Integrative single-cell meta-analysis reveals disease-relevant vascular cell states and markers in human atherosclerosis

Jose Verdezoto Mosquera[1,2], Doris Wong[1,2], Gaëlle Auguste[2], Adam W. Turner[2], Chani J. Hodonsky[2], Christian L. Lino Cardenas[3], Konstantinos Theofilatos[4], Maxime Bos[5], Maryam Kavousi[5], Patricia Peyser[6], Manuel Mayr[4], Jason C. Kovacic[7,8,9], Johan L. M. Björkegren[10,11], Rajeev Malhotra[3], Sander W. van der Laan[12], Chongzhi Zang[1,2,13], Nathan C. Sheffield[1,2,13], Clint L. Miller[1,2,13]

Affiliations:

[1]Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, Virginia, USA.

[2]Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, USA.

[3]Cardiovascular Research Center, Cardiology Division, Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA.

[4]King's British Heart Foundation Centre, King's College London, United Kingdom.

[5]Department of Epidemiology, Erasmus University Medical Center, The Netherlands.

[6]Department of Epidemiology, University of Michigan, Ann Arbor, Michigan, USA.

[7]Cardiovascular Research Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA.

[8]Victor Chang Cardiac Research Institute, Darlinghurst, New South Wales, Australia.

[9]St. Vincent's Clinical School, University of New South Wales, Sydney, New South Wales, Australia.

[10]Department of Genetics and Genomic Sciences, Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, New York, USA.

[11]Integrated Cardio Metabolic Centre, Department of Medicine, Karolinska Institutet, Huddinge, Sweden.

[12]Central Diagnostics Laboratory, Division Laboratories, Pharmacy, and Biomedical Genetics, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands.

[13]Department of Public Health Sciences, University of Virginia, Charlottesville, Virginia, USA.

Corresponding author:

Clint L. Miller, PhD
University of Virginia
Center for Public Health Genomics, MSB 3231
PO Box 800717
Charlottesville, VA 22908 USA
clintm@virginia.edu

Abstract

Atherosclerosis is a complex inflammatory process driven by plaque formation in the major elastic arteries and often leads to reduced blood flow, coronary artery disease (CAD), myocardial infarction and stroke. CAD progression involves complex interactions and phenotypic plasticity within and between distinct vascular and immune cell lineages. Several single-cell RNA-seq (scRNA-seq) studies have highlighted lineage-specific transcriptomic signatures however there remains variability on the reported cell phenotypes in humans. In this study we meta-analyzed scRNA-seq datasets across four publications to create a comprehensive map of human atherosclerosis cell diversity. We applied standardized QC, processing, and integration benchmarking to harmonize 118,578 high-quality cells for this atlas. Beyond characterizing vascular and immune cell diversity, we derived insights into smooth muscle cell (SMC) phenotypic modulation through pseudotime, transcription factor activity inference and cell-cell communication analyses. We also integrated genome-wide association study (GWAS) data to identify etiologic cell types for GWAS diseases and traits, which uncovered a critical role for modulated SMC phenotypes in CAD and coronary artery calcification. Finally, we identified candidate markers (e.g., *CRTAC1*) of synthetic and osteochondrogenic SMCs that may serve as proxies of atherosclerosis progression. Together, this represents an important step towards creating a unified cellular map of atherosclerosis to inform cell state-specific mechanistic and translational studies of cardiovascular diseases.

Introduction

Cardiovascular diseases, such as coronary artery disease (CAD), are the leading global causes of mortality and morbidity[1]. The pathological hallmark of CAD is atherosclerosis, a chronic build-up of plaque inside arterial walls, which can lead to thrombus formation and myocardial infarction (MI) or stroke[2–5]. This process involves a complex interplay of both immune and vascular cell types and cell state transitions along a continuum[6,7]. In response to injury of the inner vessel layer by oxidized low density liporoteins (ox-LDL) and immune cells, contractile smooth muscle cells (SMCs) transition to a more proliferative and migratory state[8,9]. Similarly, endothelial cells transition to a mesenchymal state in early and advanced atherosclerosis[10,11].

Recent single-cell RNA sequencing (scRNA-seq) studies have resolved the cellular diversity and gene signatures in human and murine atherosclerotic lesions[12–16] as well as in non-lesion arteries[17]. By combining lineage tracing and scRNA-seq, studies have shown that SMC readily transform into a multipotent “pioneer” cell type in response to pro-atherogenic stimuli[18–20]. However, the fate of SMCs after this transition remains controversial; a few studies generally agree that these pioneer cells can become fibroblast-like (fibromyocytes)[18] or osteogenic-like (fibrochondrocytes; FCs)[19]; other studies suggest that pioneer SMCs adopt pro-inflammatory or macrophage-like properties[8,20]. Limited sample sizes, experimental design or other technical factors could potentially confound the biological interpretation of these individual studies. Thus, there remains a need for a consensus single-cell reference[21–23], which spans atherosclerotic disease stages in humans.

Here, we harmonize and meta-analyze four single-cell studies of human atherosclerosis, encompassing both early and advanced lesion and non-lesion samples (**Fig. 1a and Supplementary Table 1**). This high-resolution atlas of 118,578 high quality cells enables the discovery of previously missed vascular and immune cell types and clarifies markers for known disease-relevant immune cells (e.g., inflammatory, and foamy macrophages). We perform integrative downstream analyses and GWAS trait enrichment to define cardiovascular traits and disease-relevant etiologic cell types and states. We further validate SMC phenotypes identified in lineage-tracing studies, reveal underrepresented SMC states from individual scRNA-seq studies and highlight *CRTAC1* as a new candidate marker of pro-calcifying SMCs and plaque stability in humans. This comprehensive map of vascular and immune cell diversity in human atherosclerosis provides a critical step towards translating mechanistic knowledge and developing more targeted interventions.

Results

Integration of lesion and non-lesion artery datasets

We sought to build a comprehensive single-cell reference that would be well-powered to further investigate complex vascular processes such as SMC phenotypic modulation. In order to avoid biases towards lymphoid (e.g., T-cells) and myeloid (e.g., macrophage) cells and achieve proper representation of mural cells (SMC and pericytes), the current version of this atlas was assembled using data from three studies, *Wirka et al* [18], *Pan et al* [19], and *Alsaigh et al* [16] profiling human atherosclerotic lesions in coronary or carotid arteries. We also included a recently published dataset of non-lesion coronary arteries[17] with the goal of spanning the continuum of CAD risk (**Supplementary Table 1**). We then established a standardized pipeline for quality control (QC) and uniform processing of the 22 raw sequencing libraries, involving removal of doublets[24] and ambient RNA [25] as well as normalization to account for variable cell sequencing depth[26,27] (**Supplementary Fig. 1 and Methods**). We visually inspected cell embeddings and observed optimal separation and cohesion of cell clusters post-filtering (**Supplementary Fig. 1**).

The choice of single-cell integration approach highly depends on the context of the individual datasets, the magnitude of batch effect and cell number. Thus, we independently evaluated tools recommended by 3 recent benchmarks[28–30] including Canonical Correlation Analysis + Mutual Nearest Neighbors (CCA + MNN), reciprocal PCA (rPCA)[26], Harmony[31] and Scanorama [32] (**Methods**). We tested a subset of the included libraries[16,18,19] and found that rPCA and Harmony outperformed the other tools in terms of running time (**Supplementary Fig. 1**). We also evaluated the effectiveness of batch removal from each approach using integration Inverse Local Simpson Index (iLISI)[31] and conservation of biological variation using the “cell type LISI” (cLISI)[31]. Finally, clustering purity was measured using silhouette coefficients, which capture elements of both sample mixing and local structure[26]. To avoid over- or under-clustering, we calculated silhouette coefficients across a wide range of resolutions (0.8-1.8). From this benchmark, we found that rPCA achieved the best balance in terms of running time, batch mixing and conservation of biological variation. This method also achieved the highest clustering purity across all tested resolutions (**Supplementary Fig. 1**). Integration of libraries

with rPCA yielded a total of 118,578 high-quality cells and 41 Louvain clusters (**Supplementary Fig. 1**).

Next, we used a combination of manual and automated annotation tools to label the broad cell (level 1) and granular subtype (level 2) compartments. We defined level 1 annotations by reprocessing and transferring cell type labels from the Tabula Sapiens (TS) vasculature single-cell atlas[33]. We found that labels were assigned with remarkably high confidence (**Supplementary Fig. 2**). These annotations were supported by the expression of well-established marker genes in corresponding level 1 clusters (**Fig. 2a-b**) and confirmed that batch effects had been properly removed while conserving biological variation.

We observed a balanced number of cells labeled as macrophages and endothelial cells across studies. However, there were slightly more SMCs in *Pan et al* and T/NK cells in *Alsaigh et al* and slight biases from small clusters (e.g., plasma cells, B cells) between studies (**Fig. 2c**). This shows that individual studies may under-represent key cell types. We also observed overrepresented fibroblasts from coronary datasets (*Wirka et al* and *Hu et al*), as expected given the intact coronary vessel wall layers compared to carotid plaques **(Fig. 2c)**. When comparing cell type frequencies across disease status, we observed a greater proportion of B cells, plasma cells and pDCs in lesion samples **(Fig. 2c-d).** Further, libraries from *Alsaigh et al* had the highest proportion of T cells among all studies, consistent with the advanced stage of the carotid lesions (**Supplementary Fig. 2**).

To further ensure the accuracy of our level 1 cell type annotations, we ran a differential expression (DE) analysis to obtain cell type gene markers **(Methods and Supplementary Table 2)**. As expected, SMCs and endothelial cells (EC) were enriched for gene ontology (GO) terms such as “muscle contraction” and “endothelium development”, respectively. We also observed SMCs enriched for terms such as “extracellular matrix organization”, likely due to the presence of phenotypically modulated SMCs that have acquired synthetic properties [9]. (**Supplementary Fig. 2, Supplementary Table 3**). In contrast, myeloid and lymphoid clusters were enriched for immune-related terms such as “antigen processing and presentation” and “regulation of T cell activation” (**Supplementary Fig. 2, Supplementary Table 3**).

Defining candidate etiologic cell types for complex traits

Next, we identified etiologic cell types enriched for atherosclerosis-related traits using our level 1 cell type annotations. Briefly, we performed stratified LD score regression (S-LDSC) analysis[34,35] using GWAS summary statistics for cardiovascular disease (CVD) and non-CVD traits as described[36–41]. To improve the specificity of GWAS enrichment per cell type, we first derived an expression specificity matrix using SCTransform-normalized counts where each value (ES_μ) represents the average of multiple differential expression metrics[42] (**Methods**). SMC and pericyte gene signatures were significantly enriched (FDR < 0.05) for CV traits such as pulse pressure, CAD, and MI (**Fig. 2e and Supplementary Table 4**). On the other hand, EC signatures were enriched for carotid plaque associations (**Fig. 2e**). Consistent with previous studies[35,43], we observed macrophages were highly enriched for Alzheimer’s disease and white

blood cell count GWAS signals. We also found high enrichment of neurons for body mass index (BMI). These findings highlight the value of integrating single-cell and human genetic data to discover atherosclerosis trait-relevant cell types, such as SMC and ECs.

Defining cell subtype heterogeneity in human atherosclerosis

Next, we surveyed the 41 clusters using a combination of automated and manual annotation **(Methods)**. Manual annotations included markers of lymphoid, myeloid and endothelial cell subtypes from the literature[21–23,44–49]. We then verified manual annotations using the CellTypist machine learning classifier [50] resulting in a more granular map of cell diversity in human atherosclerosis **(Fig. 3a).** We summarize some of the most representative cell subpopulations below:

Endothelial diversity: Within the endothelial compartment, we identified cells highly expressing classical endothelial markers (*PECAM1*, *CLDN5)* relative to neighboring clusters (**Fig. 3a, Supplementary Table 5**). Expression of homeostatic EC marker genes such as *RAMP2*[51]) led to the annotation of this cluster as “Intimal ECs”. We also identified a cluster of cells marked by the upregulation of vasa vasorum genes such as *ACKR1*[52] and angiogenesis-related genes such as *AQP1 and FABP4* [53–55] . Adjacent to pro-angiogenic ECs, we identified a cluster marked by elevated expression of chemokine and adhesion molecules (*SELE*, *CCL2*,) (**Fig. 3a-b, Supplementary Table 5**), likely reflecting a pro-inflammatory state[56]. EndoMT ECs[57] were defined by the expression of ECM genes (*COL1A2, FN1)* and contractile genes **(Supplementary Table 5)**. Finally, we defined a small subcluster of lymphatic ECs based on the expression of *LYVE1* and *CCL21*[58] (**Fig. 3a-b, Supplementary Table 5**).

Myeloid diversity: We identified a subset of myeloid cells, inflammatory macrophages, which express known markers of inflammation (*IL1B*, *TNF*), characteristic of the polarized M1 macrophage state (**Fig. 3a-b, Supplementary Table 5**). We defined foamy macrophages marked by high expression of lipid metabolism and lipoprotein uptake genes such as *APOE* and *FABP5,* along with a reduced inflammatory profile[15,21,23] (**Fig. 3a-b, Supplementary Table 5**). We also identified resident macrophages (*LYVE1, FOLR2*), classical monocytes (*S100A8*, *S100A9, LYZ*), and conventional dendritic cells (*CD1C*, *CLEC10A*)[59–61] (**Fig. 3a-b, Supplementary Table 5**). Importantly, we resolved critical smaller myeloid populations overlooked by previous individual human scRNA-seq datasets, including plasmacytoid dendritic cells (pDCs)[62,63] and neutrophils (*NAMPT*, *S100A8*) (**Fig. 3a-b**). Consistent with myeloid cell infiltration during atherosclerosis, we found that monocytes, foamy macrophages, and other myeloid populations were substantially more prevalent in libraries from lesions (**Fig. 2c**).

Lymphoid diversity: Within the largest lymphoid cell compartment (T/NK), we identified Natural Killer (NK) and several subpopulations of T cells. First, we identified a cluster of NK cells defined by expression of *XCL1*, *NKG7* and *GNLY* **(Fig 2a-b)**. Expression of *CD69,* a classical early activation marker of lymphocytes[64], suggests this cluster contains activated NK cells **(Fig. 2a, Supplementary Fig. 2 and Supplementary Table 5)**. We then defined populations of CD8 T cells based on expression of *CD8A* and *CD8B* (**Fig. 2b and Supplementary Table 5**) and

found two *CD8+* clusters with strong expression of chemokines and granzymes (*CCL5*, *GZMK*) displaying variable expression of *CD69* and *IL7R* **(Supplementary Fig. 5)**, suggesting the presence of early activated and memory/naive CD8 cytotoxic T cells (CTLs)[65,66]. We also found an adjacent *CD8+* cluster expressing *CD3E*, *NKG7*, granzymes and *GNLY* lacking *CD69* expression **(Supplementary Fig. 2)** which suggests terminally differentiated CTLs[15]. Additionally, we found 3 clusters of T cells lacking *CD8* expression and showing subtle but noticeable expression of the TFs $ROR_{\gamma}t$ and GATA3 **(Supplementary Fig. 2)**, hallmark drivers of Th17 and Th2 helper cells, respectively[67]. We also found a T cell cluster with a more concentrated expression of FoxP3, a master driver of regulatory T cells (T_{reg})[67] **(Supplementary Fig. 2)**. Finally, we defined clusters of B cells (*CD79A*, *CD79B*) and plasma cells (*IGLC2*, *IGHM*, *JCHAIN*). While all lymphoid populations showed larger frequencies in lesions, we found that B cells, plasma cells and pDCs were highly depleted in non-lesion libraries (**Supplementary Fig. 2**).

