Utrecht Universiteit

UMC Utrecht

Major Research Project
MSc in Bioinformatics and Biocomplexity

Computational deconvolution of atherosclerotic plaques cell composition and its clinical association

Examiner:

Dr. Michal Mokry

Student:

Gemma Bel Bordes (1988018)

Daily supervisor:

Lotte Slenders

November 2022

Abstract

Atherosclerotic plaques are highly complex and heterogeneous, with a large number of extracellular components and cell types involved. Traditional analysis, like histology assessment of the plaques, have been useful to characterize their vulnerability, leading to a binary risk stratification of patients. However, there is a need for other analysis to assess the plaque composition at higher resolution. Single-cell (sc) RNA sequencing (RNA-seq) studies have identified the cell composition of human carotid plaques, but the costs of this technology forces patient cohorts to be too small for clinical studies. Recently, deconvolution has been introduced to infer the cell type proportions of samples with bulk RNA-seq data using a cell type reference derived from the sc RNA-seq data. We aimed to perform deconvolution of human carotid plaques from a large cohort of patients that underwent endarterectomy, to get their cell type proportions and associate these with sex (with linear regression) and atherosclerosis severity, including symptoms (with logistic regression) and future events after surgery (with Cox proportional hazard model). We first benchmarked different deconvolution procedures regarding the data processing and the method selection. We detected that reducing the number of cell types from the sc reference improved the deconvolution performance, but all methods resulted in very different cell abundances. However, proportions from macrophages and smooth muscle cells (SMCs) were consistent among methods and, more importantly, with plaque histology. We finally observed female plaques to have a tendency to host a greater SMC percentage (β = 1.86 [-0.25, 3.96], p = 0.084), and macrophages being associated with symptomatic plaques (odds ratio = 1.02 [1.01, 1.03], p = 0.003) and major cardiovascular events after endarterectomy (hazard ratio = 1.02 [1.00, 1.03], p = 0.012). Our results suggest that obtaining deconvolved cell proportions of plaques is not straightforward for all the cell types, but macrophage and SMC proportions were consistent. While macrophages were did not differ significantly from male to female plaques, the latter showed an increased SMC content. Macrophages, in turn, might constitute a risk factor for atherosclerosis severity, and more research on the subtypes that are responsible for this should follow.

Keywords: Deconvolution, benchmarking, cell composition, carotid plaques, macrophages, sex differences, atherosclerosis severity.

Layman's Summary

Cardiovascular diseases, affecting the heart or the blood vessels, are currently the leading cause of death globally. Among these diseases, there is atherosclerosis, which will likely affect most of us when we get old, and at younger ages if you are a man. During the years, fat accumulates in our arteries and this will become a risk, since the artery can get blocked or a piece of these accumulation can travel somewhere else in our body and cause damage. Also, the artery composition will change and there will be cells that are not normally present and can cause inflammation, for example. Healthy arteries are built with muscle cells, and in a diseased artery we will also find macrophages. Macrophages are cells that normally try to destruct harmful organisms and can initiate an immune response. Depending on this composition, researchers stated that there were patients more vulnerable than others, meaning that they can experience more problems related to the disease. And in general, men tend to have more vulnerable arteries. We can identify this changes by inspecting under a microscope a piece of the artery, but we think this is not precise enough because there are too many players in atherosclerosis. So, we need other methods to investigate this composition.

In the past years, we have been using technologies that measure the RNA of a mixture of cells from these arteries, but also of individual cells. RNA is similar to DNA, but while DNA is identical in all your cells, RNA is different. If we measure the RNA of individual cells from the artery, we can then label each cell with a cell type, so for each cell type we have a pattern. But measuring cells individually is expensive; it is cheaper to get the RNA from all cells in that piece of the artery, but then we do not know which cells we have. We said from the individual cells we have the RNA pattern of each cell type, so can we try to find these patterns in the RNA from the mixture? Yes, we can, and this is called deconvolution. With deconvolution we can get the percentage of each cell type in the mixture, in our case, the piece of artery.

A lot of researchers have come up with new deconvolution methods, and we used some to obtain the proportion of cells in the arteries from patients with atherosclerosis, who previously had surgery to get a piece of the diseased artery removed. We saw that the proportion of cell types really depended on the method we chose, which was inconvenient. But we detected agreement of these methods on some cell types, including muscle cells and macrophages, suggesting we could trust on the proportions defined by the deconvolution methods. And, we also validated this proportions with inspection of these arteries under the microscope. So, these two cell types seem to be interesting. We wanted to see if arteries from women were different from arteries from men. If we only focus on our two interesting cell types, we detected that female arteries were more muscular, they had more muscle cells. Macrophage content was similar for all the arteries. But, we found that macrophages were more abundant in arteries from patients that had had symptoms before the surgery. We also found that macrophages are a risk factor, because patients whose arteries had more macrophages, were more likely to have other kind of cardiovascular symptoms after the surgery.

Our results can help us understand why women and men are different when it comes to atherosclerosis. And we could also start thinking about new therapies that try to reduce the content of macrophages, which might be dangerous. But deconvolution is still a recent type of analysis. So, we might need to validate what we found with other analysis that are more hands-on and widely accepted in the research community.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Atherosclerotic plaque and its composition

Atherosclerosis is a chronic, systemic and inflammatory disease leading the cause of morbidity and mortality in Western countries but currently spreading worldwide [1]. It has its origins in the artery walls and, despite the tunica intima being the very first layer to be affected, it ends up modifying structurally and functionally all three layers of the vessel. Progressive accumulation of fatty and fibrous elements at specific sites of such arteries starts creating a lesion, known as atherosclerotic plaque, that can become more complex with time due to its calcification or haemorrhage. Moreover, lesion complexity rises with the infiltration of immune cells present in the blood stream in response to the retention of lipids, with T cells and monocytes deriving to (foam) macrophages as the main actors.

Advanced plaques can not only narrow the lumen but also rupture or erode causing a thrombotic event, finally leading to a threatening reduction in blood flow. Whereas the former mechanism was thought to be the main predictor of atherosclerosis severity for a long time [2–4], the latter concept of plaque destabilization, as well as inflammation status, has overcome the narrowing degree as the main risk factor [5–7]. Although erosion of the plaques remains to be fully understood, plaques that are prone to rupture, also referred to as vulnerable plaques, are known to have a large lipid core, thin fibrous cap and signs of inflammation [8]. These same characteristics have been associated with symptomatic plaques [9, 10], and other histological features like internal haemorrhage and vessel density have been proven to worsen the clinical outcome after surgical removal of the plaque [11]. Therefore, plaque composition itself must be considered a potential predictor of major cardiovascular events (MACE) and a knowledge key for understanding the mechanism of atherosclerosis progression. Of note, its composition might differ also between females and males, whose plaques generally present more vulnerability signs [12–16] whereas (young) female plaques are more likely to erode [17, 18].

Plaque composition, however, can be difficult to determine at high resolution given its high complexity and heterogeneity. On the one hand, matrix elements present at the lesion like calcium and lipid levels are worthy to be studied and imaging techniques, including magnetic resonance imaging, computerised tomography and ultrasounds, are used to non-invasively identify such components [16, 19, 20]. Alongside, assessment of post-mortem plaques with staining methods have also given insight on this matter, allowing a classification of plaques into different phenotypes [8] and the analysis of histology characteristics in different groups of patients, as mentioned above [9–12, 14, 15, 17]. Nonetheless, scoring based on visual inspection of the stained plaque is time-consuming, should be only performed by trained pathologists and inevitably underlies some source of bias due to intraobserver and interobserver variability [21]. On the other hand, even though such techniques have been used to identify some major cell type content like smooth muscle cells (SMCs) and macrophages, with ACTA2 and CD68 staining respectively, they are far from fully and precisely quantifying the cell composition of atherosclerotic plaques, since now these are known to host an important number of cell populations [22, 23].

1.2 RNA sequencing in atherosclerosis research

Since the establishment of RNA sequencing (RNA-seq) as a powerful technique to uncover biological processes, by means of studying gene expression in tissues, atherosclerosis research has benefit from it [24]. Comparing the transcriptome of plaques with different characteristics is important to find critical genes and proteins involved in plaque formation or progression for a given scenario. For example, whereas studies focusing on stable against unstable plaques will provide insight on the genes that might provoke unstabilization [24], examining male against female plaques will disentangle whether there might exist some sex dependent biological processes [25]. Recently, RNA-seq analysis of 654 human carotid plaques uncovered five distinct transcriptomic profiles, that could lead to a new plaque phenotype classification [26]. However, conventional bulk RNA-seq (i.e., RNA-seq of a whole tissue) only captures the average gene expression from all the cells present in the sample, making expression values be dominated by the most populated cell types. Plaque composition, in turn, will influence the expression levels, but there is no direct way to identify the cell type that is actually responsible for a signal.

This lack of resolution at the single-cell level is solved by single-cell (sc) RNA-seq, which provides gene expression of individual cells, allowing the identification of different cell types within the tissue. Such identification, in contrast to immunostaining or flow cytometry, does not rely on prior knowledge of plaque composition, allowing unexpected cell types to be discovered as well. In the past years, thanks to sc RNA-seq, knowledge on which cell (sub)types are present in atherosclerotic plaques, in both human and mice, has been expanded and reviewed elsewhere [27]. These concluded in the recognition of multiple types of SMCs, endothelial cells (ECs) and a large list of immune cell subtypes, belonging to macrophages, T cells, natural killer cells (NK) and B cells. When analyzing the proportion of cells that fall into one cell type, discrepancies are found in regards of the dominance of macrophages over T cells within the human plaque [23, 28]. However, such disagreement can be simply technical rather than biological, since the chosen tissue digestion process to select individual cells before sequencing, might be biased towards certain cell types [29]. Thus, proportions that one can infer from scRNA-seq do not need to represent the true picture.

1.3 Computational deconvolution of cell composition

Given that bulk RNA-seq gives average expression values directly influenced by the cell ratios present in the sample, and scRNA-seq obtain the expression pattern of each cell type that populates a sample, the integration of data coming from both sides, should be sufficient to infer the cell type proportions of a tissue. This concept is computationally adopted by the so-called deconvolution methods, which have been proliferating over the last two decades [28]. Most of these methods rely on the linear assumption formulated as:

$$T = C \times P$$

where T = bulk RNAseq expression values, C = cell type-specific expression values (i.e., given by the scRNA-seq data) and P = cell type proportions. Even though a group of these techniques do not make use of scRNA-seq data as the reference, requiring instead a list of gene markers, these are not addressed in this project.

From simpler to more complex implementation and assumptions upon which the methods are built, a selection of computational deconvolution techniques being used to infer the cell composition are: non-negative least squares (NNLS) [30], MuSiC [31], Bisque [32], CIBERSORTx [33] and Scaden [34]. NNLS simply fits a linear regression model but forces all its coefficients to be non-negative, which translates into only obtaining positive (or zero) proportions for all the cell types present in the scRNA-seq data. An extension to this method is found in both MuSiC and Bisque. While MuSiC corrects NNLS fractions by weighting those genes which expression is consistent in all the cells from the same cell type and show low variance across different subjects; Bisque transforms bulk data before applying NNLS so it resembles its scRNA-seq counterpart. In contrast, CIBERSORTx and Scaden do not rely on linear regression models and they instead apply a support vector regression model and a neural network, respectively. Despite the great development made in this field, there is still no ”one-size fits all” method and differences simply based on the methods architecture and the processing of the data, will influence the final cell type proportions derived from a tissue [35, 36]. Some of these methods might also work poorly when cell types share too similar expression profiles [37].

1.4 Project design

Taking together the importance of atherosclerotic plaque composition, both for understanding better the disease progression and for potentially stratifying patients based on their risk, and the difficulties found to accurately quantify the cellular content, we derived our research question:

- Can we computationally deconvolve cell type proportions of human carotid plaques and associate these with clinical features like sex and atherosclerosis severity?

To address this, we will divide the project into two parts (visually summarized in Fig 1):

1. Benchmarking to find the best deconvolution procedure, based on the previous data processing and the deconvolution method choice. Five deconvolution methods (i.e., NNLS, MuSiC, Bisque, CIBERSORTx and Scaden) will be fed by different processed data inputs that will differ on: the genes to be used from the reference, the transformation of the expression values and the clustering of the sc RNA-sequenced cells into different numbers of cell populations. To validate each procedure we will:
    - Perform deconvolution of artificial mixtures with known proportion.
    - Assess the relationship between plaque deconvolved proportions and the proportions derived from the sc data.
    - Assess the relationship between plaque deconvolved proportions and histology.
2. Clinical data association with plaque composition. For this we will:
    - Statistically assess the influence of sex on cell proportions.
    - Statistically assess the effect of cell proportions on atherosclerosis severity, focusing on symptoms and MACE after plaque removal.

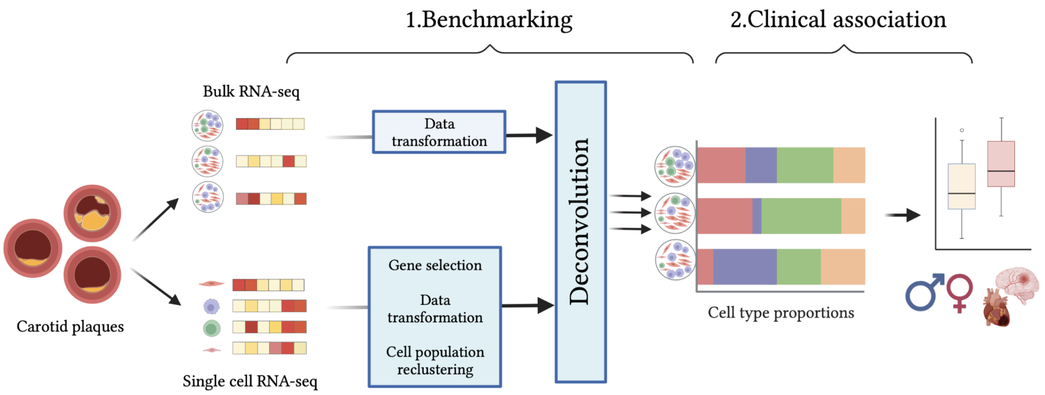

Fig. 1. Graphical design of the project. The first part of the project consists on benchmarking different deconvolution methods fed with differently processed datasets with bulk and scRNA-seq data. The second part of the project focuses on associating the cell type proportions from the chosen methodology with clinical data, including sex and atherosclerosis severity. Created with *biorender.com*

2 Methods

2.1 Study population

The data comes from the dutch Athero-Express (AE) biobank, which stores a large number of samples from patients who underwent carotid endarterectomy [38]. In particular for this project, the data selection includes plaques with RNA-seq analysis that were included between 2002 and 2016. From these, bulk RNA-seq and sc RNA-seq was performed for 656 and 46 plaques, respectively and exclusively. Associated to the bulk sequenced samples, there are baseline characteristics annotated for 654 patients, as well as a follow up to identify future MACE and histology analysis of their specimens. Besides, new bulk RNA-seq results were obtained for 28 plaques that overlapped with the 46 for which scRNA-seq was available. All study participants provided consent prior to surgery and the study followed the Declaration of Helsinki rules [39] and was approved by the Local Medical Ethical Committee.

2.1.1 Clinical features and follow up

Clinical data of the patients, like sex, was retrieved and put together with the results of other questionnaires which included questions regarding risk factors like smoking and use of medication. Symptoms were grouped into two categories: ”symptomatic”, including stroke and transient ischaemic attack, and ”asymptomatic”, including patients with negligible or minor signs, like ocular blindness for <24 hours. For the following 3 years after surgery, any MACE (i.e., vascular death, nonfatal myocardial infarction, nonfatal stroke and nonfatal aneurysm rupture) experienced by the participants were reported in the database.

2.2 AE carotid plaques

Surgically removed plaques were transferred immediately to the laboratory, where the culprit lesion (i.e., smallest lumen area) was identified, cut and prepared to be stained (for the full protocol, see [22, 38]). Sections surrounding the culprit lesion were properly stored for subsequent RNA-seq analysis.

