1 Intersecting single-cell transcriptomics and genome- 2 wide association studies identifies crucial cell 3 populations and candidate genes for atherosclerosis.

4

5

6 Lotte Slenders¹, Lennart P. L. Landsmeer¹, Kai Cui¹, Marie A.C. Depuydt², Maarten Verwer³, Joost
7 Mekke³, Nathalie Timmerman³, Noortje A.M. van den Dungen¹, Johan Kuiper⁴, Menno P.J. Winther²,
8 Koen H.M. Prange⁴, Wei Feng Ma^{5,6}, Clint L. Miller^{6,7,8}, Redouane Aherrahrou⁶, Mete Civelek^{6,9}, Gert
9 J. de Borst³, Dominique P.V. de Kleijn³, Folkert W. Asselbergs¹⁰, Hester M. den Ruijter¹⁰, Arjan
10 Boltjes¹, Gerard Pasterkamp¹, Sander W. van der Laan^{1*}, Michal Mokry^{1,10*}

11

12 **Affiliations**

13 1: Central Diagnostics Laboratory, University Medical Center Utrecht, University Utrecht, Heidelberglaan 100,
14 3584 CX, Utrecht, The Netherlands.

15 2: Leiden Academic Centre for Drug Research, Division of Biotherapeutics, Leiden University, Einsteinweg 55,
16 2333 CC Leiden, The Netherlands

17 3: Department of Vascular Surgery, University Medical Centre Utrecht, University Utrecht, Heidelberglaan 100,
18 3508 GA, Utrecht, The Netherlands

19 4: Amsterdam University Medical Centers – location AMC, University of Amsterdam, Experimental Vascular
20 Biology, Department of Medical Biochemistry, Amsterdam Cardiovascular Sciences, Meibergdreef 9,
21 Amsterdam, The Netherlands

22 5: Medical Scientist Training Program, University of Virginia, Charlottesville, VA 22908, USA

23 6: Center for Public Health Genomics, University of Virginia, Charlottesville, VA 22908, USA

24 7: Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA 22908,
25 USA

26 8: Department of Public Health Sciences, University of Virginia, Charlottesville, VA 22908, USA

27 9: Department of Biomedical Engineering, University of Virginia, Charlottesville, VA, USA

1 10: Laboratory of Experimental Cardiology, department of Cardiology, University Medical Center Utrecht,
2 University Utrecht, Heidelberglaan 100, 3508 GA, Utrecht, The Netherlands

3 *Shared last authorship

4

5 Correspondence

6 *Sander W. van der Laan, PhD*

7 Central Diagnostics Laboratory, University Medical Center Utrecht, Utrecht University

8 Heidelberglaan 100

9 3584 CX Utrecht,

10 the Netherlands

11 E: s.w.vanderlaan-2@umcutrecht.nl

12 T: +31 88 756 76 96

13

14 *Michal Mokry, MD-PhD*

15 Central Diagnostics Laboratory, University Medical Center Utrecht

16 Heidelberglaan 100

17 3584 CX Utrecht,

18 the Netherlands

19 E: m.mokry@umcutrecht.nl

20 T: +31 88 755 07 60

1 Abstract and keywords

2 Background

3 Genome-wide association studies have discovered hundreds of common genetic variants for
4 atherosclerotic disease and cardiovascular risk factors. The translation of susceptibility loci into
5 biological mechanisms and targets for drug discovery remains challenging. Intersecting genetic and
6 gene expression data has led to the identification of candidate genes. However, previously studied
7 tissues are often non-diseased and heterogeneous in cell composition, hindering accurate candidate
8 prioritization. Therefore, we analyzed single-cell transcriptomics from atherosclerotic plaques for cell-
9 type-specific expression to identify atherosclerosis-associated candidate gene-cell pairs.

10

11 Methods and Results

12 To identify disease-associated genes, we applied gene-based analyses using GWAS summary
13 statistics from 46 atherosclerotic and cardiovascular disease, risk factors, and other traits. We then
14 intersected these candidates with scRNA-seq data to identify genes specific for individual cell
15 (sub)populations in atherosclerotic plaques. The coronary artery disease loci demonstrated a
16 prominent signal in plaque smooth muscle cells (*SKI*, *KANK2*, *SORT1*) $p\text{-adj.} = 0.0012$, and
17 endothelial cells (*SLC44A1*, *ATP2B1*) $p\text{-adj.} = 0.0011$. Further sub clustering revealed genes in risk
18 loci for coronary calcification specifically enriched in a synthetic smooth muscle cell population.
19 Finally, we used liver-derived scRNA-seq data and showed hepatocyte-specific enrichment of genes
20 involved in serum lipid levels.

21

22 Conclusion

23 We discovered novel gene-cell pairs, on top of known pairs, pointing to new biological mechanisms
24 of atherosclerotic disease. We highlight that loci associated with coronary artery disease reveal
25 prominent association levels in mainly plaque smooth muscle and endothelial cell populations. We

1 present an intuitive single-cell transcriptomics-driven workflow rooted in human large-scale genetic
2 studies to identify putative candidate genes and affected cells associated with cardiovascular traits.
3 Collectively, our workflow allows for the identification of cell-specific targets relevant for
4 atherosclerosis and can be universally applied to other complex genetic diseases and traits.

5

6 **Keywords**

7 Atherosclerosis, cardiovascular disease, genome-wide association study, single-cell analysis.

1 Translational perspective

2 GWAS identified a large number of genomic loci associated with atherosclerotic disease. The
3 translation of these results into drug development and faster diagnostics remains challenging. With
4 our approach, we cross-reference the GWAS findings for atherosclerotic disease with scRNA-seq data
5 of disease-relevant tissue and bring the GWAS findings closer to the functional and mechanistic
6 studies.

7

1 Introduction

2 Since the first publication in 2005 on age-related macular degeneration¹, genome-wide association
3 studies (GWAS) soared, and to date the GWAS Catalog² registered over 5,000 publications covering
4 thousands of genetic associations. These efforts also identified hundreds of common genetic risk loci
5 for cardiovascular diseases and traits, including coronary artery disease (CAD)³, ischemic stroke
6 subtypes⁴, coronary artery calcification⁵, and carotid intima and media thickness⁶. Despite tremendous
7 socio-economic and medical progress, cardiovascular disease remains the leading cause of death⁷,
8 highlighting the need for new biomarkers and pharmaceutical targets. Previous studies convincingly
9 show that GWAS offer a solid foundation for innovative drug development⁸ as they agnostically
10 identify genetic relationships with diseases and point to potential causal mechanisms⁹.

11 Identifying the causal genes and underlying pathological mechanisms that drive the signal in genetic
12 risk loci is not straightforward. One possible approach to narrow down the lists of candidate genes is
13 to utilize gene expression data, by identifying expression quantitative trait loci (eQTLs) that
14 colocalize with genetic risk loci with expression changes in relevant tissues¹⁰. Alternatively,
15 transcriptome-wide association studies (TWAS) estimate genetic correlation between tissue-specific
16 gene expression and complex traits to predict causal relations between traits¹¹. However, these
17 powerful methods take advantage of whole-tissues collected in large biobanks and the variable cell
18 composition confounds the resulting signal. Because tissues are composed of many cell types and
19 tissue expression is averaged across all those cells, the signal from less prevalent but important cell
20 populations is obscured. In order to test potential therapeutic targets, one requires suitable functional
21 assays. This in turn requires an idea of the function and cellular expression of targets, which is
22 complicated by tissue heterogeneity. Single-cell transcriptomics enables measurement of cell-specific
23 expression, cell-to-cell expression variability, transcriptional noise, or temporal dynamics¹².

24 Intersecting GWAS data with human transcriptomic single-cell data would increase the resolution

1 both in terms of candidate gene and more importantly, it would identify the cell type most relevant for
2 the given disease.

