*Sander W. van der Laan; Lennart L.P. Landsmeer*

*UMC Utrecht | s.w.vanderlaan-2@umcutrecht.nl*

*molecular QTL analyses*

An Athero-Express Biobank Study Project

Executive Summary

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 QTLTools5 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 Athero-Express Biobank Study (AE). 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.

Table of Contents

[1. Objectives 2](#_Toc126854198)

[2. Methods 2](#_Toc126854199)

[2.1. Study participants Athero-Express Biobank Study 2](#_Toc126854200)

[2.2. Athero-Express Biobank Study 2](#_Toc126854201)

[2.2.1. DNA extraction, genotyping and imputation 2](#_Toc126854202)

[2.2.2. RNA isolation, transcriptional profiling and preprocessing 4](#_Toc126854203)

[2.2.3. DNA extraction and methylation experiment 6](#_Toc126854204)

[2.2.4. Quality control of methylation data 8](#_Toc126854205)

[3. Results 8](#_Toc126854206)

[3.1. Identification of cis-acting eQTLs in the Athero-Express Biobank Study 8](#_Toc126854207)

[3.2. Trans-acting mQTL analyses in carotid plaques 11](#_Toc126854208)

# Objectives

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

# Methods

## Study participants Athero-Express Biobank Study

The Athero-Express Biobank Study (AE, approved and registered under number TME/C-01.18) is an ongoing cohort study started in 2002(Verhoeven et al. 2004) 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(Verhoeven et al. 2004). Briefly, blood and plaque samples are obtained during surgery, and routinely stored at ‐80°C and plaque material is used for standardized (immuno)histochemical analysis8. 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 study was approved by the respective hospitals’ Ethics Committees. Only patients providing written informed consent are included and the study conform to the Declaration of Helsinki.

## Athero-Express Biobank Study

### 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 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](http://www.helmholtz-muenchen.de)). The two first batches, AEGS1 and AEGS2, were described before12. 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 AEGS312,13. 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 < 1.0×10−3), and 7) Monomorphic SNPs (< 1.0×10−6).  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 (GRCh37/hg19) using the liftOver tool from UCSC ([genome.ucsc.edu/cgi-bin/hgLiftOver](https://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/](https://imputationserver.sph.umich.edu/))14. 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 principal component analyses (PCA) and using data from the 1000G phase 3.



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

### 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 method15. 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](file:///Users/slaan3/Library/Mobile%20Documents/com~apple~CloudDocs/Genomics/Projects/%23TO_AITION/Administration/Reports/D1.4%20Report/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 replaces 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 : 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 |

Diagram

Description automatically generated

Figure : Flow diagram of sample and gene quality assessment. PCs: principal components.

#### Sample quality control

For sample covariates and subsequent sample exclusion, 2 genetics principal components (PCs) and 100 expression PCs were calculated. For expression, randomized truncated principal component analysis (PCA) estimation16 was used due to the large dataset. For the first 2 expression covariates, the sample Mahalanobis distance was calculated, and samples below the Chi2 (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.

### 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 al19. 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 3) but was not used in this study due to insufficient overlap with the genetic data.

**Diagram, schematic

Description automatically generated**

Figure 3: 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 DNAmArray19 which includes MethylAid20. *Sample relationships* were identified through correlation of methylation data derived genotypes based on work by Chen *et al.*21 and Zhou *et al.*22; 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 PLINK18, and samples with poor correlation (pi-hat ≤ 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).

### Quality control of methylation data

Quality control (QC) of the HM450k array data was performed following the workflow from the DNAmArray R-package23 ([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-package24 ([github.com/bwlewis/irlba](https://github.com/bwlewis/irlba)) and to determine the number of PCs to use for normalization. ‘Functional Normalization’25 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))26. 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 ≤ 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](http://bioconductor.org/packages/release/bioc/html/impute.html)). For analyses we also excluded probes containing SNPs or which mapped to multiple locations[9](https://paperpile.com/c/uBsHaZ/C8nOR).

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 3.

### *Cis*- and *trans*-molecular QTL analsyes

QTLtools5 was used to select the missingness threshold (10-100%) of the gene counts 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 (100-10,000x) 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 inFigure 4***.***

A picture containing chart

Description automatically generated

Figure 4: 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.

