AtheroSpectrum reveals novel macrophage foam cell gene signatures associated with atherosclerotic cardiovascular disease risk

SUPPLEMENTAL MATERIALS

Expanded Methods
Supplemental Table: 1
Supplemental Figures: 6

Expanded Methods

MESA Participants (Additional information)

MESA is a longitudinal study of subclinical ASCVD and risk factors that predict progression to clinically overt ASCVD or progression of subclinical disease.

The first clinic visits occurred from 2000 to 2002 for 6,814 participants recruited from 6 field centers across the United States, and all participants were free of clinical CVD at Exam 1. Fasting blood samples were drawn and processed using a standardized protocol and sent for measurement of lipid levels[36,37]. CVE including stroke and MI were adjudicated by a MESA committee of cardiologists, physician epidemiologists, and neurologists who provided a detailed description of the CVE adjudication process[38]. In summary, transcriptome profiles were generated from purified monocyte samples of 1269 randomly selected MESA participants of Exam 5 (April 2010–February 2012) from four MESA sites (John Hopkins University, Baltimore, MD; Columbia University, New York, NY; University of Minnesota, Minneapolis, MN; and Wake Forest University, Winston-Salem, NC). The study protocol was approved by the Institutional Review Boards of the four institutions. All participants signed informed consent[4].

GTEx

GTEx data were obtained from gtexportal.org (dbGaP Accession phs000424.v8.p2). Samples were classified as normal or atherogenic artery by histological analyses provided by GTEx. In summary, histological analysis revealing mild to severe atherosclerosis with pathological notes indicating tissues with plaque, atherosclerosis, and/or atherosclerotic plaque were considered atherogenic artery tissue (n=362); histological analysis revealing no atherosclerosis and notes which indicated “no lesions”, “no visible atherosis”, “no plaques” were considered normal artery tissue (n=537).

Multivariable regression

Gene expression levels and continuous variables (LDL-C, fasting glucose, age, MPI, etc.) of MESA participants were rescaled by standard scaler ($\frac{x-mean}{std}$) for model development. Binary and categorical variables were inputted as factors (e.g., female/male, non-smoker/smoker, non-CVE/CVE).

Multivariable logistic regressions were performed between CVE and the gene-set being investigated plus MPI and selected MESA features (sex, age, hypertension medication, lipid-lowering medication, diabetes 2003 ADA fasting glucose criteria, and diastolic blood pressure) using R (version 3.5) and the function *glm* included in R base.

Transcriptome analysis of blood monocytes

Blood monocytes of MESA participants were isolated and profiled by microarray as previously reported[4]. In summary, blood was collected in sodium heparin-containing Vacutainer CPT cell separation tubes (Becton Dickinson, Rutherford, NJ), and subsequently, monocytes were isolated with anti-CD14 antibody-coated magnetic beads, using an autoMACS automated magnetic separation unit (Miltenyi Biotec, Bergisch Gladbach, Germany). The purity of monocytes was > 90% and validated using flow cytometry. RNA was isolated from samples using the AllPrep DNA/RNA Mini Kit (Qiagen, Inc., Hilden, Germany). RNAs with RNA Integrity (RIN) scores < 9.0 were excluded from global expression microarrays. Transcriptomes of the monocytes were profiled using the Illumina HumanHT-12 v4 Expression BeadChip and Illumina Bead Array Reader.

Signaling enrichment analyses

Signaling pathway and cell function enrichment were analyzed using Ingenuity Pathway Analysis (IPA) (Qiagen) and NIH DAVID Bioinformatics Resources (david-d.ncifcrf.gov).

Calculation of PCE 2013 risk score

The PCE 2013 scores for assessing ASCVD of the MESA cohort were calculated using R according to the Pooled Cohort Equations in 2013 ACC/AHA guideline[26] for sex, race/ethnicity, age, total cholesterol levels (mg/dL), HDL-C (mg/dL), systolic blood pressure (mmHg), treatment for high blood pressure, smoking status, and diabetes.

Macrophage transcriptome profile similarity to atherosclerosis or healthy reference populations

Two sets of reference transcriptome profiles were generated as the average Reads Per Kilobase of transcript, per Million mapped reads (RPKM) of each FSG gene in artery macrophages from atherosclerosis plaques of atherogenic diet mice (n=3) or chow diet mice (n=3) (GSE116239)[20], and termed the chow-average reference ($Chow_{ref}$) and the atherosclerotic-average reference ($Athero_{ref}$). Similarities between a macrophage transcriptome profile and the two references were calculated similarly as previously reported[23] with R using a method modified from Pearson's correlation (r_{chow} and $r_{athero,}$ equations listed below).