3 We hypothesised that genetic drivers for atherosclerotic disease can be found in plaque tissue. We
4 aimed to project GWAS signals for 46 traits directly into a cell-type specific expression using single-
5 cell RNA sequencing (scRNA-seq) data derived from atherosclerotic plaques. To this end, we devised
6 a workflow that utilized gene-based testing to identify candidate genes from GWAS summary
7 statistics and calculated an enrichment score for each candidate list using expression profiles from
8 scRNA-seq driven cell populations. This approach allows us to find the enrichment in expression
9 patterns that point to specific, and potentially causal, cell populations. We report that loci associated
10 with coronary artery disease reveal prominent association in mainly plaque smooth muscle cells and
11 endothelial cells.

1 Methods

2 **Single-cell sequencing: carotid artery plaques**

3 Atherosclerotic lesions were collected from 12 female and 26 male patients undergoing a carotid
4 endarterectomy procedure. All pathological tissue was included in the Athero-Express Biobank Study
5 biobank (AE, www.atheroexpress.nl) at the University Medical Centre Utrecht (UMCU)¹³. Plaque
6 processing for single-cell sequencing is described elsewhere¹⁴. After sequencing, data was processed
7 in an R 3.5 and 4.0 environment¹⁵ using Seurat version 3.2.2¹⁶. Mitochondrial genes were excluded
8 and doublets were omitted by gating for unique reads per cell (between 400 - 10 000) and total reads
9 per cell (between 700 – 30 000). Data was corrected for sequencing batches using the function
10 SCTransform. Clusters were created with 30 principal components at resolution 0.8. Multiple
11 iterations of clustering were used to determine optimal clustering parameters. Population identities
12 were assigned by evaluating gene expression per individual cell clusters. Sub-populations of SMC
13 populations were determined by isolating the SMC and EC populations from the complete population
14 of cells mentioned above and re-assigned clusters. Clusters were created with 10 principal
15 components at resolution 1.4. New subpopulation identities were assigned by evaluating gene
16 expression per individual cell clusters. The newly defined SMC subpopulation identities were
17 projected onto the SMC populations in the complete scRNA-seq dataset of atherosclerotic lesions for
18 downstream processing.

19

20 **GWAS summary statistics**

21 GWASs summary statistics were leveraged from publicly available datasets. Mining these datasets
22 uncovered a plethora of challenges. Not all GWAS summary statistics are available for download, can
23 be requested by the authors, and/or only provide the lead SNPs. Other issues included ambiguous
24 reports of sample size, genome build or missing information such as standard errors and effect sizes.
25 We were able to partly resolve these issues by using a custom pipeline, which automates tasks such as

1 genome build liftover and SNP alignment and provides a uniform file format (**Supplemental Figure**
2 **2A**). This significantly decreased workload and improved downstream applications.

3 GWAS summary statistics were selected by means of relevance, “latest and the largest” and publicly
4 available, easily accessible and/or open-source data was given priority. Incomplete summary statistics
5 were omitted during this phase (**Supplemental Figure 2A**). A custom Python¹⁷ pipeline (see **Data**
6 **availability**) was used to pre-process GWAS summary statistics and primary quality control. This
7 primary control step was used to uniformly transform GWAS summary statistics because of a lack of
8 a standardized format. In short, our custom pipeline was built to automate effect/other allele
9 alignment with 1000 Genomes phase 3¹⁸ release with a maximum frequency distance of 25%,
10 maximum MAF of 45% for ambivalent variants and lifting genomic positions to hg19 where needed.
11 In preparation for GWAS alignment, dbSNP153 (GCF1405) was referenced to translate
12 MEGASTROKE rs-ids to genetic positions and to augment AF, ASD, NICM and TAGC with allele
13 frequency information. Standard errors for HF and NICM were recovered as follows:

$$14 \quad Z = \text{sign}(\text{beta}) * \text{abs}(\text{qnorm}(p/2))$$

$$15 \quad SE = \text{beta} / Z$$

16 Where beta is the effect size, p is the p-value of association which is quantile normalized using the
17 qnorm-function in R. SE is the standard error calculated by dividing the beta by Z. Effect sizes were
18 calculated from odds-ratios as $\text{beta} = \log(\text{OR})$ for BIP, Insomnia, IBD and MDD before entering our
19 custom pipeline.

20

21 The final selection was grouped into one of 3 categories: atherosclerotic disease and other
22 cardiovascular disease (12), risk factors (14) and other (20). Data was further processed with
23 MAGMA¹⁹ (version 1.07) for genome-wide analysis, annotation and characterisation of significant
24 hits via the FUMA²⁰ web platform (version 1.3.6, accessed June 2020, <https://fuma.ctglab.nl/>). We
25 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 minor allele frequency (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} .

Selecting cell population specific genes

DEGs were selected via the Wilcoxon sum rank test. A gene was considered a DEG if 1) 10% of cells within a cluster expressed the gene with $\log_2$ fold change of at least 0.25; 2) the gene passed the Bonferroni adjusted significance threshold of $p < 0.05$ for this test. DEGs were determined via a “one cluster vs one cluster” and “one cluster vs remaining clusters”⁵ (Supplemental Figure 2B). This method results in genes specific for a cluster compared to every other single cluster and genes specific for a cluster compared to the whole single cell reference set.

Gene score calculation

Fully processed summary statistics and cell population specific DEG were used for calculating gene scores. The background gene set ($n = 20112$ genes) was defined as all the genes that are reported and annotated by MAGMA, regardless of their respective p value or Z score with duplicates excluded. We calculated the random gene score per cell-population by overlapping the genes in this background set with the cell-population specific DEG. For every analysis, the background genes are the same per GWAS. Z-score threshold for the p values from MAGMA was set at 3 and we included a maximum of 150 candidate genes meeting this threshold (Supplemental Figure 1B). If there were more than 150 genes meeting the threshold, we only used the first 150. Calculations were made as follows: 1) Per cell population, determine the overlap of the candidate genes of a GWAS with the cell population specific DEG. 2) Score each of these genes according to average cell population expression: assign one point for expression between 1 and 2, and two points for expression > 2 and 0 points for

1 expression < 1 . Only genes that are cell population specific receive a score. Assign the points to the
 2 genes from the random candidate gene list in the same way (**Supplemental Figure 2Cii**). 3) Compare
 3 the scores from the candidate genes with the scores of the random genes via one tailed Wilcoxon
 4 Rank sum test in R and adjust the results per GWAS for multiple testing using the Benjamini-
 5 Hochberg method. Adjusted p-values smaller than 0.1 (10% false discovery rate) were considered
 6 significant (**Supplemental Figure 2Ciii**). Enrichment was defined as the fold change of the mean of
 7 the gene scores divided over the mean of the gene scores in the random set. All scripts are available
 8 on GitHub (see **data availability**).

9 Author contributions

10 L.S. drafted the manuscript and designed the figures. M.V., J.M. and N.T. collected clinical data.
 11 N.v.d.D. executed human plaque processing and general maintenance of the biobank. G.J.d.B.
 12 performed carotid endarterectomy procedures. G.P. and G.J.d.B. participated in the design of the
 13 Athero-Express Biobank Study. M.A.C.D. executed the human plaque processing, FACS and flow
 14 cytometry. L.S., M.M., and S.W.v.d.L. participated in conceptualisation, data collection and data
 15 interpretation. L.P.L.L. and S.W.v.d.L. designed and executed workflow for raw GWAS summary
 16 statistics processing. L.S. performed data analysis on scRNA-seq data and gene score calculations.
 17 W.F.M and C.M. designed and integrated the plaque scRNA-seq into plaqview. R.A. and M.C.
 18 provided data for functional validation. S.W.v.d.L., F.W.A., M.M., J.K., M.W. and G.P. participated
 19 in the conceptualization, funding and supervision of the scRNA-seq experiments. All authors
 20 provided feedback on the research, analyses and manuscript.

Results

Single-cell RNA-seq of atherosclerotic plaques reveals 18 cell populations.