Fibroblast diversity: Defining fibroblast diversity in atherosclerosis is particularly challenging given the low specificity of widely used fibroblast markers[48]. We found that most cells in this compartment express traditional fibroblast ECM markers such as *LUM* and *DCN* (**Supplemental Table 5**). We were able to dissect a subset of fibroblasts that upregulated the contractile marker *ACTA2* (**Fig. 2b**) as well as complement genes (*C3* and *C7*). This subset likely represents activated fibroblasts (myofibroblasts) known to adopt increased contractile, ECM-producing, and pro-inflammatory states in response to injury or atherosclerotic stimuli[48,57]. Of note, we identified a group of cells within the fibroblast compartment strongly expressing *APOE* in addition to the chemokine ligands *CXCL12* and *CXCL14* and complement genes, which we term APOE fibroblasts (**Supplemental Table 5**).

Characterization of SMC phenotypes in human atherosclerosis

To refine the role of SMC phenotypes in our human scRNA reference, we performed gene set enrichment of gene modules from a recent scRNA meta-analysis of murine vascular SMCs[22]. First, we subset the full atlas to include only SMCs, pericytes and a subset of fibroblasts. We then assessed enrichment of lineage-traced murine SMC gene modules on a per-cell basis using the UCell R package (**Methods**)[68]. This analysis showed a progressive loss of the murine SMC contractile signature within a portion of the human subset, coincident with a gain in the *Lgals3+* transitional gene signature (**Fig. 4a)**, supporting a transitional SMC signature in humans. Further, we detected an enriched signature of the murine calcification-promoting fibrochondrocytes distinct from non-SMC-derived fibroblasts (**Fig. 4a).** Cluster DE markers and UCell module enrichment scores were used as a guide to annotate SMCs as contractile, ECM-rich transitional SMCs, fibromyocytes and fibrochondrocytes (FCs) (**Fig. 4b-c, Supplementary Fig. 3, Supplementary Table 6**).

We observed similar proportions of contractile, transitional ECM SMCs and fibromyocytes across arterial beds and lesion status, consistent with previous reports[69]. However, FCs predominated in lesions compared to non-lesion samples (**Supplementary Fig. 3**), in line with their role in calcification. The FC annotation was further supported by higher ES_{μ} values for

SOX9 and RUNX2, known master regulators of SMC osteochondrogenic transitions[70] (**Supplementary Fig. 3**). At a global level, SMCs, transitional SMCs, fibromyocytes, and FCs were enriched for relevant biological processes thus validating our annotation approach (**Supplementary Fig. 3 and Supplementary Table 7**). Interestingly, we also identified a cluster enriched for a lipid metabolism transitional SMC gene signature (*APOE*, *APOC1, AGT)* (**Fig. 4b-c, Supplementary Fig. 3, Supplementary Table 6**), which we termed "foam-like" SMCs. These cells also expressed ECM-remodeling genes such as *TIMP1* and pro-inflammatory genes *CCL19*, *CCL2, IGFBP3*, consistent with a potential role in leukocyte recruitment[71].

Finally, we leveraged these SMC labels to dissect the disease relevance of SMC modulated phenotypes using S-LDSC. Fibromyocytes and foam-like SMCs were highly enriched for CAD heritability, while fibromyocytes were enriched for MI and subclinical CAD traits (**Fig. 4d and Supplemental Table 8**). In contrast, we observed FCs enriched for coronary artery calcification (CAC) using our recent meta-analysis summary data[72] (**Fig 4d**). This FC enrichment is consistent with our understanding of the biology of CAC, but to our knowledge has not been previously reported in any integrative single-cell and human genetic analysis.

Cell crosstalk in human atherosclerosis

We then dissected key cellular crosstalk from our level 1 and 2 annotations across lesion status using CellChat[73]. We observed strong interactions between SMCs and fibroblasts in non-lesion samples, while SMC and EC interactions with macrophages and T/NK were stronger in lesions (**Fig. 5a**). Unexpectedly we observed tumor necrosis factor alpha (TNFa) and platelet-derived growth factor (PDGF) signaling pathways were less represented in lesions **(Fig. 5b),** likely due to the higher proportion of inflammatory macrophages in non-lesion samples from *Hu et al.* While we did not find significant differences in information flow for these two pathways by lesion status (**Supplementary Table 9**), tumor-necrosis factor-like weak inducer of apoptosis (TWEAK) and osteopontin (SPP1) mediated signaling pathways were highly enriched in lesion samples **(Fig. 5b).** TWEAK mediated interactions between SMC and monocyte/DC were also greater compared to other myeloid subtypes (**Supplementary Fig. 4**). Signaling involving osteopontin (SPP1)[14,21], specifically targeted SMCs and was mostly driven by macrophage foam cell clusters (**Fig. 5c**). We subsequently used our SMC subtype annotations to further understand interactions with myeloid populations (**Fig. 5d**). Among pathways where SMC subtypes were noted as signaling targets, we found TGF-β, NOTCH, PDGF, granulin (GRN), vascular cell adhesion molecule (VCAM), TWEAK and SPP1 signaling (**Supplementary Table 10**). Given TWEAK and SPP1 enrichment in lesions, we focused on these pathways and observed more TWEAK mediated interactions between contractile/transitional SMCs with distinct myeloid subtypes in lesion samples (**Supplementary Fig. 4**). We also found that incoming SPP1 mediated signals from foamy macrophage cells (foamy mac1) specifically targeted contractile and transitional SMCs (**Fig. 5e**). Finally, using ligand-receptor contribution analyses, we found that cells expressing genes encoding SPP1 ligand preferentially signal via the heterodimeric ITGA8/ITGB1 receptor (**Fig. 5f**).

Modeling SMC gene expression across pseudotime

Current evidence suggests that SMCs transition into fibromyocytes/FCs through an *Lgals3*+ transitional state[19,20] . We modeled SMC de-differentiation via pseudotime analysis using Monocle 3[74], in which we defined *MYH11*-expressing contractile SMCs as the starting point of phenotypic modulation (**Fig. 6a**). This inferred trajectory revealed a branchpoint where transitional SMCs could adopt either a fibromyocyte or FC fate (**Fig. 6a**). In addition, we observed more FCs in lesion samples towards the latter pseudotime stages, consistent with calcification in advanced lesions (**Fig. 6a-b**). Using differential expression and Louvain community analysis (**Methods**), we identified modules specific to transitional SMC (Modules 5 and 10), fibromyocytes (Module 4) and FCs (Module 9) (**Supplementary Fig. 5**). Transitional SMC modules harbored genes involved in early SMC investment in atherosclerotic lesions (e.g., *LGALS3*)[20], as well as cell division and proliferation (e.g., *TUBA1B* and *SIRT6*)[75] and ECM remodeling (e.g. *KRT8* and *SRARC*) (**Supplementary Fig. 5**). As expected, fibromyocyte module 4 included known markers (e.g., *FN1*, *VCAN*, *COL4A1/2*, *PDGFRB*) (**Supplementary Fig. 5**). In contrast, the FC module 9 harbored chrondrocyte related genes such as *BMP4* [76], *WISP2*, and *SPRY1*[77] in addition to known ECM genes *LUM* and *DCN.*

Next, we modeled the expression dynamics of our DE genes using cubic spline interpolation across pseudotime. As expected, expression of canonical SMC contractile markers, *MYH11* and *CNN1* sharply decreased across pseudotime (**Fig 6c**), whereas *ACTA2* and *TAGLN* persisted longer (**Supplementary Fig. 5**). Interestingly some fibromyocyte markers such as *FN1, AEBP1 and LTBP1* showed a steady increase with adoption of the transitional state (**Fig. 6c, upper panel**), while genes such as *PDGFRB,* were increased later suggesting a distinct role in the fibromyocyte state (**Fig. 6c, upper panel**). In parallel, we inspected FC markers from our previous DE analysis (**Supplementary Table 6**) as well as genes from module 9 such as *MMP2*. We observed a steady increase in expression of *COL1A2* and *MMP2*, whereas *IBSP*, *CRTAC1* and *COMP* were increased at later pseudotime stages, presumably as transitional SMCs adopt a FC fate (**Fig. 6c, lower panel**).

TF activity inference analysis

We next investigated the upstream transcriptional factors driving cell specific expression changes using TF activity inference with VIPER[78] and the DoRothEA collection of well-curated and stable human regulons[79]. This analysis revealed known regulators of fibromyocytes and FCs such as TCF21 and SOX9 (**Fig. 6d**) in addition to AP-1 (e.g., JUN, FOSL), TEAD, ETV and ETS factors (**Supplemental Fig. 5**)[80]. Interestingly, we observed increased regulon activity of the TGF-β signaling mediator SMAD3 in fibromyocytes and FCs compared to contractile and transitional SMCs (**Fig. 6d**). To confirm these results we interrogated our previously published coronary artery snATAC-seq data[80], analyzed using ArchR[81]. Besides confirming increased accessibility of AP1 factors, we found that accessible regions in the ECM-rich SMC cluster were specifically and highly enriched for SMAD3 motifs compared to contractile SMCs (**Supplemental Fig. 5**). This suggests that SMAD3 activity is critical as SMCs transdifferentiate towards more synthetic phenotypes.

CRTAC1 as candidate marker of FC and atherosclerosis progression

Given that arterial calcification is a major risk factor for clinical cardiovascular disease events [82,83], we further explored specific markers of FCs. Using the top FC marker genes, we found a previously unreported marker of this cell type, cartilage acidic protein (*CRTAC1)* expressed 3-fold higher relative to other SMC clusters (**Fig. 4c, Supplementary Fig. 3, Supplemental Table 6**). *CRTAC1* has been previously established as a specific marker for human chondrocytes during ossification[76,84], and has been implicated in osteoarthritis[85]. In the Genotype Tissue Expression (GTEx) database, *CRTAC1* showed highly arterial-specific expression compared to other tissues (**Supplementary fig. 6**). We also observed expression in SMCs enriched for murine FC gene signatures along with the calcification marker, *IBSP*[20,70] (**Fig. 4a and Fig. 7a**). By comparing all co-expressed genes in contractile SMCs and FCs, we found that *CRTAC1* was positively correlated with known ECM markers (e.g., *DCN* and *LUM*) and other osteochondrogenic markers (e.g., *COL1A2*, *SOX9*) and negatively correlated with canonical SMC markers *(e.g., MYH11, CNN1)* (**Fig. 7b**). This suggests that expression of this gene in SMCs is associated with loss of the SMC contractile phenotype and gain of pro-calcification gene programs[86,87].

To further validate *CRTAC1* as a potential marker of human atherosclerosis, we queried our human coronary bulk RNA-seq and proteomics data as well as public datasets. Using coronary arteries from a cohort of 45 individuals, we found that *IBSP* and *CRTAC1* were significantly upregulated in lesions compared to non-lesion samples (**Fig. 7c**). We also observed a significant increase in CRTAC1 protein abundance in lesion samples (**Fig. 7d**). Consistently, *CRTAC1* was upregulated in unstable relative to stable carotid plaques in a published dataset [88] (**Fig. 7e**), suggesting a potential role for *CRTAC1* in calcification underlying plaque stability. Next, we queried the Stockholm-Tartu Atherosclerosis Reverse Network Engineering Task (STARNET) gene regulatory networks across seven cardiometabolic tissues[89]. *CRTAC1* was identified as a significant key driver within its co-expression module (**Supplementary table 11 and 12**), which was highly associated with CAD genes, C-reactive protein (CRP), LDL cholesterol (**Fig. 7f**), and enriched for “ossification” and “extracellular matrix organization” GO terms, further suggesting a link to calcification (**Supplementary table 13**).

To complement these expression-based analyses, we queried our coronary artery snATAC-seq data. Differential analysis of chromatin accessibility in SMC and modulated SMC peaks revealed increased chromatin accessibility near known fibromyocyte and FC markers (*TCF21*, *RUNX2* and *SOX9)* (**Supplementary Fig. 6**). We also observed increased accessibility near *CRTAC1* and within a portion of modulated SMCs which overlapped increased *SOX9* but not *RUNX2* accessibility-derived gene activity **(Supplementary Fig. 6)**, suggesting a chondrogenic rather than osteoblastic transition. Together these findings strongly support *CRTAC1* as a new marker for atherosclerosis development and SMC modulation in humans.

Discussion

In this study we generated the first comprehensive single-cell transcriptomic atlas of human atherosclerosis, encompassing 22 sequencing libraries **(Supplementary Fig. 2)** from 4 different

studies[16–19], which included data from atherosclerotic coronary and carotid arteries. After performing extensive QC and benchmarking of batch correction methods, we harmonized 118,578 high-quality cells for annotation and analysis. Using both automated cell annotation[26,50] and manual curation we defined major vascular and immune cells and subtypes, many of which have not been previously identified from human atherosclerosis scRNA-seq studies. We further performed heritability enrichment to define disease-relevant cell types, as well as cell-cell communication and gene regulatory network inference to gain insights into atherosclerosis mechanisms. Finally, we provide external and internal validation for a new marker of SMC-derived FCs, *CRTAC1*.

We showcase the discovery potential of this scRNA-seq reference by investigating SMC phenotypic modulation in humans and identifying etiologic cell types in disease. Defining SMC phenotypes in human lesions has been previously achieved by transferring labels from mouse lineage tracing scRNA-seq studies[18–20]. However, this restricts SMC annotations to murine-defined labels, which may limit the discovery of human-specific disease markers. We addressed this by instead using lineage-traced murine SMC genes for per-cell enrichment analysis of our >30,000 mural cells prior to subclustering and differential marker discovery. Beyond corroborating known SMC phenotypes reported in murine studies[18–20], this helped us uncover rare transcriptomic SMC clusters including a “foam-like” state. This provides unbiased support of previous *in vitro* and *ex vivo* studies reporting a SMC-derived foam-like phenotype upon exposure to lipoproteins[90,91] and in human lesions[92]. The lower abundance of these foam-like SMCs in previous scRNA-seq studies could be due to their high sensitivity to single-cell digestion protocols. These cells expressed lipid metabolism genes (e.g., *APOE*, *APOC1)* but no other traditional macrophage markers, in line with previous findings[92]. Their expression of ECM genes such as *TIMP1* suggest SMC-derived foam cells may acquire a unique gene signature from their monocyte-derived counterparts.

Our granular SMC annotations were also critical to define etiologic SMC phenotypes for cardiovascular diseases and traits. Previous work from our group and others[93–95] has established a substantial contribution of SMCs towards CAD risk. By leveraging larger cell numbers, we further separated the SMC signal to prioritize fibromyocytes and foam-like SMCs underlying cardiovascular diseases. Supporting the emerging role of fibromyocytes in plaque stability[18] we demonstrate enrichment for these cells in CAD and MI risk. These heritability analyses also linked SMC-derived FCs to coronary artery calcification, an established pathological hallmark of subclinical and advanced atherosclerotic lesions[96].