The FSG were first adapted to the macrophage transcriptomes, resulting in (sub)sets of FSG (FSG_{sub}) that were expressed in the macrophages to be computed. To emphasize the changes in gene expression during macrophage foam cell development under various conditions, we performed a gene set centering adjustment to focus on fold-change difference of each individual gene in all samples within a study. The average expression level of a given gene (S_{avg}) was calculated as the mean of this gene in each dataset (S) in the whole study. S_{ctr} is calculated as the centered value of this gene: let S be the original expression level, then the centered expression S_{ctr} will be:

$$S_{ctr} = S - S_{avg}$$

Accordingly, the two references, $Chow_{ref}$ and $Athero_{ref}$, were adjusted by centering each gene expression level as described, and termed $Chow_{ctr}$ and $Athero_{ctr}$. The similarities between a macrophage transcriptome profile (gene expression levels in RPKM/FPKM/UMI, etc.) and the centered references were calculated as the following: i represents a gene ID among the FSG set; n is the number of genes in FSG_{sub}.

$$r_{chow} = \frac{\sum_{i=1}^{n} S_{ctr\,i} \times Chow_{ctr\,i}}{\sqrt{\sum_{i=1}^{n} S_{ctr\,i}^2}\sqrt{\sum_{i=1}^{n} Chow_{ctr\,i}^2}}$$

and

$$r_{athero} = \frac{\sum_{i=1}^{n} S_{ctr\,i} \times Athero_{ctr\,i}}{\sqrt{\sum_{i=1}^{n} S_{ctr\,i}^2}\sqrt{\sum_{i=1}^{n} Athero_{ctr\,i}^2}}$$

Adjusted correlation values for each comparison are provided as r_{chow} and r_{athero} for each macrophage transcriptome profile. We calculate r_{chow} and r_{athero} for each cell in a single-cell RNA sequencing (scRNA-seq) data set or bulk RNA-seq profile in a given experiment.

Screening of foam cell signature genes (FSG)

AtheroSpectrum computes two indices, the Macrophage Polarization Index (MPI) and the Macrophage-Derived Foam cell Index (MDFI). MPI was inherited from MacSpectrum and was calculated as previously described[23]; the only difference is that in AtheroSpectrum the MPI was scaled 0 to 100, instead of -50 to 50 as in MacSpectrum.

To create MDFI, we first screened 500 FSG. For identification of foam cell signature genes (FSG), we conducted differential expression of genes in macrophages from atherosclerotic

plaques of atherogenic diet mice (n=3) vs. artery macrophages from chow diet mice (n=3) (GSE116239)[20] using edgeR package. From the 1118 genes that were significantly different (p<0.05) between atherogenic and chow diet mice, we excluded any genes in our previously reported 500 polarization signature genes (PSG) or 435 activation-induced macrophage differentiation signature genes (AMDSG), as those genes are also involved in macrophage inflammation or activation-induced differentiation (e.g., maturation from monocyte to macrophage)[23]. From this gene set, we took the top 500 as the FSG. The 500 FSG represented the most significantly changed genes in macrophage foam cell development. Further information for macrophage transcriptome profile similarity analyses are detailed in Expanded Methods.

Machine learning-powered CVD risk signature gene identification

To identify ASCVD risk signature genes from the 2209 genes enriched in the pathological foaming program, based on characteristics of the dataset we designed a 7-step, self-optimizing machine learning procedure, which we named EPIC (Exploration system of Process-Incorporated features in Cells), as follows:

Every step consists of 20 batches. Step 1: For each batch, we randomly selected 5,000 gene sets from the 2209 gene pool; each gene set contained randomly selected genes with a range of 25–30. Multivariable logistic regression modeling was conducted using each gene set plus MPI and selected MESA features to predict CVE incidence. The gene set that produced the highest AUC (with the 5-fold cross validation strategy as previously described) in MESA-set1 in model testing was kept as the candidate of this batch. 15 out of the 20 batch candidate gene sets that had the highest AUCs were selected and pooled, which was used as the gene pool in the next step.

Step 2: Same as step 1, except that the outcome of step 1 was used as the gene pool, and the top 12 of 20 batch candidate gene sets were pooled and used for step 3.