We have expanded our previously published¹⁴ scRNA-seq datasets of human atherosclerotic plaque cells in a cohort which now includes 38 patients and identified 18 cell populations amongst 6,191 cells. These expanded datasets allowed us to map the genetic loci not only to main cell types (e.g., smooth muscle cells (SMCs)) but also to more specific cell subtypes (e.g., synthetic, contractile or transitioning SMCs) - increasing resolution compared to the previous dataset. By combining known and recently published sets of markers we identified clusters of lymphocytes, macrophages, mast cells, endothelial cells (ECs), and SMCs (**Figure 1A**). This data is available for interactive exploration via the online tool Plaqview²¹ accessible on www.plaqview.com. Notably, the expression of selected, widely studied GWAS targets for coronary artery disease (*EDN1*, *HDAC9*, *NOS3*, *VEGFA*, *FNI*, *APOE*, *APOB*, *SMAD3*, *PCSK9* and *LDLR*) clearly demonstrates cell-specific patterns (**Figure 1B**). This underscores the need for a systematic analysis using large-scale GWAS summary statistics and motivated us to devise an intuitive two-stage approach.

Gene-based analyses of atherosclerotic diseases and cell population specific genes.

First, we aimed to identify trait-associated genes using GWAS summary statistics by aggregating per-variant statistics and to calculate an empirically derived p-value for a given gene. Thus, gene-trait associations would be agnostically derived and solely based on genetic signals. To do this we collected and partitioned GWAS summary statistics for 46 traits across 3 categories: “atherosclerotic disease” (4), “cardiometabolic traits” (22) and “other” (20) (**Supplemental Table 1, Supplemental Figure 1A**). From these GWASs, 40 were performed on population with European ancestry, and the remaining 6 had mixed or other ancestries. Next, to identify genes associated with individual traits we performed gene-based testing using MAGMA¹⁹ per individual GWAS. Genes mapped by gene-based testing are referred to as *candidate genes* (**Supplemental Figure 2A, Supplemental Data**). Second, we aimed for an unbiased selection of genes that would best represent each cell population in

a given dataset (**Supplemental Figure 2B**). Per tissue, these genes were selected through differential gene expression and include genes that are overexpressed in specific cell populations compared to all other cells (one vs. all) or compared to any individual cell cluster (one vs. one)¹⁴. This approach eliminates ubiquitously expressed genes and genes with low expression without the need of a predefined list (**Supplemental Figure 4**) and identifies a similar number of cell-specific genes in each population. Cell population specific genes will be further referred to as *differentially expressed genes* (DEGs). For cell populations in the atherosclerotic plaque, the average number of DEG was 461, with the lowest 307 for B Cells II and highest 678 DEGs for T Cells III.

Plaque cell populations show specific enrichment of GWAS candidate genes.

To test our hypothesis that the genetic drivers for atherosclerotic disease can be found in plaque tissue, we applied our workflow on cardiovascular candidate genes and the DEG sets from individual cell types present in the atherosclerotic plaque (**Supplemental Figure 3A, C, Supplemental Figure 4, Supplemental Table 2, Supplemental Data**). The signals we observed were independent of the total number of DEGs in each cell population (**Supplemental Figure 3C**), and did not show a bias towards GWAS sample size (**Supplemental Figure 1C**) or trait heritability (**Supplemental Figure 1D**). We found cell-specific enrichment for different GWAS in the atherosclerotic disease category, which was mainly observed in ECs and SMCs. Stroke candidate genes were enriched in EC populations, Regulatory T Cells and B Cells II. (**Figure 3A**).

Smooth muscle cells are predominantly associated with coronary artery disease genes.

By highlighting the overlap between plaque specific DEGs and candidate genes for CAD, we observed a significant enrichment in SMCs (p-adj. = 0.0012, 3.1 fold enrichment) and ECI (p-adj. = 0.0011, 2.7 fold enrichment) populations (**Figure 2A**). All p values and fold enrichment values are listed in **Supplemental Table 2**. This uncovered well studied GWAS loci such as *APOE*, *COL4A1* and *COL4A2*. The remainder of the overlap is driven by *KANK2* and *SORT1* in SMCs, *SLC44A1* and

1 *ATP2B1* in ECI (**Figure 2B**). *KANK2* belongs to the KANK protein family which plays a role in
2 cytoskeletal function. *SKI* expression was present in both cell populations (**Figure 2C**). This proto-
3 oncogene inhibits TGF-beta signalling²². Interestingly, while using the single cell transcriptomics
4 enabled us to pinpoint specific cell (sub)populations, these potential targets exhibited ubiquitous
5 expression in bulk transcriptomics data from GTEx tissue samples (**Figure 2D**). Notably,
6 CD14+CD68+ Macrophages III showed an enrichment of 4.3 for CAD with genes *APOE*, *SKI* and
7 *HNRNPUL1* but did not meet the significance threshold (p-adj. = 0.19).

8

9 To verify our results we used an independent scRNA-seq dataset of carotid artery plaques by Pan *et*
10 *al.*²³ (**Supplemental Figure 3B, D**). The cell population labels broadly correspond between datasets
11 (**Supplemental Figure 3E**). The patterns for enrichment are similar between datasets (**Supplemental**
12 **Figure 3A, B**). We found that CAD, coronary artery calcification and plaque presence candidate
13 genes were enriched in the Fibrochondrocyte cell population (**Figure 3B**), which corresponds to
14 ACTA2+ SMCs cells in our dataset. For CAD, the enriched was driven by genes *CDKN2A*, *COL4A1*,
15 *COL4A2*, *APOC1*, *APOE*, and *CXCL12*. Similar to the genes found in our datasets in SMCs and ECs
16 (**Supplemental Table 2**).

17

18 **Endothelial cells are enriched for ischemic stroke-associated genes**

19 Atherosclerotic disease is one of the major culprits underlying stroke, with carotid plaques being the
20 main driver of large artery stroke. We subsequently investigated overlap of GWAS-associated genes
21 for stroke subtypes any stroke (AS) and any ischemic stroke (IS) and found they were predominantly
22 enriched in ECI (AS: p-adj. = 0.011, 2.7 fold enrichment; IS: p = 0.077, 2.5 fold enrichment) and ECII
23 (AS: p-adj = 1.9×10^{-5} , 3.9 fold enrichment ; IS: p-adj = 0.023, 3.2 fold enrichment) subtypes (**Figure**
24 **3A**). Overlapping genes between AS and cell populations include *ESAM*, *LMNA* and *SLC44A2*.
25 (**Figure 3B**) Soluble ESAM (endothelial cell adhesion molecule) levels have been correlated to

1 myocardial infarction and heart failure²⁴. *Lmna* deficient mouse models develop cardiovascular
2 disease at an accelerated pace due to premature ageing²⁵. Of note, expression of these genes was not
3 specific to vascular derived tissue of GTEx when using bulk transcriptomics data, suggesting that our
4 findings are cell population and tissue specific (**Figure 3C**). No significant enrichment is found for IS
5 in SMCs, but the cell population specific signal in ECI and ECII is largely different from AS (**Figure**
6 **3D**).

7 Large Artery Stroke (LAS) is a subtype of IS. Strikingly, we see enrichment in CD3+ Regulatory T
8 Cells and CD79A+ B Cells II (plasma cells) (p-adj = 0.026 and p-adj = 0.021 resp.) populations for
9 LAS (**Figure 3E**). Suggesting the involvement of the adaptive immune system on the development of
10 the LAS phenotype.

11

12 **Transitional SMCs enriched for atherosclerotic disease candidate genes.**

13 Plaque ACTA2+ Smooth muscle cells can be further divided into three sub-populations with a distinct
14 phenotype; Synthetic Smooth Muscle Cells, Contractile Smooth Muscle Cells and a smaller cell
15 population consisting of SMCs with an intermediate phenotype (Transitional Smooth Muscle Cells)
16 (**Figure 4A, B, Supplemental Figure 5A, Supplemental Table 2**). To explore if we can narrow
17 down the signal from SMCs into a specific SMC subtype, we subjected this subset of plaque cells to
18 our workflow (**Supplemental Figure 5B, C**). This revealed that the majority of the signal can be
19 found in Transitional SMCs. This population is predominantly enriched for CAD (p-adj. = 0.026 2.4
20 fold enrichment) , carotid intima-media thickness (cIMT) (p-adj. = 0.027 , 2.4 fold enrichment) and
21 plaque presence (p-adj. = 0.081, 2.3 fold enrichment) (**Figure 4C**). Intima-media thickness is a
22 commonly used tool to measure atherosclerosis²⁶. Signal for LAS was present in Transitional and
23 Synthetic SMCs with unique hits for each cell population (**Figure 4D**).