Though it has been shown that both fibromyocytes and FCs originate from SMCs[18–20,97], these two ECM-rich phenotypes are postulated to play opposing roles in plaque stability[9] and their exact lineage relationship is not well understood. Murine studies have previously suggested fibromyocytes are progenitors of FCs[71]. Using pseudotime analysis we revealed a branchpoint where transitional SMCs could adopt either fibromyocyte or FC fates. This does not preclude the possibility that fibromyocytes could be primed to become FCs as suggested by *Cheng et al*. Our TF activity inference results showed high SOX9 in FCs, but we also observed a gradient of decreasing TCF21 and increasing SMAD3 activity from fibromyocytes to FCs, hinting at a

fibromyocyte-FC transition. Due to limitations of pseudotime techniques, however, we note that additional lineage-tracing experiments will be needed to address the plausibility of these two lineage paths.

SMAD3 has been previously defined as a causal CAD GWAS gene and increased expression levels are associated with elevated disease risk[93,98], yet the precise role of this gene in human SMC modulation is less understood. Previous studies suggest that SMAD3 antagonizes atheroprotective TCF21 modulation activity, constraining SMCs from migrating to the lesion and fibrous cap[98]. Moreover, recent SMC-specific *Smad3* KO mice resulted in increased proportions of FCs at the expense of fibromyocytes[71]. Interestingly we observed increased SMAD3 motif accessibility in ECM-rich SMCs (fibromyocytes and FCs) compared to contractile SMCs using our combined human scRNA-seq based regulons and snATAC data. Although the specific role of SMAD3 in osteochondrogenic SMC transitions has not been completely elucidated, TGF-β signaling during atherosclerosis has been linked to increased calcification[99] and Smad3 shown to stimulate chondrogenesis in mesenchymal stem cells by enhancing Sox9 transcriptional activity[100,101]. Given the versatility of SMAD3 co-activators and targets, we speculate that this TF might play different roles along different stages of SMCs transition towards ECM-rich phenotypes, but additional mechanistic studies will be required to fully address this possibility.

Our single-cell reference also provides further insights into the FC gene signature in atherosclerotic lesions **(Fig. 6b and Supplementary Fig. 3)**. Differential gene expression in SMC subtypes identified *CRTAC1* as a top, previously unreported FC marker. Our single-cell analysis as well as in-house bulk coronary RNA-seq and proteomics data showed elevated CRTAC1 in coronary lesions, which also correlated with osteochondrogenic markers such as *COMP* and *IBSP*[84]. While CRTAC1 was also upregulated in unstable carotid plaques[88], its role in plaque rupture is unknown. More exhaustive ex vivo and in vivo functional characterization is required to pinpoint its role in diverse calcification phenotypes (e.g. micro/puncate vs. macro/sheet-like)[102–104] and plaque stages (e.g. thin-cap fibroatheromas vs. fibrocalcific plaques).

There are known limitations of this study worth noting. Although non-lesion samples, as denoted throughout this study, had no discernible lesions according to available histology data or clinical CAD diagnosis[17], these arteries were extracted from patients with dilated cardiomyopathies (DCM). Some of the most striking differences in cell type frequency we observed across lesion status included a strong enrichment of foamy macrophages, monocytes, B cells, plasma cells, pDCs and FCs in lesion compared to non-lesion libraries. This is consistent with the strong immune landscape and development of calcification with atherosclerosis progression. Also, while the higher representation of inflammatory macrophages in non-lesion libraries **(Fig. 3c)** is unexpected, this could be potentially attributed to events such as subclinical diffuse intimal thickening. We also acknowledge the potential over-simplification of our lesion status in our group comparisons, which is less refined than in murine studies. Nonetheless, this represents a valuable step forward given the limited metadata from published arterial samples in both healthy and diseased individuals.

Taken together, we provide a comprehensive map of cell diversity in human atherosclerosis (**Supplementary Fig. 7**). With newly generated large-scale single-cell datasets, there will be a need to address the variability of reported phenotypes and create a unified map of human atherosclerosis. As demonstrated through GWAS meta-analyses, integrative analyses hold promise in capturing more robust and subtle signals. We expect this atlas will represent the first iteration of future references to build upon scientific discoveries and help annotate new single-cell multi-omic datasets. Ultimately this will catalyze mechanistic and translational studies and contribute towards developing novel therapeutic strategies for CAD.

Figures and Legends

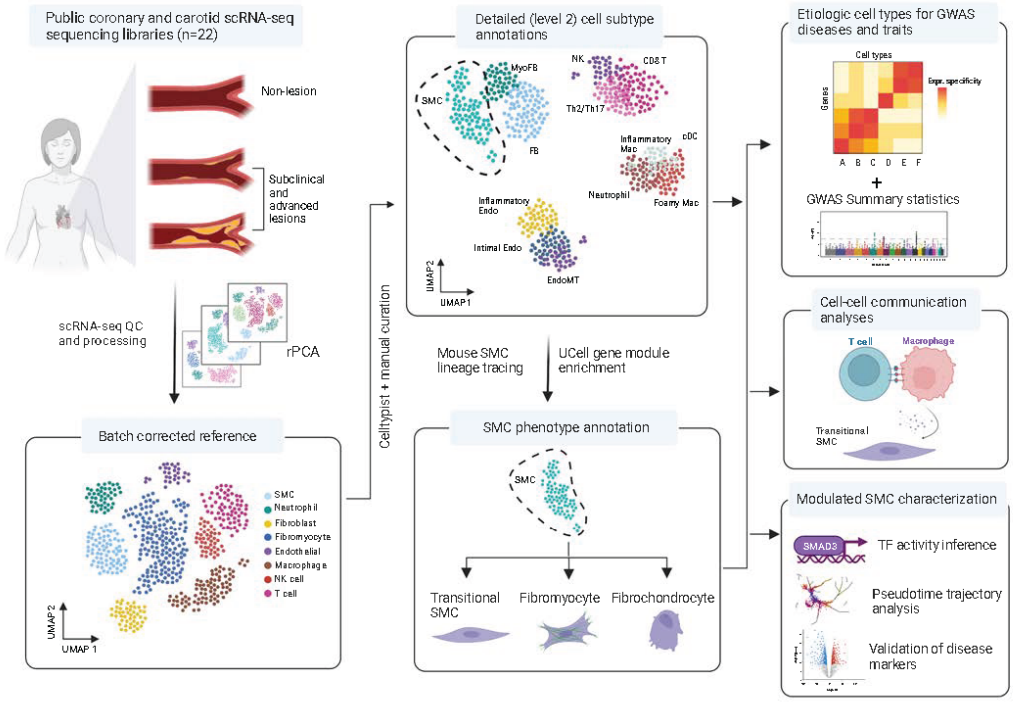


Fig. 1. General workflow of the study. Briefly, we collected human atherosclerosis scRNA-seq libraries across four publications. Three of them originated from atherosclerotic lesions of varying stages[16,18,19] while one harbored samples that had no CAD diagnosis or discernable lesions[17]. We devised a pipeline for rigorous QC and processing of each scRNA library and benchmarked four state-of-the-art batch-correction methods to find the integration approach that best fit the included data. We then used Transfer learning as well as machine learning classifiers and literature markers to define broad (level 1) and more granular (level 2) cell type annotations across vascular and immune lineages. We also leveraged murine lineage-traced smooth muscle cell (SMC) gene modules to identify modulated SMC populations in human data. In addition to cell communication analyses and integration of GWAS data for identification of etiologic cell types in disease, we further characterized modulated SMC phenotypes through pseudotime inference, TF activity predictions and identification of candidate novel human-specific gene markers. Details of scRNA-seq libraries QC and processing can be found in **Methods** and Supplementary figures.

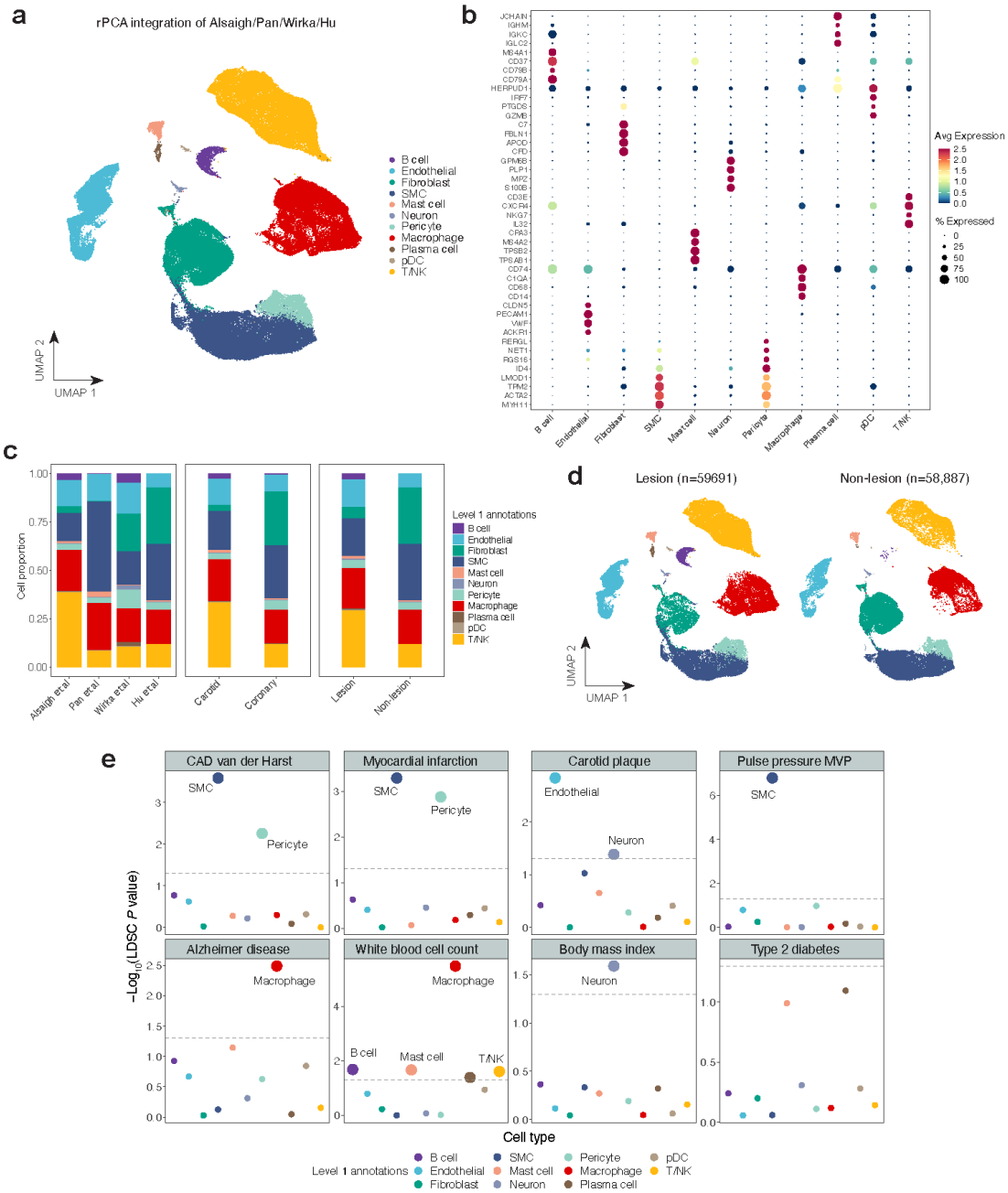


Fig. 2. Integration of single cell data conserves major cell compartments in atherosclerosis. (**a**) UMAP representation of 118,578 cells based on rPCA integration of 22 sequencing libraries. Dot colors depict broad cell lineage labels (level 1) defined through Transfer learning with the TS vasculature subset as annotation reference. (**b**) Dot plot of top five marker genes SCTransform-normalized expression by major cell lineage compartment. Dot size represents the portion of cells expressing the gene per level 1 compartment. (**c**) Stacked bar plot showing the distribution of level 1-annotated cells across included publications, arterial beds (coronary, carotids), and lesion status (lesion, non-lesion). (**d**) Distribution of level 1-annotated cells across lesion status in UMAP space. (**e**) Stratified LD Score Regression (S-LDSC)

analyses prioritizing the contribution of level 1-annotated cell type to cardiovascular and non-cardiovascular GWAS traits. LDSC analysis was carried out using a gene expression specificity matrix generated with CELLEX [42]. Large circles depict cells that passed the cutoff of FDR < 5% at -log10(*P*) = 1.301.

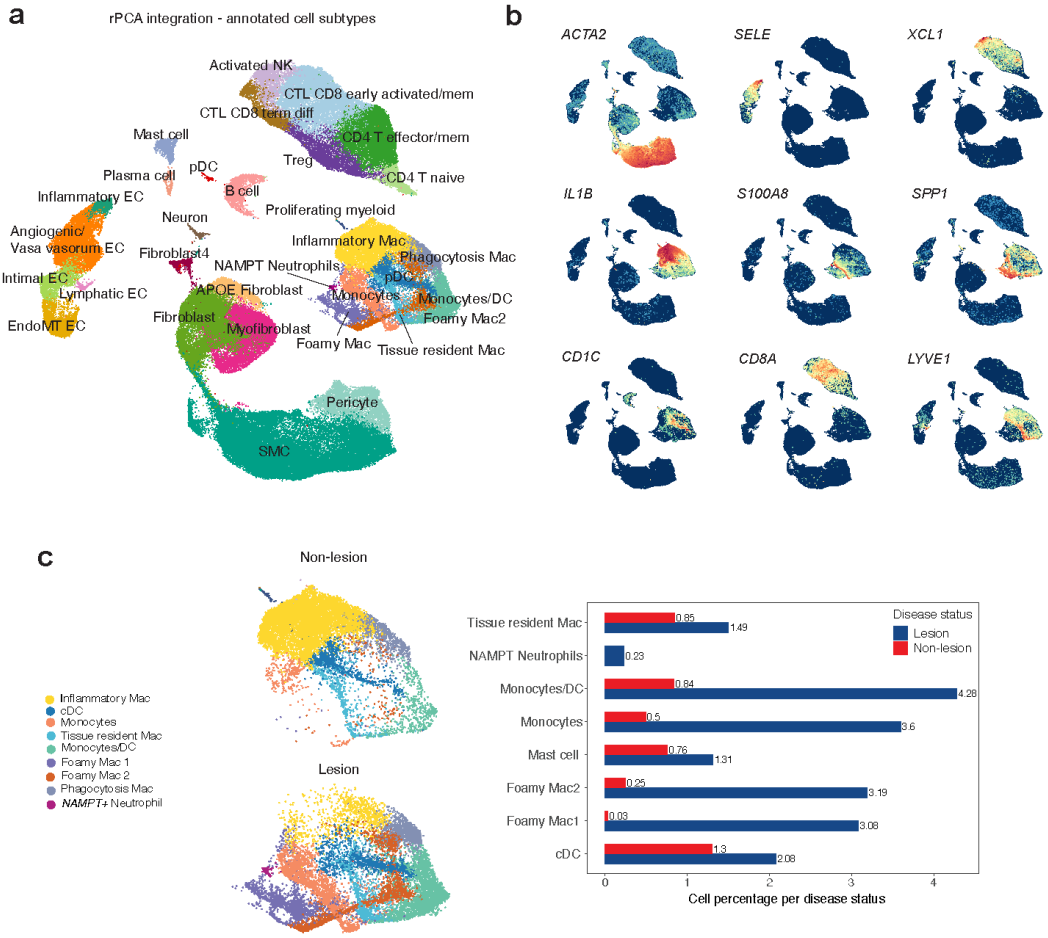


Fig. 3. Human atherosclerosis cell subpopulations (level 2) and distribution of myeloid subtypes across disease status. (**a**) UMAP representation of cell subtypes (level 2 labels) within the largest level 1 cell compartments (T/NK, Macrophages, Endothelial, Fibroblast). Level 2 labels were defined using a combination of the CellTypist classifier and survey of the literature. (**b**) UMAP plots of genes delineating immun and non-immune cell subtypes. SCTransform-normalized gene expression is indicated by color. (**c**) UMAP and bar plot of level 2 Myeloid cell subtypes according to lesion status. Frequencies for each subtype shown in the bar plot are normalized to the total number of cells in each condition (lesion n=59691; non-lesion n=58887) and shown as percentages.