A similar approach was taken to calculate the number methylation covariates. Due to the number of CpG sites a random subset of 1 in 20 methylation sites was taken. QTLtools was used to calculate methylation PCs using the filtered genetic data as described above and QTLtools was run in a cis permutation mode (100-10,000x) with a 1Mb window. Next, in steps of 5 up to 395 covariates were included and 2 genetic PCs to obtain genome-wide adjusted results; the number of significant hits flattened off at 40 methylation PCs.

Ultimately, for the final *cis*- and *trans*-molQTL mapping we employed the TensorQTL(Taylor-Weiner et al. 2019) package on an Nvidia RTX6000 GPU. We included 2 genetic PCs and 45 expression PCs for the eQTL analyses and used the previously determined missingness threshold (50%) for the gene counts, and 2 genetic PCs and 40 methylation PCs for the mQTL analyses; age and sex were not included. For final QTL mapping, the above generated VCF-files (MAF > 0.03, INFO > 0.4) were converted to PLINK18 BED-file format using the default settings of PLINK2 (2.0.0a2lm, <https://www.cog-genomics.org/plink/2.0/>) and as required by TensorQTL.

Depending on the file size, results were saved in parquet- or txt-format.

# Results

## 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 pemperical < 0.05 (Figure 7).

**A picture containing graphical user interface

Description automatically generated**

Figure : Genome-wide *cis*-acting eQTL results at pemperical < 0.05.

We previously showed smoking and biological sex has a profound effect on plaque molecular content27,28. 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).

**Timeline

Description automatically generated with low confidence**

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

**Chart

Description automatically generated**

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

**Diagram

Description automatically generated with medium confidence**

Figure : Sex-interaction eQTL-effect at AOPEP.

## 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.

**Chart, scatter chart

Description automatically generated**

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

References

Baglama, J., and L. Reichel. 2005. “Augmented Implicitly Restarted Lanczos Bidiagonalization Methods.” *SIAM Journal of Scientific Computing* 27 (1): 19–42.

Chen, Yi-An, Mathieu Lemire, Sanaa Choufani, Darci T. Butcher, Daria Grafodatskaya, Brent W. Zanke, Steven Gallinger, Thomas J. Hudson, and Rosanna Weksberg. 2013. “Discovery of Cross-Reactive Probes and Polymorphic CpGs in the Illumina Infinium HumanMethylation450 Microarray.” *Epigenetics: Official Journal of the DNA Methylation Society* 8 (2): 203–9.

Das, Sayantan, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E. Locke, Alan Kwong, Scott I. Vrieze, et al. 2016. “Next-Generation Genotype Imputation Service and Methods.” *Nature Genetics* 48 (10): 1284–87.

Delaneau, Olivier, Halit Ongen, Andrew A. Brown, Alexandre Fort, Nikolaos I. Panousis, and Emmanouil T. Dermitzakis. 2017. “A Complete Tool Set for Molecular QTL Discovery and Analysis.” *Nature Communications* 8 (May): 15452.

Fortin, Jean-Philippe, Aurélie Labbe, Mathieu Lemire, Brent W. Zanke, Thomas J. Hudson, Elana J. Fertig, Celia Mt Greenwood, and Kasper D. Hansen. 2014. “Functional normalization of 450k methylation array data improves replication in large cancer studies.” *Genome Biology* 15 (11): 503.

Halko, Nathan, Per-Gunnar Martinsson, and Joel A. Tropp. 2009. “Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions.” *ArXiv [Math.NA]*. arXiv. http://arxiv.org/abs/0909.4061.

Hartman, Robin J. G., Marten A. Siemelink, Saskia Haitjema, Koen F. Dekkers, Lotte Slenders, Arjan Boltjes, Michal Mokry, et al. 2021. “Sex-Dependent Gene Regulation of Human Atherosclerotic Plaques by DNA Methylation and Transcriptome Integration Points to Smooth Muscle Cell Involvement in Women.” *BioRxiv*. https://doi.org/10.1101/2021.01.28.428414.