2.2.1 Histological assessment of carotid plaques

Stainings for alpha-smooth muscle actin (ACTA2) and CD68 were performed to identify the presence of SMC and macrophages, respectively. These were binned into ”no”, ”minor”, ”moderate” and ”major” categories, and converted to numbers from 0 to 3. SMC to macrophage ratios were determined based on the same stainings and quantification was binned into “macrophage dominant””, “equal” or “SMC dominant”, which were again converted to numbers from 0 to 2. Plaque phenotype was determined based on fat percentage of the total lesion size area (plaques with <10% (fat score 1), 10-40% (fat score 2) and >40% (fat score 3) fat were categorized as ”fibrous” (converted to number 3), ”fibro-atheromatous” (converted to 2) and ”atheromatous” (converted to 1), respectively). Also, picrosirius red was stained for collagen (binned from 0 to 3) and immunohistochemical staining for CD34 was performed to assess vessel density, defined as a continuous variable (refer to [38] for the details). Besides, for the 46 plaques with sc RNA-seq data, CD3 and CD34 stainings were also performed and binned to numerical categories.

2.2.2 Bulk RNA sequencing

The whole RNA-seq protocol is described by Mokry et al. [26], but the main methodologies implemented were the following: CEL-seq2 [40] for library preparation, which is a 3' assay and adds unique molecular identifiers (UMIs); Bioanalyzer (Agilent) for both RNA and cDNA quality, the latter also assessed by Qubit fluorometric quantification (Thermo-Fisher); and Illumina Nextseq500 for sequencing. The sequencing output was aligned with BWA (0.7.13) to the human reference (Ensembl 84).

2.2.3 From bulk RNA-seq to different mixture inputs

We only included in the analysis protein-coding genes that were well annotated with HGNC names and we excluded all ribosomal genes. Moreover, in case that several reads mapped to the same gene, we only kept the read with the highest count. Since the UMIs used for the RNA-seq were 6 bp long, we set a maximum for counts to 4095 and we corrected them for UMI saturation so as *corrected_count = -4096*(ln(1-(raw_count/4096)))*. Once corrected, the original 656 plaques were subset from the 28 plaques that overlapped with the scRNA-seq dataset. While for the latter we did not implement any other processing step, the count matrix with the original 656 carotid plaques was kept in the linear scale (i.e. as it is) and transformed to the logarithmic (*ln(raw_count+1)*) and the squared root scale, creating 3 different inputs for the bulk data.

2.2.4 Single cell RNA sequencing

For 46 plaque samples, the removed specimen was used for scRNA-seq and the steps followed regarding the sample handling, the sequencing protocol and the reads processing are described elsewhere [22]. In short, the sample was enzymatically digested to obtain viable cells that were isolated by fluorescence-activated cell sorting. Single cells were then sequenced with Illumina Nextseq500 by first following again the CEL-Seq2 protocol for library preparation. Resulting counts were processed with Seurat [41], by removing doublets and low-quality cells, as well as mitochondrial and ribosomal genes. This resulted in 4948 cells and 20111 genes in total, which were annotated with HGNC symbols. In order to map each of these cells to a cell type, tSNE was applied to transformed data to obtain clusters, which were associated to a cell population by thoroughly analyzing their gene expression. This resulted in the identification of 20 cell types, including: inflammatory macrophages, resident macrophages, foam cells, dendritic cells, monocytes, FOXP3+ T cells, CD3+ T cells (I, II, III, IV, V and VI), natural killer (NK) cells (I and II), mastocytes, plasma B cells, switched-memory B cells, ECs (I and II) and SMCs.

2.2.5 From single cell RNA-seq to different reference inputs

For the purpose of deconvolution, we processed the counts matrix partly adopting the steps followed by Avila et al. [35]. Filtering out genes expressed in less than 5% of the cells was proposed by Avila et al. but we took it as a new experimental setting, with 2 resulting matrices including and excluding such genes (12264 out of 20111 genes). Then, we excluded, for both, those cells belonging to a low populated cell type (<50 sequenced cells), namely monocytes and plasma B cells. As in the previous section, these 2 matrices were kept in the linear scale, log-transformed and square root-transformed, resulting in 6 different inputs for the scRNA-seq data.

Besides the original clustering of the cells into 20 different cell types, from which two were already excluded in the previous step, we reclustered again the cells into lower resolution cell population to reduce the complexity of the deconvolution process, resulting in two new scenarios: reduction to 9 and 6 cell populations. For the 9-clustering, we clustered together all CD3+ T cells, all NK cells, all ECs and also one bigger cluster included foam cells, resident macrophages and resident macrophages (i.e., macrophage group); while all the other cell types remained independently, namely FOXP3+ T cells, SMCs, switched-memory B cells, mastocytes and dendritic cells. Whereas for the 6-clustering, we added the grouping of T and NK cells into one population and dendritic cells were also included in the previous macrophage group. From this step, we multiply by 3 the number of outputs created above accounting the different gene filtering and transformations (6x3 = 18 different inputs). Despite not being part of the initial workflow but a consequence of further results, we finally clustered the cell into 8 populations, namely dendritic cells, ECs I, ECs II, macrophages, mastocytes, SMCs, switched-memory B cells, and T and NK cells.

2.2.6 Single cell RNA proportions

For each of the 46 patients included in the sc RNA-seq dataset, we computed the sc-derived cell type proportions as the total number of cells belonging to each cell type divided by the total number of sequenced cells. In order to analyze the quality of such proportions, we computed the Kendall rank correlation coefficient between ECs, SMCs, macrophage and T cell proportions with CD34, ACTA2, CD68 and CD3 staining scores respectively.

2.3 Generation of artificial pseudo-bulk mixtures

The AE sc RNA-seq dataset including all the genes, in linear scale and with 18 cell types-annotation was used to artificially generate mixtures of cells, referred to as pseudo-bulk mixtures. All the individual cells were divided into the training and the test dataset (50%:50% split), making sure all patients were represented in both sets (i.e., there is at least one sequenced cell from each patient in both training and test set). Then, using the training set, we created 700 pseudo-bulk mixtures. For that, we first established random but uniform proportions for each cell type, which added up to 1, and then, we calculated the number of cells to be taken per cell type (total number of cells was 100). The mixtures were finally built by adding the counts of the randomly selected cells, and the initial cell type proportions were saved as the true proportions.

Finally, the test dataset with the remaining cells was again stored following the 3 different clustering strategies, from 18 cell types to 9 and 6 cell clusters. These 3 test sets, as well as the 700 pseudo-bulk mixtures, were finally transformed to obtain the data in the linear, log and square root scale. True proportions for each clustering scenario were obtained by adding the simulated proportions of the cell types that were merged.

2.4 External sc RNA-seq reference

An external sc RNA-seq dataset by Pan et al. [42] including 3 human carotid plaques was also considered for validation purposes. These data originally included 17818 genes and 8867 sequenced cells, for which 13 cell clusters were identified. In order to follow a similar process workflow as for the AE sc RNA-seq data, we excluded the lowest populated and undefined cell cluster labelled as *cluster 13*. We also merged the three defined macrophage types (annotated as 1, 2 and 3) into a single macrophage cluster. Therefore, after grouping this population, 10 cell clusters were kept for this dataset

2.5 Deconvolution scenarios to benchmark

Five deconvolution methods were tested, using sc RNA-seq data as the reference and bulk RNA-seq data as the mixtures for which the cell composition must be assessed. The selected methods included three regression-based approaches, namely NNLS [30], MuSiC [31] and Bisque [32], CIBERSORTx[33], which implements a support vector regression, and Scaden [34], which works with a complex neural network. Moreover, different processing pipelines were implemented in this step regarding the genes selection, the sc RNA-seq data clustering, and the bulk and sc RNA-seq data transformation, using different sourced datasets. The pipeline to decide the final methodology was:

1. Decision on sc RNA-seq gene selection (i.e., using all genes or filtering out genes present in less than 5% of the cells).
2. Decision on data transformation (i.e., using linear, log or square root scale).
3. Decision on number of cell clusters in the reference dataset (i.e., all 18 cell types or a reduced number of populations).
4. Decision on deconvolution method (i.e., NNLS, MuSiC, Bisque, CIBERSORTx or Scaden).

All deconvolution settings that were used for this study are summarized in Table 1 (note that in each pipeline step, some decisions were taken based on the results and therefore not every single combination of datasets was used as an input). Due to the requirements and/or architecture of the deconvolution methods, deconvolution using other than linear data (i.e., log and squared root scale) was only performed with the three regression-based methods.

Table 1. Summary of the deconvolution scenarios used for each benchmarking step. Step 1 for gene selection, 2 for data transformation, 3 for cell clustering and 4 for deconvolution method. AE refers to Athero-Express data and fields with (*) are new scenarios introduced due to the results of the previous step.

Sc RNA-seq (reference)				**Bulk RNA-seq (mixture)**		
Dataset	**Gene selection**	**Data trans-formation**	**Clustering (number of populations)**	**Dataset**	**Data trans-formation**	**Bench-marking step**
AE	All genes	Linear	18	AE (656 plaques)	Linear	1
AE	Genes >5% cells	Linear	18	AE (656 plaques)	Linear	1
AE	All genes	Linear	18	AE (656 plaques)	Linear	2
AE	All genes	Log	18	AE (656 plaques)	Log	2
AE	All genes	Square root	18	AE (656 plaques)	Square root	2
Splitted test sc (section 2.3)	All genes	Linear	18	Pseudo-bulk (section 2.3)	Linear	2
Splitted test sc (section 2.3)	All genes	Linear	9	Pseudo-bulk (section 2.3)	Linear	2

Splitted test sc (section 2.3)	All genes	Linear	6	Pseudo-bulk (section 2.3)	Linear	2
Splitted test sc (section 2.3)	All genes	Log	18	Pseudo-bulk (section 2.3)	Log	2
Splitted test sc (section 2.3)	All genes	Log	9	Pseudo-bulk (section 2.3)	Log	2
Splitted test sc (section 2.3)	All genes	Log	6	Pseudo-bulk (section 2.3)	Log	2
Splitted test sc (section 2.3)	All genes	Square-root	18	Pseudo-bulk (section 2.3)	Square-root	2
Splitted test sc (section 2.3)	All genes	Square-root	9	Pseudo-bulk (section 2.3)	Square-root	2
Splitted test sc (section 2.3)	All genes	Square-root	6	Pseudo-bulk (section 2.3)	Square-root	2
Splitted test sc (section 2.3)	All genes	Linear	18	Pseudo-bulk (section 2.3)	Linear	3
Splitted test sc (section 2.3)	All genes	Linear	9	Pseudo-bulk (section 2.3)	Linear	3
Splitted test sc (section 2.3)	All genes	Linear	6	Pseudo-bulk (section 2.3)	Linear	3
AE	All genes	Linear	8*	AE (656 plaques)	Linear	4
AE	All genes	Linear	8*	AE (28 plaques, matching sc)	Linear	4
Pan et. al. (ref)	All genes	Linear	10	AE (656 plaques)	Linear	4

2.6 Deconvolution

For NNLS and MuSiC, the MuSiC package in R was used and all the data inputs were converted to an expression set (R object with the count matrix and the metadata associated to each sample). We used the function *music_props* and we took the *Est.prop.allgene* results as the NNLS proportions, and the *Est.prop.weighted* as the MuSiC proportions. The same expression set was used for Bisque (BisqueRNA package), which was also implemented in R with the function *ReferenceBasedDecomposition.*

On the other hand, for CIBERSORTx we made use of their web portal (*https://cibersort.stanford.edu*), which requires text files for both reference and mixture data. From the former, a signature matrix is built, which will then be the input next to the mixture file for the *Impute Cell Fractions* step. Another kind of text files were created for Scaden deconvolution. As specified, we normalized the reference counts by the median library size and next, we created individual input files accounting the counts matrix and the cells annotation for each patient in the reference, as well as a general mixture file. These were the inputs for Scaden's publicly available Python scripts, which were run following this order: *simulate*, *process*, *train* and *predict.*

Of note, for the second part of the project (i.e., clinical association), we made some adjustments to the Scaden workflow for the sake of reproducibility, since the *simulate* step produced different outputs every time we ran the code. This step randomly takes cells from the reference dataset to create pseudo-bulk mixtures that subsequently train the neural network. To ensure reproducibility, we added a fixed random seed to the *bulk_simulator.py* script (see supplementary code, section 7.3). And we finally ran the whole Scaden process 10 times using 10 different seeds and took the average of all these proportions.

2.7 Evaluation of deconvolution scenarios

Since there was no ground truth associated to the bulk RNA-sequenced plaques, we could only visualize and biologically reason the differences we see between different settings and methods. However, we addressed the evaluation by (1) using other sources of data, (2) analyzing the correlation of results coming from different methods with each other, and (3) associating histological features annotated to the 654 plaques to their deconvolved proportions.

First, in order to assess the performance of the deconvolution of pseudo-bulk mixtures and the 28 plaques which had matching bulk and sc RNA-seq data, we computed the normalized root-mean-square error and the Pearson correlation between the deconvolved and the true proportions for each cell type and globally. While for pseudo-bulk analysis, the true proportions were taken from the artificial generation step of the pseudo-bulk mixtures, for the matching plaques we defined the true proportions as the proportions directly derived from the sc-RNA seq dataset.

For the final step of the benchmarking, we also computed the Pearson correlation between the proportions coming from all five deconvolution methods individually for each cell population. Moreover, Pearson correlations were computed between the proportions obtained using the AE reference and the Pan et. al. reference for overlapping cell populations.

Also, for the scenario above, we combined histological annotation of the plaques with deconvolved proportions to compute the Kendall rank correlation coefficient between SMC and macrophage proportions with related histology scores (i.e., ACTA2 and CD68 staining, plaque phenotype, fat and collagen), and Pearson correlation between EC proportions and vessel density.

2.8 Statistics on plaque composition associated with sex and atherosclerosis severity

For this part of the project, we made use of the cell type composition predicted by Scaden, from which we took the average proportions obtained by 10 runs using 10 different seeds (see section 2.6), and these were multiplied by 100 to get the percentages, to make the statistics more interpretable. The reference used was the AE sc RNA-seq dataset with 8 cell clusters, using all the genes in the dataset and the counts in linear scale. Deconvolved proportions were merged with clinical data for 654 patients.

Statistical analysis to determine sex differences in cell abundances were performed using univariate linear regression. Structural cells were defined as SMCs and both EC populations, whereas immune cells were defined as the remaining cell populations. Baseline characteristics of the 654 patients included were stratified by sex in Table S2.

Baseline characteristics for these patients were again stratified by symptoms (642 patients, Table S3) and MACE (649 patients, Table S4). Foucusing on macrophage content, we created univariate and multivariate models to determine whether it held predictive power. A logistic model was built taking symptoms as the outcome and macrophage proportion (and confounders in case of the multivariate analysis) as the predictors. Regarding future MACE, we used Cox proportional hazard models to study the event-free time related to the macrophage content. For this last analysis, we took the macrophage proportion as a continuous variable (i.e., as it is) but also as a ternary categorical variable, dividing the patients based on macrophage proportion tertiles for the sake of visualizing this association. As confounders, we included those variables that had a significant ($p < 0.05$) at the baseline level. In an attempt to unravel which macrophage subtype was giving the signal for these last associations, we performed the same deconvolution described above but unfolding the macrophage cluster (i.e., considering resident, inflammatory and foam macrophages as independent clusters), and we repeated the same statistical analysis.

In order to assess the same associations directly on the bulk AE RNA-seq data (with the 656 plaques), the same statistical tests were performed but substituting cell proportions for marker gene expression levels. The selected marker genes were: ACTA2 for SMCs, CD79A for switched-memory B cells, CD68 and CD14 for macrophages in general, ABCA1 for foam macrophages, IL1B for inflammatory macrophages, and LYVE1 for resident macrophages. Previously, the bulk counts were normalized and transformed to counts per million.

2.9 Software versions

R v4.1.2; Seurat v4; MuSiC v0.2.0; BisqueRNA v1.0.5; Python v3.8.5; Scaden v1.1.2.

3 Results

3.1 Benchmarking

To benchmark different deconvolution settings we focused both in data processing (assessing gene selection, data transformation and clustering of the sc RNA-seq reference) and the method selection (including NNLS, MuSiC, Scaden, CIBERSORTx and Scaden).