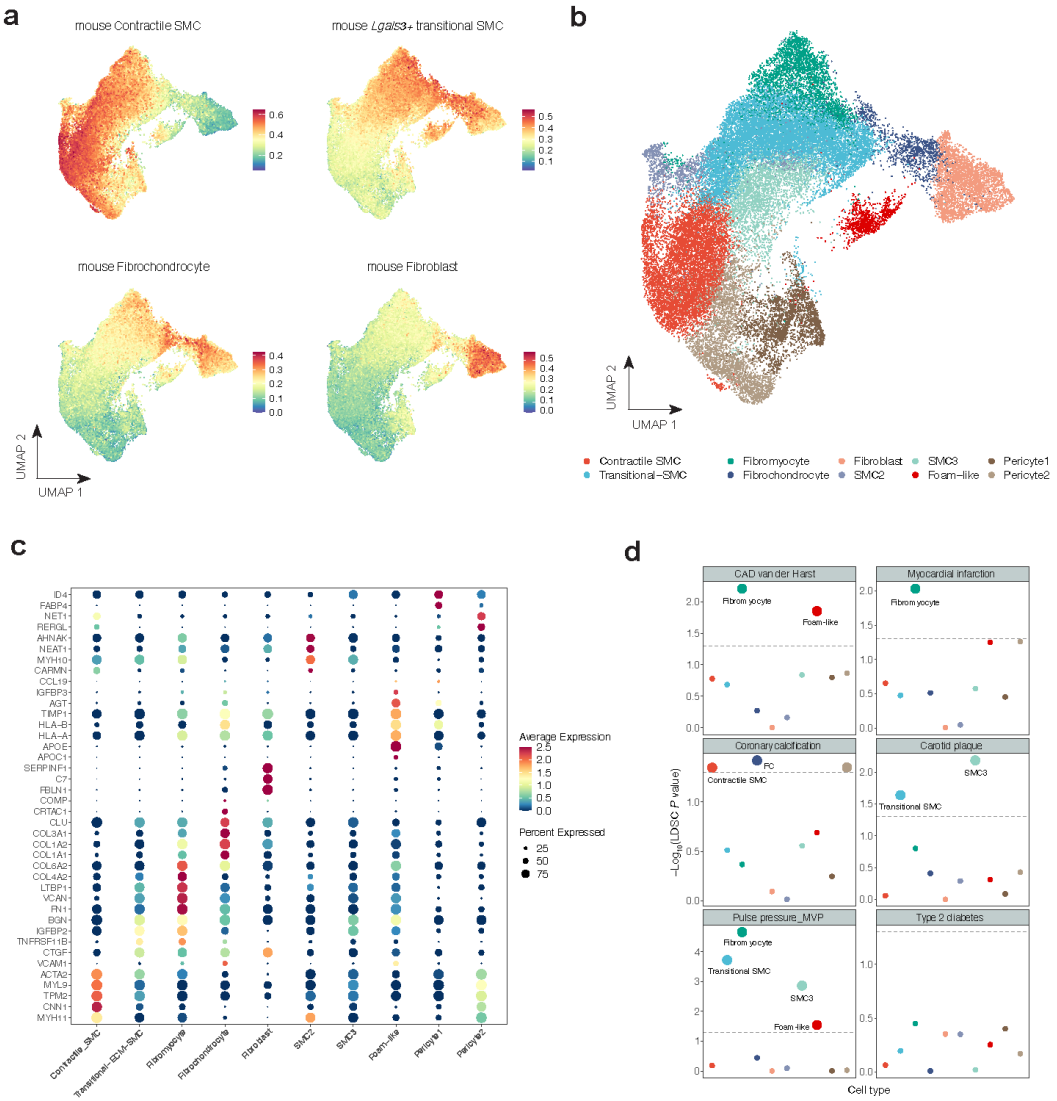


Fig. 4. Characterization of etiologic SMC phenotypes for cardiovascular traits and diseases. (**a**) UCell enrichment of meta-analyzed SMC murine gene modules (Contractile, Lgals3+ transitional, Fibrochondrocytes) and non-SMC-derived fibroblasts in the level 1 SMC compartment as well as a Fibroblasts. UCell[68] scores were calculated for each cell based on the Mann-Whitney U statistic where higher scores depict a higher enrichment for the tested gene signature. (**b**) UMAP embeddings of subclustered cells described in (**a**) SMC level 2 labels in addition pericytes and a subset of Fibroblasts. Labels were defined using UCell scores as reference for SMC differentiation state in addition to DE markers from Louvain clusters at a resolution=0.9. (**c**) Dot plot representing top marker genes SCTransform-normalized expression for SMC level 2 labels. Dot size represents the portion of cells expressing the gene. (**d**) Stratified LD score regression (S-LDSC) analyses prioritizing the contribution of SMC phenotypes, Pericytes and Fibroblasts to cardiovascular GWAS traits. Type 2 diabetes was used as a negative control in this analysis. LDSC was carried out using a gene expression specificity matrix for SMC clusters generated with CELLEX[42]. Large circles depict cells that passed the cutoff of FDR < 5% at -log10(*P*) = 1.301.

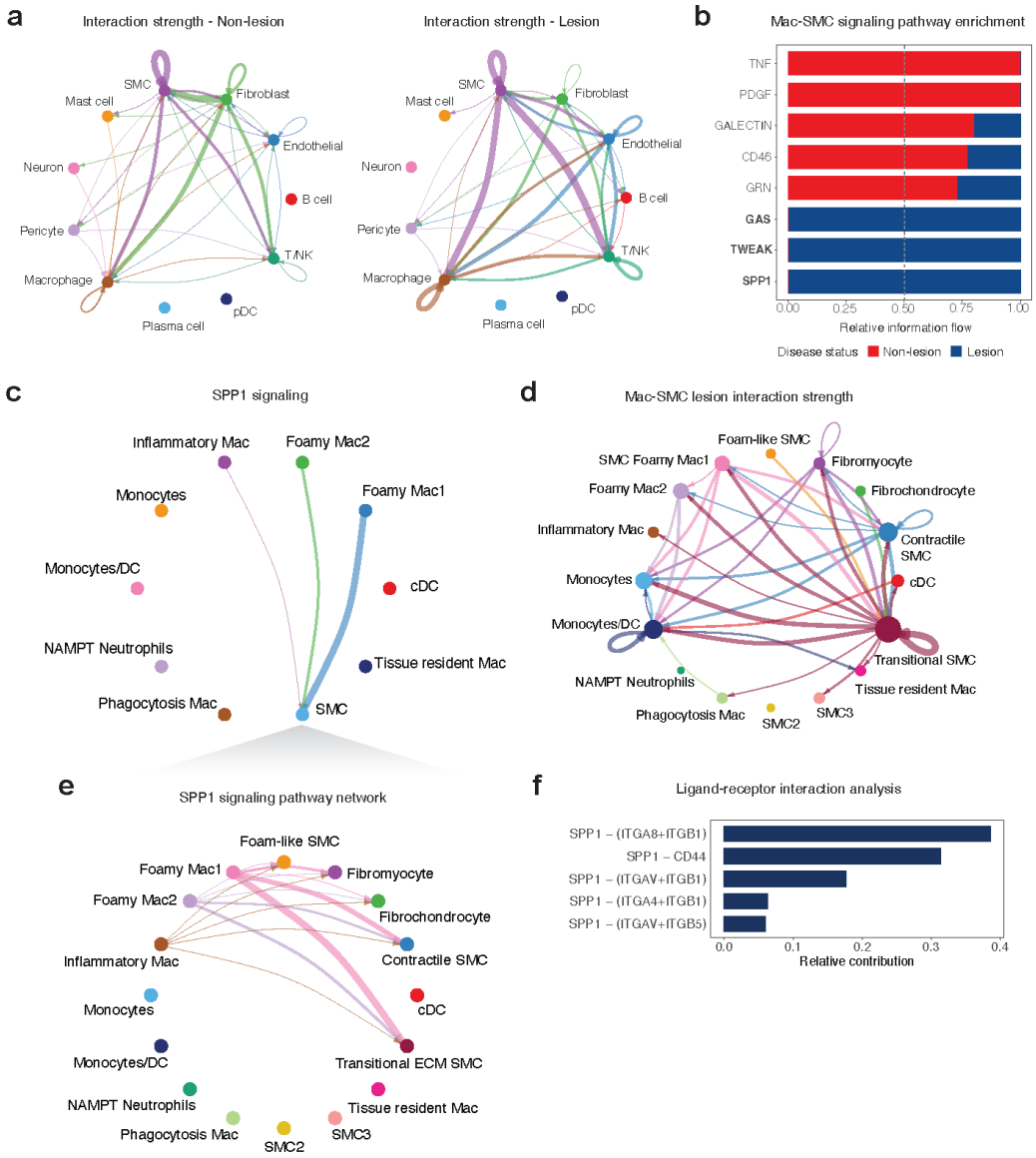


Fig. 5. Summary of cell crosstalk in human atherosclerosis. (**a**) Circle plots depicting aggregated cell-cell communication network for level 1-labeled cell compartments leveraging the CellChat[73] human database. Interactions considered include secreted signaling, ECM-receptor and cell-cell contacts. Interactions were calculated separately across disease status (lesion vs non-lesion). Top 30% of interactions are shown in the plot. (**b**) Stacked bar plot showing conserved and disease status-specific signaling pathways. Signaling enrichment is based on changes on pathways information flow (defined by the sum of communication probability among all pairs of cell groups in the inferred network or total weights in the network). Pathways in bold denote those that showed statistically significant ($P < 0.05$) enrichments in each disease condition. (**c**) Circle plot depicting sources and targets for SPP1 signaling using level 2 labels for myeloid cells and level 1 SMC labels. (**d**) Circle plot showing the aggregated cell-cell communication network for level 2 Myeloid and SMC labels. Top 15% of interactions are shown in the plot. (**e**) Circle plot depicting SPP1 signaling sources and targets for level 2

Myeloid and SMC labels. (**f**) Bar plot showing the relative contribution of each ligand-receptor pair for SPP1 signaling. Width of the edges in the circle plot depicts the weight/strength of the interactions in (**a**,**c**-**e**).

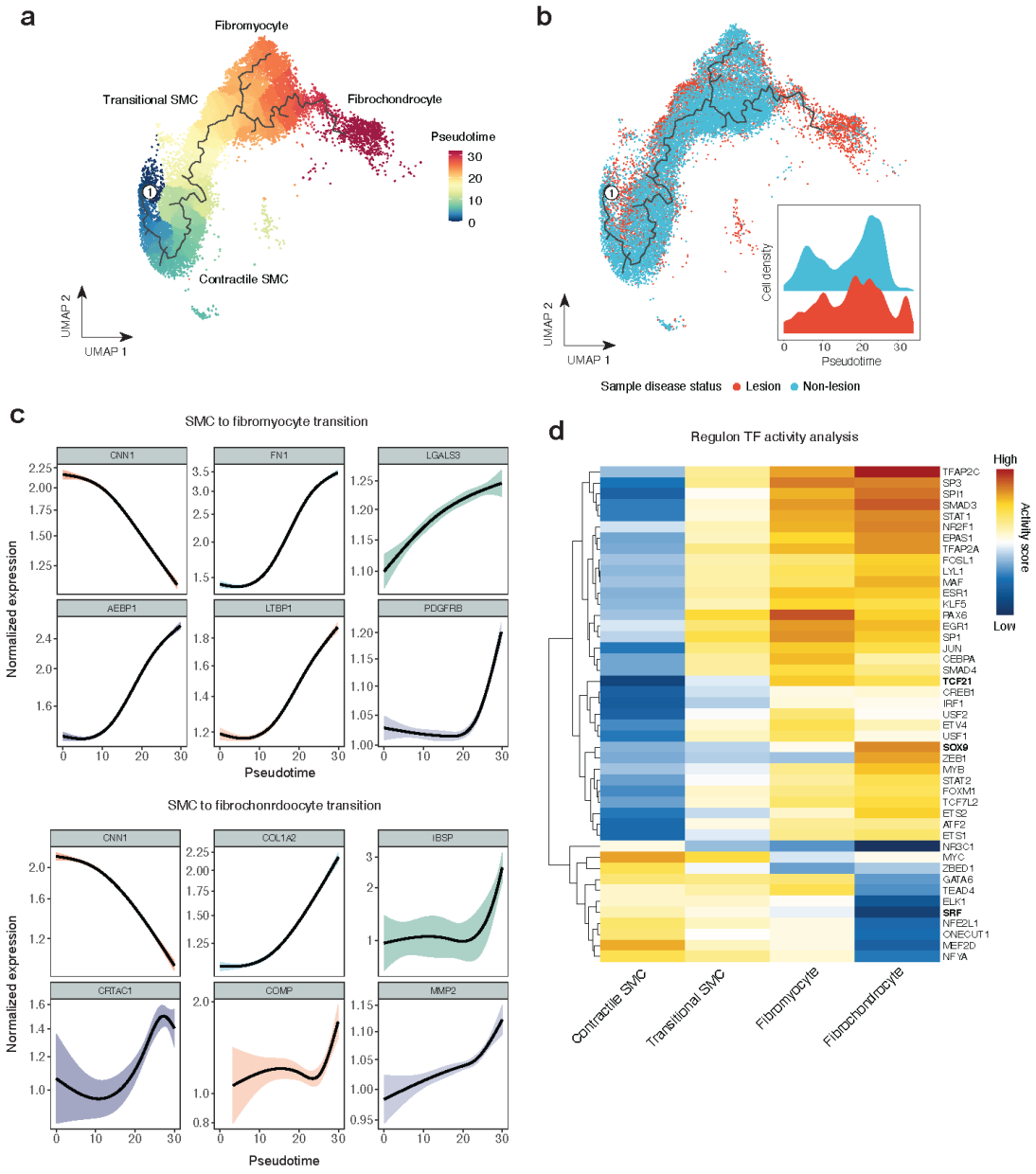


Fig. 6. Pseudotime and TF inference activity for ECM-rich SMC phenotypes. (**a**) UMAP embeddings showing supervised pseudotime trajectory from Contractile to modulated SMCs calculated with Monocle 3[74]. SMC phenotypes for this analysis included contractile, transitional SMCs, fibromyocytes and fibrochondrocytes (FCs). The numbered circle depicts the root of the trajectory, which was defined as the subset of Contractile SMCs with highest *MYH11* expression. (**b**) Pseudotime trajectory with SMCs grouped according to lesion status. Inset plot depicts the density of cells from lesions and non-lesion libraries across pseudotime. (**c**) Cubic spline interpolation of SCTransform-normalized gene expression as a function of pseudotime. Genes plotted include hits from Monocle 3 and Seurat DE tests (FDR < 0.05). DE genes from SMC to fibromyocyte trajectory: *FN1*, *LGALS3*, *AEBP1*, *LTBP1*, *PDGFRB.* DE genes from SMC to FC trajectory: *COL1A2*, *IBSP*, *CRTAC1*, *COMP*, *MMP2*. (**d**) Transcription factor (TF) activity

prediction with VIPER[78] based on DoRothEA[79] regulons for contractile and ECM-rich SMC phenotypes. Only regulons with confidence scores A-C (based on the number of supporting evidence) were used for this analysis. Highly variable TFs were selected for plotting and scale indicates relative predicted activity per TF.