24 To test the robustness of these results we have repeated the analysis with the other public dataset (Pan
25 *et al.*). For CAD, the overlap between Transitional SMCs and corresponding “Fibrochondrocytes” in

1 Pan et al. (**Supplemental Figure 5DE**) showed significant enrichment and again identified a largely
2 overlapping set of commonly studied GWAS loci *COL4A1*, *COL4A2*, *APOC1*, *APOE*, and *CXCL12*.

3 **Functional assessment of smooth muscle cell associated candidates.**

4 To demonstrate the potential role of newly identified gene-cell pairs in atherosclerosis-relevant
5 process we selected 3 genes enriched in the SMC population of plaques. *SKI*, *KANK2* and *EDNRA*
6 are CAD candidate genes and currently understudied in relation to atherosclerotic disease in this cell
7 population. Making them possible interesting targets for further functional studies. To investigate the
8 role of these genes in SMCs, we functionally characterized SMCs from the ascending aortas in 151
9 healthy and diverse donors for cellular migration, proliferation and calcification²⁷. *KANK2* gene
10 expression is correlated with proliferation under different stimuli (p-adj. < 0.1), and was negatively
11 correlated with calcification in quiescent cells (Calcification Osteo: p-adj = 0.0058; Calcification (Pi):
12 p-adj 0.069) (**Figure 5A, Supplemental Table 5**). Under all stimuli, the response was positive,
13 except upon TGFβ1 exposure. *SKI* expression positively correlates with calcification in proliferative
14 cells (p-adj = 0.069) but has a negative response towards TGFβ1 induced proliferation (p-adj = 0.069,
15 R -0.21) (**Figure 5A, Supplemental Table 5**). *EDNRA* was negatively correlated with proliferation
16 response to IL1β (p-adj. = 0.038) and calcification (p-adj = 0.062) in non-proliferative cells and
17 showed a positive correlation in proliferative cells for migration (p-adj = 0.069) (**Figure 5A,**
18 **Supplemental Table 5**).

19

20 **Lipid related risk factors enriched in liver hepatocyte populations.**

21 Causal mechanisms for atherosclerosis have been linked to lipid metabolic processes active in the
22 liver and blood glucose levels. This offered us the opportunity to assess the robustness of our
23 approach across tissues and diseases. We acquired published scRNA-seq datasets of human liver
24 cells²⁸ and human pancreatic Langerhans islet cells²⁹ to identify possible cell- and disease-specific
25 targets.

1 Since one of the major liver functions is lipid metabolism, we expected strong signals for candidate
2 genes involved in triglycerides levels. In line with our expectations, these GWASs show indeed the
3 strongest signal in hepatocytes 3,4 and 5 driven by several apolipoproteins (**Supplemental Figure**
4 **6A, C, Supplemental Table 2**).

5 Langerhans islet cells overall showed weak enrichments for all atherosclerosis related GWASs
6 (**Supplemental Figure 5B, D**). Signal for CAD in endothelial and mesenchymal populations overlap
7 with hits in comparable cell populations in plaque (**Supplemental Table 2**). We could detect
8 significant overlap of type 2 diabetes associated genes with Beta cells (p-adj. = 0.001) driven by *INS*,
9 *WFS1* and *FTO* amongst others (**Supplemental Table 2**).

10

11 Altogether, we show that candidate genes for atherosclerotic disease are significantly enriched in EC
12 and SMC populations of atherosclerotic plaque, notably in transitional SMCs. Stroke related genes
13 were enriched in ECs and lipid loci were found enriched in liver hepatocytes. Pancreatic beta cells
14 were enriched for BMI candidate genes. Interestingly, only limited signals were observed in the cells
15 of the adaptive immune system suggesting the genetic component for atherosclerosis is rooted in the
16 structural cells of the vessel. We selected 3 candidate genes enriched in SMCs for functional testing in
17 cells of ascending aortas for cellular calcification, proliferation and migration that showed that
18 *KANK2*, *SKI* and *EDNRA* expression correlated with calcification, migration and proliferation in
19 vascular SMCs.

20

1 Discussion

2 Post-GWAS analyses aiming to identify candidate genes for translation into clinical care are not
3 straightforward. The thousands of variants associated with complex traits, including cardiovascular
4 diseases, have made it abundantly clear that many genes contribute to disease risk and progression.
5 For instance, GWAS for coronary artery disease have identified 163 loci linking to hundreds of
6 candidate genes³. Further efforts like chromatin interaction mapping linked almost 300 additional
7 gene targets for atherosclerotic disease via distal chromatin interactions¹⁴. This adds another layer to
8 candidate gene prioritization. Selecting causal GWAS loci for functional testing can be deduced with
9 vast knowledge of the trait or disease, but this remains speculative. There are numerous ways to
10 accelerate the path to translational research. The intersection of loci found amongst different GWAS
11 related to the same trait or disease can highlight some of the primary genes involved but, the cellular
12 resolution is still lacking^{4,30}.
13 Overlapping GWAS hits with pathways or within gene networks offers more context to study the
14 genes. However, this is highly dependent on knowledge of the respective pathways and networks, and
15 locating the relevant key players in the network is not indisputable³¹.
16 Colocalization of tissue-specific eQTLs^{10,32,33} with GWAS data provides another alternative that can
17 be used to estimate if there are overlapping causal variants in both datasets. For example, a recent
18 study on type 2 diabetes leveraged publicly available eQTL data and a separate pancreatic dataset to
19 find targets for glucose- and insulin-related trait loci¹⁰. In concordance to our data, trait associated
20 genes are expressed in pancreatic islet tissue. However cell population specific signals are lacking and
21 eQTL signals can be confounded by variability in tissue cell composition. Currently, efforts are being
22 made to study eQTLs in a single-cell space, with single-cell quantification of gene expression in
23 disease relevant tissue. With this method it is possible to detangle the effect of SNPs on the single cell
24 and patient specific level and construct personalized gene regulatory networks^{12,34}.

1

2 Another straightforward approach to identify the causative genes is overlapping the expression

3 profiles of different tissues and organs to GWAS candidate genes. Utilizing bulk RNA sequencing

4 data applicable to the GWAS helps to find targets that are directly acting in the affected tissue. This

5 powerful approach has one major drawback: the data is composed of multiple cell populations and

6 this signal is difficult to deconvolute into individual cell populations. It is especially limiting when

7 looking for cells that are present in almost all tissues – like endothelial cells or smooth muscle cells..

8 We show here that these bulk RNA tissues approaches can be misleading due to lack of specific

9 expression patterns for relevant GWAS loci in specific tissue, like vascular wall tissue (**Figure 2E**).

10 On the other hand, the overlap with single cell transcriptomics can detect significant overlaps in

11 distinct cell types such as ECs.

12 Here we present a possible approach to the prioritization problem. We projected GWAS loci directly

13 into single-cell transcriptomics datasets derived from disease-relevant tissue. This allowed us to

14 identify both known and novel gene targets together with their associated cell populations. Notably,

15 the CAD-associated gene *SKI* is significantly enriched in SMC and EC populations (**Figure 2A**). *SKI*

16 inhibits TGF-beta signalling²², which plays a major role in atherosclerotic disease progression by

17 controlling processes such as cell proliferation and matrix formation. Consequently, *SKI* increases

18 lesion size in LDR-/- mice³⁵. Another identified gene - *KANK2* plays a role in cytoskeletal function

19 by reducing the cell's ability to migrate by activating *TALIN*³⁶. This gene is specifically expressed in

20 the SMC-population of plaques. Possible impairment of cell mobility by *KANK2* could lead to

21 inefficient SMC migration towards the cap resulting in an unstable plaque phenotype. Interestingly,

22 our functional analysis of *KANK2* showed a positive association between gene expression and cell

23 migration in quiescent and proliferative cells of SMCs from ascending aortas (R = 0.15 and 0.2 resp.)