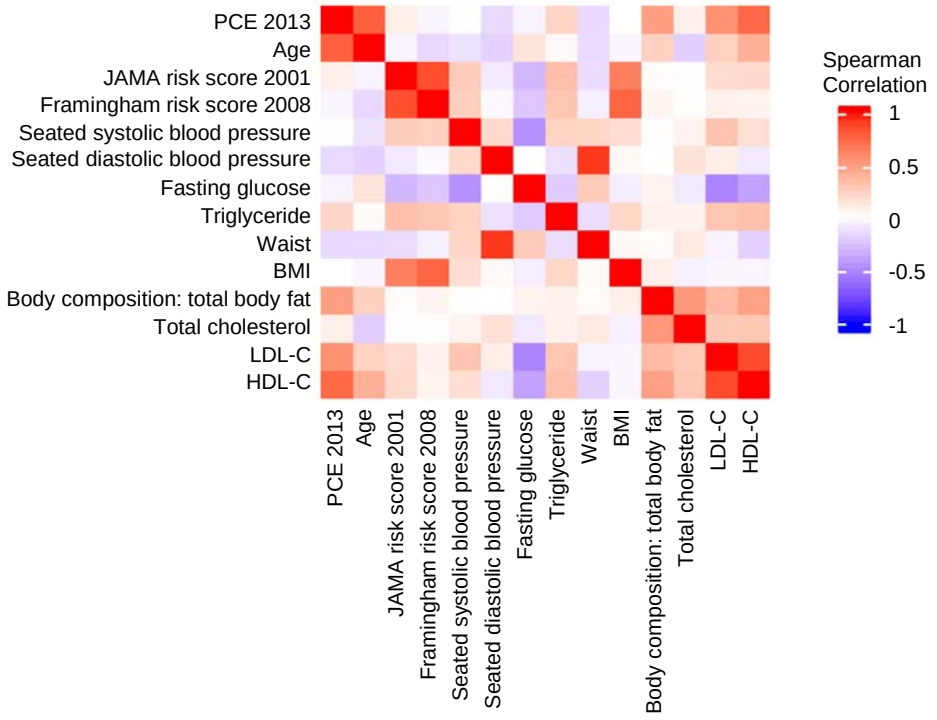
Steps 3–7 were conducted using the same procedures, and the top 8, 5, 3, 2, and 1 candidate gene sets were selected and pooled at the end of each step, respectively.

The outcome gene set of step 7 was the final ASCVD risk signature genes identified by our machine learning procedure, which included 30 genes.

Supplementary Table S1. Ethnic distribution of participants from the MESA cohort (Exam 5, n=1207).

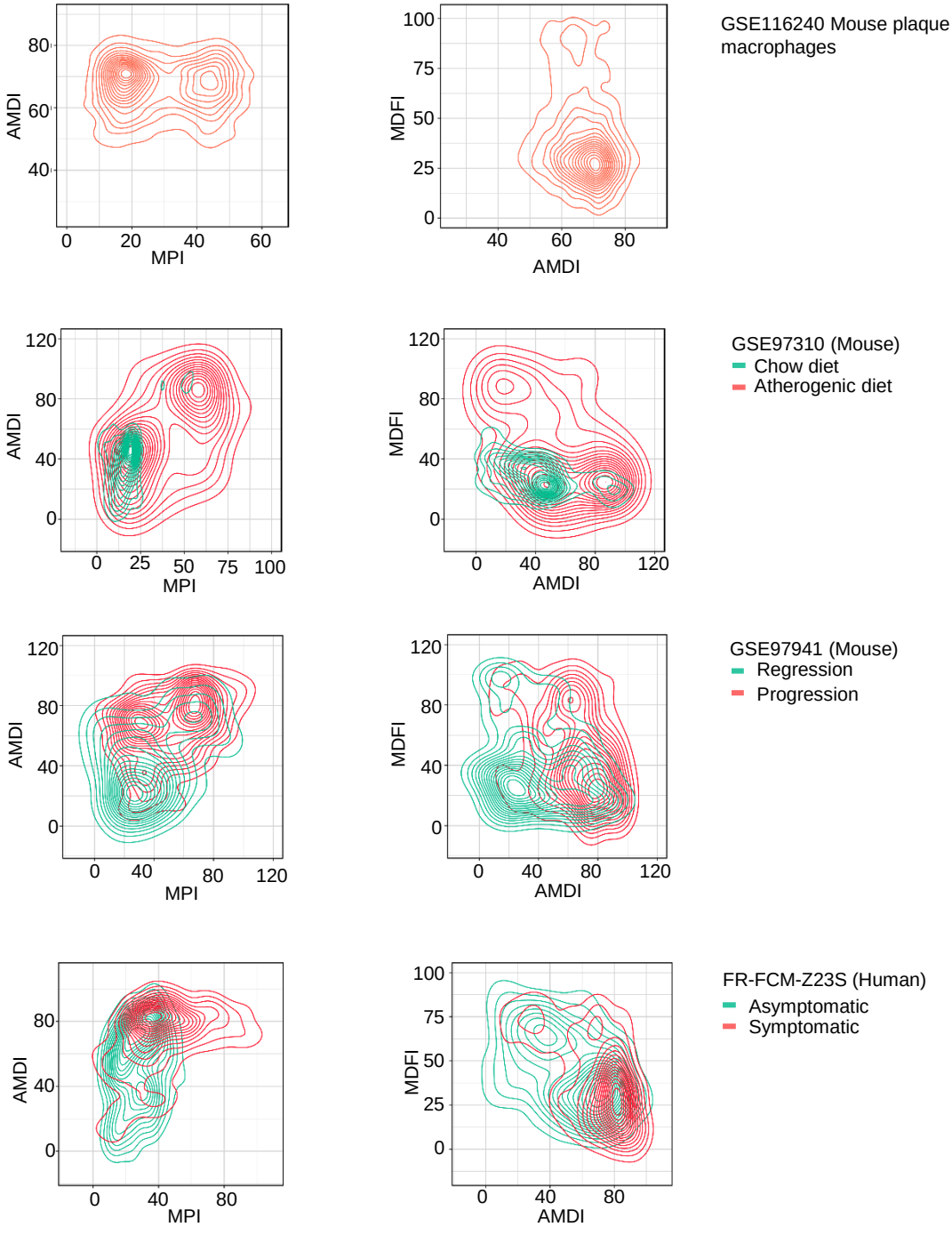
	Female		Male	
	No CVE	CVE	No CVE	CVE
White, Caucasian	246	35	247	51
Black, African-American	130	8	91	15
Hispanic	169	22	151	42
Chinese	0	0	0	0