3.1.1 Filtering out low-detected genes from the reference dataset has a minor impact on the deconvolution but removes marker genes of some cell types from the reference

Setting the study of Avila Cobos et. al. [35] as the start point for the study, we see that a commonly adopted step previous to deconvolution is to remove undetectable genes from the reference, defined as genes present in less than 5% of the sequenced cells. However, we believe this step might be risky with AE data, since we have a high number of cell types in our reference dataset (i.e., AE sc RNA-seq data), even after removing the lowest populated cell types, namely plasma B cells and monocytes for the sake of deconvolution quality. We observed that for the remaining 18 cell types (Fig 2A), 11 of these do not reach the 5% of sequenced cells threshold (Fig 2B). Investigating genes that would be filtered out with this step, we detected marker genes for some of these 11 cell types, like CD79A for switched-memory B cells (Fig S1). In contrast, when performing deconvolution on the AE bulk RNA-sequenced plaques, taking this reference with all genes and only detectable genes, we saw that, in general, differences due to this filtering are minor compared to the differences in proportions due to the method choice (Fig 2C). Since this first benchmarking step seems to be little relevant, we just decided to work with all the genes in the reference from now on for the sake of cautiousness.

3.1.2 Data transformation has a significant effect in deconvolved results, being linear scale preferable

To assess the effect data transformation has on the methods built upon linear assumptions (i.e., NNLS, MuSiC and Bisque), we performed deconvolution using AE data in linear scale, but also in logarithmic and square-root scale, which could strengthen the linear relationship between highly skewed variables. We detected huge differences on deconvolved proportions emerging from differently transformed data (Fig 3A), forcing us to take a closer look to this effect. For that, we artificially built 700 pseudo-bulk mixtures with random but uniformly cell type proportions (Fig S3A; see Methods) by adding up the counts of 100 cells from half of the AE sc RNA-seq dataset. The other half and the pseudo-bulk mixtures were kept in linear scale and log and square-root transformed, and finally used as the input for new deconvolutions (Fig S3B depicts the deconvolved proportions for the linear case). Since the next benchmarking step considers different clustering levels of our AE sc RNA-seq dataset (i.e., different number of cell populations), we ensure this is not masking the transformation effect by repeating these experiments with the reference annotated with 18, 9 and 6 cell clusters (Fig 4). In all cases, MuSiC and NNLS deconvolved proportions of the pseudo-bulk mixtures were notoriously more accurate when the data was kept in the linear scale, but this behaviour was absent with Bisque, which showed a similar accuracy regardless of the data transformation (Fig 3B). Even though this last analysis does not include AE (real) bulk data, we believe the results could be translated to this other scenario and, therefore, linear data is kept for further analysis.

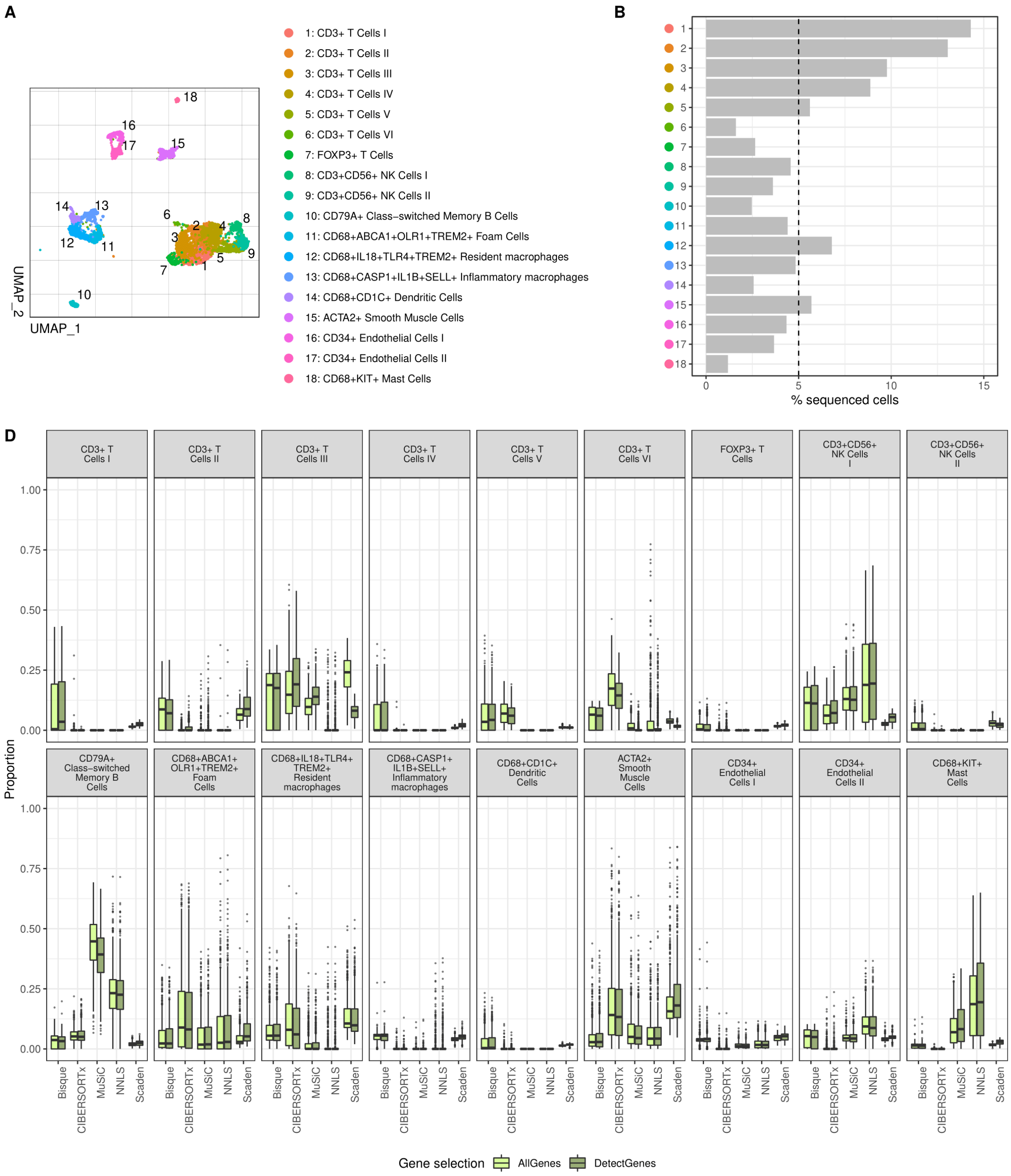

Fig. 2. Benchmarking 1st step: gene selection. A) UMAP visualization of 18 cell types identified in plaques from Athero-Express (AE). B) Percentage of sequenced cells per cell type. The dotted line corresponds to the 5% of total sequenced cells. C) Boxplot with the deconvolved cell type proportions of 656 AE plaques using 5 deconvolution methods and the reference using all genes and using only detectable genes.

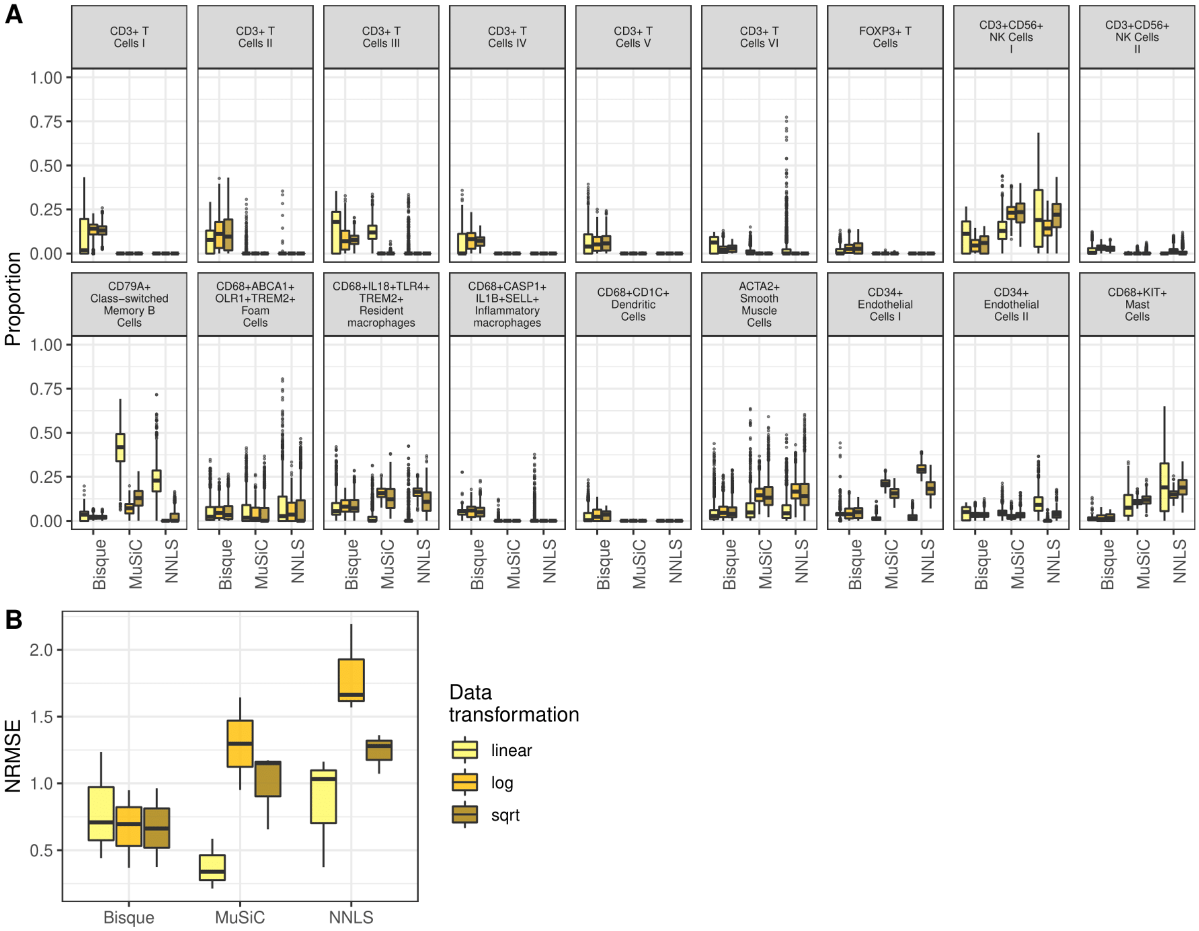

Fig. 3. Benchmarking 2nd step: data transformation. A) Boxplot with the deconvolved cell type proportions of 656 plaques from Athero-Express using 3 deconvolution methods built upon linear assumptions. B) Normalized root-mean-square error (NRMSE) between the deconvolved and the known proportions from 700 pseudo-bulk mixtures. Each box contains the NRMSE value from the deconvolution performed with the reference annotated with 18, 9 and 6 clusters.

3.1.3 Balance between resolution, accuracy and biology results in clustering the AE sc RNA-seq dataset into 8 cell populations

High similarity of cell types, in terms of their gene expression profiles, leads to a more complex deconvolution process [37]. Thus, we hypothesized that reclustering the cell populations identified in the AE sc RNA-seq data to a lower resolution level, would reduce this complexity. Instead of only considering the original 18 cell types-clustering, we grouped some of these cell types into bigger clusters, based on their similarities (Fig S4A, Fig S2), to obtain new references. Originally, the reduction was performed to 9 cell populations (taking all the cell types independently and grouping all CD3+ T cells, NK cells, ECs and macrophages) (Fig S4B) and 6 cell populations (taking the last clustering, but also grouping T and NK cells altogether, and adding dendritic cells to macrophages to obtain a CD68+ cluster) (Fig S4C).

Since such modifications only affect the sc RNA-seq data, we assessed the clustering effect focusing on pseudo-bulk mixtures (Fig S3) instead of the AE bulk RNA-sequenced samples. As mentioned in the section above, we performed the deconvolution of these pseudo-bulk mixtures using the other half of the AE sc RNA-seq dataset that was not used for building such artificial data. Next, we computed two metrics between the deconvolved and known proportions for the three clustering scenarios (Fig 4), namely correlation and normalized root-mean-square error (NRMSE). In some cases, decreasing the number of cell populations, and therefore the resolution, is beneficial (i.e., NRSME is lower and the correlation is higher). The most notorious improvement is detected in the grouping of T and NK cells, so we decide to keep it. In contrast, the metrics for ECs do not seem to change significantly and since having these two subtypes can be useful for further analysis (e.g., endothelial to mesenchymal transition studies), we will leave them independently. And finally, CD68+ cells clustering show a weak improvement, but going back to the biology and the possibilities we have with our data (e.g., macrophage histological assessment), we keep the whole macrophages group but leaving out dendritic cells, as a balance between accuracy of the deconvolution and resolution of cell populations. These decisions converge to a reference clustered into 8 cell populations: T and NK cells, switched-memory B cells, macrophages, SMCs, ECs I, ECs II and mastocytes.

3.1.4 Bisque proportions show the highest similarity to sc RNA-seq-derived proportions, but these are not in agreement with histology

Next, to obtain more insight into which method gives us the most accurate deconvolved proportions in real AE data, we performed deconvolution on 28 bulk RNA-seq samples from which sc RNA-seq data was also present. We hypothesized that real plaque cell composition would be, at least, comparable to the proportion of sequenced cells for each plaque. Therefore, from these 28 plaques we obtained both deconvolved proportions and sequenced cell proportions for the final 8 cell populations (Fig S5). Considering the latter as the true proportions, we computed the NRMSE and correlation globally and individually for each cell population (Fig 5A). Overall, Bisque outperforms the other methods in this analysis, particularly showing an improvement in the case of T and NK cells. However, proportions of sequenced cells are not always translated to the true composition due to bias in the digestion step. This might be the case for our sc RNA-seq data since, overall, its derived cell proportions showed no agreement with histology, when comparing CD3 score to T and NK cells, CD34 score to ECs, CD68 to macrophages and dendritic cells, and ACTA2 to SMC proportions (Fig 5B).

3.1.5 Despite absolute proportions highly differing between methods, macrophage and SMCs proportions show agreement with histology and great correlation among all methods

Going back to our original AE cohort with 656 bulk RNA-sequenced plaques, we deconvolved their proportions for the 8 cell populations defined in the AE reference. However, such proportions highly differ according to the method that was used (Fig 6A). While dendritic cells and ECs have low and similar proportions for all the methods, the other cell clusters show a high variability regarding the method choice. For example, Bisque suggests T and NK cells to be the most predominant cell population, whereas with CIBERSORTx this population is almost absent. MuSiC proportions are high for switched-memory B cells, which shows a lower abundance in all other cases, even with NNLS, which is the most similar method in terms of architecture. Finally, Scaden results in a balanced predominance of T and NK cells, macrophages and SMCs.