24 (**Supplemental table 5**) – suggesting more complex mechanisms. Follow-up experiments creating

25 gene-deficient primary cell lines derived from atherosclerotic plaques to study the molecular

26 mechanistic of those genes can provide more information of its role in atherosclerosis³⁷. Notably, the

1 CAD-associated locus in the vicinity of *KANK2* also contains *LDLR* gene (**Figure 5B** bottom). Even
2 though *LDLR* involvement in atherosclerosis is widely accepted, we were able to show that *KANK2* is
3 another strong candidate in this region as it also shows a strong correlation with SMC proliferation
4 (**Figure 5A**). This also demonstrates that our approach can identify genes that are located more further
5 from the lead SNP but are positioned within same broader regulatory domain (**Figure 5B**) - as based
6 on Hi-C data from IMR90 fibroblast cells³⁸ (**Figure 5B** top).

7
8 Currently 32 stroke loci have been implicated by GWAS studies⁴. From the new main stroke loci
9 found by Malik *et al.*, only *SLC44A2* was significantly enriched in scRNAseq populations. However,
10 we identified other potential targets in SMCs, ECs and in B cells and regulatory T Cell populations
11 (**Figure 3A**). One of the targets, *ESAM*, was not yet directly linked to atherosclerosis, but known to
12 enhance vascular permeability in selective tissues³⁹ and could therefore contribute to the influx of
13 cholesterol and inflammatory cells in the arteries contributing to the stroke phenotype. Another target,
14 Gs-coupled receptor calcitonin receptor-like receptor (*CALCRL*) is expressed in both ECI and ECII
15 populations. Activation of this receptor leads to an anti-inflammatory response in endothelial cells.
16 Subsequently *CALCRL* deficient mice present with more advanced atherosclerotic lesions⁴⁰. A recent
17 study by Örd *et al.* researched the enrichment of CAD, ischemic stroke and seven other
18 cardiometabolic GWAS signals in scATAC-seq peaks of carotid artery plaques. Similarly, they report
19 enrichment of CAD and stroke loci in SMCs and ECs and total cholesterol loci in macrophages⁴¹.

20 Remarkable is the absence of significant enrichment in the cells of the adaptive immune system of the
21 plaque for CAD. Out of six T cell populations, four exhibited the enrichment of 0.00 and
22 CD3+CD8A+ T Cells II and II showed enrichment of 1.3 and 0.37 respectively, indicating that there
23 is no or very little overlap between CAD candidate genes and cell population-specific DEGs. This
24 suggests that the genetic component of atherosclerotic disease is not rooted in the adaptive immune
25 system but rather in the structural cells of the vessel and influences the integrity of the vascular wall
26 or plaque stability.

1 Our workflow can be generalized to other single-cell transcriptomics datasets and complex traits. We
 2 applied it to scRNAseq-data from tissues with expected causal involvement in atherosclerotic disease:
 3 human liver cells and human pancreatic Langerhans Islet cells. GWASs studying lipid levels were
 4 associated with hepatocyte subsets but in particular subsets Hep3, Hep4 and Hep5 (**Supplemental**
 5 **Figure 6A**). It suggests that these risk factors for atherosclerosis have a stronger biological foundation
 6 in lipid metabolism in the liver- rather than local lipid metabolism in the plaque. Langerhans Islet
 7 cells show less prominent signals for our selected GWAS, suggesting that the genetic mechanisms
 8 involved in atherosclerosis do not have a local component in the pancreas (**Supplemental Figure 6B**).

9 Little to no signal can be observed for GWASs that have no relation to the cells composing the
 10 atherosclerotic plaque, such as height, cancers and behavioural related traits such as bipolar disorder
 11 and neuroticism (**Supplemental Figure 3A**). Yet, these GWASs have ample candidate genes and
 12 considerable GWAS sample size - suggesting the specificity of our approach. (**Supplemental Figure**
 13 **1B, C, Supplemental Table 1**). Similarly, no signal for BMI is found in plaque cell populations,
 14 intuitively confirming that the mechanisms underlying obesity cannot be accounted for in plaque
 15 tissue⁴².

16

17 This research, however, is subject to several limitations. In our workflow, a bias may be introduced by
 18 low-powered GWAS, or the definition of a gene, e.g. ± 50 kb around the transcription start and end
 19 sites, but it is not affected by the tissue expression. Only in our second step, we introduce a
 20 prioritization step for tissue expression and cell populations.
 21 In highly heterogeneous tissues, such as atherosclerotic plaque, multiple genes can be shared amongst
 22 the same class of cell populations (e.g. all endothelial cells), whilst they are lacking in others. To
 23 resolve this problem and to keep such genes as cell type-specific, we implemented an approach that
 24 can unambiguously uncover cell-specific genes without the need of a predefined list of ubiquitously
 25 expressed genes. Although this method is inclusive, the balance between the sufficient number of
 26 differentially expressed genes and their specificity can be adapted depending on the scRNA-seq

dataset in question.
The depth and quality of the scRNAsequencing is a another limiting factor. The amount and kind of RNA that can be detected and favors the most abundant genes with highest mRNA counts per cell. As a result certain gene classes, like transcription factors or signaling molecules are more difficult to detect. At the same time, cell types like neutrophils and foam cells in atherosclerotic plaque are difficult to capture by single cell transcriptomics and might be underrepresented or completely missed. Digestion protocols, tissue quality and sample handling also affect the quality and/or cell composition of scRNA-seq data.

In conclusion, we systematically projected GWAS summary statistics for 46 traits directly into single-cell transcriptomic profiles across 18 distinct cell populations derived from plaques. We devised a two-step workflow that 1) leverages large-scale human genetic data to prioritize trait-associated genes, and 2) calculates a cell-type specific enrichment score. We uncovered 11 potential cell population specific targets for CAD and provided functional evidence for *KANK2*, *SKI* and *EDNRA* in smooth muscle cells. Additionally, we verified the robustness and transportability of our approach by confirming the tissue- and trait-specific enrichment of circulating lipid genes in liver and glycemic disorder genes in pancreatic tissue. The principal strength of our framework is the cell-specific resolution it offers for disease associated genes mapped to GWAS loci, thus putting the emphasis on prioritization of candidates (instead of adding to an ever growing candidate-list) and accelerating the path to functional testing.

Data availability

Custom scripts used for this publication are available on GitHub <https://github.com/CirculatoryHealth/gwas2single>. Raw summary statistics can be accessed through their original publications (**Supplemental Table 4**). Quality controlled and/or FUMA-MAGMA processed summary statistics are available for download at Zenodo

(<https://zenodo.org/record/4823040>). Raw human carotid artery plaque scRNAseq data can be accessed via <https://dataverse.nl/dataverse/atheroexpress>. The plaque scRNAseq data after quality control, i.e. counts, scores, etc., used in this paper can be found at <https://doi.org/10.34894/TYHGEF>. Other data supplements can be found in the **Supplementary Data section**.

Funding

This work was supported by grants from: Fondation Leducq ['PlaqOmics' 18CVD02 to G.P. and C.L.M.], the National Institutes of Health [R01HL148239 and R00HL125912 to C.L.M.], the Netherlands CardioVascular Research Initiative of the Netherlands Heart Foundation [CVON 2011/B019 and CVON 2017-20 supporting S.W.v.d.L.], Generating the best evidence-based pharmaceutical targets for atherosclerosis [GENIUS I&II supporting S.W.v.d.L.], Interuniversity Cardiology Institute of the Netherlands [ICIN, 09.001] supporting S.W.v.d.L.], American Heart Association Transformational Project Award [19TPA34910021 supporting R.A. and M.C.].