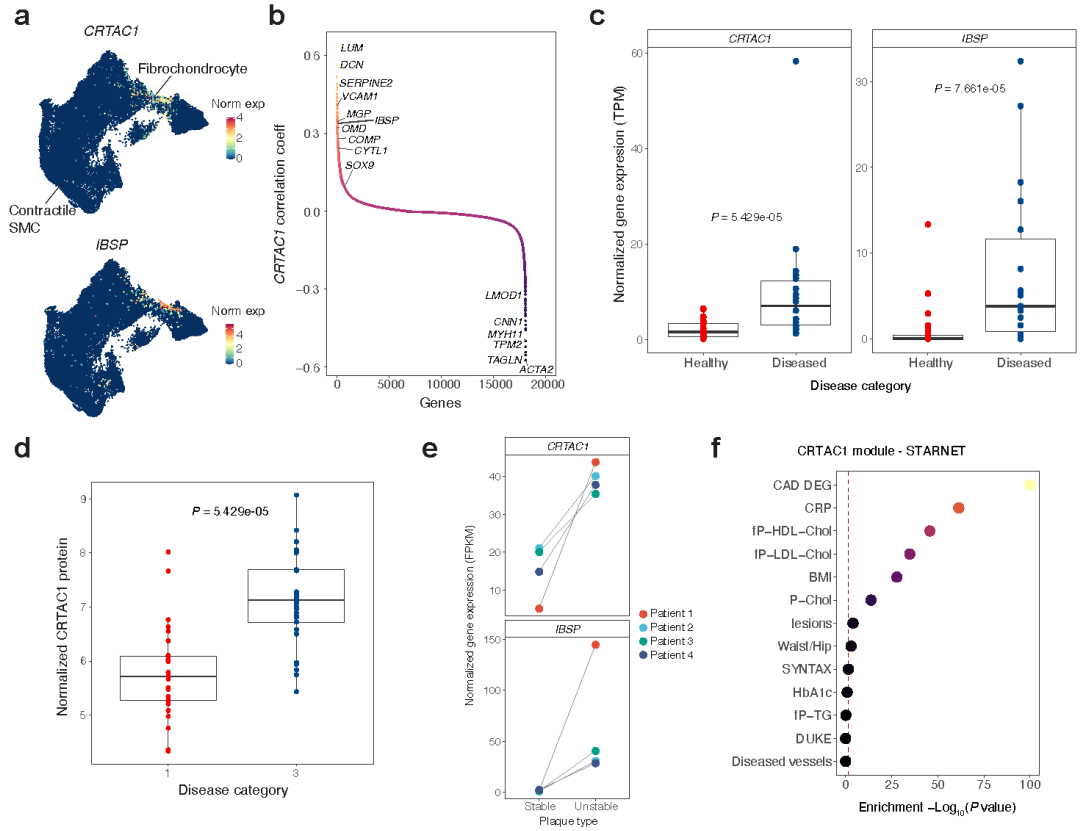
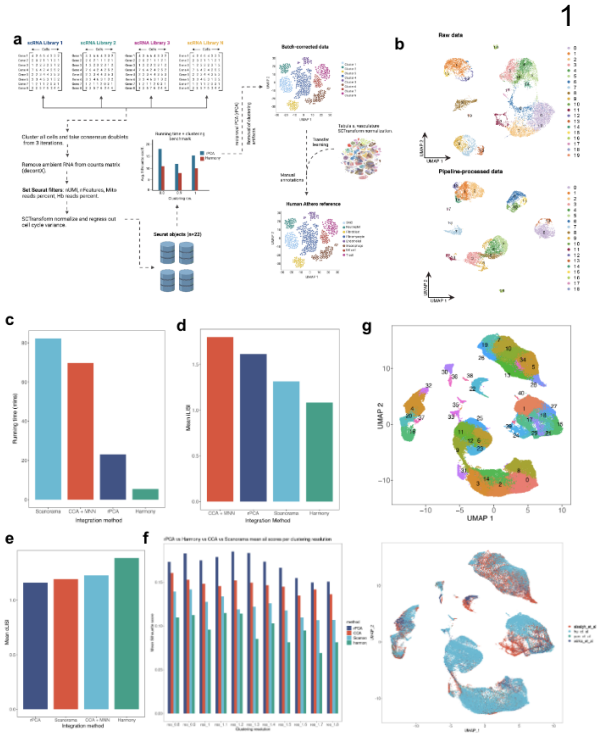
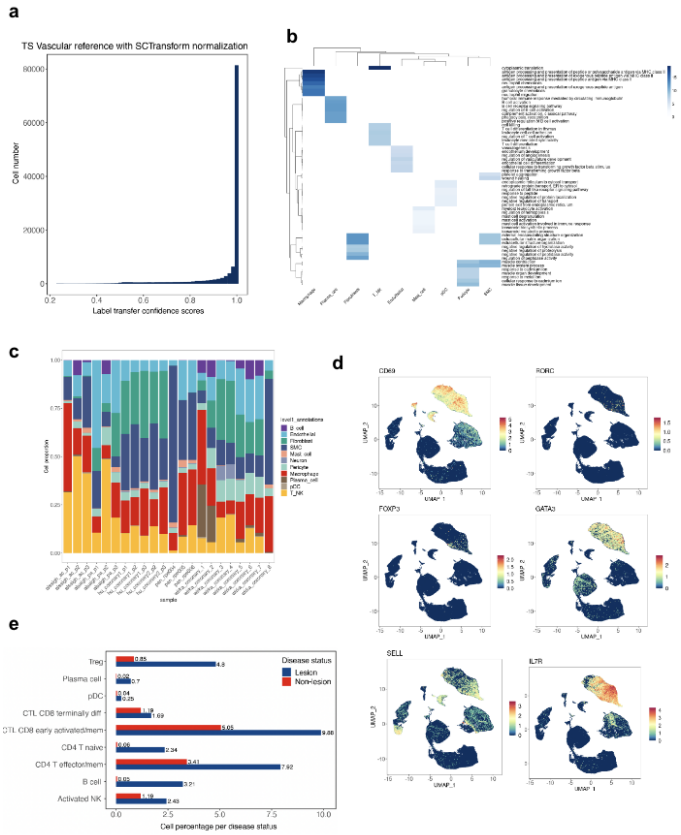


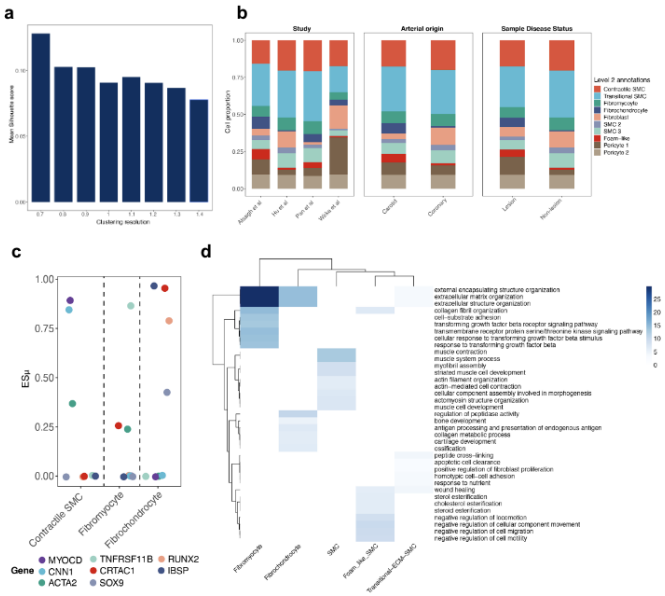
Fig. 7. CRTAC1 as a novel candidate marker of atherosclerosis progression. (**a**) UMAP embeddings for SCTransformed normalized expression of *CRTAC1*. UMAP of cells highly enriched for the murine FC gene signature is shown as reference for the location of human FCs. *IBSP* is used as a control marker of calcification. (**b**) Pearson correlation plot of *CRTAC1* with every other gene across SMCs and FC clusters. Selected examples of canonical contractile and ECM-related genes regulated during SMC modulation are shown. (**c**) Bulk RNA-seq expression of *CRTAC1* and *IBSP* in coronary arteries from healthy (n=27) and diseased samples (n=21). Data points represent normalized expression counts (TPMs). *P* values were calculated using a non-parametric Wilcoxon rank sum test. (**d**) Log-normalized protein expression of CRTAC1 in category 1 (n=27) and 3 (n=29) disease samples. For details in definition of disease category **(Methods)**. *P* value was calculated using a parametric unpaired Student's T-test. Boxplots in (**c**) and (**d**) represent the median and the inter-quartile (IQR) range with upper (75%) and lower (25%) quartiles shown, and each dot represents a separate individual. (**e**) Dot plot showing normalized expression (FPKMs) from a public RNA-seq dataset of human fresh carotid lesions. Dots of the same color represent matched patient (n=4) samples of stable and unstable plaque regions (stable, n=4; unstable, n=4). (**f**) Clinical trait enrichment for *CRTAC1*-containing module in a subclinical mammary artery in STARNET gene regulatory network datasets. Pearson's correlation *P* values (gene-level) were aggregated for each co-expression module using a two-sided Fisher's exact test. Case/control differential gene expression (DEG) enrichment was estimated by a hypergeometric test.



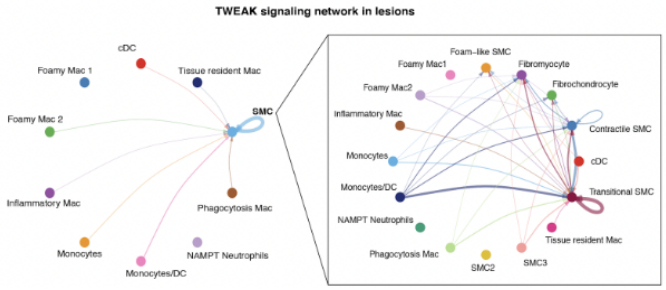
Supplementary Fig. 1. scRNA processing pipeline and integration benchmark metrics. (**a**) Workflow for standardizing processing of each scRNA-seq library and integration. Doublets were first identified and removed using scDblFinder[105]. Upon doublet removal, ambient RNA was removed using decontX[25]. The decontaminated matrix was then used for downstream filtering of cells based on 1) number of detected genes 2) number of UMIs 3) percentage of reads mapping to mitochondrial genome 4) percentage of reads mapping to hemoglobin genes using Seurat [26]. Libraries were normalized using SCTransform[27] integrated using four approaches (CCA + MNN, rPCA [26], Harmony [31], Scanorama [32]). PCA embeddings from each approach were then used for measuring LISI scores and silhouette coefficients. Finally, rPCA was used to harmonize the 22 included sequencing libraries and level 1 labels were added using Transfer learning with the Tabula Sapiens Vasculature subset as reference. (**b**) UMAP embeddings of a representative library before and after going through the scRNA-seq processing pipeline. (**c**) Running time of each of the four integration approaches tested. The Y axis shows time in minutes. (**d**) Mean integration LISI (iLISI) scores calculated for each integration approach. Higher iLISI scores depict improved mixing and batch removal. (**e**) Mean cell type LISI (cLISI) calculated for each integration approach. Lower cLISI scores represent increased biological conservation. (**f**) Mean silhouette coefficients calculated for each integration approach. Silhouette coefficients were calculated using euclidean distances across a range of clustering resolutions (0.8-1.8) to determine optimal clustering resolutions. Silhouette scores range from (-1, 1) where higher scores depict improved clustering quality or purity. PCA embeddings (30 PCs) were used for calculation of metrics in (**d**-**g**). Visual inspection of batch removal through UMAPs for the 41 Louvain clusters generated after integration with rPCA and cells grouped by the “Study” variable. For additional details on processing and benchmark see **Methods**.



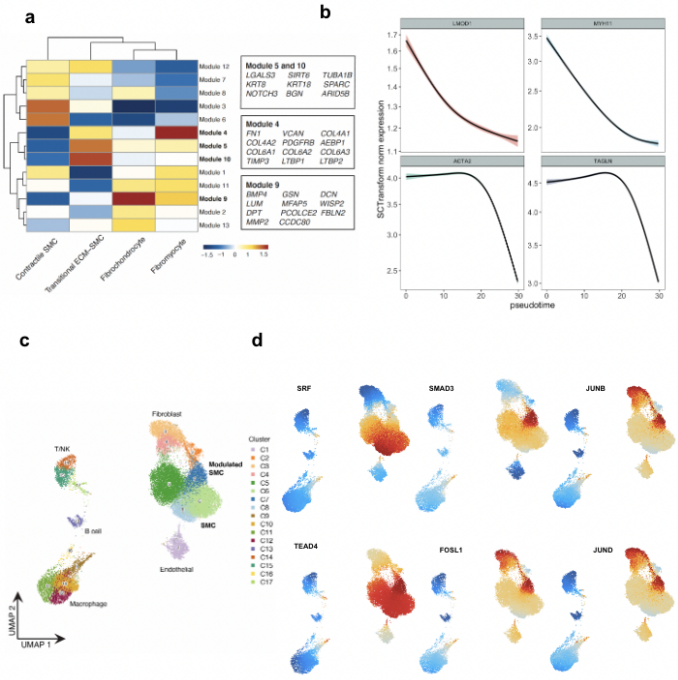
Supplementary Fig. 2. Additional characterization of level 1 and level 2 cell type annotations. (**a**) Confidence scores from predicted labels using the Seurat Transfer learning classifier with the TS vasculature reference. Confidence scores range from 0-1 where higher scores refer to unambiguous calls. (**b**) Gene set enrichment analysis (GSEA) for level 1 annotated cell types. This analysis was carried out with gProfiler2[106] and the top seven significantly enriched terms (FDR <0.05) were selected for plotting. (**c**) Bar plot showing the distribution of level 1 annotated cell types across the 22 sequencing libraries included in this study. (**d**) UMAPs of SCTransform-normalized expression of genes defining key T cell states and subtypes (*CD69*: early activation; *RORC*: Th_{17} cells; *GATA3*: Th_2 cells; *SELL* and *IL7R*: memory/naive T cells. (**e**) UMAP and bar plot of level 2 Lymphoid cell subtypes according to lesion status. Frequencies for each subtype shown in the bar plot are normalized to the total number of cells in each condition (lesion n=59691; non-lesion n=58887) and shown as percentages.