Next, we aimed to analyze the correlation between cell type proportions coming from the 5 methods with related histology scores for 654 plaques. However, due to the lack of stainings for other markers, we could only focus on macrophages (in relation to CD68, fat, ACTA2:CD68 and plaque phenotype), SMC

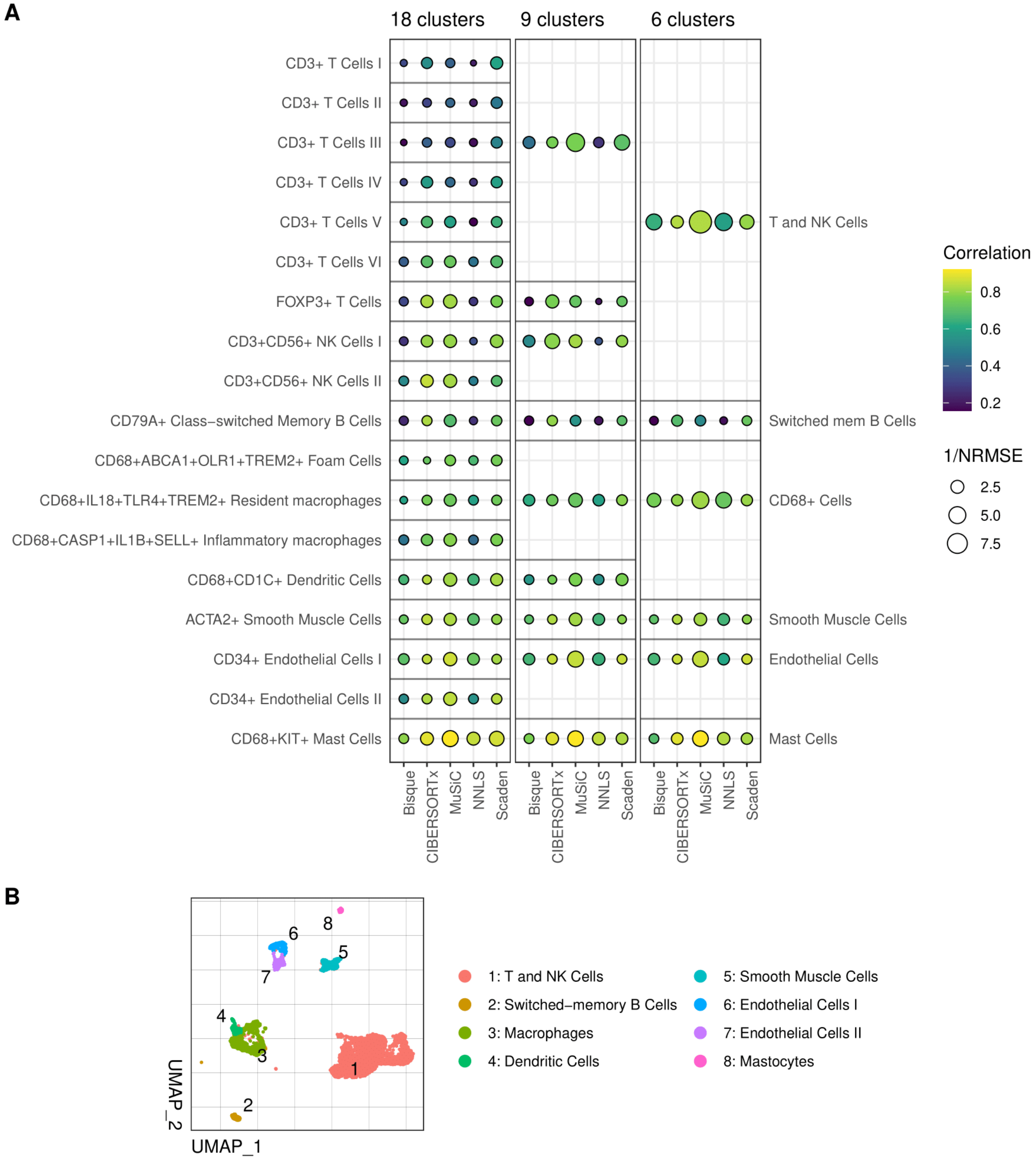

Fig. 4. Benchmarking 3rd step: cell clustering. A) Pearson correlation and normalized root-mean-square error (NRMSE) between deconvolved and known proportions of 700 pseudo-bulk mixtures, using the reference with 18, 9 and 6 cell populations to reduce the resolution. The dot colour accounts for the correlation level, whereas its size increases when NRMSE decreases. B) Final clustering of the Athero-Express single-cell RNA sequencing reference into 8 cell populations.

(in relation to ACTA2, collagen, ACTA2:CD68 and plaque phenotype) and EC proportions (in relation to vessel density). All the results are summarized in Table 2. For macrophages and SMCs, all methods proportions obtained a similar and significant correlation with histology (visualized also in Fig S6). In contrast, comparisons between ECs and vessel density did not show common results among the different

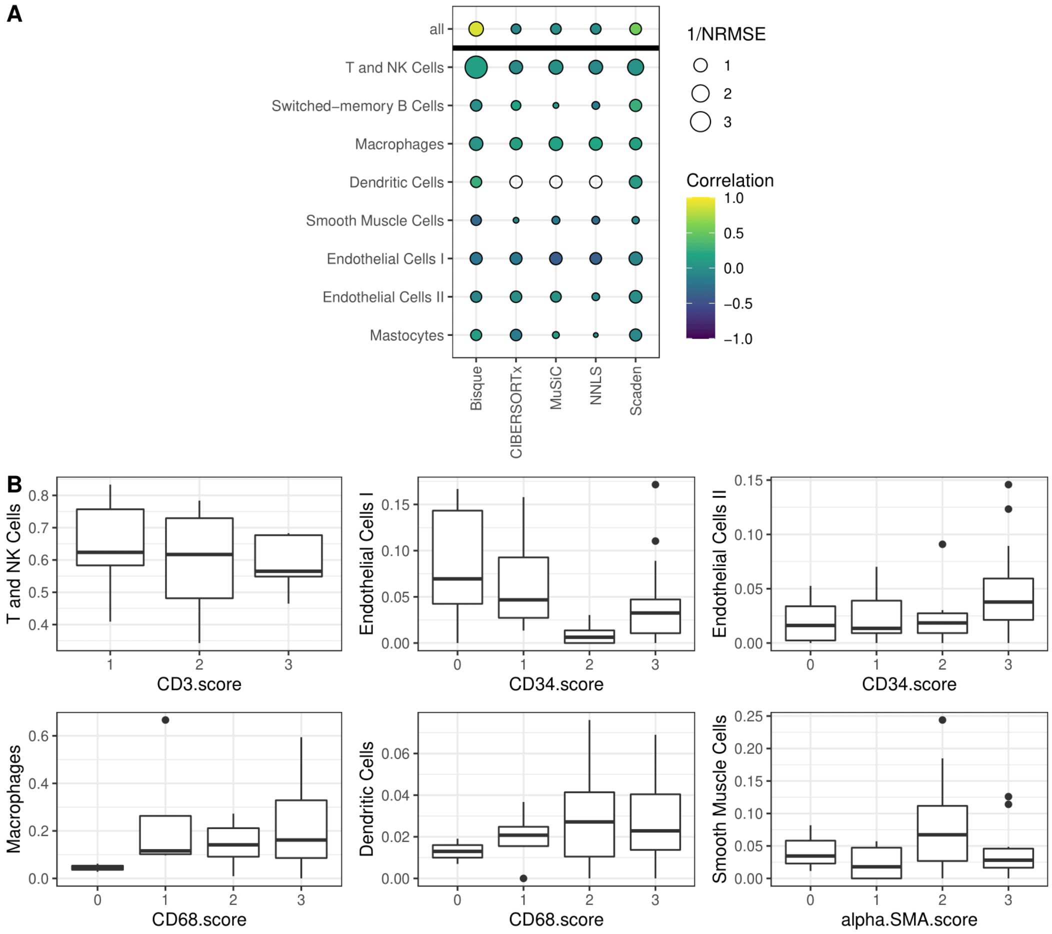

Fig. 5. Deconvolution of plaques with both bulk and sc RNA-seq data. A) Pearson correlation and normalized root-mean-square error (NRMSE) between deconvolved and sc-derived proportions of 28 AE plaques, globally (all) and individually for each cell type. White dots appear in cases where correlation could not be computed (i.e., standard deviation is zero. In these cases, due to proportions being zero for all plaques). B) Boxplot with the sc-derived proportions for the 46 plaques in the reference against the histology score for T and NK cells (CD3.score), ECs (CD34.score), macrophages and dendritic cells (CD68.score) and SMCs (alpha.SMA.score).

methods and, except for the case of EC I proportions with NNLS, there were no significant results.

Given the discrepancies in proportions (Fig 6A) but the similarity of results obtained above for certain cell types regardless of the method choice (Table 2), we hypothesized the following: even if the absolute proportions might not be blindly trusted; given a cell population, the trend of its proportions to be lower or higher for certain plaques might be kept for all the methods. These cell populations could therefore be trusted for stratifying plaques or patients based on their proportions. In order to assess this, for each of the 8 cell clusters, we computed the pair-wise Pearson correlation between the proportions deconvolved with the 5 methods (Fig 6B). From this, we detected a strong pair-wise correlation for macrophages and SMCs, whereas the other cell populations showed a weak agreement among the methods. Furthermore, macrophage proportions of the same plaques, but deconvolved using totally different sc RNA-seq data as reference ([42], including 3 human carotid plaques), highly correlated with those obtained with the AE reference (Table S1).

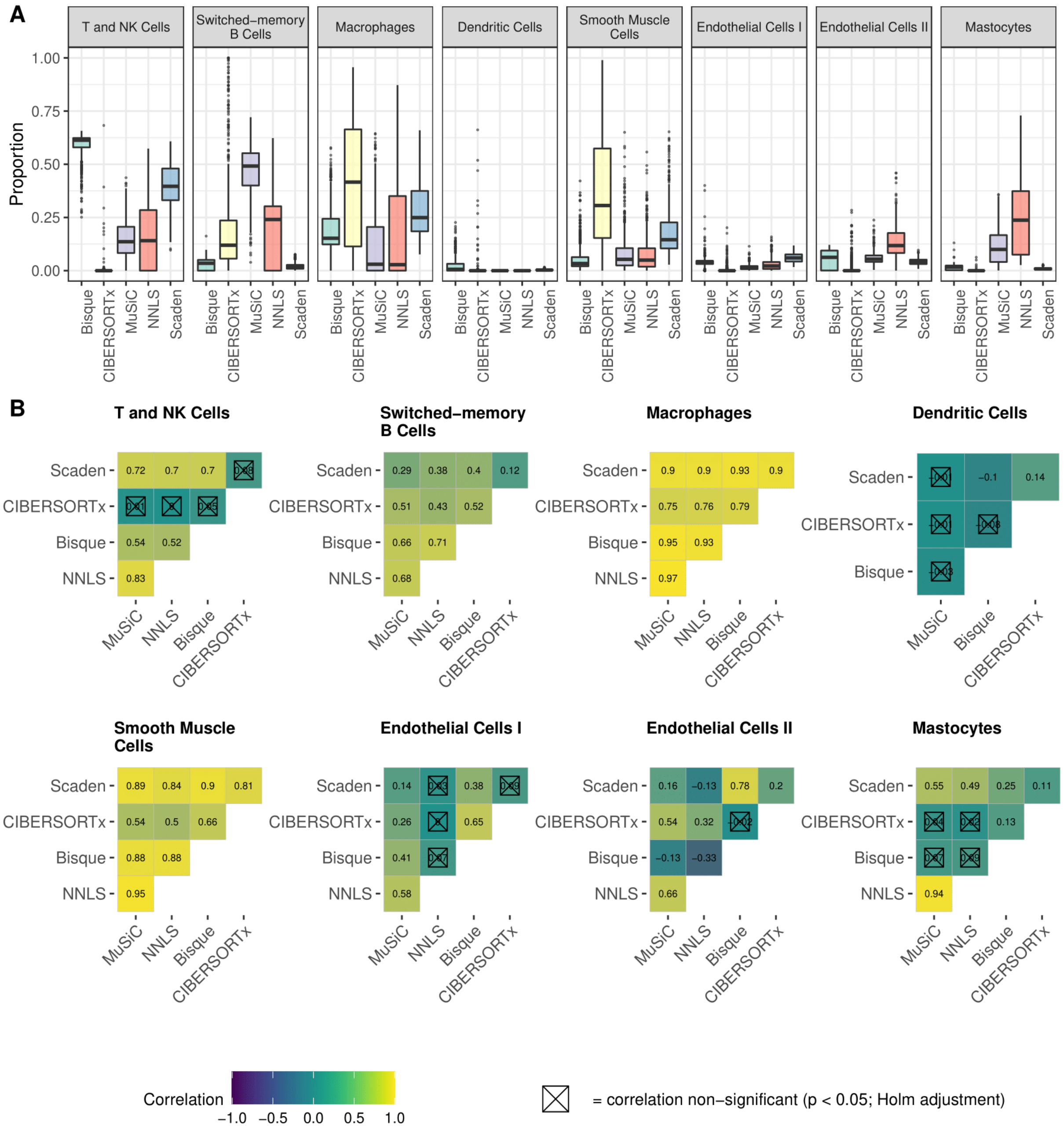

Fig. 6. Deconvolution of Athero-Express plaques for 8 cell populations (656 plaques). A) Boxplot with the deconvolved cell type proportions from 5 deconvolution methods. B) Pair-wise correlations per cell type between the proportions obtained by the 5 methods. A method is not shown if the correlation could not be computed (i.e., standard deviation is zero).

Table 2. Deconvolved proportions compared to histology. Associations between deconvolved proportions and related histological features were computed with Kendall rank correlation for categorical features and with Pearson correlation for continuous features (vessel density). P-values are shown with symbols (*** p<0.01; ** p<0.05; * p<0.1). Plaque phenotypes were converted to scores from 1 to 3, from atheromatous to fibrous.

Cell population	**Histology score**	**Bisque**	**CIBERSORTx**	**MuSiC**	**NNLS**	**Scaden**
Macrophages	CD68	0.094 (***)	0.096 (***)	0.089 (***)	0.09 (***)	0.1 (***)
	Fat	0.132 (***)	0.132 (***)	0.113 (***)	0.105 (***)	0.114 (***)
	ACTA2: CD68	-0.115 (***)	-0.134 (***)	-0.096 (***)	-0.088 (***)	-0.1 (***)
	Plaque phenotype	-0.123 (***)	-0.127 (***)	-0.1 (***)	-0.088 (***)	-0.1 (***)
Smooth muscle cells	ACTA2	0.12 (***)	0.113 (***)	0.085 (***)	0.08 (***)	0.142 (***)
	Collagen	0.054 (*)	0.07 (**)	0.021 ( )	-0.001 ( )	0.066 (**)
	ACTA2: CD68	0.165 (***)	0.199 (***)	0.11 (***)	0.1 (***)	0.195 (***)
	Plaque phenotype	0.118 (***)	0.151 (***)	0.051 (*)	0.036 ( )	0.132 (***)
Endothelial cells I	Vessel density	-0.037 ( )	0.008 ( )	0.065 ( )	0.114 (***)	-0.032 ( )
Endothelial cells II	Vessel density	-0.015 ( )	0.044 ( )	0.014 ( )	0.003 ( )	0.01 ( )

3.2 Clinical data association

In order to find associations between plaque composition and clinical features, we proceeded with the proportions provided with Scaden (see discussion, section 4). Instead of proportions from 0 to 1, we will work with percentages from 0 to 100 to make results easier to interpret. Our main interest is to assess sex differences in plaque composition, as well as the relationship with atherosclerosis severity, in terms of clinical symptoms and future MACE following endarterectomy.

3.2.1 Sex differences: female plaques show an increase in SMC content and a significant reduction of switched-memory B cells

From 654 plaques (Table S2), we assessed with univariate analysis differences in plaque composition due to the sex of the patient (Table 3). We observed a significantly lower content of switched-memory B cells in female plaques (β = -0.29 [-0.54, -0.05], p = 0.019) and a trend towards a higher proportion of SMCs (β = 1.86 [-0.25, 3.96], p = 0.084). The latter observation also held true for the ratio between SMC and macrophage content (0.13 [-0.016, 0.282], p = 0.079), even though macrophage proportions alone did not show any significant association. In addition, when we compared the ratio between all the structural (i.e., SMCs and ECs) and immune cells (i.e., the rest of populations), there was an increase in women but it did not reach the significance level (β = 0.046 [-0.02, 0.11], p = 0.136).

Table 3. Sex differences in plaque composition. β coefficients, presented with the 95% confidence interval (CI), and p-values are calculated with univariate linear regression between sex and the percentage of cells.

Cell population	**β-coefficient [95% CI] (female)**	**p-value**
T and NK cells	-0.50 [-2.19, 1.02]	0.473
Switched-memory B cells	-0.29 [-0.54, -0.05]	0.019
Macrophages	-0.64 [-0.31, 0.18]	0.605
Dendritic cells	-0.02 [-0.09, 0.04]	0.462
Smooth muscle cells	1.86 [-0.25, 3.96]	0.084
Endothelial cells I	-0.09 [-0.42, 0.25]	0.606
Endothelial cells II	-0.13 [-0.33, 0.07]	0.198
Mastocytes	-0.10 [-0.31, 0.11]	0.373
Ratio smooth muscle cells/macrophages	0.13 [-0.016, 0.282]	0.079
Ratio structural cells/immune cells	0.046 [-0.02, 0.11]	0.136

3.2.2 An increased total macrophage content associates with atherosclerosis severity, but macrophage subtypes show different atheroprotective and atheroprogressive behaviours

In order to understand the role of plaque composition in atherosclerosis severity, we only focused on the macrophage population and its relationship with symptoms and future MACE based on the results obtained in the first part of the project (i.e, showing that macrophage deconvolved proportion can be trusted for patient stratification). We also aimed to disentangle which macrophage subtype was causing the association signal by performing deconvolution with the same reference, but separating the macrophage cluster into the 3 original subtypes, namely foam cells, resident macrophages and inflammatory macrophages (Fig S7).