Conflict of interest

The Authors declare that there is no conflict of interest.

Acknowledgements

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from: FUMA GTEx Portal on 12-07-2020.

We thank Sonya A. MacParland and Mauro J. Muraro for providing their scRNA-seq data on human liver and pancreatic islet cells at our request.

- 1 We are thankful for the support of the ERA-CVD program ‘druggable-MI-targets’ (grant number:
- 2 01KL1802), the EU H2020 TO_AITION (grant number: 848146) and the Leducq Fondation
- 3 ‘PlaqOmics’ (grantnumber: 18CVD02).

References

1. Klein, R. J. *et al.* Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science* **308**, 385–389 (2005).
2. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
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).
7. Khera, A. V. & Kathiresan, S. Genetics of coronary artery disease: discovery, biology and clinical translation. *Nat. Rev. Genet.* **18**, 331–344 (2017).
8. Kamb, A., Harper, S. & Stefansson, K. Human genetics as a foundation for innovative drug development. *Nat. Biotechnol.* **31**, 975–978 (2013).
9. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).

- 1 10. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes.
2 *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).
- 3 11. Mancuso, N. *et al.* Integrating Gene Expression with Summary Association Statistics to
4 Identify Genes Associated with 30 Complex Traits. *Am. J. Hum. Genet.* **100**, 473–487
5 (2017).
- 6 12. Petretto, E. Single cell expression quantitative trait loci and complex traits. *Genome*
7 *Med.* **5**, 72 (2013).
- 8 13. Verhoeven, B. A. N. *et al.* Athero-express: differential atherosclerotic plaque expression
9 of mRNA and protein in relation to cardiovascular events and patient characteristics.
10 Rationale and design. *Eur. J. Epidemiol.* **19**, 1127–1133 (2004).
- 11 14. Depuydt, M. A. C. *et al.* Microanatomy of the Human Atherosclerotic Plaque by Single-
12 Cell Transcriptomics. *Circ. Res.* **127**, 1437–1455 (2020).
- 13 15. R Core Team (2020). — European Environment Agency.
14 [https://www.eea.europa.eu/data-and-maps/indicators/oxygen-consuming-substances-in-](https://www.eea.europa.eu/data-and-maps/indicators/oxygen-consuming-substances-in-rivers/r-development-core-team-2006)
15 [rivers/r-development-core-team-2006](https://www.eea.europa.eu/data-and-maps/indicators/oxygen-consuming-substances-in-rivers/r-development-core-team-2006).
- 16 16. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.* **20**, 257–272
17 (2019).
- 18 17. van Rossum, G. Python tutorial, Technical Report CS-R9526. in (1995).
- 19 18. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation.
20 *Nature* **526**, 68–74 (2015).
- 21 19. Leeuw, C. A. de, Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-
22 Set Analysis of GWAS Data. *PLOS Comput. Biol.* **11**, e1004219 (2015).
- 23 20. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and
24 annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).

- 1 21. Ma, W. F. *et al.* Single-cell RNA-seq analysis of human coronary arteries using an
2 enhanced workflow reveals SMC transitions and candidate drug targets. *bioRxiv*
3 2020.10.27.357715 (2021) doi:10.1101/2020.10.27.357715.
- 4 22. Suzuki, H. *et al.* c-Ski inhibits the TGF- β signaling pathway through stabilization of
5 inactive Smad complexes on Smad-binding elements. *Oncogene* **23**, 5068–5076 (2004).
- 6 23. Pan, H. *et al.* Single-Cell Genomics Reveals a Novel Cell State During Smooth Muscle
7 Cell Phenotypic Switching and Potential Therapeutic Targets for Atherosclerosis in
8 Mouse and Human. *Circulation* **142**, 2060–2075 (2020).
- 9 24. Park, M. *et al.* Soluble endothelial cell selective adhesion molecule and cardiovascular
10 outcomes in patients with stable coronary disease: A report from the Heart and Soul
11 Study. *Atherosclerosis* **243**, 546–552 (2015).
- 12 25. Hamczyk, M. R. & Andrés, V. Vascular smooth muscle cell loss underpins the
13 accelerated atherosclerosis in Hutchinson-Gilford progeria syndrome. *Nucleus* **10**, 48–
14 54 (2019).
- 15 26. O'Leary, D. H. & Polak, J. F. Intima-media thickness: a tool for atherosclerosis imaging
16 and event prediction. *Am. J. Cardiol.* **90**, 18L–21L (2002).
- 17 27. Aherrahrou Redouane *et al.* Genetic Regulation of Atherosclerosis-Relevant Phenotypes
18 in Human Vascular Smooth Muscle Cells. *Circ. Res.* **127**, 1552–1565 (2020).
- 19 28. MacParland, S. A. *et al.* Single cell RNA sequencing of human liver reveals distinct
20 intrahepatic macrophage populations. *Nat. Commun.* **9**, (2018).
- 21 29. Muraro, M. J. *et al.* A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst.*
22 **3**, 385–394.e3 (2016).
- 23 30. Global Lipids Genetics Consortium. Discovery and refinement of loci associated with
24 lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).

- 1 31. Nguyen, T.-M., Shafi, A., Nguyen, T. & Draghici, S. Identifying significantly impacted
2 pathways: a comprehensive review and assessment. *Genome Biol.* **20**, 203 (2019).
- 3 32. Shadrina, A. S. *et al.* Prioritization of causal genes for coronary artery disease based on
4 cumulative evidence from experimental and in silico studies. *Sci. Rep.* **10**, 10486 (2020).
- 5 33. Ongen, H. *et al.* Estimating the causal tissues for complex traits and diseases. *Nat.*
6 *Genet.* **49**, 1676–1683 (2017).
- 7 34. van der Wijst, M. *et al.* The single-cell eQTLGen consortium. *eLife* **9**, e52155 (2020).
- 8 35. Potì, F., Ceglarek, U., Burkhardt, R., Simoni, M. & Nofer, J.-R. SKI-II – a sphingosine
9 kinase 1 inhibitor – exacerbates atherosclerosis in low-density lipoprotein receptor-
10 deficient (LDL-R^{-/-}) mice on high cholesterol diet. *Atherosclerosis* **240**, 212–215 (2015).
- 11 36. Sun, Z. *et al.* Kank2 activates talin, reduces force transduction across integrins and
12 induces central adhesion formation. *Nat. Cell Biol.* **18**, 941–953 (2016).
- 13 37. Buono, M. F. *et al.* The changing landscape of the vulnerable plaque: A call for fine-
14 tuning of preclinical models. *Vascul. Pharmacol.* 106924 (2021)
15 doi:10.1016/j.vph.2021.106924.
- 16 38. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals
17 principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
- 18 39. Duong Cao Nguyen *et al.* Interference With ESAM (Endothelial Cell-Selective Adhesion
19 Molecule) Plus Vascular Endothelial-Cadherin Causes Immediate Lethality and Lung-
20 Specific Blood Coagulation. *Arterioscler. Thromb. Vasc. Biol.* **40**, 378–393 (2020).
- 21 40. Nakayama, A. *et al.* Disturbed flow-induced Gs-mediated signaling protects against
22 endothelial inflammation and atherosclerosis. *JCI Insight* **5**, (2020).
- 23 41. Örd, T. *et al.* Single-Cell Epigenomics and Functional Fine-Mapping of Atherosclerosis
24 GWAS Loci. *Circ. Res.* **129**, 240–258 (2021).

