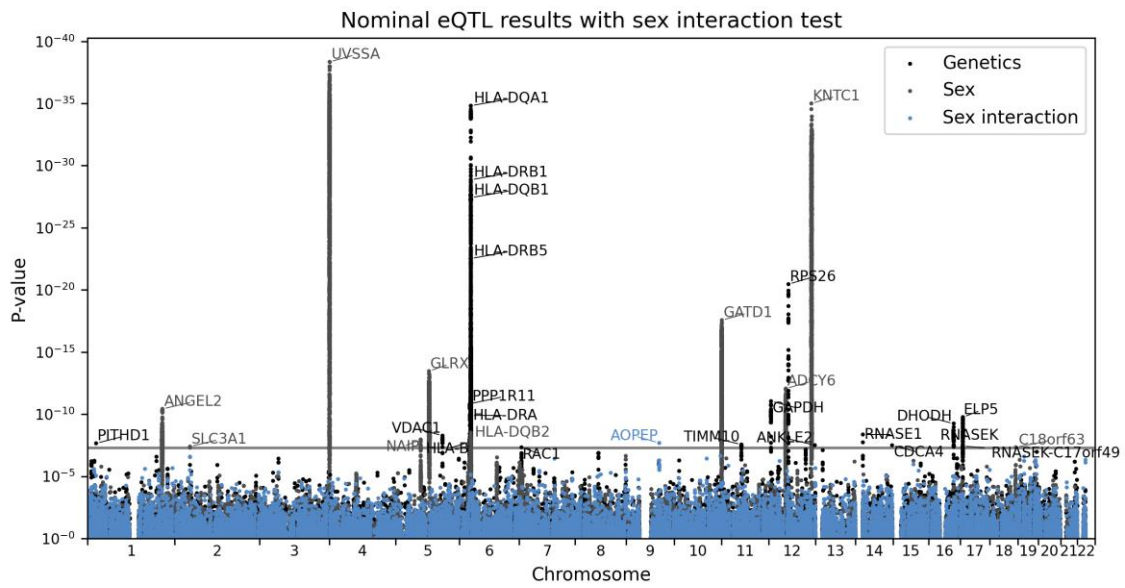


**Figure 7:** Genome-wide *cis*-acting eQTL results at  $p_{\text{empirical}} < 0.05$ .

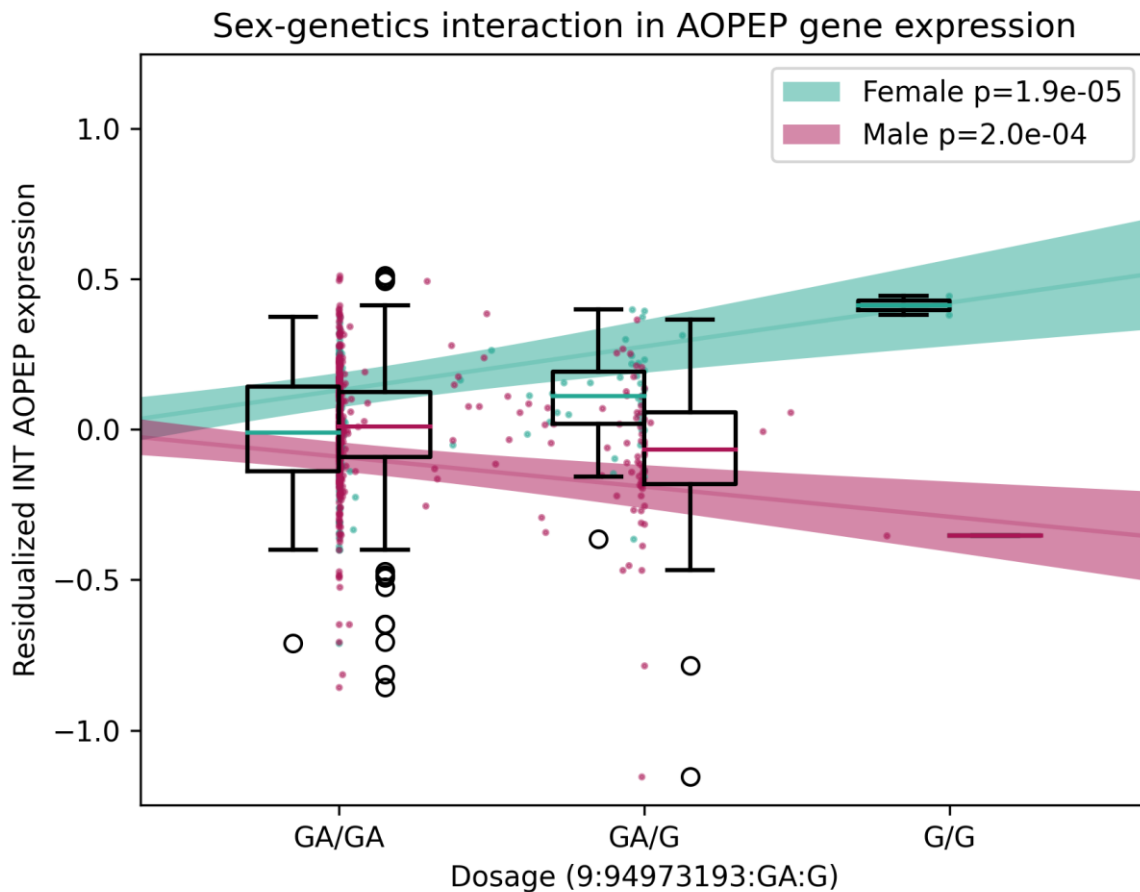
We previously showed smoking and biological sex has a profound effect on plaque molecular content<sup>27,28</sup>. Therefore, we applied interaction-analyses of smoking and sex, respectively, on SNP to gene expression. Current smoking showed no significant interaction effect (Figure 8), but sex did (Figure 9). We found one eQTL-eGene pair showing significant sex-interaction where the G-allele is associated with increased *AOPEP* expression in plaques from females, and decreased expression in males (Figure 10).