CVE= Cardiovascular event

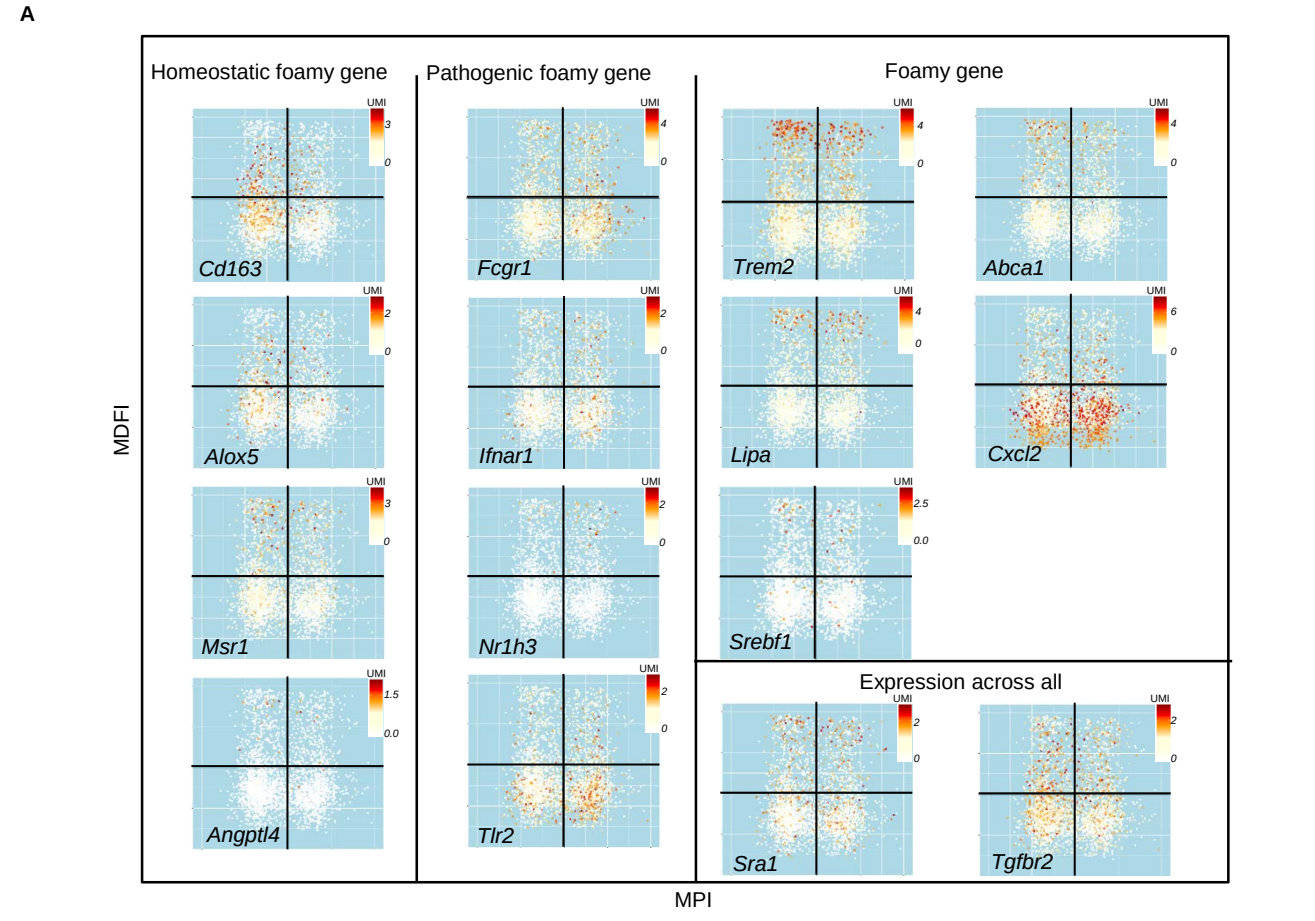


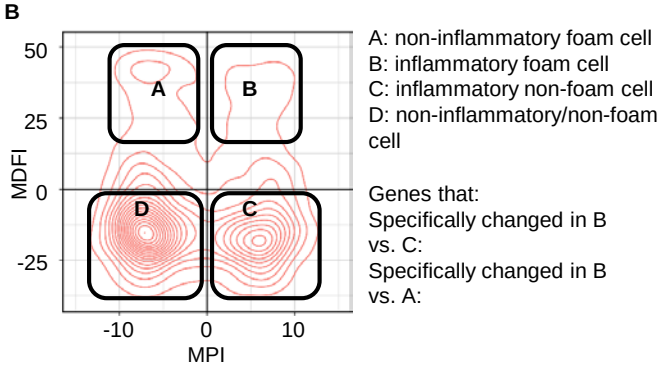
Supplemental Fig. S1. Spearman correlation between variables of the 1207 MESA participants.

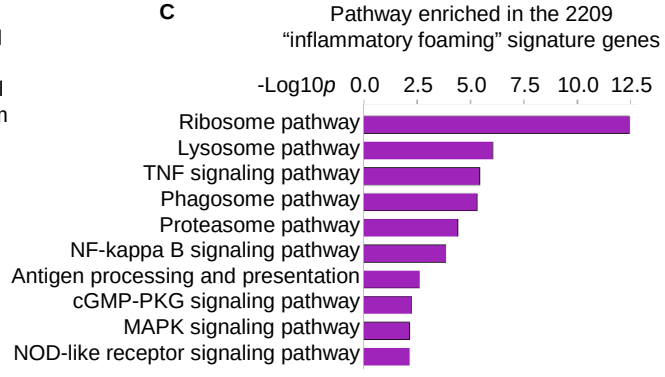
Two dimensional depiction of macrophage/foam cell distribution by MPI, AMDI, and MDFI of MacSpectrum and AtheroSpectrum