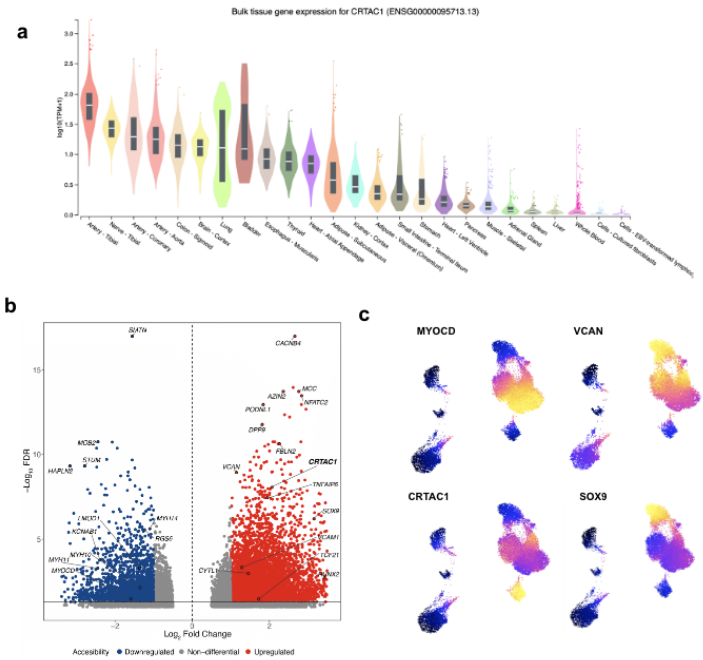
Supplementary Fig. 3. Additional characterization of SMC subtypes. (**a**) Barplot depicting silhouette analysis using PCA embeddings for reference subset (SMCs, Pericytes and Fibroblasts). As before, silhouette coefficients were calculated across a range of resolutions to find optimal parameters for subclustering. Silhouette scores range from (-1, 1) where higher scores depict improved clustering quality or purity. (**b**) Stacked bar plot showing the distribution of level 2 annotations for SMCs (as well as Pericytes and Fibroblasts included in the subclustering step across studies, arterial beds (Coronary or Carotid) and sample disease status (lesion and non-lesion). (**c**) Dot plot showing ES_{μ} values for canonical contractile markers (*MYOCD*, *CNN1*, *ACTA2*) as well as synthetic (*TNFRSF11B*) and osteochondrogenic markers (*RUNX2*, *SOX9*, *IBSP*) of SMC modulation for Contractile and ECM-rich SMC phenotypes (Fibromyocytes and FCs). ES_{μ} values were plotted from a gene expression specificity matrix generated with CELLEX[42]. For additional details on ES_{μ} values see **Methods**. (**d**) Gene set enrichment analysis (GSEA) for level 2 annotated SMCs. This analysis was carried out with gProfiler2[106] and the top nine significantly enriched terms (FDR < 0.05) were selected for plotting.



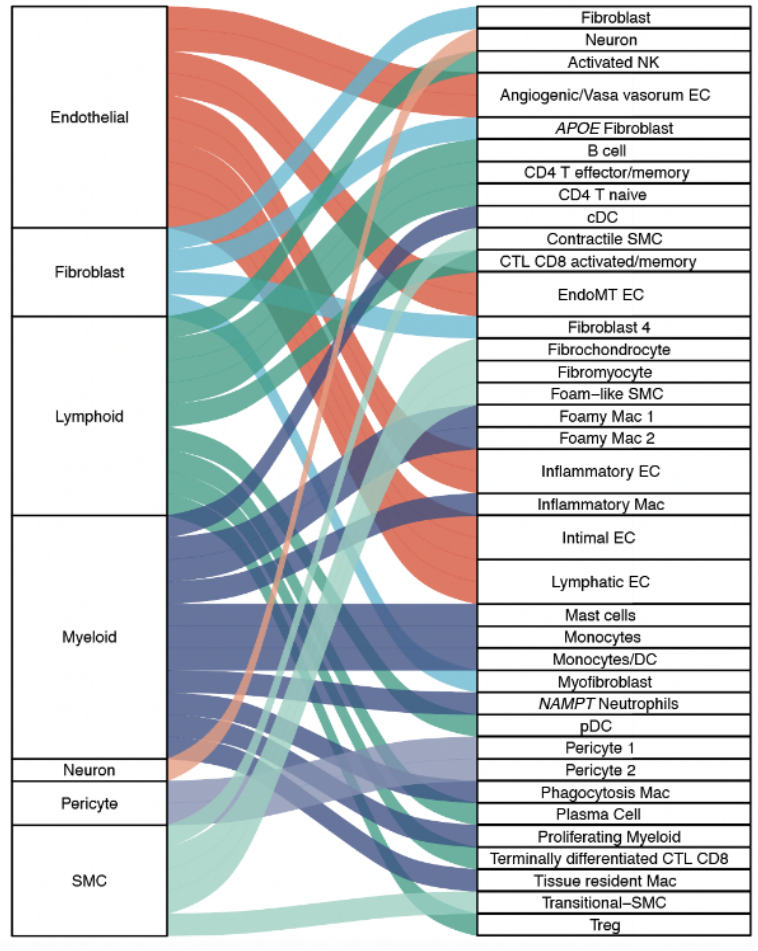
Supplementary Fig. 4. TWEAK signaling for myeloid and SMC cell types. Circle plot depicting sources and targets for TWEAK signaling using level 2 labels for myeloid cells and level 1 SMC labels. Circle plot within the square depicts TWEAK signaling using level 2 labels for myeloid and SMC labels. Width of the edges depicts weight or strength of the interaction based on calculated communication probability between a pair of cell types. Interactions were calculated using the CellChat[73] human database.



Supplementary Fig. 5. Differential gene expression through pseudotime and snATAC-seq TF activity inference. (**a**) Heatmap of varying gene module expression as Contractile SMCs transition into ECM-rich phenotypes (fibromyocytes and FCs). Differential genes across pseudotime were calculated using graph autocorrelation analysis with Monocle 3 and then grouped into modules using Louvain community analysis. Color scale represents aggregated expression of genes in each module across the above-mentioned SMC phenotypes. Boxes (right) list key genes found in each module. (**b**) Cubic spline interpolation of SCTransform-normalized expression of canonical contractile markers (*LMOD1*, *MYH11*, *ACTA2*, *TAGLN*) as a function of pseudotime. (**c)** UMAP and Louvain clustering of coronary arteries snATAC-seq data. Each dot represents an individual cell colored by cluster assignment. Cell type labels in bold represent Contractile and ECM-rich modulated SMC populations as defined in *Turner et al* [80]. (**d**) UMAPs of ChromVAR TF motif accessibility deviation scores for factors shown as highly variable in previous SMC analysis with DoRothEA regulons (SRF, TEAD4, SMAD3, FOSL1, JUN).



Supplementary Fig. 6. External validation for CRTAC1 activity in arteries and SMCs. (**a**) Violin plot showing Expression expression across GTEx tissues (sorted according to normalized expression). The Y axis shows log-normalized expression values (TPMs) and violin plot shows median as well as inter-quartile (IQR) range with upper (75%) and lower (25%) quartiles. Dots represent outliers. (**b**) Volcano plot of differential accessibility analysis comparing ECM-rich Modulated to Contractile SMCs. Analysis was carried out using a Wilcoxon test as implemented in ArchR[81]. Peaks with significant differences at FDR ≤ 0.05 and $\log_2$ fold change > 1 were colored red (Modulated SMC upregulated) and blue (Modulated SMC downregulated). Each dot represents a differentially accesible region. Regions were annotated with the nearest protein coding genes using GenomicRanges[107] and key contractile and modulated SMC genes are shown. (**c**) UMAP plots of snATAC-seq cells colored according to accessibility-derived gene scores (**Methods**) for canonical contractile factors (MYOCD) as well as modulated SMC ECM and osteochondrogenic markers (*VCAN*, *SOX9*, *CRTAC1*). *CRTAC1* gene scores are elevated within a position of the Modulated SMC cluster and overlap with higher SOX9 activity, suggesting a chondrogenic transition.



Supplementary Fig. 7. Summary of cell type diversity in human atherosclerosis. Riverplot depicting the relationship between level 1 cell compartments and level 2 cell subtypes for vascular and immune lineages. This plot was generated using the ggalluvial R package.

Methods

Ethics statement

Details regarding data collection for scRNA-samples included in this meta-analysis can be found in each publication[16–19]. Collection of coronary artery samples for bulk RNA-seq and proteomics data generation described in this manuscript complies with ethical guidelines for human subjects research under approved Institutional Review Board (IRB) protocols at Stanford University (no. 4237 and no. 11925) and the University of Virginia (no. 20008), for the procurement and use of human tissues and information, respectively.

QC and normalization of scRNA-seq sequencing libraries

Raw count matrices from each library across the 4 studies were downloaded from GEO and Zenodo **(Data Availability, Supplementary Table 1)**. Processing for each of the 22 sequencing libraries was standardized in the following manner: Each library was loaded into the R programming environment (v.4.0.3) using Seurat[26] (v.4.1.0). For each library we did a first pass of clustering with SCTransform normalization[27] without removing low-quality cells.

In order to remove doublets, we referred to a recent benchmark of doublet-removal tools[108] and chose the scDblFinder R package[105] (v.1.4.0) given its superior accuracy compared to other tools. Seurat objects for each library were converted to SingleCellExperiment objects and used as input to generate artificial doublets using the cluster-based modality of scDblFinder. Briefly, scDblFinder creates a K-Nearest Neighbors graph using the union of real cells and artificial doublets and estimates the density of artificial doublets in the neighborhood of each cell. Since artificial-doublet generation approaches tend to display slight variance across different runs, we only kept consensus doublets from 3 iterations of the above-described process. Cell-barcodes that were marked as doublets were then removed from each raw counts matrix.

Ambient RNA contamination is a key issue during 10x protocols and can negatively impact clustering and extraction of gene markers. To filter out reads from ambient RNA, we ran DecontX[25] within the celda R package (v1.6.1) in doublets-filtered raw counts matrices using default parameters. The decontaminated raw count matrices output by DecontX were then added into each Seurat object. We then set quality filters to keep cells that had 1) >= 200 and <= 4000 uniquely expressed genes 2) >= 200 and <= 20000 UMIs 3) <= 10% of reads mapped to the mitochondrial genome; cells with high percentages of reads mapped to mitochondrial genomes are considered to be low quality as this indicates cell membrane breaches and 4) <= 5% of reads mapped to hemoglobin genes since these cells likely depict contaminating erythrocytes as done in *Alencar et al*.

Raw count matrices were then normalized using SCTransform[27] with parameter (vst.flavor=v2), which accounts for sequencing depth variability across cells. This omits the need for heuristic steps such as log-transformation and it has been shown to improve variable gene selection, dimensionality reduction and differential expression[27]. To avoid clustering results confounded by

cell cycle state, cell cycle variance was regressed out during SCTransform normalization. We then carried out dimensionality reduction of the normalized counts matrix using PCA. The first 30 principal components (PCs) were used as input for clustering in Seurat, which relies on a Shared-Nearest-Neighbors (SNN) and Louvain community detection approach. We then applied Uniform Manifold Approximation and Projection (UMAP) non-linear dimensionality reduction using the first 30 PCs. UMAP embeddings were used for visualization of Louvain clustering results. Processed matrices were then stored as seurat objects for batch-correction.

Integration benchmarking and building the reference

In order to harmonize processed sequencing libraries, we selected the following methods recommended from three recent benchmarks[28,29,109] of single-cell transcriptomic data integration: Canonical Correlation Analysis + Mutual Nearest Neighbors (CCA + MNN), reciprocal PCA (rPCA)[26] (v.4.1.0), Harmony[31] (v.1.0) and Scanorama[32] (v.1.7.1). We focused on four different metrics to choose a method: running time, efficiency of batch effect removal as denoted by the integrative Inverse Local Simpson Index (iLISI), conservation of biological variation using the “cell type” LISI (cLISI) and clustering purity measured by silhouette coefficients. The silhouette score provides a measure of how well each cell has been classified by measuring how similar it is to its own cluster (cohesion) compared to other clusters (separation). For the benchmark, we used a subset of the data including 3 studies: *Wirka et al*, *Alsaigh et al* and *Pan et al*. Libraries from these studies were integrated as follows:

CCA + MNN: we created a list of selected Seurat objects and then selected 3000 highly variable genes. Integration with those variable genes was done using the PrepSCTIntegration(), FindIntegrationAnchors() and IntegrateData() functions. The batch-corrected expression matrix was then used for PCA dimensionality reduction, creation of the shared-nearest-neighbors (SNN) graph using 30 PCs and visualization with UMAP embeddings.

Harmony: libraries were first stored into a list and highly variable genes extracted using the function SelectIntegrationFeatures(). Libraries were merged into a single seurat object, and the list of highly variable genes was used for PCA dimensionality reduction. We used the first 30 PCs as input for RunHarmony() from the harmony package (v1.0), setting sequencing libraries (sample column in metadata) as the variables to correct for batch effects. Harmony embeddings were used for subsequent generation of the SNN graph, Louvain clustering and visualization with UMAP by setting reduction=”Harmony” within the FindNeighbors() and RunUMAP() Seurat functions and using the first 30 PCs.

rPCA: we created a list of processed Seurat objects and extracted the 3000 most highly variable genes using SelectIntegrationFeatures(). We then ran PCA across each library using the 3000 variable genes, identified integration anchors using FindIntegrationAnchors() setting reduction=”rpca” and harmonized datasets using IntegrateData(). As done for CCA, the batch-corrected expression matrix was then used for PCA dimensionality reduction, creation of the shared-nearest-neighbors (SNN) graph using 30 PCs and Louvain clustering followed by visualization with UMAP embeddings.

Scanorama: We used the reticulate R package (v.1.18) to import the Scanorama python module (v.1.7.1) into the R environment. We created a list with seurat objects containing the datasets to be integrated and stored normalized SCTransform-normalized counts and gene names for each dataset into a new list. We then batch corrected the data using the function using the correct() function from the Scanorama package setting the following parameters (return_dimred=TRUE and reurn_dense=TRUE). The batch-corrected expression matrix output by correct() was used to create a new Seurat object and Scanorama-produced dimensionality reduced embeddings were inserted into the Seurat object using the CreateDimReducObject() function. Scanorama embeddings were subsequently used to create a shared-nearest-neighbors (SNN) graph for Louvain clustering and for visulaization with UMAP using the first 30 PCs.

Running time measurements:Running times for each integration task were then measured using base R Sys.time() functions. Sys.time() was defined at the beginning and the end of each integration task and then the time difference was calculated as end_time - start_time.

Silhouette analysis: Here we measured the quality or “goodness” of resulting clusters using the silhouette coefficient. For silhouette analyses, we extracted PCA embeddings from seurat objects with CCA+MNN, rPCA, Harmony and Scanorama integration outputs keeping the first 30 PCs. We then used these embeddings to compute an Euclidean distance matrix. Cluster IDs for each cell were obtained iteratively across a range of clustering resolutions (0.8-1.8) and Euclidean distance matrices were used to calculate silhouette width values using the cluster R package (v.2.1.0). The purpose of using the above range was to control for the clustering granularity parameter and to identify a range of clustering resolutions that would not lead to over- or under-clustering of the data.