**Figure 8:** Genome-wide *cis*-acting eQTL results from smoking-interaction analyses.



**Figure 9:** Genome-wide cis-acting eQTL results from sex-interaction analyses.



**Figure 10:** Sex-interaction eQTL-effect at AOEPEP.

### 4.3. Trans-acting mQTL analyses in carotid plaques

After QC (see methods) we mapped thousands of trans-acting mQTLs across genome (Figure 11). Most notably were variants on chromosome 3, 5, and 22 that have effects on methylation sites at almost all other chromosomes implying key regulatory effects.

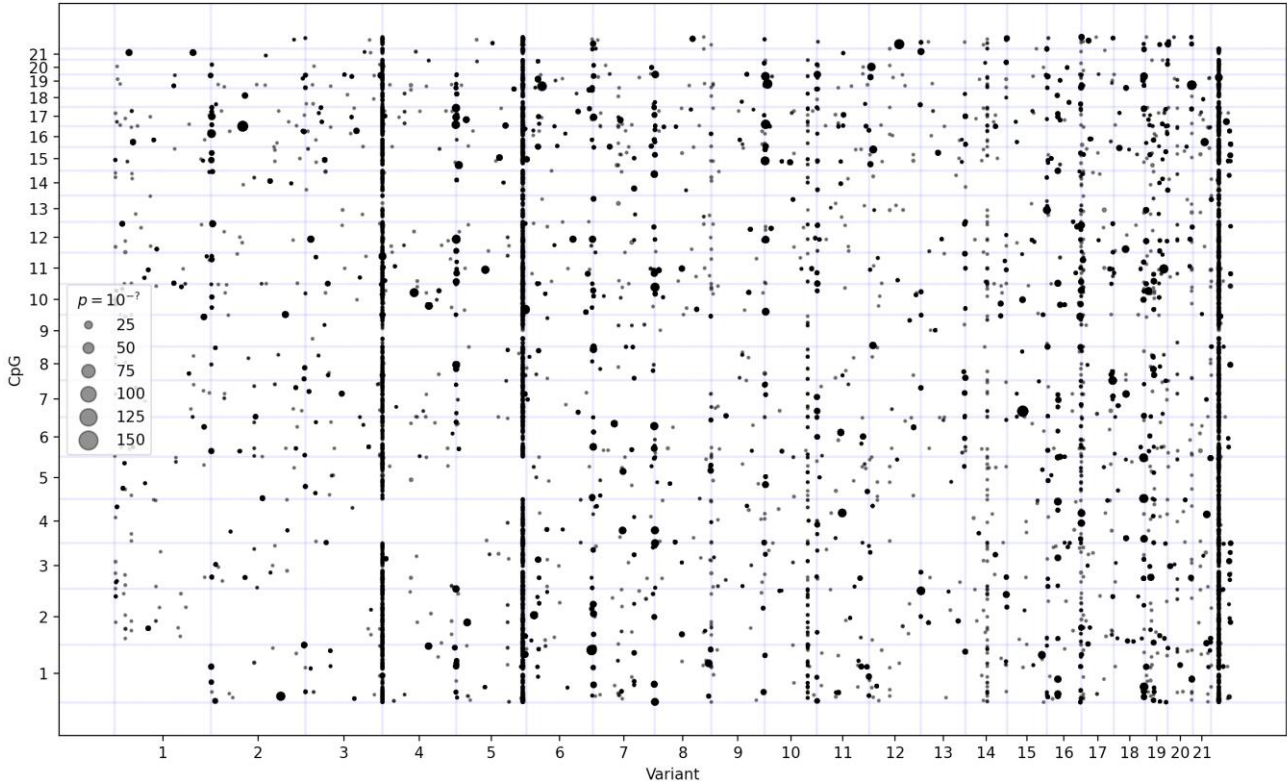


Figure 11: Trans-acting mQTL results in the Athero-Express Biobank Study.