Hashimshony, Tamar, Naftalie Senderovich, Gal Avital, Agnes Klochendler, Yaron de Leeuw, Leon Anavy, Dave Gennert, et al. 2016. “CEL-Seq2: Sensitive Highly-Multiplexed Single-Cell RNA-Seq.” *Genome Biology* 17 (1): 77.

Iterson, Maarten van. 2016. *Quality Control, Probe/Sample Filtering and Normalization of Infinium HumanMethylation450 BeadChip Data: “The Leiden Approach.”* https://doi.org/10.5281/zenodo.158908.

Iterson, Maarten van, Davy Cats, Paul Hop, BIOS Consortium, Bastiaan T. Heijmans, and Oliver Stegle. 2018. “OmicsPrint: Detection of Data Linkage Errors in Multiple Omics Studies.” *Bioinformatics*  34 (12): 2142–43.

Iterson, Maarten van, Elmar W. Tobi, Roderick C. Slieker, Wouter den Hollander, René Luijk, P. Eline Slagboom, and Bastiaan T. Heijmans. 2014. “MethylAid: visual and interactive quality control of large Illumina 450k datasets.” *Bioinformatics*  30 (23): 3435–37.

Laan, Sander W. van der, Hassan Foroughi Asl, Pleunie van den Borne, Jessica van Setten, M. E. Madeleine van der Perk, Sander M. van de Weg, Arjan H. Schoneveld, et al. 2015. “Variants in ALOX5, ALOX5AP and LTA4H Are Not Associated with Atherosclerotic Plaque Phenotypes: The Athero-Express Genomics Study.” *Atherosclerosis* 239 (2): 528–38.

Laan, Sander W. van der, Marten A. Siemelink, Saskia Haitjema, Hassan Foroughi Asl, Ljubica Perisic, Michal Mokry, Jessica van Setten, et al. 2018. “Genetic Susceptibility Loci for Cardiovascular Disease and Their Impact on Atherosclerotic Plaques.” *Circulation. Genomic and Precision Medicine* 11 (9): e002115.

Lammeren, Guus W. van, Hester M. den Ruijter, Joyce E. P. Vrijenhoek, Sander W. van der Laan, Evelyn Velema, Jean-Paul P. M. de Vries, Dominique P. V. de Kleijn, et al. 2014. “Time-Dependent Changes in Atherosclerotic Plaque Composition in Patients Undergoing Carotid Surgery.” *Circulation* 129 (22): 2269–76.

Laurie, Cathy C., Kimberly F. Doheny, Daniel B. Mirel, Elizabeth W. Pugh, Laura J. Bierut, Tushar Bhangale, Frederick Boehm, et al. 2010. “Quality control and quality assurance in genotypic data for genome-wide association studies.” *Genetic Epidemiology* 34 (6): 591–602.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, et al. 2007. “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses.” *American Journal of Human Genetics* 81 (3): 559–75.

Siemelink, Marten A., Sander W. van der Laan, Saskia Haitjema, Ian D. van Koeverden, Jacco Schaap, Marian Wesseling, Saskia C. A. de Jager, et al. 2018. “Smoking Is Associated to DNA Methylation in Atherosclerotic Carotid Lesions.” *Circulation. Genomic and Precision Medicine* 11 (9): e002030.

Taylor-Weiner, Amaro, François Aguet, Nicholas J. Haradhvala, Sager Gosai, Shankara Anand, Jaegil Kim, Kristin Ardlie, Eliezer M. Van Allen, and Gad Getz. 2019. “Scaling Computational Genomics to Millions of Individuals with GPUs.” *Genome Biology* 20 (1): 228.

Verhoeven, Bart A. N., Evelyn Velema, Arjan H. Schoneveld, Jean Paul P. M. de Vries, Peter de Bruin, Cees A. Seldenrijk, Dominique P. V. de Kleijn, et al. 2004. “Athero-Express: Differential Atherosclerotic Plaque Expression of MRNA and Protein in Relation to Cardiovascular Events and Patient Characteristics. Rationale and Design.” *European Journal of Epidemiology* 19 (12): 1127–33.

Zhou, Wanding, Peter W. Laird, and Hui Shen. 2017. “Comprehensive Characterization, Annotation and Innovative Use of Infinium DNA Methylation BeadChip Probes.” *Nucleic Acids Research* 45 (4): e22.