- 1 42. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity
2 biology. *Nature* **518**, 197–206 (2015).
- 3 43. Roselli, C. *et al.* Multi-ethnic genome-wide association study for atrial fibrillation. *Nat.*
4 *Genet.* **50**, 1225–1233 (2018).
- 5 44. Aragam, K. G. *et al.* Phenotypic Refinement of Heart Failure in a National Biobank
6 Facilitates Genetic Discovery. *Circulation* (2018)
7 doi:10.1161/CIRCULATIONAHA.118.035774.
- 8 45. Evangelou, E. *et al.* Genetic analysis of over 1 million people identifies 535 new loci
9 associated with blood pressure traits. *Nat. Genet.* **50**, 1412–1425 (2018).
- 10 46. Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci
11 associated with smoking behavior. *Nat. Genet.* **42**, 441–447 (2010).
- 12 47. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using
13 high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513
14 (2018).
- 15 48. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body
16 mass index in ~700000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–
17 3649 (2018).
- 18 49. Jansen, I. E. *et al.* Genome-wide meta-analysis identifies new loci and functional
19 pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).
- 20 50. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum
21 disorder. *Nat. Genet.* **51**, 431–444 (2019).
- 22 51. Stahl, E. A. *et al.* Genome-wide association study meta-analysis identifies seven new
23 rheumatoid arthritis risk loci. *Nat. Genet.* **42**, 508–514 (2010).

- 1 52. Okbay, A. *et al.* Genetic variants associated with subjective well-being, depressive
2 symptoms, and neuroticism identified through genome-wide analyses. *Nat. Genet.* **48**,
3 624–633 (2016).
- 4 53. Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with
5 educational attainment. *Nature* **533**, 539–542 (2016).
- 6 54. Savage, J. E. *et al.* Genome-wide association meta-analysis in 269,867 individuals
7 identifies new genetic and functional links to intelligence. *Nat. Genet.* **50**, 912–919
8 (2018).
- 9 55. Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine
10 the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).
- 11 56. Nalls, M. A. *et al.* Identification of novel risk loci, causal insights, and heritable risk for
12 Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet*
13 *Neurol.* **18**, 1091–1102 (2019).
- 14 57. Demenais, F. *et al.* Multiancestry association study identifies new asthma risk loci that
15 colocalize with immune-cell enhancer marks. *Nat. Genet.* **50**, 42–53 (2018).
- 16 58. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel
17 disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986
18 (2015).
- 19 59. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci.
20 *Nature* **551**, 92–94 (2017).
- 21 60. Schumacher, F. R. *et al.* Association analyses of more than 140,000 men identify 63
22 new prostate cancer susceptibility loci. *Nat. Genet.* **50**, 928–936 (2018).

1 61. van Rheenen, W. *et al.* Genome-wide association analyses identify new risk variants
2 and the genetic architecture of amyotrophic lateral sclerosis. *Nat. Genet.* **48**, 1043–1048
3 (2016).

4 62. Zheng, H.-F. *et al.* Whole-genome sequencing identifies EN1 as a determinant of bone
5 density and fracture. *Nature* **526**, 112–117 (2015).

6

7

Figure legends

Figure 1. Finding suitable therapeutic targets from GWAS with scRNA-seq.

A) UMAP visualisation of 6,191 cells divided into 18 cell populations derived from 38 human carotid artery plaques. B) Heatmap of frequently studied CAD GWAS hits: *EDN1*, *HDAC9*, *NOS3*, *VEGFA*, *FNI*, *APOE*, *APOB*, *SMAD3*, *PCSK9* and *LDLR* in atherosclerotic tissue shown in cells depicted in cell populations shown in A.

Figure 2. GWAS candidate genes show cell-specific enrichment in atherosclerotic plaque cell populations.

A) Plot representing overlaps of GWAS candidate genes on cardiovascular disease with genes specifically expressed in cell populations from human atherosclerotic plaque. Colour intensity indicates enrichment over random data and size indicates adjusted p value (Benjamini-Hochberg). B) Genes for CAD underlying the overlaps in SMC and ECI populations. Only significant overlaps are plotted. C) Dotplot depicting expression of genes from B in the respective cell populations. D) Average of normalized expression (zero mean across samples) of genes from CAD in SMCs from A in 30 tissues from GTEx.

Figure 3. GWASs related to stroke show substrate in Endothelial cells.

A) Plot representing overlaps of GWAS candidate genes on stroke with genes specifically expressed in cell populations from human atherosclerotic plaque. Colour intensity indicates enrichment over random data and size indicates adjusted p value (Benjamini-Hochberg). B) Heatmap depicting expression of genes from any stroke from A in the respective cell populations. Only significant overlaps are plotted, symbols represent the cell populations with the significant overlaps. C) Average of normalized expression (zero mean across samples) of genes from any stroke in from B in 30 tissues from GTEx. D) Genes for any ischemic stroke underlying the overlaps in ECI and ECII populations. Only significant overlaps are plotted. E) Genes for large artery stroke underlying the overlaps in T Reg and B cells II populations. Only significant overlaps are plotted.

1

2 **Figure 4. Analysis of 3 sub-populations of atherosclerotic Smooth Muscle Cells.** A) UMAP
3 visualisation of 3 sub cell populations of SMCs in atherosclerotic plaque. B) Dotplot depicting
4 expression of synthetic and contractile SMC markers in the 3 sub-populations. C) Plot representing
5 overlaps of GWAS candidate genes on cardiovascular disease and cardiometabolic traits with genes
6 specifically expressed in the subset of SMCs from human atherosclerotic plaque. Colour intensity
7 indicates enrichment over random data and size indicates adjusted p value (Benjamini-Hochberg) D)
8 Plot depicting the number of unique and intersecting candidate genes from Large artery stroke
9 enriched in Transitional SMCs and Synthetic SMCs.

10 **Figure 5. SMC associated candidates influence migration and proliferation *in vitro*.** A)
11 Correlation of *KANK2*, *SKII* and *EDNRA* expression non-proliferative and proliferative cells. B) Plot
12 depicting the Hi-C interactions around the main SNP associated with *KANK2* (top) and regional
13 association plot (bottom) in Hi-C data derived from IMR90 fibroblasts. AUC: Area under curve; Pi :
14 Inorganic Phosphate; DHS: DNaseI hypersensitivity sites; TAD: Topologically associating domain.

15 **Supplemental Figure 1. Overview of the 46 selected GWAS summary statistics.** A) Visual
16 representation of 46 GWAS across 3 categories: atherosclerotic disease, cardiometabolic traits, and
17 others. B) Barplot depicting the n candidate genes per GWAS with Z score cutoff 3. Horizontal line
18 indicates the threshold of max candidate genes (150) included in the gene score calculations. C)
19 Barplot depicting the sample size per GWAS. D) Genetic heritability as calculated using LD Score
20 Regression. LDSC: Linkage Disequilibrium Score Regression.

21

22 **Supplemental Figure 2. Schematic overview of *in-silico* workflow.** A) GWAS summary statistic
23 processing. i) GWAS summary statistics undergo quality control and are transformed into a uniform
24 file format. ii) Individual GWAS summary statistics are processed followed by gene based testing by
25 MAGMA, resulting in a list of associated genes per GWAS. B) Yielding scRNA-seq cell population

1 specific genes. i) scRNA-seq data undergo quality control. ii) Differentially expressed genes are
2 acquired via a “one vs. one” or “one vs. all” method. C) Gene score calculations. i) Generate a
3 random GWAS dataset from all other GWASs and subject them to step i. ii) Where DEGs and
4 candidate genes overlap, the size of the overlap can be adjusted. Remaining genes will be scored and
5 ranked according to average expression in the scRNA-seq data, and the scores for the random data. iii)
6 Enrichment over random genes is calculated. Results have resolution on a single gene basis.

7

8 **Supplemental Figure 3. Gene score results for 46 GWAS in atherosclerotic plaque.** A) Plot
9 representing overlaps of GWAS candidate genes on cardiovascular disease and cardiometabolic traits
10 with genes specifically expressed in cell populations from human atherosclerotic plaque from
11 AtheroExpress biobank study. Colour intensity indicates enrichment over random data and size
12 indicates p value (Wilcoxon rank sum test). B) Equivalent to A, but cell populations from
13 atherosclerotic plaques from Pan *et al.*. C, D) Bar graph representing the numbers of DEGs per cell
14 population. E) Heatmap showing similarity between the two different plaque datasets as determined
15 with SingleR (log2 transformed data).