Plaques were categorized as symptomatic (including those patients with stroke or transient ischaemic attack) or asymptomatic (including those patients with no or minor symptoms, like amaurosis fugax) for 642 patients (Table S3). We built logistic regression models between macrophages and symptoms to obtain the odds ratios (OR) and these are summarized in Table 4. For the deconvolved proportions of macrophages as one cell population, multivariate analysis (corrected for age and body mass index) between these and the presence of symptoms showed a positive association (OR = 1.02 [1.01, 1.04], p = 0.001). On the other hand, when assessing the macrophage subtypes individually, we detected this positive association with symptoms in the resident subtype (OR = 1.02 [1.01, 1.05], p = 0.008), while inflammatory macrophages showed the opposite behaviour (i.e., a protective role) (OR = 0.86 [0.74, 0.99], p = 0.038). For this scenario, analysis with foam cells were no significant.

Table 4. Association between macrophage content and symptoms. Odds ratios (OR), presented with the 95% confidence interval (CI), and p-values are calculated with univariate and multivariate logistic regression between macrophage percentage and symptoms. Two references were used for the deconvolution (white: 8 cell populations [1 single macrophage cluster], grey: same except for 3 subtypes of macrophages). Multivariate models were adjusted for age and body mass index.

Macrophage population	**Univariate**		**Multivariate**	
	OR [95% CI]	**p-value**	**OR [95% CI]**	**p-value**
Macrophage (single cluster)	1.02 [1.01, 1.03]	0.003	1.02 [1.01, 1.04]	0.001
Foam cells	1.03 [0.99, 1.08]	0.199	1.03 [0.99, 1.08]	0.180
Resident macrophages	1.02 [1.00, 1.05]	0.018	1.02 [1.01, 1.05]	0.008
Inflammatory macrophages	0.86 [0.75, 0.99]	0.040	0.86 [0.74, 0.99]	0.038

MACE were annotated for 649 patients (Table S4) during the 3 years following the endarterectomy procedure. Univariate and multivariate Cox proportional hazard regression models were used to calculate the hazard ratio (HR) for macrophage populations. For the sake of visualization and interpretability, we divided the patients into 3 categories regarding their macrophage percentage. All the results are summarized in Table 5 and the models based on tertiles are also depicted in Fig 7. Univariate analysis showed the following. When macrophages were deconvolved as a whole population, there was a positive association between macrophages and MACE: patients whose plaques had high macrophage content, compared to those with low content, were significantly at higher risk of undergoing a MACE within the next 3 years after surgery (HR = 1.84 [1.09, 3.11], p = 0.023). Besides, refined deconvolution with the 3 macrophage subtypes revealed plaques with higher foam cells content to be at higher risk of MACE (HR = 2.08 [1.21, 3.58], p = 0.008). The same association was found for resident macrophages (HR = 1.77 [1.05, 2.97], p = 0.031), but this was not applicable to inflammatory macrophages. However, the significance levels of the models for macrophages as a single cluster and resident macrophages individually were diminished in multivariate analysis (p from <0.05 to <0.1), when correcting for age, diabetes, hypertension and high-density lipoprotein (HDL) levels.

Finally, since we provided evidence that proportions of macrophages, as a whole cluster, could be trusted but this was not assessed for the 3 subtypes case, we wanted to evaluate whether the same associations held true when using different sc RNA-seq data as a reference. The reference from Pan et. al. ([42]) was also used for deconvolution taking the 3 macrophage subtypes individually (annotated as 1, 2 and 3). Interestingly, macrophage 3, which were found to be enriched for inflammatory markers (data not shown), showed again a negative association with symptoms (Table S5). And, even though we could

Table 5. Association between macrophage content and major cardiovascular events. Hazard ratios (HR), presented with the 95% confidence interval (CI), and p-values are calculated with Cox proportional hazard models between macrophage percentage and symptoms. Two references were used for the deconvolution (grey: 8 cell populations [1 single macrophage cluster], white: same except for 3 subtypes of macrophages). Multivariate models were adjusted for age, hypertension, diabetes and high-density lipoprotein levels.

Macrophage population	**Continuous or tertile (% interval)**	**Univariate**		**Multivariate**	
		HR [95% CI]	**p-value**	**HR [95% CI]**	**p-value**
Macrophage (single population)	Continuous	1.02 [1.00, 1.03]	0.012	1.02 [0.99, 1.03]	0.067
	Low tertile (8.22 to 25.22%)	1 (ref)		1 (ref)	
	Intermediate tertile (25.22 to 39%)	1.29 [0.74, 2.26]	0.370	1.46 [0.76, 2.81]	0.253
	High tertile (39 to 75.9%)	1.84 [1.09, 3.11]	0.023	1.78 [0.94, 3.37]	0.075
Foam cells	Continuous	1.05 [1.02, 1.09]	0.004	1.06 [1.01, 1.12]	0.001
	Low tertile (0.35 to 1.47%)	1 (ref)		1 (ref)	
	Intermediate tertile (1.47 to 3.49%)	1.54 [0.88, 2.72]	0.134	1.68 [0.59, 3.34]	0.135
	High tertile (3.49 to 40.6%)	2.08 [1.21, 3.58]	0.008	2.25 [1.15, 4.42]	0.018
Resident macrophages	Continuous	1.02 [1.00, 1.05]	0.070	1.01 [0.99, 1.04]	0.384
	Low tertile (3.03 to 15.5%)	1 (ref)		1 (ref)	
	Intermediate tertile (15.5 to 23%)	1.21 [0.69, 2.10]	0.511	1.41 [0.71, 2.76]	0.316
	High tertile (23 to 55.2%)	1.77 [1.05, 2.97]	0.031	1.76 [0.57, 3.29]	0.075
Inflammatory macrophages	Continuous	1.10 [0.93, 1.30]	0.263	1.05 [0.86, 1.28]	0.657
	Low tertile (0.13 to 0.94%)	1 (ref)		1 (ref)	
	Intermediate tertile (0.94 to 1.82%)	1.25 [0.80, 2.14]	0.416	1.34 [0.74, 2.60]	0.382
	High tertile (1.82 to 6.71%)	1.46 [0.68, 2.46]	0.155	1.55 [0.64, 2.98]	0.186

not identify the other two subtypes as resident or foam-like, we obtained a significant positive association for one subtype, while the other remained non significant, suggesting the same behaviour obtained with our data. However, none of the associations with MACE reached the significant level (Table S6).

3.2.3 Classical marker gene expression levels also reveal some clinical associations but these are weaker, pointing towards an added value of the deconvolution

The last question to address is whether deconvolution holds an added value with respect to simpler analysis relying only on bulk data. We aimed to recover the same associations found in this section by interrogating the expression of gene markers in the bulk RNA-seq data at a univariate level. On

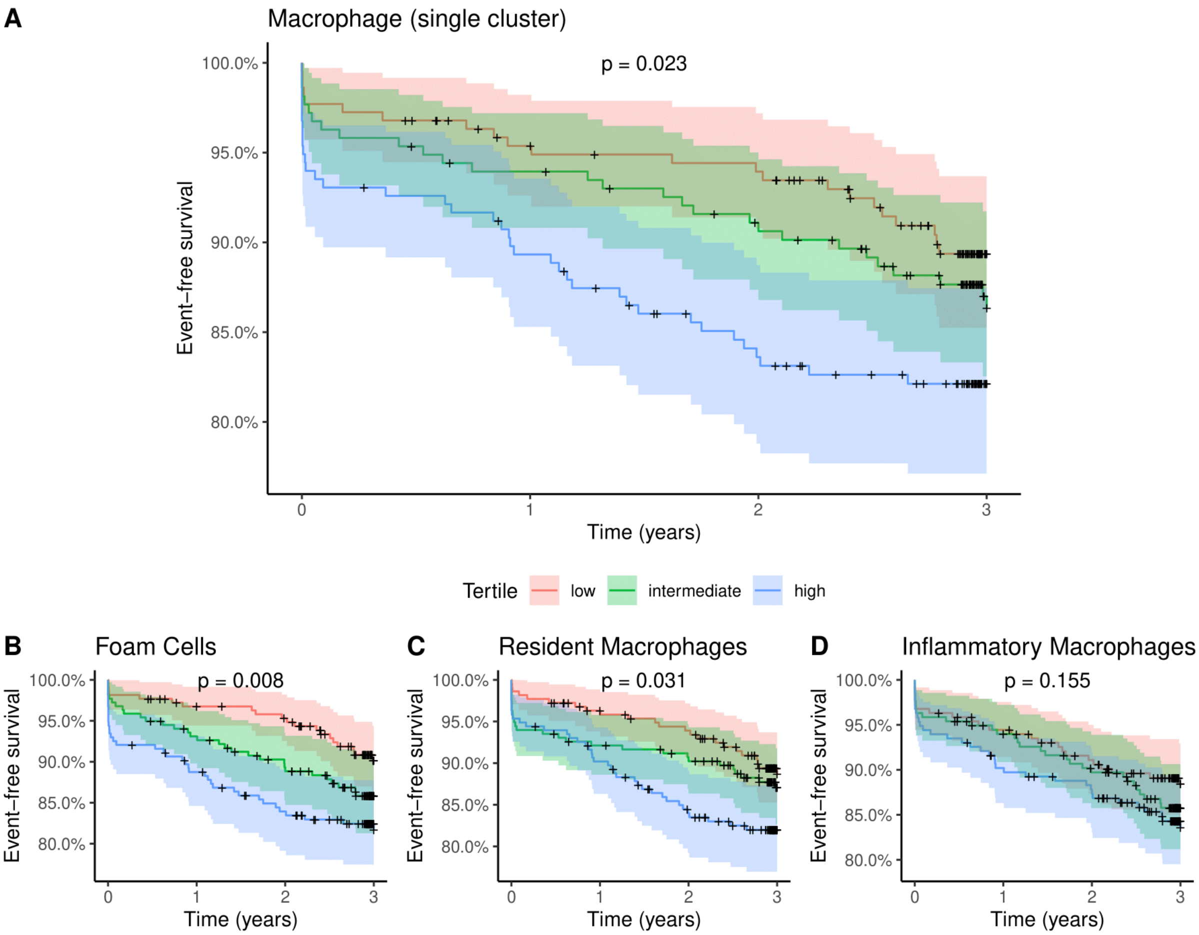

Fig. 7. Cox proportional hazard models for macrophage content associated with major cardiovascular events. The percentages tested were derived from the deconvolution using the reference with 8 cell populations (A: Macrophage as a single population) and the reference with the 3 macrophage subtypes (B: foam cells, C: resident macrophages, D: inflammatory macrophages). P-values obtained by the univariate association between the lowest to highest tertile.

the one hand, linear regression models were built to associate sex with expression levels of CD79A, for switched-memory B cells, and ACTA2, for SMCs. While CD79A showed a positive but weak trend towards female plaques (β = -1.32 [-2.88, 0.23], p = 0.095), ACTA2 association with sex had no statistical power. On the other hand, CD68 and CD14 were chosen as macrophage general markers, while ABCA1 was selected for foam macrophages, IL1B for inflammatory-like and LYVE1 for resident-like. Logistic regression models with symptoms revealed significant relationships with CD14 (OR = 1.0013 [1.0003, 1.0024], p = 0.014) and ABCA1 (OR = 1.0025 [1.0003, 1.0048], p = 0.031). Associated with MACE, we found weak trends with both general markers: CD68 (HR = 1.003 [0.997, 1.007], p = 0.083) and CD14 (1.001 [0.999, 1.002], p = 0.096).

Table 6. Association of marker genes expression and clinical features. Coefficients presented with the 95% confidence interval (CI) and the p-value from linear regression (association with sex), log regression (association with symptoms) and cox hazard proportional model (association with MACE). Expression levels in counts per million. OR: odds ratio, HR: hazard ratio.

	β-coefficient [95% CI] (sex: female)	**p-value**
ACTA2	44.5 [-72.27, 161.21]	0.455
CD79A	-1.32 [-2.88, 0.23]	0.095
	OR [95% CI] (symptoms)	**p-value**
CD68	1.00252 [1.000, 1.007]	0.201
CD14	1.0013 [1.0003, 1.0024]	0.014
ABCA1	1.0025 [1.0003, 1.0048]	0.031
IL1B	1.0152 [0.993, 1.040]	0.201
LYVE1	1.0089 [0.963, 1.059]	0.711
	HR [95% CI] (MACE)	**p-value**
CD68	1.003 [0.997, 1.007]	0.083
CD14	1.001 [0.999, 1.002]	0.096
ABCA1	1.002 [0.999, 1.004]	0.151
IL1B	1.015 [0.986, 1.039]	0.226
LYVE1	1.016 [0.984, 1.075]	0.571

4 Discussion

Composition of atherosclerotic plaques is not only known to be rich and heterogeneous, but also to be influenced by patient characteristics, like sex, and to influence atherosclerosis manifestation. Traditional methods to study the composition of plaques might be falling behind as single-cell analysis unravel increasing number of cells playing a role in these plaques. Recently, the concept of computationally imputing tissue composition has been a subject undergoing intense study. Known as deconvolution, it uses sc RNA-seq data as a reference to impute the cell type proportions of tissues with bulk gene expression. However, in literature, studies that propose new deconvolution algorithms still outnumber those that make use of deconvolution as part of their research workflow. This is probably caused by the lack of proper data, since the tissue origin of both sc and bulk RNA-seq datasets should be similar [43]. Second of all, there is not yet a gold standard to perform this analysis, leaving researchers alone with the choice of data processing and method selection. Here, we aimed to use sc and bulk RNA-seq results coming from human carotid atherosclerotic plaques to retrieve their cell proportions. In order to understand how data processing and the selection of the method can influence our results, and therefore be able to select the best approach, we first performed a benchmarking accounting different deconvolution scenarios and methods (including NNLS, MuSiC, Bisque, CIBERSORTx and Scaden). Next, with the final proportions, we investigated associations with sex and atherosclerosis severity.

Initially, we wanted to follow the first steps of a popular deconvolution study [35]. As a quality step, genes of the sc reference that could not be detected in more than 5% of the sequenced cells were removed. However, we were reluctant of this approach and we showed that, even though deconvolved proportions did not change significantly, this can be dangerous for sc datasets like ours. If a large number of cell types is identified, there will be cell types with a number of sequenced individual cells below this threshold and, therefore, marker genes would get lost. Like Avila-Cobos et al. concluded for artificial pseudo-bulk analysis [35], we observed a big effect of data transformation on the deconvolution of our plaques for some methods. Bisque was not as affected as MuSiC and NNLS, probably owing to its inherent transformation of the bulk data to resemble the sc properties [32]. While square-root and log transformed data would give us more plausible results for some cell types (MuSiC linear results showed an undesired high proportion of switched-memory B cells), analysis on pseudo-bulk mixtures created from our sc reference, determined a better performance with linear data, in accordance with their results and other studies [35, 43].

We found that reducing the number of cell types in the sc reference by grouping similar clusters together, improved the deconvolution performance on artificial mixtures. Such clustering-based improvement was specially notorious for T and NK cells, which were in turn, the group of cells whose transcriptomic profiles showed a higher correlation. This agrees with the spillover effect suggested by Sturm et al. owing to the multicolinearity of some cell types [37]. Of note, MuSiC was built taking this similarity problem into account and it initially infers the proportion of similar clusters of cell types [31]; but, interestingly, we can see it is one of the methods that benefits more from the reclustering we performed. Results from a very recent research challenge on tumour deconvolution (DREAM) also showed that while most of the published methods could predict well proportions for the major cell types, these worsen when cell subtypes were considered [44]. Worth to mention, another benchmarking study focused on brain data, concluded that subtypes should only be included either if their abundance is higher than 2% or if they are not too transcriptomically similar [36]. Nevertheless, the former condition is met for some subtypes we grouped together (like resident macrophages). Still, our reduction of the sc dataset into eight distinct cell populations was done as a balance between accuracy and resolution. Unfortunately, such decision goes against our willing to unravel plaque composition at high resolution, which was one of the needs we wanted to cover with deconvolution.