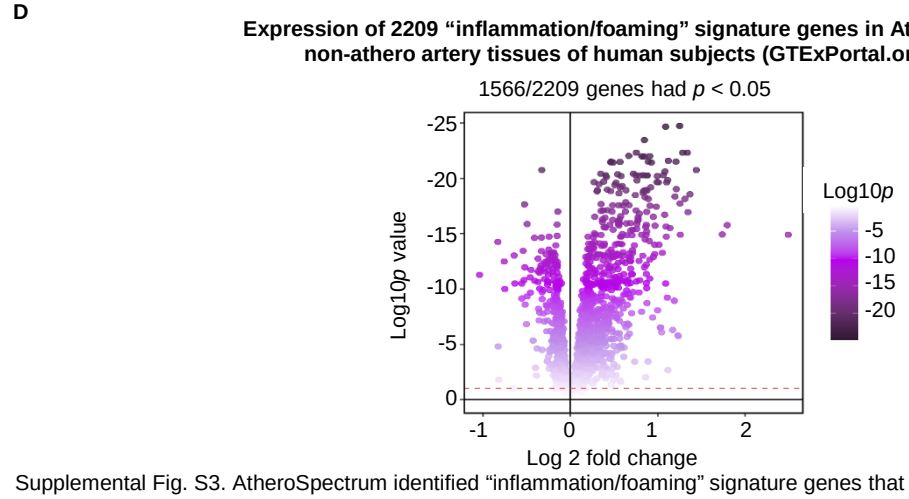


Supplemental Fig. S2. Two dimensional depiction of murine atherosclerotic-plaque macrophages, atherogenic vs. control (mice), progression vs. regression (mice), and symptomatic vs. asymptomatic (human) atherosclerosis conditions by MPI (Macrophage Polarization Index), AMDI (Activation-induced Macrophage Differentiation Index), and MDFI (Macrophage-Derived Foam cell Index) of MacSpectrum and AtheroSpectrum, which is an extension of our recent original program MacSpectrum that was designed to fine-map monocyte/macrophage activation/inflammatory features in complex tissue settings or diseases.