Calculation of LISI scores: Briefly, iLISI scores are a measure of the diversity within each cell neighborhood on a K-nearest-neighbor (KNN) graph. Higher iLISI scores depict increased mixing of batches within a cell neighborhood and therefore suggest improved removal of batch effects. For each of the integration methods described above we extracted PCA embeddings (30 first PCs) from the corresponding integrated Seurat object. We then created a data frame with each row corresponding to one cell and columns depicting batch variables (“Study”). We then computed iLISI scores for each cell using the compute_lisi() function from the lisi R package[31] (v.1.0). Mean iLISI values were plotted and compared across different integration methods. cLISI scores, in turn, are considered a metric that measures conservation of biological variation. With the assumption that each cluster should generally harbor cells from the same type, we created a dataframe with each row corresponding to each cell and a column depicting Louvain cluster identities. cLISI scores for each cell were calculated and plotted as described above.

Integration of scRNA libraries and additional quality control: Upon determining the appropriate integration approach for the datasets of interest, we used rPCA to harmonize the 22 sequencing libraries as described in the above section. Upon integrating libraries, we reduced dimensionality of the data using PCA. A SNN graph was constructed using 50 nearest

neighbors and the first 30 PCs as input. Clusters were identified using the above graph with a resolution of 1, which was within the range of higher mean silhouette coefficients from the previous benchmark. Gene markers for each cluster were identified using PrepSCTMarkers() and the Wilcoxon Rank Sum test as implemented in the FindAllMarkers() function from Seurat (v.4.0). We considered genes that were expressed in at least 25% of the clusters being compared (one cluster vs all others) and that had a logfc.threshold=0.25. Genes fulfilling that criteria in addition to having multiple-testing adjusted P-values <= 0.05 were considered as differential cluster markers. Upon inspection of the gene signatures of each cluster, we found two small clusters comprising 432 cells (0.36% of cells in the integrated reference) expressing markers of multiple major lineages, which likely represent residual doublets and thus were removed from the reference. Upon removing these residual doublets, cells were re-clustered using the above-described parameters. Finally, inspection of cluster markers expression in UMAP space allowed us to identify and remove UMAP artifacts (e.g., cells with Natural Killer signatures within candidate fibroblast clusters). The 306 cells (0.25% of cells in the reference) comprising these artifacts were removed to obtain the final iteration of the reference. This step was necessary to ensure robustness of cell type and subtype annotations as well as other downstream analyses.

Cell type annotations

To annotate cell types in the integrated reference, we used a systematic approach to define broad labels (level 1) as well as more granular cell subtype labels (level 2).

Level 1 annotations: To define broad cell type partitions, we accessed public data from the Tabula Sapiens (TS) consortium (https://tabula-sapiens-portal.ds.czbiohub.org/organs). To improve the specificity of annotations, we downloaded the vasculature subset of this transcriptomic atlas. Upon downloading the TS vaculature h5ad file, this dataset was converted into a Seurat-compatible format using the SeuratDisk R package (v.0.0.0.9019). To match the normalization workflow described in the scRNA sequencing library processing section, we extracted the TS vasculature raw counts matrix and normalized gene expression data using SCTransform. We then applied Seurat's reference-based transfer learning (using FindIntegrationAnchors() and TransferData() to annotate cells in our meta-analyzed reference. In this case, the TS vasculature seurat object with author-provided cell type annotations was defined as reference for label transfer. Confidence scores of predicted labels ranging from 0-1 (where 1 indicates that labels were annotated in a fully unambiguous manner) were extracted from the output of TransferData() and are shown in the Data Supplement. Gene markers for level 1 annotations were obtained using the PrepSCTMarkers() and FindAllMarkers() functions from Seurat (v.4.1.0) setting the following thresholds: logFC=0.25 and min.pct=0.25

Level 2 annotations for endothelial, fibroblasts and immune cells: To define more granular cell subtypes for the meta-analyzed data, we used a combination of automated and manual annotations. We first annotated cell subtypes for endothelial, myeloid and lymphoid lineages using markers from atherosclerosis murine scRNA meta-analyses of SMCs and immune cells as well as relevant human atherosclerosis scRNA studies[21–23,44–49]. Annotations using curated

markers from the literature were corroborated with the assistance of experts at UVA. To further confirm and inspect immune cell subtype annotations in our reference, we used a logistic-regression with stochastic gradient-descent framework implemented by the command-line tool CellTypist [50]. CellTypist leverages a database of 20 different tissues and 19 reference datasets with a focus on myeloid and lymphoid cells. Specifically, we applied CellTypist low-hierarchy classifiers (using the Immune_All_Low.pkl and Immune_All_AddPIP.pkl models which harbor 90 and 101 cell types, respectively) to our SCT-normalized reference counts matrix using both default settings as well as the majority voting classifier. Gene markers for level 2 annotations were obtained using the PrepSCTMarkers() and FindAllMarkers() functions from Seurat (v.4.1.0) setting the following thresholds: logFC=0.25 and min.pct=0.25

Level 2 annotations for SMCs: To explore SMC diversity in human atherosclerosis, we subset the main meta-analyzed reference to include only the pericyte-SMC-fibroblast level1 partitions. This subset was then reclustered using Seurat (v.4.0) with a resolution of 0.9 based on an additional silhouette width benchmark. Next, gene modules (encompassing top markers from differential expression analyses) specific to contractile (n=50), *Lgals3*+ pioneer (n=50), and fibrochondrocyte (n=50) SMC phenotypes were extracted from a recent SMC lineage-traced murine scRNA meta-analysis. We also extracted a non-SMC-derived fibroblast module (n=50) as a negative enrichment control. Genes in each module were ranked by Log2FC and then converted to human homologs nomenclature and filtered to keep those with a one-to-one orthology relationship using custom wrapper functions with the biomaRt R package[110] (v.2.46). We then calculated the enrichment of murine gene modules on individual cells within the pericyte-SMC-Fibroblast human subset using the UCell R package[68] (v1.3.1) . In addition to the enrichment of murine gene modules, we also obtained gene markers for each of the 17 SNN-derived clusters using the PrepSCTMarkers() and FindAllMarkers() functions from Seurat (v.4.1.0) setting the following thresholds: logFC=0.25 and min.pct=0.1. Final annotations for SMC subtypes were derived based on the UCell enrichment scores along UMAP coordinates and cluster markers.

LD score regression analyses

LDSC for SMC level 1 cell type annotations: Integration of scRNA and GWAS summary statistics was performed using the LDSC wrapper within the CELLECT python pipeline[42]. First, we created a gene expression specificity matrix for level 1 annotations using the SCTransform-normalized expression matrix as input for the CELLEX python pipeline[42]. Shortly, gene expression specificity values (ES_μ) output by CELLEX are derived using four different expression specificity metrics (Differential expression T-statistic, Gene enrichment score, Expression proportion and Normalized specificity index) and they represent a score that a gene is specifically expressed on a given cell type (level 1 annotation).

We downloaded GWAS summary statistics for: CAD (van der Harst et al)[36]; Myocardial infarction[37]; carotid intima-media thickness[38], carotid artery plaques[38], diastolic blood pressure, systolic blood pressure and pulse pressure from the Million Veterans Program [39]; Alzheimer disease[40]; type 2 diabetes (UK Biobank)[41]; body mass index (UK Biobank)[41]; White blood cell

count (UK Biobank)[41]. UK Biobank summary statistics were downloaded from https://alkesgroup.broadinstitute.org/UKBB/.

We used custom R scripts (https://github.com/MillerLab-CPHG/Human_athero_scRNA_meta) as well as the provided mtag_munge.py python script (https://github.com/pascaltimshel/ldsc/tree/d869cfd1e9fe1abc03b65c00b8a672bd530d0617) to convert GWAS summary statistics to a format compatible with that of the CELLECT S-LDSC wrapper. We then performed S-LDSC with the gene expression specificity matrix for level 1 annotations across the above described GWAS studies using the established CELLECT snakemake workflow as shown in https://github.com/perslab/CELLECT/wiki/CELLECT-LDSC-Tutorial.

LDSC for SMC level 2 cell type annotations: We proceeded to subset the whole meta-analyzed reference Seurat object to include only cells along the pericyte-SMC-Fibroblast partitions. Metadata of this subset were used to generate the gene expression specificity matrix for level 2 annotations. In addition to GWAS studies described above, we also included summary statistics from our recent Coronary Artery Calcification (CAC) multi-ancestry GWAS meta-analysis[72]. Munging of GWAS summary statistics and subsequent S-LDSC analyses were performed as described above.

Cell communication analyses

Cell communication analyses were carried out using the Cellchat R package[73] (v.1.5.0). We selected the CellChat human database (Interactions considered include secreted signaling, ECM-receptor and cell-cell contacts). First, we extracted SCTransform-normalized counts from the integrated Seurat object. For the first round of analyses, we separated cells from each disease status (lesion and non-lesion) and grouped them according to level 1 labels. We created a 'Cellchat' object for matrices from each disease status using the createCellChat() function. We subsequently identified overexpressed genes in each condition using the identifyOverExpressedInteractions(). Communication probabilities were estimated with computeCommunProb() and aggregated cell communication networks calculated with the aggregateNet() function. We then merged lesion and non-lesion cellchat objects using the mergeCellChat() function. In order to identify pathways between Myeloid cells and SMCs that were enriched in each condition compared to the other, we input the merged 'Cellchat' object to the function rankNet() with parameters (mode="comparison, sources.use="Macrophage", targets.use="SMC"). Significantly enriched pathways were denoted as those with $P<0.05$. To further explore differentially enriched pathways with increased granularity, we created a new 'CellChat' object using normalized counts from Macrophages and SMCs from lesions and grouped them using their respective level 2 annotations. We computed communication probabilities and aggregated cell communication networks as described above. Circle plots for specific signaling pathways were generated with the netvisualAggregate() function. The top 30% of interactions (based on interaction weights/strength from computed communication probability) were used for plotting interactions between level 1-annotated cell types. Given that we had a

larger number of cell types when deriving networks with level 2 labels, we chose to plot the top 15% of interactions.

Pseudotime analyses for SMCs

Cells within the pericyte-SMC-fibroblast axis were subset to contain only contractile SMCs, transitional-ECM-SMCs, fibromyocytes and fibrochondrocytes. Single cell transcriptomic pseudotime analyses were performed using monocle3[74] (v1.0.0). Given that gene expression within this subset was normalized, the SCTransform-normalized expression matrix and corresponding metadata were extracted from the corresponding seurat object. Metadata and SCT counts were used to create a 'cell_data_set' object. To preserve clustering structure from previous analyses, we also extracted PCA/UMAP embeddings, cluster IDs and cell type annotations from the processed seurat object and inserted those into the corresponding slots of the 'cell_data_set' object. A trajectory was then inferred using the learn_graph() and order_cells() functions setting contractile SMCs with the highest expression of *MYH11* as the root of the trajectory. DEG across the trajectory were calculated with grapth_test() and grouped into modules using the find_gene_modules() function. To model gene expression dynamics across pseudotime, we extracted pseudotime assignment values for each cell in the trajectory as well as SCTransform-normalized expression values and cell type annotations from the 'cell_data_set' object. We then wrote a custom script to plot gene expression changes as a function of pseudotime where we applied cubic spline interpolation to expression values using the geom_smooth() function with parameters (method="lm", formula = y ~ splines::ns(x, 3)).

TF activity inference using DorothEA regulons

For inference of TF activity, we also used a subset of the main reference only including SMCs, transitional SMCs, fibromyocytes and FCs. We downloaded a collection of curated TF regulons from the DoRothEA R package[79] (v.1.8.0). We accessed human regulons using the dorothea_hs() function and only kept those with A, B and C confident scores for a more accurate prediction of regulon activity on each cell. Confidence scores had been previously defined based on the number of supporting evidence for each regulon[79]. TF activities for each cell were then estimated with the R package VIPER (v.1.24.0)[78] providing the list of filtered regulons and the processed seurat object as input. Mean TF activities were then calculated across the SMC annotations of interest and the most variable TFs were selected for plotting.

Human coronary artery tissue procurement

Freshly explanted hearts from orthopedic heart transplant recipients were obtained at Stanford University under approved Institutional Review Board (IRB) protocols with the respective informed consents. Hearts were arrested in cardioplegic solution and rapidly transported from the operating room to adjacent laboratory on ice. The proximal 5-6 cm of three major coronary arteries (LAD, LCX, RCA) were dissected from the epicardium, trimmed of surrounding adipose,

rinsed in cold PBS and snap-frozen in liquid nitrogen. Human coronary artery tissue biospecimens were also obtained at Stanford University from non-diseased donor hearts rejected for orthotopic heart transplantation and processed following the same protocol as hearts for transplant. Reasons for rejected hearts included size incompatibility, risk for cardiotoxicity or comorbidities. Tissues were de-identified and clinical and histopathology information was used to classify ischemic, non-ischemic hearts and lesion and non-lesion containing arteries. All normal arteries originated from hearts with left ventricular ejection fraction (LVEF) greater than 50%. Frozen tissues were transferred to the University of Virginia through a material transfer agreement and IRB approved protocols.

Coronary artery snATAC-seq tissue processing and data analysis

Coronary artery samples processing and nuclei isolation for snATAC: We performed snATAC-seq on four coronary artery samples per day. Nuclei isolation was done using a similar protocol to Omni-ATAC, which has been optimized for frozen tissues. Using approximately 50 mg of tissue per sample, we set the iodixanol gradient and then carefully took the band containing the nuclei. Nuclei was then added to 1.3 ml of cold Nuclei Wash Buffer (10 mM Tris-HCl (pH 7.4), 10 mM NaCl, 3 mM MgCl2, 1% BSA, 0.1% Tween-20) in a 1.5-ml Lo-Bind microcentrifuge tube. The tube was then inverted five times, gently mixed by pipetting and contents were then passed through a 40-µm Falcon cell strainer (Corning) into a new 1.5-ml Lo-Bind microcentrifuge tube (Eppendorf). Nuclei were then pelleted by centrifugation for 5 min at 500g at 4C and supernatant removed. Nuclei were then resuspended in 100 µl of the Nuclei Buffer provided with the 10X snATAC kit. Nuclei concentration was measured using Trypan blue (Thermo Fisher) and the Countess II instrument (Thermo Fisher).

snATAC-seq library preparation, sequencing and data quality control: We used the 10X Genomics Chromium Single Cell ATAC kit for all snATAC-seq experiments. snATAC-seq libraries were shipped on ice to the Genome core facility at the Icahn School of Medicine at Mount Sinai (New York, NY, USA) for sequencing on an Illumina NovaSeq 6000. Resulting FASTQ files were preprocessed using the 10x Genomics Cellranger pipeline (CellRanger ATAC v1.2.0) using the hg38 reference genome and default parameters. Samples from different patients were preprocessed separately. Cellranger outputs were used to filter low-quality cells with the ArchR pipeline[81] (v.1.0.2) as follows: TSS enrichment > 7, unique number of fragments > 10000 and a doublet ratio < 1.5.