Even after all the data processing decisions, we still detected an important disagreement between all deconvolution methods results on real bulk data. This was not found, at least that noticeably, when we worked with artificial pseudo-bulk mixtures. However, this is not surprising given that these mixtures were created by a strict linear calculation (i.e., adding the counts of individual cells from the reference dataset); and we know that most of the methods are inherently assuming this linear association between the sc and bulk RNA-seq data. Thus, this difference on artificial and real data scenarios suggests that the relationship between our data is more complex rather than strictly linear. If we investigate methods behaviour and architecture individually, we detect the following. (I) Bisque results in a large abundance of T and NK cells and it is the method with better agreement with the proportions derived from the sc reference. Actually, Bisque implementation considers sc proportions in one of its main steps [32], but we have shown that our sc proportions might not be trusted. Therefore, Bisque shoud be rejected. (II) CIBERSORTx is a widely used tool and it is not built under the strict linear assumption. However, it generally suggested the absence of cell populations other than SMCs, macrophages and switched-memory B cells in the plaques. (III) While NNLS and MuSiC are theoretically very similar, MuSiC give weights to genes that are consistent across different plaques [31]. This weighting did not seem to affect the results of artificial-data analysis and MuSiC performed better than other methods, but it made switched-memory excessively outnumber all the other populations on real plaques data. (IV) Scaden, which uses a neural network to infer the proportions [34], resulted in a plausible composition of the plaques with T and NK cells, macrophages and SMCs as the main populations [23, 28]. Furthermore, in line with the suggested false strict linear relationship we might have between our bulk and sc data, we believe that more complex architectures (e.g., Scaden) are needed for the sake of robustness, which seems to be improved with neural network-based methods [34, 45].

While most studies acknowledge differences between methods, they mainly focus on selecting the method that correlates best with known proportions [35, 36, 45, 46], but they do not address if method results correlate with each other. We showed evidence that rather than relying on a single method, it might be wiser to only rely on some cell types. Even if the absolute proportions scored by different methods might differ, we know that for macrophages and SMCs, proportions correlated between all the methods. This suggests that, based on these abundances, all methods would stratify plaques (or patients) the same way, making the associations method-independent. Also, for these two cell populations we also found agreement with previous histology assessment and, for macrophages, we also detected a nice correlation with proportions deconvolved with another reference dataset. Therefore, not only might macrophage associations be method-independent but also reference-independent, which is key since the selection of reference is known to highly impact the results [36, 43, 47]. It has been mentioned that deconvolution methods might be chosen based on the cell type of interest [37], but this interestingly suggests the opposite. The cell type from the reference we can trust better might be chosen based on the methods agreement. This finding about SMCs and macrophages is more consistent than relying on Scaden for all the cell type proportions.

For the clinical part of the project, we aimed to detect sex differences on plaque composition based on Scaden deconvolution results for these 8 reduced cell populations. With univariate analysis, we found a trend for female plaques to be richer in SMCs and poorer in immune cells, specially showing a significant reduction of switched memory B cells compared to male plaques. The association with SMCs strengthen the concept of fibrous cap thinning in males [48], who normally present vulnerable plaques [12–14]. Therefore, we also expected a higher macrophage abundance in male plaques, but this was not found in our analysis. However, this association could be hindered by the clustering of all macrophages, which included foam and inflammatory macrophages on top of resident macrophages, with the former group being particularly associated with sex regarding Sangiorgi et al. results [14]. The significant reduction of switched-memory B cells in female plaques is interesting, but we have to be careful, since we did not find agreement of these proportions with other deconvolution methods. Still, peripheral switched and unswitched-memory B cells were found to be higher in patients with less secondary events after endarterectomy [49]; and patients who benefit more from this surgery are males [50]. This might open a line to be further explored, but we can not establish conclusive results regarding this specific cell type.

Narrowing the analysis to the macrophage population, we detected a relationship between its content and atherosclerosis severity. A higher macrophage abundance was significantly associated with symptoms (including stroke and TIA) and MACE within 3 years after endarterectomy, suggesting macrophages to be a risk factor. Macrophages have been defined as hallmarks of the vulnerable plaque [8] and have been frequently localized at plaque rupture sites [51]. On the other hand, vulnerable plaque characteristics are more typical in patients undergoing more severe symptoms [8, 10] and worse outcome after surgery [11]. However, this last study by Hellings et al., with histological comparison of plaques also from the AE cohort, could only establish this association for intraplaque hemorrhage, but not for macrophage infiltration. Rather than regarding this disagreement as an opposite direction, we believe deconvolution might overcome the power of histology assessment in this case. Nevertheless, the rough classification of macrophages into pro-inflammatory M1 and anti-inflammatory M2 unraveled different behaviours, suggesting M2-like to boost atherosclerosis regression [52, 53]. Also, De Gaetano et al. concluded with gene and protein expression analysis, that while M1 markers where higher expressed in symptomatic plaques, M2 markers where higher expressed in asymptomatic plaques [54]. For this reason, we obtained the deconvolved proportions of the macrophage subtypes individually, namely resident, inflammatory and foam macrophages. Nevertheless, if we regard our inflammatory subtype as the M1 group, and resident subtype as the M2 macrophages, our results do not match with De Gaetano et al. [54], and neither did analysis on our bulk RNA-seq expression levels. Our deconvolved data suggests that (I) resident macrophages are more abundant in symptomatic plaques and in patients with future MACE; (II) foam cells are a risk factor for MACE; and (III) inflammatory macrophage content is lower in symptomatic plaques. Therefore, our defined inflammatory macrophages seem to hide an atheroprotective role. Still, we might think about another option. We are assuming plaque composition is predictive of the symptoms, but it has also been assessed that these clinical manifestations trigger a remodeling of the plaque [55], therefore changing plaques composition. Thus, associations with symptoms could be targeted bidirectionally.

Clinical associations with plaque compositions could have also been targeted at the gene expression level analyzed from the bulk RNA-seq data. However, we detect an added value of deconvolution, since some associations could be found, but less consistently than what we evaluated with deconvolved results. CD14, as a general macrophage marker, was associated with symptoms, as well as ABCA1 as a foam cell marker. However, we would also expect the same for CD68, but this was not found. In terms of sex differences, ACTA2 as a marker of SMCs showed no association, but interestingly, CD79A, targeting switched memory-B cells was lower expressed in female plaques.

Finally, we must mention some of the limitations of our study. On the one hand, since we did not have known cell proportions of the plaques to compare the deconvolved proportions with, we had to find other approaches. We used the deconvolution of artificially created pseudo-bulk mixtures to assess the effect of data transformation and reference clustering; but these artificial data is not similar to the real bulk RNA-seq data we have from the plaques. When working with real mixtures, we wanted to compare deconvolved with sc-derived proportions, but this was discarded owing to a poor reliability of the latter. This left us with the comparison between deconvolved proportions and histology, but we only had histology scores targeting macrophages, SMCs and (undirectly) ECs. On the other hand, clinical associations found for cell types other than SMCs and macrophages, might be taken carefully since for the rest of populations we could not show evidence on proportions being as reliable (either in terms of histology or in terms of deconvolution methods agreement). The same holds true for associations with macrophages subtypes.

Near-future steps to be taken must consider a sanity check of the deconvolved proportions. Still in the computational side, we might perform deconvolution using another sc RNA-seq reference from carotid plaques coming from a cohort with a similar number of patients to ours, or we could also compare it with results from the deconvolution of DNA methylation data. Nevertheless, we will still need a validation with approaches like stainings for other cell populations markers that were not gathered in the AE data, flow cytometry or fluorescence activated cell sorting analysis. For a longer-term future, if the associations found here can be validated, new therapies could aim to lower those macrophage subtypes that associate with atherosclerosis severity. And, if macrophage content is indeed predictive of MACE following the entarterectomy, we could consider to systematically send removed plaques to sequence to get their deconvolved proportions, and based on this analysis, adapt patients care.

5 Conclusions

Deconvolution of atherosclerotic plaques based on RNA-seq data is not straightforward. To make it easier, we should reduce the number of cell types identified in the sc RNA-seq reference. Even though all deconvolution methods obtained very different cell abundances, the relative proportions for macrophages and SMCs were method-independent (and reference-independent for macrophages), suggesting an equal stratification of the patients based on these proportions. Female plaques showed a tendency to have greater SMC content, but no sex differences were found based on macrophages abundance. Macrophages, in turn, were associated with atherosclerosis severity and results suggest that its subtypes might have different behaviours on this association (i.e., atheroprotective vs atheroprogressive), but this remains to be validated with further analysis.

6 References

1. Libby, P. The changing landscape of atherosclerosis. *Nature* **592,** 524–533 (2021).
2. Nobuyoshi, M. *et al.* Progression of coronary atherosclerosis: Is coronary spasm related to progression? *Journal of the American College of Cardiology* **18,** 904–910 (1991).
3. Alderman, E. L. *et al.* Five-year angiographic follow-up of factors associated with progression of coronary artery disease in the Coronary Artery Surgery Study (CASS). *Journal of the American College of Cardiology* **22,** 1141–1154 (1993).
4. Rumberger, J. A. Coronary Artery Disease: A Continuum, Not a Threshold. *Mayo Clinic Proceedings* **92,** 323–326 (2017).
5. Libby, P. *et al.* Atherosclerosis. *Nature Reviews Disease Primers* **5,** 1–18 (2019).
6. Falk, E., Shah, P. K. & Fuster, V. Coronary Plaque Disruption. *Circulation* **92,** 657–671 (1995).
7. Lee, R. T. & Libby, P. The Unstable Atheroma. *Arteriosclerosis, Thrombosis, and Vascular Biology* **17,** 1859–1867 (1997).
8. Virmani, R., Kolodgie, F. D., Burke, A. P., Farb, A. & Schwartz, S. M. Lessons From Sudden Coronary Death. *Arteriosclerosis, Thrombosis, and Vascular Biology* **20,** 1262–1275 (2000).
9. Verhoeven, B. *et al.* Carotid atherosclerotic plaques in patients with transient ischemic attacks and stroke have unstable characteristics compared with plaques in asymptomatic and amaurosis fugax patients. *Journal of Vascular Surgery* **42,** 1075–1081 (2005).
10. Redgrave, J. N. E., Lovett, J. K., Gallagher, P. J. & Rothwell, P. M. Histological Assessment of 526 Symptomatic Carotid Plaques in Relation to the Nature and Timing of Ischemic Symptoms. *Circulation* **113,** 2320–2328 (2006).
11. Hellings, W. E. *et al.* Composition of Carotid Atherosclerotic Plaque Is Associated With Cardiovascular Outcome. *Circulation* **121,** 1941–1950 (2010).
12. Hellings, W. E. *et al.* Gender-associated differences in plaque phenotype of patients undergoing carotid endarterectomy. *Journal of Vascular Surgery* **45,** 289–296 (2007).
13. Wendorff, C. *et al.* Carotid Plaque Morphology Is Significantly Associated With Sex, Age, and History of Neurological Symptoms. *Stroke* **46,** 3213–3219 (2015).
14. Sangiorgi, G. *et al.* Sex-related differences in carotid plaque features and inflammation. *Journal of Vascular Surgery* **57,** 338–344 (2013).
15. Vrijenhoek, J. E. P. *et al.* Sex Is Associated With the Presence of Atherosclerotic Plaque Hemorrhage and Modifies the Relation Between Plaque Hemorrhage and Cardiovascular Outcome. *Stroke* **44,** 3318–3323 (2013).

16. Van Dam-Nolen, D. H. K. *et al.* Sex Differences in Plaque Composition and Morphology Among Symptomatic Patients With Mild-to-Moderate Carotid Artery Stenosis. *Stroke* **53,** 370–378 (2022).

17. Farb, A. *et al.* Coronary Plaque Erosion Without Rupture Into a Lipid Core. *Circulation* **93,** 1354–1363 (1996).

18. Yahagi, K., Davis, H. R., Arbustini, E. & Virmani, R. Sex differences in coronary artery disease: Pathological observations. *Atherosclerosis* **239,** 260–267 (2015).

19. Daghem, M., Bing, R., Fayad, Z. A. & Dweck, M. R. Noninvasive Imaging to Assess Atherosclerotic Plaque Composition and Disease Activity: Coronary and Carotid Applications. *JACC: Cardiovascular Imaging* **13,** 1055–1068 (2020).

20. Kolossváry, M., Szilveszter, B., Merkely, B. & Maurovich-Horvat, P. Plaque imaging with CT—a comprehensive review on coronary CT angiography based risk assessment. *Cardiovascular Diagnosis and Therapy* **7** (2017).

21. Hellings, W. E. *et al.* Intraobserver and interobserver variability and spatial differences in histologic examination of carotid endarterectomy specimens. *Journal of Vascular Surgery* **46,** 1147–1154 (2007).

22. Depuydt, M. A. *et al.* Microanatomy of the Human Atherosclerotic Plaque by Single-Cell Transcriptomics. *Circulation Research,* 1437–1455 (2020).

23. Winkels, H. *et al.* Atlas of the immune cell repertoire in mouse atherosclerosis defined by single-cell RNA-sequencing and mass cytometry. *Circulation Research* **122,** 1675–1688 (2018).

24. McQueen, L. W. *et al.* Next-Generation and Single-Cell Sequencing Approaches to Study Atherosclerosis and Vascular Inflammation Pathophysiology: A Systematic Review. *Frontiers in Cardiovascular Medicine* **9,** 1–19 (2022).

25. Hartman, R. *et al.* Sex-dependent gene regulation of human atherosclerotic plaques by DNA methylation and transcriptome integration points to smooth muscle cell involvement in women. *Atherosclerosis* **331,** e217 (2021).

26. Mokry, M. *et al.* Transcriptomic-based clustering of advanced atherosclerotic plaques identifies subgroups of plaques with differential underlying biology that associate with clinical presentation. *medRxiv,* 2021.11.25.21266855 (2021).

27. Slenders, L., Tessels, D. E., Laan, S. W. V. D. & Pasterkamp, G. The Applications of Single-Cell RNA Sequencing in Atherosclerotic Disease. **9,** 1–12 (2022).

28. Fernandez, D. M. *et al.* Single-cell immune landscape of human atherosclerotic plaques. *Nature Medicine* **25,** 1576–1588 (2019).

29. Fernandez, D. M. & Giannarelli, C. Immune cell profiling in atherosclerosis: role in research and precision medicine. *Nature Reviews Cardiology* **19,** 43–58 (2022).

30. Mullen, K. M. & van Stokkum, I. H. M. *nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS)* R package version 1.4 (2012).

31. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature Communications* **10** (2019).

32. Jew, B. *et al.* Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nature Communications* **11,** 1971 (2020).

33. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology* **37,** 773–782 (2019).

34. Menden, K. *et al.* Deep learning–based cell composition analysis from tissue expression profiles. *Science Advances* **6,** eaba2619 (2020).

35. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nature Communications* **11** (2020).

36. Sutton, G. & Poppe, D. Comprehensive evaluation of human brain gene expression deconvolution methods. *Nature Communications,* 2020.06.01.126839 (2022).

37. Sturm, G. *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **35,** i436–i445 (2019).

38. Verhoeven, B. A. N. *et al.* Athero-express: Differential atherosclerotic plaque expression of mRNA and protein in relation to cardiovascular events and patient characteristics. Rationale and design. *European Journal of Epidemiology* **19,** 1127–1133 (2004).

39. Association, W. M. World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects. *Bulletin of the World Health Organization* **79,** 373–374 (2001).

40. Hashimshony, T. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome biology* **17,** 1–7 (2016).

41. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* (2021).

42. Pan, H. *et al.* Single-Cell Genomics Reveals a Novel Cell State during Smooth Muscle Cell Phenotypic Switching and Potential Therapeutic Targets for Atherosclerosis in Mouse and Human. *Circulation,* 2060–2075 (2020).

43. Schelker, M. *et al.* Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nature Communications* **8,** 1–12 (2017).

44. White, B. S. *et al.* Community assessment of methods to deconvolve cellular composition from bulk gene expression. *bioRxiv,* 2022.06.03.494221 (2022).

45. Lin, Y. *et al.* DAISM-DNNXMBD: Highly accurate cell type proportion estimation with in silico data augmentation and deep neural networks. *Patterns* **3,** 100440 (2022).

46. Jin, H. & Liu, Z. A benchmark for RNA-seq deconvolution analysis under dynamic testing environments. *Genome Biology* **22,** 1–23 (2021).