16

17 **Supplemental Figure 4. Heatmap with the DEG distribution of carotid atherosclerotic plaque.**
18 Average expression of cell-specific DEGs in cell populations from human atherosclerotic plaque.

19

20 **Supplemental Figure 5. Smooth Muscle Cell population can be divided into three sub-**
21 **populations.** A) Pathway analysis of DEGs specific for synthetic and contractile SMCs. B) Plot
22 depicting the number of unique and intersecting DEGs per sub-population. C) Plot representing
23 overlaps of GWAS candidate genes on cardiovascular disease and cardiometabolic traits with genes
24 specifically expressed in the subset of SMCs from human atherosclerotic plaque. Colour intensity
25 indicates enrichment over random data and size indicates p value (Wilcoxon rank sum test). D)

1 Heatmap showing similarity between SMC sub-populations from Athero-Express and Pan *et al.* as
2 determined with SingleR (log2 transformed data). **E)** Plot depicting the number of unique and
3 intersecting candidate genes from CAD in Fibrochondrocytes and Transitional SMCs.

4

5 **Supplemental Figure 6. Gene score results for 46 GWAS in liver and Langerhans islets.** **A)** Plot
6 representing overlaps of GWAS candidate genes on cardiovascular disease and cardiometabolic traits
7 with genes specifically expressed in cell populations from human liver from MacParland *et al.*.
8 Colour intensity indicates enrichment over random data and size indicates p value (Wilcoxon rank
9 sum test). **B)** Equivalent to **A**, but in cell populations from Langerhans islets from Muraro *et al.*. **C, D)**
10 Bar graph representing the numbers of DEGs per cell population.

11

1 Appendices

1 Text tables

2

1 Figures

1 Supplementary files

2 Supplementary methods

3 **PlaqView**

4 To facilitate the sharing and exploration of our dataset and uncovering of potential therapeutic
 5 candidates, we integrated our plaque data into PlaqView²¹: www.plaqview.com. PlaqView is a
 6 standalone, interactive, and reproducible web-based tool to explore single-cell RNA-sequencing data
 7 of cardiovascular tissue material. Within PlaqView, users can quickly query gene expression, cell
 8 trajectories, and explore drug targets within the context of the single-cell data. On PlaqView,
 9 researchers can explore our dataset, download high-quality graphical outputs without the need for
 10 prior coding experience. PlaqView is built on R and Shiny, and its source code is publicly available at
 11 <https://github.com/MillerLab-CPHG/PlaqView>.

12

13 **Estimation of SNP-heritability and genetic correlation**

14 We estimated the SNP-heritability due to genotyped and imputed SNPs (□g, the proportion of
 15 phenotypic variance casually explained by common SNPs) of 46 atherosclerotic, cardiometabolic, and
 16 other traits using high-definition likelihood (HDL) and LD score regression (LDSC). HDL was
 17 performed using Rpackage "HDL" v1.3.9 (<https://github.com/zhenin/HDL>) and LDSC was
 18 performed using LDSC software v1.0.1 (<https://github.com/bulik/ldsc>). In LDSC analysis, GWAS
 19 summary statistics were harmonized by the *munge sumstats* procedure. In HDL analysis,
 20 we estimated the SNP-based narrow-sense heritability and those with SNPs with less than 90%
 21 overlap with the HDL reference panel (UKB_imputed_SVD_eigen99_extraction) were removed.

22

23 **Single cell sequencing: plaque, liver and pancreas**

Sequencing data from human carotid artery plaques²³, liver²⁸ and pancreas²⁹ can be accessed through their original publications (**Supplemental Table 4**). Prior to downstream processing, reads from liver and pancreas were filtered for UGDH-AS1, PGM2P2, KCNQ1OT1, MALAT1, PGM5P2, MAB21L3, EEF1A1, and ERCC spike-ins. Original cell population identities were preserved and no additional filtering on cells and genes was performed in all cases.

Functional assay of smooth muscle cells

Vascular smooth muscle cells (SMCs) were isolated from aortic explants of anonymous donors (n=136) through the UCLA heart transplant program as described previously²⁷. Cells from additional 15 donors were purchased from Lonza and PromoCell. They were maintained in Smooth Muscle Cell Basal Medium (SmBM, CC-3181, Lonza) supplemented with Smooth Muscle Medium-2 SingleQuots Kit (SmGM-2, CC-4149, Lonza) (complete media). They were cultured in complete media (containing 5% FBS) until 90% confluence. Media was then switched to either serum-free media to mimic the quiescent state of SMCs in healthy arteries or continued to culture in complete media to mimic the proliferative state of SMCs in atherosclerotic arteries for 24 hours. Total RNA was extracted with Qiagen miRNEasy kit. Sequencing libraries were prepared with the Illumina TruSeq Stranded mRNA Library Prep Kit and were sequenced to ~100 million read depth with 150 bp paired-end reads at the Psomogen sequencing facility. The reads with low average Phred scores (<20) were trimmed using Trim Galore. They were then mapped to the hg38 version of the human reference genome using the STAR Aligner in two-pass mode to increase the mapping efficiency and sensitivity. Only the uniquely mapped read pairs were retained. Gene expression was quantified by calculating the transcripts per million (TPM) for each gene using RNA-SeQC based on GENCODE v32 transcript annotations. Proliferation, migration, and calcification were also measured in these cells. The amount of calcification was quantified in media containing high inorganic phosphate or osteogenic stimuli. These two media formulations recapitulate different aspects of arterial calcification that occur in advanced stages of atherosclerosis. The proliferation was quantified in control media and media

1 containing platelet-derived growth factor-BB (PDGF-BB), transforming growth factor β 1 (TGF- β 1),
2 or interleukin-1 β (IL-1 β). Finally, the amount of migration was quantified to a PDGF-BB gradient
3 over 24 hours using a modified Boyden chamber assay. Overall, 12 quantitative phenotypes related to
4 calcification, proliferation, and migration were performed. Correlation between gene expression and
5 phenotypes were calculated using Pearson Correlation Coefficient in R.

6

- 1 Supplementary table list
- 2 Supplemental Table 1
- 3 Supplemental Table 2 (access via Zenodo, see data availability)
- 4 Supplemental Table 3 (access via Zenodo, see data availability)
- 5 Supplemental Table 4
- 6 Supplemental Table 5
- 7

Workflow

GWAS categories

atherosclerotic
and cardio-
vascular
disease
(n = 12)

risk factors
(n = 14)

other
(n = 20)

GWAS
summary
statistics

gene-
based
test

Candidate
genes

DEG

parameter
toggles

enrichment
calculation

exploration

target
gene(s)

Functional
assesment

**Carotid artery
plaques** *Athero-
Express, Pan et
al. 2020*

Liver *MacParland
et al. 2018*

**Langerhans
Islets** *Muraro
et al. 2016*

Determine DEGs
per tissue per cluster

Role of CAD target genes in SMCs

CAD genes enriched
in smooth muscle
cells, specifically
Transitional smooth
muscle cells of
atherosclerotic plaque

SKI, KANK2 & EDNRA
new and under-
studied targets in
smooth muscle cells

Target gene expression
correlated with
migration, proliferation
and calcification
in vascular SMCs

medRxiv preprint doi: <https://doi.org/10.1101/2021.11.23.21266487>; this version posted November 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

A

B

A

B

C

D

A
i

ii

B
i

ii

C
i

ii

iii

Expression

Identity

- CD3+CD4+ T Cells I
- CD3+CD4+ T Cells II
- CD3+ Regulatory T Cells
- CD3+CD8A+ T cells I
- CD3+CD8A+ T Cells II
- CD3+CD8A+ T cells III
- ACTA2+ Smooth Muscle Cells
- CD34+ Endothelial Cells I
- CD34+ Endothelial Cells II
- CD14+CD68+ Macrophages I
- CD14+CD68+ Macrophages II
- CD14+CD68+ Macrophages III
- NCAM1+ Natural Killer Cells
- CD79A+ B Cells I
- CD79A+ B Cells II
- KIT+ Mast Cells
- Mixed Cells I
- Mixed Cells II

A

B

C

D