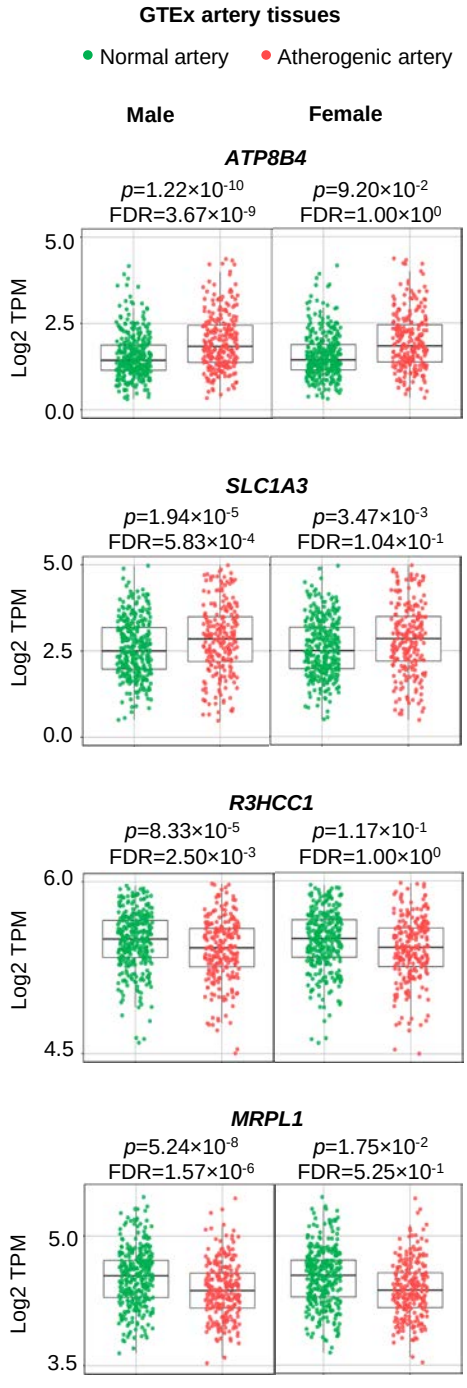




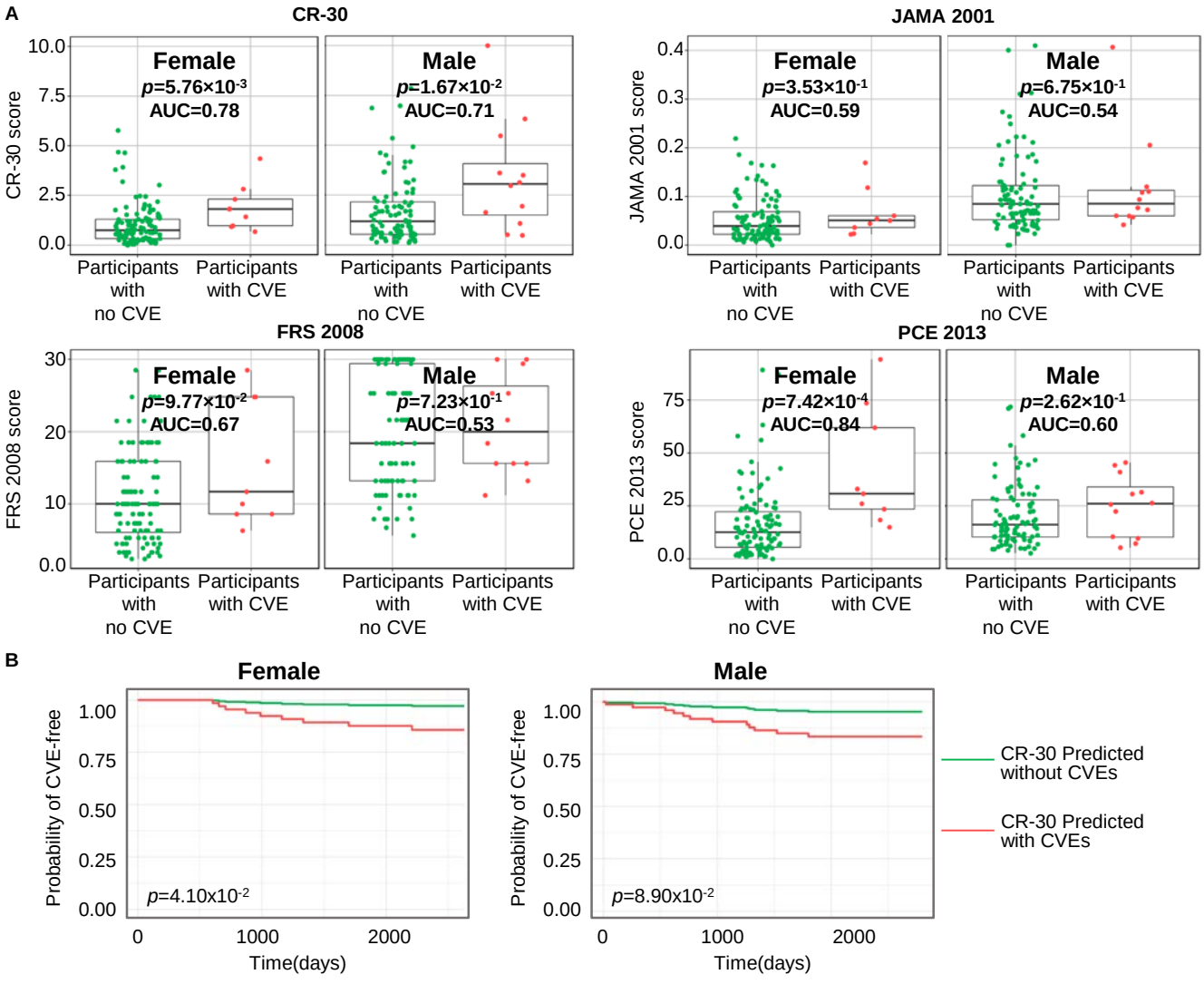




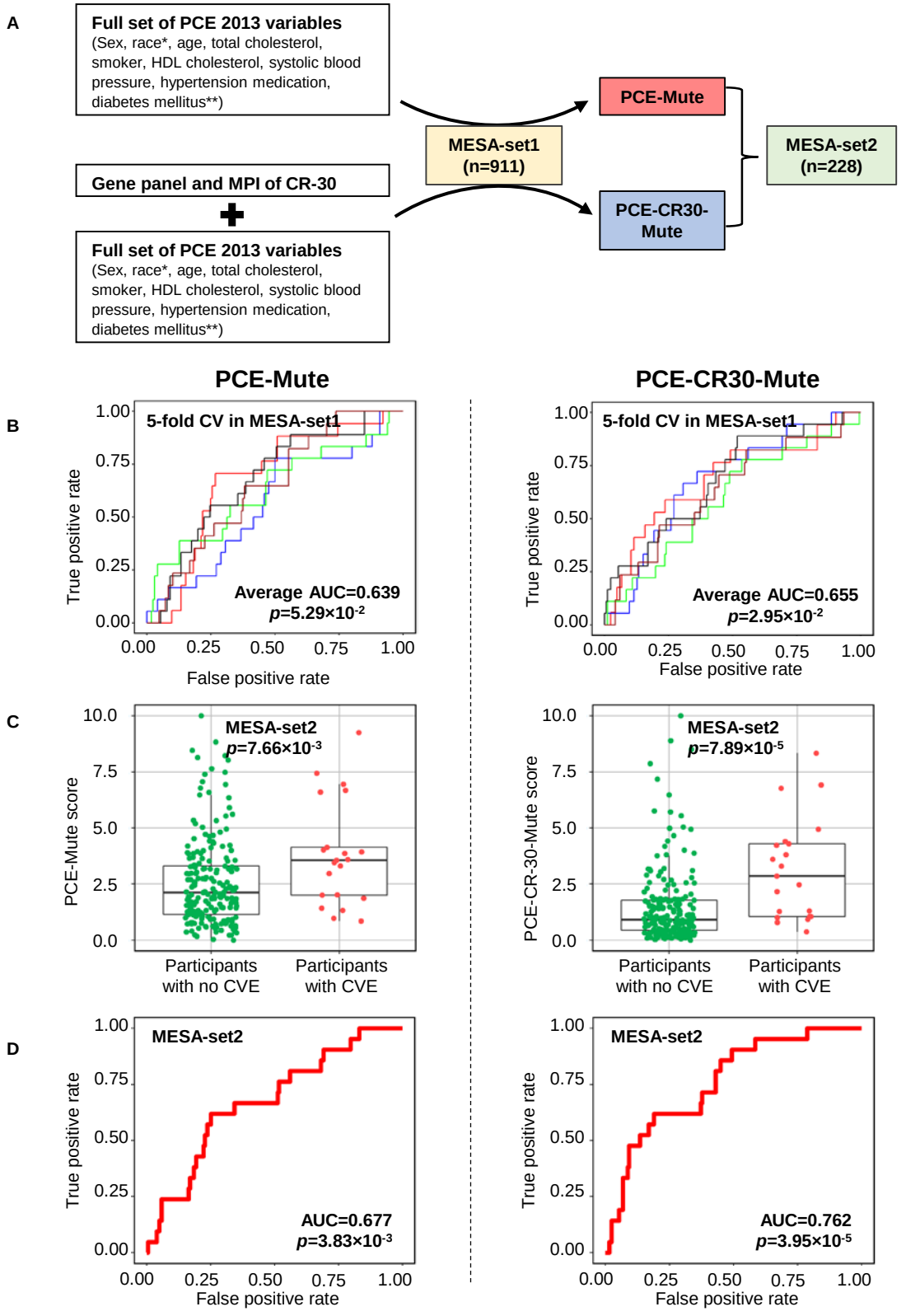
Supplemental Fig. S3. AtheroSpectrum identified "inflammation/foaming" signature genes that are significantly altered in patients' arteries. A, AtheroSpectrum gene expression levels (UMI) with 4 distinct patterns in GSE116240: homeostatic foaming genes, pathogenic foaming genes, foaming genes (inflammation+noninflammation), and genes patterns in all populations. B, AtheroSpectrum identified 2209 "inflammation/foaming" signature genes that are specifically enriched in the inflammatory foaming macrophage sub-population. C, The top 10 pathways enriched in the 2209 "inflammatory foaming" signature genes. A major portion of the 2209 genes are significantly changed. D, Genes in atherosclerotic vs. non-atherosclerotic artery tissues of human subjects deposited at GTEx portal.org. Samples were categorized using pathologists' annotation for each sample provided by GTEx. *p* values were calculated by Mann-Whitney U test with false discovery rate (FDR) adjustment.