47. Vallania, F. *et al.* Leveraging heterogeneity across multiple datasets increases cell-mixture deconvolution accuracy and reduces biological and technical biases. eng. *Nature communications* **9,** 4735 (2018).

48. Man, J. J., Beckman, J. A. & Jaffe, I. Z. Sex as a Biological Variable in Atherosclerosis. *Circulation Research,* 1297–1319 (2020).

49. Meeuwsen, J. A. L. *et al.* High Levels of (Un)Switched Memory B Cells Are Associated With Better Outcome in Patients With Advanced Atherosclerotic Disease. *Journal of the American Heart Association* **6,** e005747 (2017).

50. Sarac, T. P. *et al.* Gender as a primary predictor of outcome after carotid endarterectomy. eng. *Journal of vascular surgery* **35,** 748–753 (2002).

51. Kolodgie, F. D. *et al.* Localization of Apoptotic Macrophages at the Site of Plaque Rupture in Sudden Coronary Death. *The American Journal of Pathology* **157,** 1259–1268 (2000).

52. Moore, K. J., Sheedy, F. J. & Fisher, E. A. Macrophages in atherosclerosis: A dynamic balance. *Nature Reviews Immunology* **13,** 709–721 (2013).

53. Barrett, T. J. Macrophages in Atherosclerosis Regression. *Arteriosclerosis, Thrombosis, and Vascular Biology* **40,** 20–33 (2020).

54. De Gaetano, M., Crean, D., Barry, M. & Belton, O. M1- and M2-Type Macrophage Responses Are Predictive of Adverse Outcomes in Human Atherosclerosis. eng. *Frontiers in immunology* **7,** 275 (2016).

55. Peeters, W. *et al.* Carotid Atherosclerotic Plaques Stabilize After Stroke. *Arteriosclerosis, Thrombosis, and Vascular Biology* **29,** 128–133 (2009).

7 Supplementary data

7.1 Supplementary figures

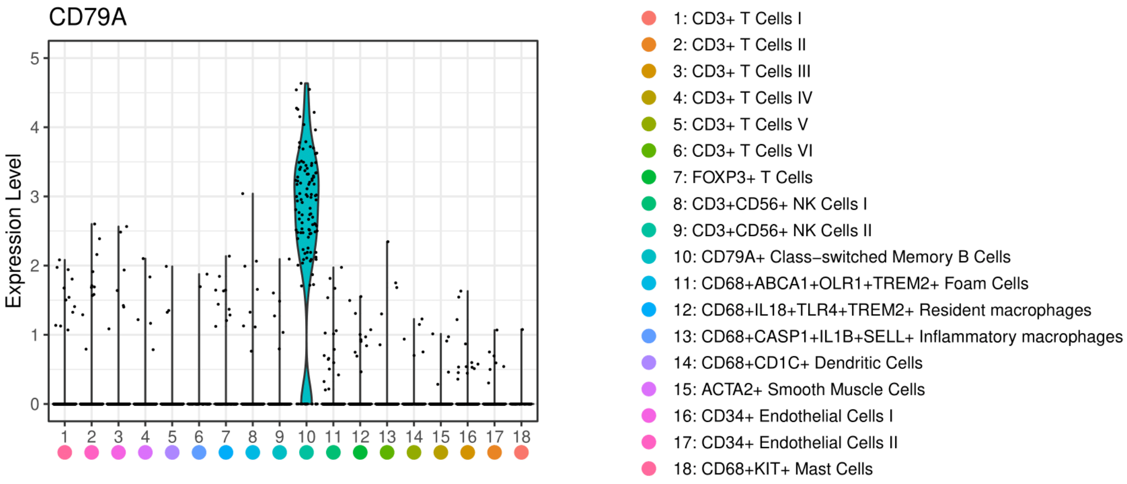

Fig. S1. Switched-memory B cells gene marker. Violin plot of CD79A expression in 18 cell types of the Athero-Express single-cell RNA sequencing dataset.

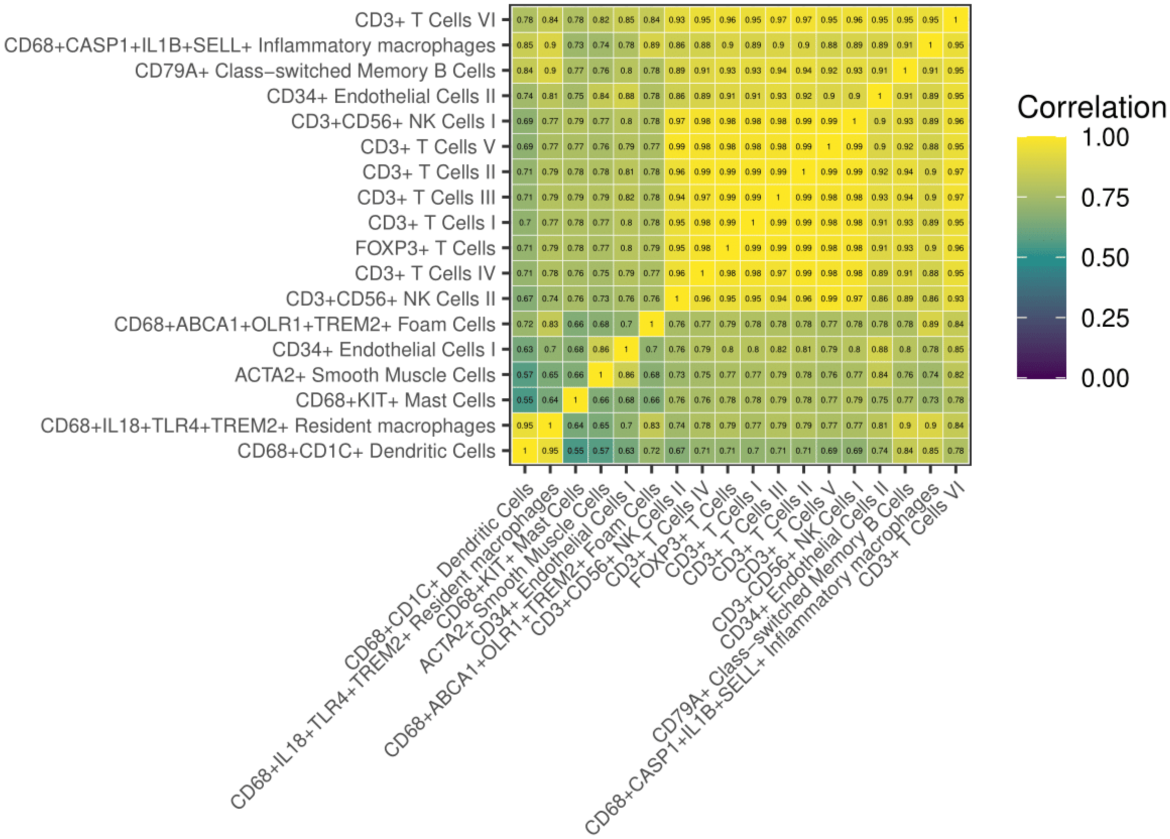

Fig. S2. **Cell type signature correlation.** Pearson correlation plot for the average gene expression signature of each cell type in the Athero-Express single-cell RNA sequencing dataset.

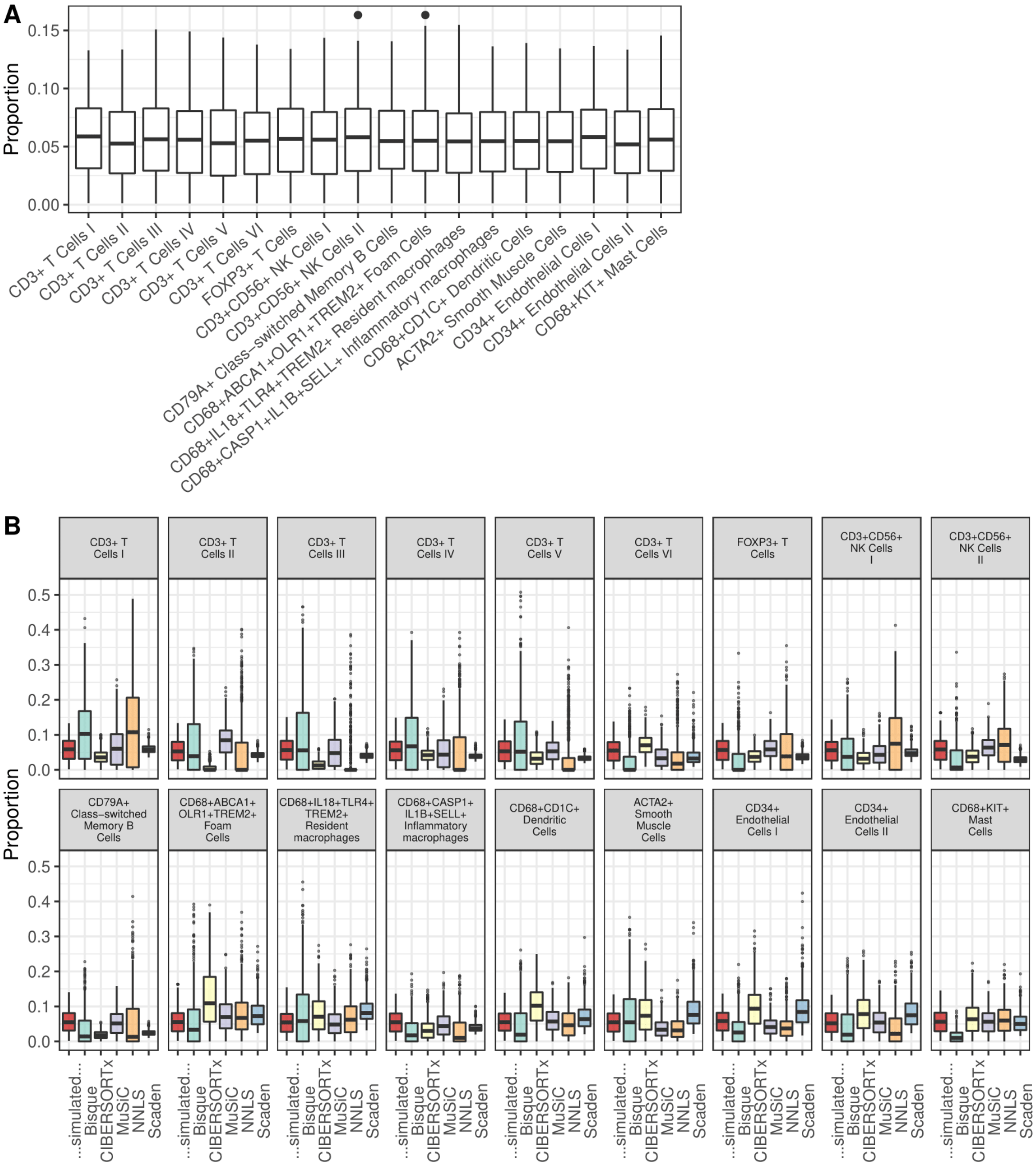

Fig. S3. Pseudo-bulk mixtures proportions. A) Boxplot of the simulated proportions for 700 pseudo-bulk mixtures created by adding the counts of 100 randomly selected cells of the Athero-Express single-cell RNA sequencing (AE sc RNA-seq) training dataset (50% split). The cell type proportions were established previously by a random but uniform distribution. B) Boxplot of the pseudo-bulk mixtures deconvolved proportions using the AE sc RNA-seq test dataset (50% split). In red, known proportions.

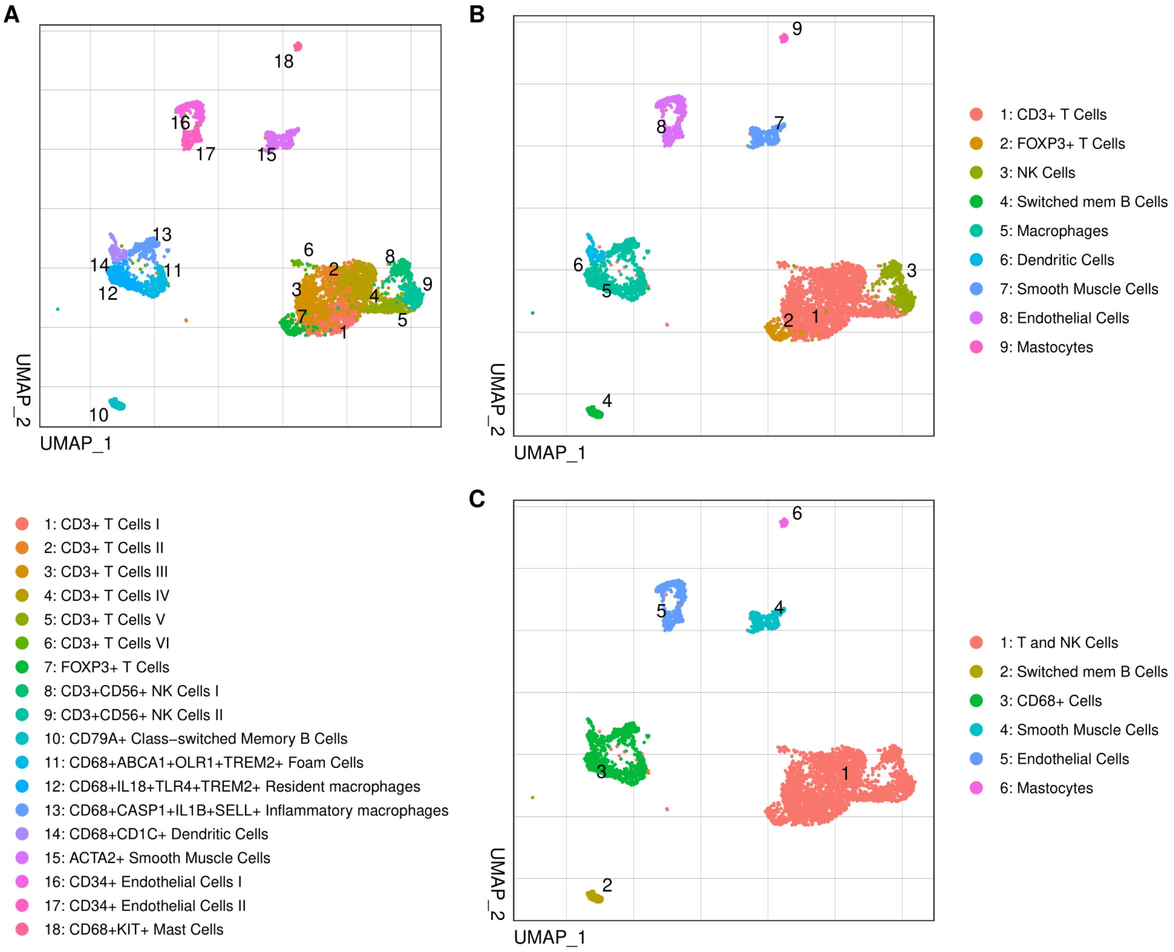

Fig. S4. Athero-Express single-cell RNA sequencing data clustering. UMAP visualization of A) Original clustering of 18 cell populations. B) Reduction to 9 cell populations. C) Reduction to 6 cell populations.