## 5. Conclusions and Future perspectives

### 5.1. Conclusions: with respect to deliverable.

#### D1.4 Molecular QTL analyses in relevant healthy and diseased cells and tissues (M30).

Achieved. With respect to the deliverable we achieved the objective. We identified 14,284 unique eQTL-eGene pairs at  $p < 0.05$  and 951 *cis*-acting eQTLs after permutation testing (Figure 7). We also identified thousands of mQTLs in carotid plaques (Figure 11).

### 5.2. Future perspectives

The following is a future perspective how we envision the results from D1.4 can be use in work package 2 and 3 in *TO\_AITION*, but also beyond the scope of the current project.

Now that we have identified eQTLs in both TVS and AE, and mQTLs in AE, we will proceed with integrative analyses in collaboration with the *TO\_AITION* partners at UVA. Here we will study the pleiotropic effects of depression- and cardiovascular disease associated genetic variants on gene expression and DNA methylation in arterial tissues (including plaques), infer causal pathways to identify prospective biomarkers of disease, and determine driver nodes through causal network analyses.

Driving variants, which are derived from specific tissues and cells, are used individually and in aggregate (in polygenic models) as instruments for Mendelian randomisation. These instruments will be validated in summary statistics from genome-wide association studies (GWAS) of depression phenotypes, cardiovascular outcomes, relevant risk factors, and molecular and physiological biomarkers, to examine the pleiotropic effects with gene regulation and transcription. 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.

We will also construct multiplex disease hypergraphs in WP2 which enables the prediction of driver nodes for co/multimorbid disease manifestations. Driver nodes may not be the most central or highly connected nodes, but they have the largest impact on the dynamics of the entire network and can be a first indication of potential causality. We seek to add further biological insights by anchoring our causal inference in genomic variation. Longitudinal cohorts which have collected dense genomic and molecular phenotypic information about individuals contain very large volumes of high dimensional data. The combination of genomic data with this broad range of phenotypes in the context of longitudinal studies is a rich substrate for causal inference. The inclusion of genomic data along with molecular phenotypes and longitudinal clinical trajectories in these representations allows for application of methods such as Mendelian randomization for causal inference through multiple 'omic' and clinical layers. In this way, pathways within the data structure can be mapped from genetic variants, through the plasma proteome and metabolome, to clinical biomarkers, events and disease progression. Pathways within the inevitably complex biological networks involved in these cardiovascular disease and depression are unlikely always to be linear. Through incorporating MR with machine learning methods, the pleiotropy of driver nodes in a network can be navigated. For example, a method using a mixture-of-experts framework addresses the challenge of horizontal pleiotropy of MR genetic instruments and generates estimates of causality between nodes in an a biologically interpretable way<sup>29</sup>. These analyses are planned for Q1-Q3 of 2023.

## 6. 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 (Annex 1) 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<sup>7</sup> 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.