Supplemental Fig. S4. ATP8B4, SLC1A3, R3HCC1, and MRPL1, which are significantly associated with CVEs, had significantly different expression levels in atherosclerotic vs. non- atherosclerotic artery tissues of both female and male subjects deposited at GTEx portal.org(female: 202 non- atherosclerotic samples, 115 atherosclerotic samples, male: 335 non-atherosclerotic samples, 247 atherosclerotic samples). *p* value s in C and D were calculated by Mann-Whitney U test with false discovery rate (FDR) adjustment.



Supplemental Fig. S5. CR-30 effectively depicted CVD risk in females and males. A, 10y risk scores (JAMA 2001), Framingham 2008 risk score (FRS 2008), and the 2013 Pooled Cohort Equation for ASCVD risk score (PCE 2013) in MESA-set2 for females (n=121) and males (n=107). *p* values were calculated by Mann-Whitney U test. B, Probability of survival (CVE-free) since monocyte collection (Exam 5) in female (n=121) and male (n=107) MESA-set2 participants who were predicted to have CVE or not by CR-30 scores was calculated by Cox regression with Wald test for *p* values.



Supplemental Fig. S6. Model sensitivity test. To evaluate if the pathogenic gene-set in CR-30 can provide prediction power to a model, a sensitivity test was performed using the strategy (A) to generate PCE-Mute and PCE-CR30-Mute models. B, ROC plot of 5-fold cross-validation in MESA-set1 for PCE-Mute and PCE-CR30-Mute; Different colored curves represent individual fold's result. *p* values were calculated by Mann-Whitney U test using the "verification" R package. C,D, PCE-Mute and PCE-CR30-Mute scores and ROC plots for MESA-set2 participants (n=228). *p* values were calculated by Mann-Whitney U test. *, **: variables defined as in PCE 2013: *race (white and non-white) and **diabetes (Yes or No).