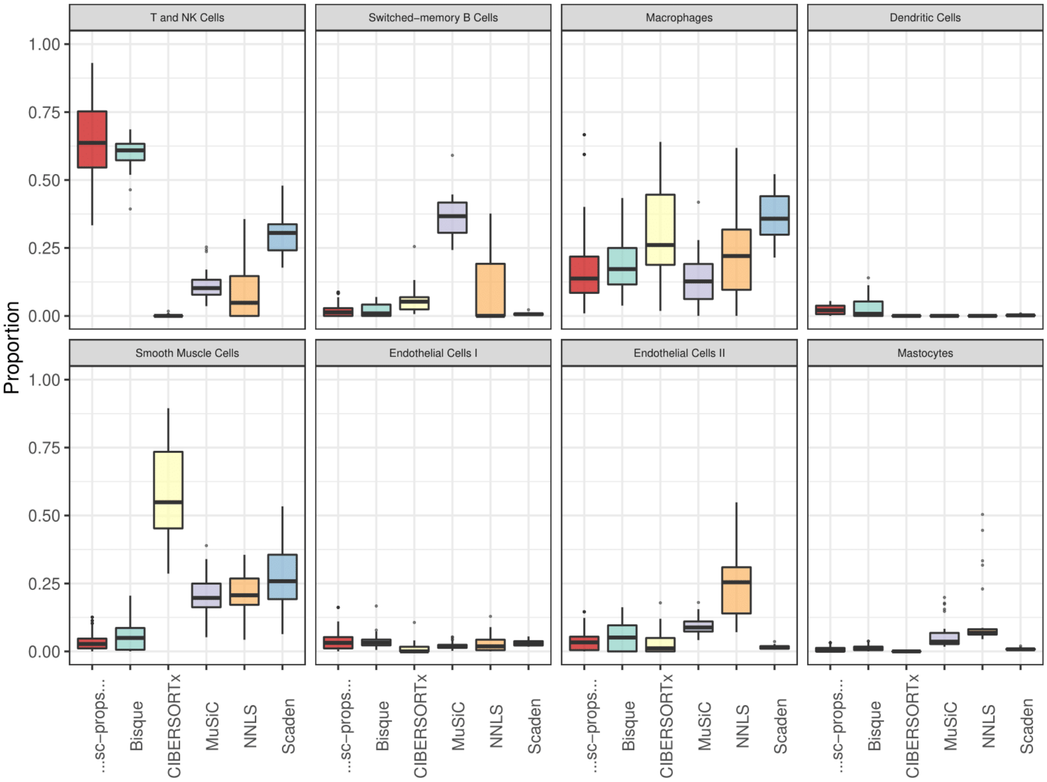

Fig. S5. Plaques with both bulk and single-cell RNA sequencing (sc RNA-seq) data from Athero-Express (AE). A) Boxplot with the deconvolved cell type proportions of 28 AE plaques which were also part of the sc RNA-seq dataset. In red, proportions derived from the sc RNA-seq dataset (proportion of sequenced cells per cell type). B) Boxplot with the deconvolved proportions against the histology score for T and NK cells (CD3.score), ECs (CD34.score), macrophages and dendritic cells (CD68.score) and SMCs (alpha.SMA.score).

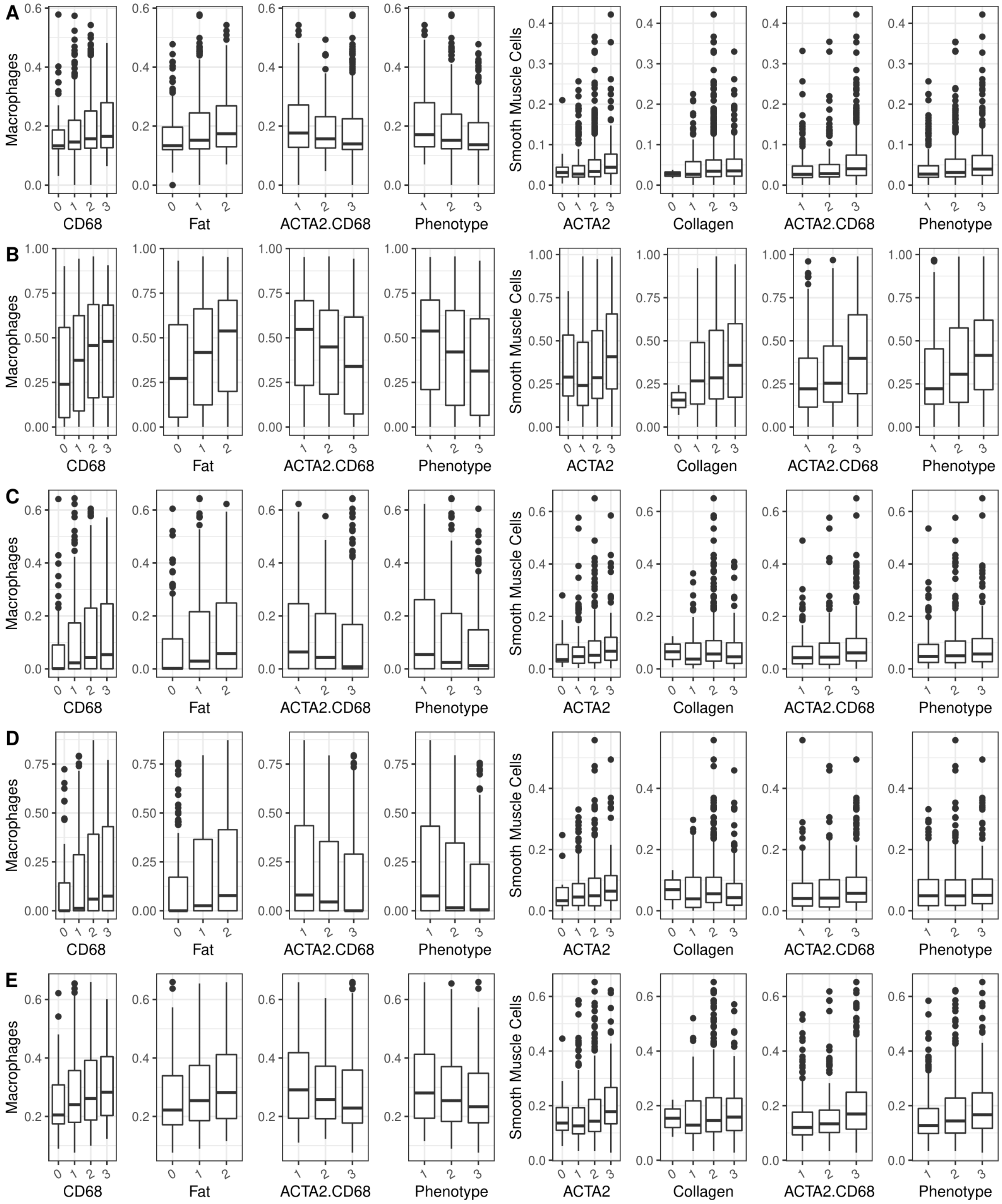

Fig. S6. Deconvolved proportions with histology association. Boxplot of the deconvolved macrophage and smooth muscle cells proportions of 654 AE plaques by histology scores, with the method A) Bisque, B) CIBERSORTx, C) MuSiC, D) NNLS and E) Scaden. Plaque phenotype 1-3 goes from atheromatous-fibrous.

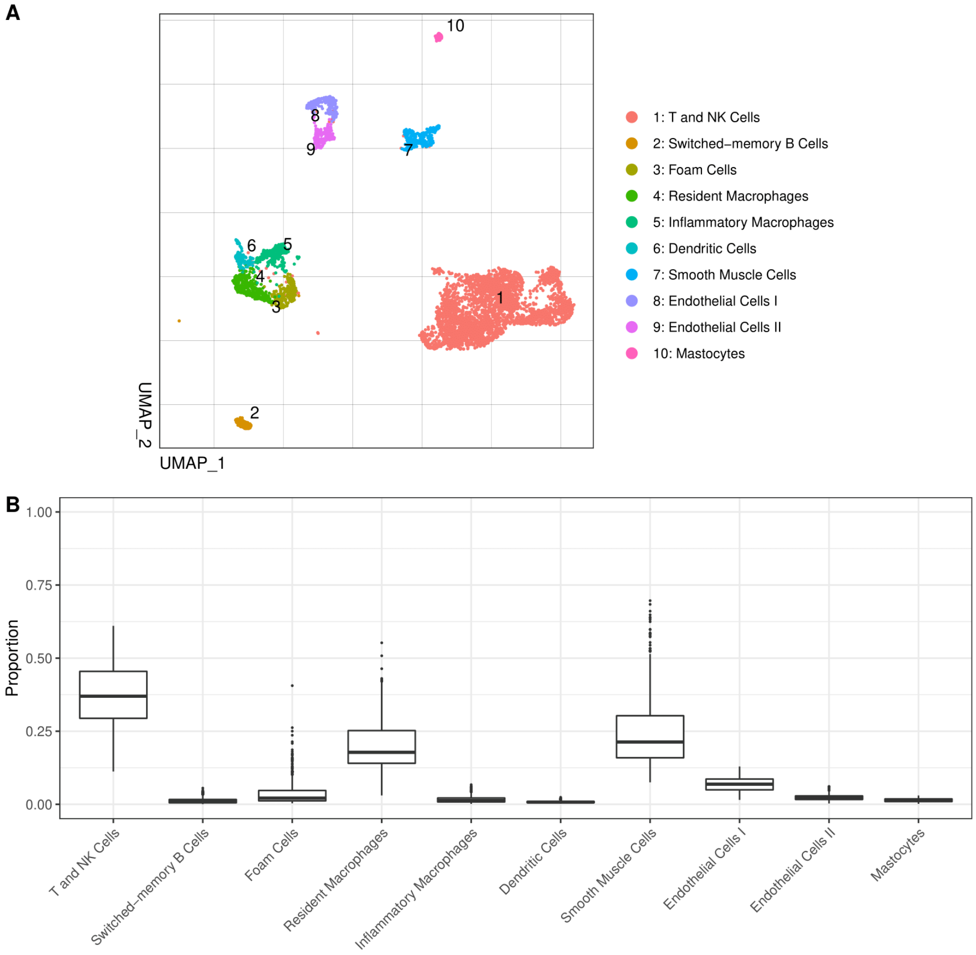

Fig. S7. Deconvolved proportions with macrophage population divided into 3 clusters. A) UMAP visualization of the cell populations. B) Boxplot with the deconvolved proportions of the populations in (A) for 656 Athero-Express plaques.

7.2 Supplementary tables

Table S1. Deconvolved proportions using Athero-Express and Pan et al. [42] references. Pearson correlation between the proportions of similar cell type proportions identified in both single-cell RNA sequencing references. P-values are shown with symbols (*** p<0.01; ** p<0.05; * p<0.1).

Cell population	Bisque	CIBERSORTx	MuSiC	NNLS	Scaden
T Cells	0.499 (***)	0.087 (**)	0.263 (***)	0.027 ( )	0.081 (**)
Macrophages	0.873 (***)	0.958 (***)	0.932 (***)	0.913 (***)	0.943 (***)
Smooth Muscle Cells	0.683 (***)	0.504 (***)	0.709 (***)	0.599 (***)	0.484 (***)
Endothelial Cells I	0.783 (***)	0.678 (***)	0.35 (***)	0.153 (***)	0.344 (***)
Endothelial Cells II	0.154 (***)	-0.033 ( )	-0.044 ( )	0.075 (*)	0.384 (***)
Mastocytes	0.738 (***)	0.107 (***)	0.304 (***)	0.124 (***)	0.854 (***)

Table S2. Baseline characteristics of the Athero-Express patients stratified by sex. P-values are calculated using t-tests for continuous variables (presented as mean (SD)) and chi-squared tests for categorical variables (presented as N (%)). LDL: Low-Density Lipoprotein, HDL: High-Density Lipoprotein, BMI: Body Mass Index, CAD: Coronary Artery Disease

Baseline characteristics	Female (n = 169)	Male (n = 485)	p-value
Smoker = yes (%)	71 (44.1)	160 (33.3)	0.017
Diabetes = yes (%)	35 (20.7)	108 (22.3)	0.754
Hypertension = yes (%)	144 (85.2)	426 (87.8)	0.456
Symptoms = symptomatic (%)	113 (68.1)	324 (68.1)	1.000
Age (mean (SD))	68.254 (9.509)	68.623 (8.705)	0.644
BMI (mean (SD))	26.685 (4.565)	26.600 (3.490)	0.811
Total cholesterol (mean (SD))	4.919 (1.340)	4.579 (1.199)	0.011
Triglycerides (mean (SD))	1.578 (0.887)	1.638 (0.955)	0.564
LDL (mean (SD))	2.842 (1.076)	2.754 (1.031)	0.455
HDL (mean (SD))	1.285 (0.417)	1.093 (0.345)	<0.001
CAD history = no (%)	126 (75.0)	308 (63.5)	0.009

Table S3. Baseline characteristics of the Athero-Express patients stratified by symptoms. P-values are calculated using t-tests for continuous variables (presented as mean (SD)) and chi-squared tests for categorical variables (presented as N (%)). LDL: Low-Density Lipoprotein, HDL: High-Density Lipoprotein, BMI: Body Mass Index, CAD: Coronary Artery Disease, MACE: MAjor Cardiovascular Event.

Baseline characteristics	**Asymptomatic (n = 205)**	**Symptomatic (n = 437)**	**p-value**
Sex = male (%)	152 (74.1)	324 (74.1)	1.000
Smoker = yes (%)	80 (39.2)	146 (34.3)	0.262
Diabetes = yes (%)	44 (21.5)	95 (21.7)	1.000
Hypertension = yes (%)	181 (88.3)	379 (86.7)	0.669
Age (mean (SD))	67.000 (8.561)	69.320 (8.956)	0.002
BMI (mean (SD))	27.034 (3.907)	26.407 (3.687)	0.056
Total cholesterol (mean (SD))	4.716 (1.267)	4.650 (1.244)	0.608
Triglycerides (mean (SD))	1.637 (0.812)	1.614 (0.998)	0.817
LDL (mean (SD))	2.861 (1.083)	2.743 (1.028)	0.297
HDL (mean (SD))	1.155 (0.402)	1.137 (0.358)	0.630
CAD history = no (%)	127 (62.0)	298 (68.3)	0.131
MACE = no (%)	179 (87.3)	370 (85.6)	0.655

Table S4. Baseline characteristics of the AE patients stratified by major cardiovascular events (MACE). P-values are calculated using t-tests for continuous variables (presented as mean (SD)) and chi-squared tests for categorical variables (presented as N (%)). LDL: Low-Density Lipoprotein, HDL: High-Density Lipoprotein, BMI: Body Mass Index, CAD: Coronary Artery Disease.

Baseline characteristics	**No MACE (n = 561)**	**MACE (n = 88)**	**P-value**
Sex = male (%)	410 (73.1)	72 (81.8)	0.107
Smoker = yes (%)	195 (35.2)	35 (41.2)	0.343
Diabetes = yes (%)	110 (19.6)	30 (34.1)	0.003
Hypertension = yes (%)	481 (85.7)	84 (95.5)	0.019
Symptoms = symptomatic (%)	370 (67.4)	62 (70.5)	0.655
Age (mean (SD))	68.173 (8.858)	70.989 (8.490)	0.005
BMI (mean (SD))	26.512 (3.717)	27.266 (4.071)	0.094
Total cholesterol (mean (SD))	4.709 (1.261)	4.446 (1.131)	0.121
Triglycerides (mean (SD))	1.612 (0.913)	1.709 (1.083)	0.451
LDL (mean (SD))	2.805 (1.045)	2.650 (1.015)	0.291
HDL (mean (SD))	1.157 (0.370)	1.057 (0.396)	0.054
CAD history = No (%)	379 (67.6)	51 (58.6)	0.129

Table S5. Association between symptoms and macrophage content obtained with Pan et al. reference [42]. Odds ratios (OR), presented with the 95% confidence interval, and p-values are calculated with univariate and multivariate logistic regression between macrophage percentage and symptoms. Multivariate models were adjusted for age and body mass index.

Macrophage population	**Univariate**		**Multivariate**	
	OR [95% CI]	**p-value**	**OR [95% CI]**	**p-value**
Macrophage 1	1.02 [1.00, 1.04]	0.062	1.02 [1.00, 1.05]	0.040
Macrophage 2	1.17 [0.78, 1.90]	0.492	1.08 [0.65, 1.86]	0.777
Macrophage 3	0.80 [0.67, 0.95]	0.009	0.77 [0.64, 0.93]	0.006

Table S6. Association between MACE and macrophage content obtained with Pan et al. reference [42]. Hazard ratios (HR), presented with the 95% confidence interval, and p-values are calculated with Cox proportional hazard models between macrophage percentage and symptoms. Multivariate models were adjusted for age, hypertension, diabetes and high-density lipoprotein levels.

Macrophage population	**Univariate**		**Multivariate**	
	HR [95% CI]	**p-value**	**HR [95% CI]**	**p-value**
Macrophage 1	1.00 [0.97, 1.03]	0.999	1.00 [0.97, 1.04]	0.848
Macrophage 2	1.19 [0.82, 1.73]	0.362	1.16 [0.63, 2.14]	0.632
Macrophage 3	1.15 [0.93, 1.41]	0.199	1.15 [0.90, 1.47]	0.278

7.3 Supplementary code

Developed code for this project can be found on the following repository: https://github.com/gemmabb/DeconvolutionAtheroscleroticPlaques.