## 7. References

1. Nica, A. C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* **6**, e1000895 (2010).
2. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
3. Ongen, H. *et al.* Estimating the causal tissues for complex traits and diseases. *Nat. Genet.* (2017) doi:10.1038/ng.3981.
4. Stunnenberg, H. G., International Human Epigenome Consortium & Hirst, M. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* **167**, 1145–1149 (2016).
5. Delaneau, O. *et al.* A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* **8**, 15452 (2017).
6. Oksala, N. *et al.* ADAM-9, ADAM-15, and ADAM-17 are upregulated in macrophages in advanced human atherosclerotic plaques in aorta and carotid and femoral arteries--Tampere vascular study. *Ann. Med.* **41**, 279–290 (2009).
7. 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).
8. van Lammeren, G. W. *et al.* Time-Dependent Changes in Atherosclerotic Plaque Composition in Patients Undergoing Carotid Surgery. *Circulation* **129**, 2269–2276 (2014).
9. Raitoharju, E. *et al.* A comparison of the accuracy of Illumina HumanHT-12 v3 Expression BeadChip and TaqMan qRT-PCR gene expression results in patient samples from the Tampere Vascular Study. *Atherosclerosis* **226**, 149–152 (2013).
10. Arloth, J., Bader, D. M., Röh, S. & Altmann, A. Re-Annotator: Annotation pipeline for microarray probe sequences. *PLoS One* **10**, e0139516 (2015).
11. Ge, S. X., Jung, D. & Yao, R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics* **36**, 2628–2629 (2020).
12. 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).
13. Laurie, C. C. *et al.* Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet. Epidemiol.* **34**, 591–602 (2010).
14. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
15. Hashimshony, T. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).
16. Halko, N., Martinsson, P.-G. & Tropp, J. A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *arXiv [math.NA]* (2009).
17. Taylor-Weiner, A. *et al.* Scaling computational genomics to millions of individuals with GPUs. *Genome Biol.* **20**, 228 (2019).
18. 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).
19. van der Laan, S. W. *et al.* Variants in ALOX5, ALOX5AP and LTA4H are not associated with atherosclerotic plaque phenotypes: the Athero-Express Genomics Study. *Atherosclerosis* **239**, 528–538 (2015).
20. van Iterson, M. *et al.* MethylAid: visual and interactive quality control of large Illumina 450k datasets. *Bioinformatics* **30**, 3435–3437 (2014).
21. Chen, Y.-A. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8**, 203–209 (2013).
22. Zhou, W., Laird, P. W. & Shen, H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res.* **45**, e22 (2017).

23. van Iterson, M. *Quality control, probe/sample filtering and normalization of Infinium HumanMethylation450 BeadChip data: "The Leiden Approach."* (2016). doi:10.5281/zenodo.158908.
24. Baglama, J. & Reichel, L. Augmented Implicitly Restarted Lanczos Bidiagonalization Methods. *SIAM J. Sci. Comput.* **27**, 19–42 (2005).
25. Fortin, J.-P. *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol.* **15**, 503 (2014).
26. van Iterson, M. *et al.* omicsPrint: detection of data linkage errors in multiple omics studies. *Bioinformatics* **34**, 2142–2143 (2018).
27. Siemelink, M. A. *et al.* Smoking is Associated to DNA Methylation in Atherosclerotic Carotid Lesions. *Circ Genom Precis Med* **11**, e002030 (2018).
28. Hartman, R. J. G. *et al.* Sex-dependent gene regulation of human atherosclerotic plaques by DNA methylation and transcriptome integration points to smooth muscle cell involvement in women. *bioRxiv* 2021.01.28.428414 (2021) doi:10.1101/2021.01.28.428414.
29. Hemani, G. *et al.* Automating Mendelian randomization through machine learning to construct a putative causal map of the human phenome. *bioRxiv* (2017) doi:10.1101/173682.

## 8. Annex 1

### Plan Overview

#### A Data Management Plan created using DMPonline

#### Title: Athero-Express Biobank Study

Creator: Sander W. van der Laan

Principal Investigator: Sander W. van der Laan, Gerard Pasterkamp

Data Manager: Saskia Haitjema

Project Administrator: Sander W. van der Laan, Gerard Pasterkamp

Affiliation: Other

Template: UMC Utrecht DMP

ORCID iD: 0000-0001-6888-1404

ORCID iD: 0000-0001-5345-1022

#### Project abstract:

In clinical practice, biological markers are not available to routinely assess the progression of atherosclerotic disease or the development of restenosis following endarterectomy or catheter-based interventions. Endarterectomy procedures provide an opportunity to study mechanisms of restenosis and progression of atherosclerotic disease since atherosclerotic tissue is obtained. Athero-Express is an ongoing prospective study, initiated in 2002, with the objective to investigate the etiological value of plaque characteristics for long term outcome. Patients are included who undergo an endarterectomy of the carotid artery. At baseline blood

is withdrawn, patients fill in an extensive questionnaire and diagnostic examinations are performed. Atherosclerotic plaques are freshly harvested, immunohistochemically stained and examined for the presence of macrophages, smooth muscle cells, collagen and fat. Parts of the atherosclerotic plaques are freshly frozen to study protease activity and protein and RNA expressions. Patients undergo a duplex follow up to assess procedural restenosis (primary endpoint) at 3 months, 1 year and 2 years. Secondary endpoints encompass major adverse cardiovascular events. In the future, the creation of this biobank with atherosclerotic specimen will allow the design of cross-sectional and follow up studies with the



objective to investigate the expression of newly discovered genes and proteins and their interaction with patients and plaque characteristics in the progression of atherosclerotic disease. Objective is to include 1000-1200 patients in 5 years. In January 2004, 289 patients had been included. It is expected that 250 patients will be included yearly. To date around 3500 patients were included.

ID: 74856

Start date: 01-04-2002

Last modified: 27-01-2023