Dimensionality reduction, clustering of snATAC-seq data and generation of gene activity scores: Fragment files for each of the 41 patients were used to generate ArchR arrow files. The genome was then divided into 500bp windows and then fragments within each window were used to generate a tile matrix (28316 cells x ~ 6 million tiles). Iterative latent semantic index (LSI) was then used to reduce dimensionality of the tile matrix. We checked for batch effects using Harmony (v.1.0) and did not observe major differences in the data clustering structure (clusters driven by individual samples). We then used the first 30 components output by LSI for running non-linear dimensionality reduction (UMAP). Subsequent cell clustering was performed using the SNN modularity optimization-based algorithm as implemented in Seurat (v.4.1.0).

Chromatin accessibility (defined as the number of fragments within each tile) within gene bodies as well as proximally/distally from the TSS was used to infer gene expression by means of a gene activity score model. In this model, the number of fragments inside tiles of gene bodies are considered as well as surrounding tiles. To account for the activity of putative distal regulatory elements, an exponential weighting function is applied where tiles that reside further from genes TSS are assigned lower weights. Additionally, this model imposes gene boundaries to minimize the contribution of unrelated regulatory elements to a specific gene score.

Differential accessibility for SMC phenotypes: Using cell type groupings defined in our recenet publication[80], peaks were called for each 'pseudo bulk' sample (reads from each cell type cluster were combined as a new sample) using the addReproduciblePeakSet() function in ArchR (with parameters cutOff = 0.05, extendSummits = 250). Peaks called during this analysis had a width of 500 bp. Regions with differential accessibility between Modulated and Contractile SMCs were identified using a Wilcoxon-test as implemented in ArchR. The threshold for differential peak significance was set at FDR <= 0.05 and Log2 fold change > 1, resulting in a total of 5681 significantly upregulated peaks and 2121 downregulated peaks. For differential peak annotation, protein coding gene coordinates were extracted with ensembldb[111] (v.2.14.04) and EnsDb.Hsapiens.v86 (v.2.99.0). Upregulated and downregulated peaks were annotated with the nearest protein coding gene using GenomicRanges[107] (v1.42.0). This annotation was validated using the R package ChIPseeker[112] (v1.26.0) along with TxDb.Hsapiens.UCSC.hg38.knownGene (v.3.10.0).

TF motif enrichments: Enriched TF motifs for each cell type were predicted using the addMotifAnnotations() function in ArchR. Z deviation scores for each TF were then estimated with the chromVAR R package[113] (v.1.12.0).

Coronary artery calcification GWAS meta-analysis data

The GWAS meta-analysis for coronary artery calcification (CAC) was conducted on 16 cohorts including 26,909 participants of European ancestry and 8,867 participants of African ancestry. CAC scores were calculated from computed tomography imaging at baseline, or first examination as described[72]. Genotyping quality control, imputation (1000 Genomes Phase 3), and variant filtering was performed as described. A joint meta-analysis of all available CAC GWAS was performed using a fixed-effects meta-analysis in METAL, using sample size weighted SNP p-values. The summary statistics from each study were combined using an inverse variance weighted meta-analysis.

Pearson correlation calculations and gene set enrichment analyses

Normalized counts for cell types of interest were extracted from the corresponding Seurat object. Matrices were transposed to define genes as variables and then we calculated pairwise Pearson correlations for a gene of interest (e.g., *CRTAC1*) with all of the other genes across the cell types of interest using apply() and cor.test() functions with parameters (method="pearson")

from the stats R package (v.4.0.3). *P*-values from the correlations were then adjusted for multiple testing using the Benjamini Hochberg correction as implemented in the R stats package with the p.adjust() function with parameters (method="fdr").

For gene set enrichment analyses, we calculated DE genes as described in the above section. We ranked genes by log2 fold change values (log2FC) and extracted the top 100 hits per cell annotation. We then use the gost() function within the R gProfiler2 package[106] (v.0.2.1) with parameters (order=TRUE) to weight genes according to their log2FC values. We then selected significant GO:BP ontology terms (FDR < 0.05) and ranked them according to their adjusted *P*-values for plotting using custom functions from our scRNA_processing_utils.R script (https://github.com/MillerLab-CPHG/Human_athero_scRNA_meta). We found that the top GO:BP terms for fibrochondrocytes were highly redundant. Therefore, we used the gosemsim package[114] (v2.16.1) and a custom script adapted from (https://github.com/YuLab-SMU/clusterProfiler/blob/master/R/simplify.R) in order to calculate semantic similarity between GO:BP terms. We removed highly redundant terms accordingly.

Gene expression analysis in coronary artery datasets

RNA Extraction, QC, library construction and sequencing: Total RNA was extracted from frozen coronary artery segments using the Qiagen miRNeasy Mini RNA Extraction kit (catalog #217004). Approximately 50 mg of frozen tissue was pulverized using a mortar and pestle under liquid nitrogen. Tissue powder was then further homogenized in Qiazol lysis buffer using stainless steel beads in a Bullet Blender (Next Advance) homogenizer, followed by column-based purification. RNA concentration was determined using Qubit 3.0 and RNA quality was determined using Agilent 4200 TapeStation. Samples with RNA Integrity Number (RIN) greater than 5.5 and Illumina DV_{200} values greater than 75 were included for library construction. Total RNA libraries were constructed using the Illumina TruSeq Stranded Total RNA Gold kit (catalog #20020599) and barcoded using Illumina TruSeq RNA unique dual indexes (catalog # 20022371). After re-evaluating library quality using TapeStation, individually barcoded libraries were sent to Novogene for next generation sequencing. After passing additional QC, libraries were multiplexed and subjected to paired end 150 bp read sequencing on an Illumina NovaSeq S4 Flowcell to a median depth of 100 million total reads (>30 G) per library.

RNA-seq processing and analysis: The raw passed filter sequencing reads obtained from Novogene were demultiplexed using the bcl2fastq script. The quality of the reads was assessed using FASTQC and the adapter sequences were trimmed using trimgalore. Trimmed reads were aligned to the hg38 human reference genome using STAR[115] (v.2.7.3a) according to the GATK Best Practices for RNA-seq. To increase mapping efficiency and sensitivity, novel splice junctions discovered in a first alignment pass with high stringency, were used as annotation in a second pass to permit lower stringency alignment and therefore increase sensitivity. PCR duplicates were marked using Picard and WASP was used to filter reads prone to mapping bias. Total read counts and Transcripts per million normalization (TPM) for both genes and isoforms was calculated from individual bam files using the RSEM (https://deweylab.github.io/RSEM/README.html) rsem-calculate-expression command with the

paired-end option and gencode version 32 as a reference[116]. *CRTAC1* and *IBSP* expression were plotted comparing ischemic vs normal disease classification. All normal samples were from donor hearts with no evidence of atherosclerosis.

Coronary artery proteomics data generation and analysis

Tissue processing: Frozen human coronary artery segments were shipped in 1.5 mL microcentrifuge tubes to King's College London (London, United Kingdom). First, extracellular matrix (ECM) and associated ECM proteins were isolated from the frozen coronary artery samples using a 3-step extraction procedure (NaCl, SDS, and guanidine-HCl). Diced coronary artery tissue was vortexed thoroughly in 0.5 M NaCl, 0.1% SDS, and 4 M guanidine-HCl, sequentially. All 3 extracts were collected, and we subsequently used the guanidine extract for mass spectrometry and ECM protein analysis. Next, to remove glycans attached to ECM proteins, we used deglycanation enzymes (Heparinase II (Sigma-Aldrich H6512-10UN), Chondroitinase ABC (Sigma-Alrich C3667-5UN), Keratanase (G6920-5UN)) and a glycoprotein deglycosylation kit (Merck catalog #362280). We then used Water-18O (97% atom) to label N-linked glycosylation sites. After deglycosylation the ECM protein samples (n=150) underwent denaturing, reduction, alkylation, precipitation, and overnight trypsin digestion. We purified the resultant ECM fragments with AssayMAP C18 cartridges (Agilent) on an Agilent Bravo AssayMAP robot. We analyzed the purified peptide samples using nanoflow liquid chromatography tandem mass spectrometry (LC-MS/MS). We performed data-dependent analysis (DDA) (on the top 15 ions in each full MS scan) using a nanoflow LC system (Dionex UltiMate 3000 RSLC nano) coupled to a high-resolution accurate-mass Orbitrap mass analyzer (Q Exactive HF, Thermo Fisher Scientific).

Proteomics data analysis: We used the Thermo Scientific Proteome Discoverer software (v.2.3) to search the raw proteomic data files against the human database (UniProtKB/Swiss-Prot version 2019_01, containing 20,349 protein entries) using the Mascot server (version 2.6.0, Matrix Science). We measured protein abundance in each sample using label-free quantitation (LFQ). Since the data was generated with the guanidine-HCl extract, we focused analysis on matrisome-related proteins. Data was analyzed according to the King's College London pipeline and processing protocol[117,118]. Data was normalized according to the total ion intensity and subsequently scaled to remove batch effects. We filtered out proteins with more than 30% missing values. For the remaining missing values, we performed imputation with the K-nearest neighbor (KNN) impute algorithm. To tune the parameter k of the KNN-impute method we experimentally tested the Euclidean distance of the imputed values compared to the real ones for 100 randomly selected values, testing for k=2 until 20. The optimal k value was set to 5 according to this procedure and this was applied to impute all the remaining missing values. Values were then displayed in Log2 scale.

Disease categories: Disease status of coronary artery segments was determined as previously described[80]. Briefly, samples containing no evidence of atherosclerosis were included in category 1. Samples that are lesion-free and have no evidence of atherosclerosis, however the

patient has evidence of ischemic heart disease were included in category 2. Samples that have evidence of atherosclerosis due to presence of lesion were included in category 3.

STARNET regulatory networks and clinical trait enrichment analysis

Based on STARNET[89] multi-tissue bulk RNA-seq data, tissue specific and cross-tissue co-expression modules were inferred using WGCNA[119]. Enrichment for clinical traits was computed by aggregating Pearson's correlation *P* values by co-expression module using Fisher's method. Enrichment for DE genes was calculated with the hypergeomteric test using DESeq2 called genes (30% change, FDR <0.01) adjusting for age and sex covariates. The gene regulatory network for *CRTAC1* co-expressed genes was inferred using GENIE3[120]. Weighted key driver analysis was then applied to identify hub or highly influential genes in the regulatory network using the Mergeomics R package[121].

Data Availability

Raw count matrices included in this study were accessed through GEO and Zenodo. Raw count matrices for *Wirka et al* [18], *Pan et al* [19], *Alsaigh et al* [16] were obtained through the following accession numbers: *Wirka et al* (GSM3819856, GSM3819857, GSM3819858, GSM3819859, GSM3819860, GSM3819861, GSM3819862, GSM3819863); *Alsaigh et al* (GSM4837523, GSM4837524, GSM4837525, GSM4837526, GSM4837527, GSM4837528); *Pan et al* (GSM4705589, GSM4705590, GSM4705591). Raw count matrices from *Hu et al* [17] were obtained from Zenodo (https://zenodo.org/record/6032099#.Y1RDa-zMITU). The corresponding accession numbers can also be found in **Supplementary Table 1**.

Bulk RNA-seq data from human carotid lesions[88] was accessed through GEO with the accession number GSE120521.

Code Availability

Code used for processing of raw count matrices, integration benchmark and other downstream analyses can be found in the following Github repository: https://github.com/MillerLab-CPHG/Human_athero_scRNA_meta

Acknowledgements

This work was supported by grants from: the National Institutes of Health (grant numbers R01HL148239 and R01HL164577 to C.L.M.; R35GM133712 to C.Z.; R01HL142809 and R01HL159514 to R.M.; K01HL164687 to C.L.L.C.; R01HL125863 to J.L.M.B; R01HL130423, R01HL135093 and R01HL148167-01A1 to J.C.K.), the American Heart Association (grant number 20POST35120545 to A.W.T.; AHA909150 to J.V.M.; A14SFRN20840000 to J.L.M.B.), the Swedish Research Council and Heart Lung Foundation (grant number 2018-02529 and

20170265 to J.L.M.B.), the Fondation Leducq (grant number 'PlaqOmics' 18CVD02 to C.L.M., S.W.vdL., J.L.M.B., and M.M.), the European Union H2020 TO_AITION grant number 848146 (to SWvdL) and the Single-Cell Data Insights award from the Chan Zuckerberg Initiative, LLC and Silicon Valley Community Foundation (to C.L.M., C.Z. and S.W.vdL.).

Author contributions

C.L.M. supervised research primarily related to the study. M.M., J.C.K., J.L.M.B., R.M., N.C.S., C.Z., and S.W.vdL. jointly supervised research secondarily related to the study. J.V.M. and C.L.M. conceived and designed the experiments. J.V.M., G.A., A.W.T., K.T., and C.L.L.C. performed the experiments. J.V.M. performed the statistical analyses. J.V.M., A.W.T., C.J.H., and K.T. analyzed the data. M.B., M.K., P.P., M.M., J.C.K., J.L.M.B. and S.W.vdL. contributed reagents/materials/analysis tools. J.V.M., D.W., G.A., A.W.T., C.J.H., N.C.S. and C.L.M. wrote the paper.

Competing Interests

J.L.M.B. is a shareholder in Clinical Gene Network AB who have a vested interest in STARNET. S.W.vdL. has received funding from Roche for unrelated work. C.L.M. has received funding from AstraZeneca for an unrelated project. R.M. has received funding from Angea Biotherapeutics and Amgen and serves as a consultant for Myokardia/BMS, Renovacor, Epizon Pharma, and Third Pole, all unrelated to the current project. J.C.K. is the recipient of an Agilent Thought Leader Award, which includes funding for research that is unrelated to the current project. The other authors declare no competing interests.

References

1. Tsao, C. W. *et al.* Heart Disease and Stroke Statistics-2022 Update: A Report From the American Heart Association. *Circulation* **145**, e153–e639 (2022).
2. Cassar, A., Holmes, D. R., Rihal, C. S. & Gersh, B. J. Chronic Coronary Artery Disease: Diagnosis and Management. *Mayo Clinic Proceedings* vol. 84 1130–1146 Preprint at https://doi.org/10.4065/mcp.2009.0391 (2009).
3. Khera, A. V. & Kathiresan, S. Genetics of coronary artery disease: discovery, biology and clinical translation. *Nat. Rev. Genet.* **18**, 331–344 (2017).
4. Stary, H. C. *et al.* A Definition of Advanced Types of Atherosclerotic Lesions and a Histological Classification of Atherosclerosis. *Circulation* vol. 92 1355–1374 Preprint at

https://doi.org/10.1161/01.cir.92.5.1355 (1995).

5. Virmani, R., Kolodgie, F. D., Burke, A. P., Farb, A. & Schwartz, S. M. Lessons from sudden coronary death: a comprehensive morphological classification scheme for atherosclerotic lesions. *Arterioscler. Thromb. Vasc. Biol.* **20**, 1262–1275 (2000).

6. Libby, P. Inflammation in atherosclerosis. *Nature* **420**, 868–874 (2002).

7. Libby, P. The changing landscape of atherosclerosis. *Nature* **592**, 524–533 (2021).

8. Shankman, L. S. *et al.* KLF4-dependent phenotypic modulation of smooth muscle cells has a key role in atherosclerotic plaque pathogenesis. *Nature Medicine* vol. 21 628–637 Preprint at https://doi.org/10.1038/nm.3866 (2015).

9. Espinosa-Diez, C., Mandi, V., Du, M., Liu, M. & Gomez, D. Smooth muscle cells in atherosclerosis: Clones but not carbon copies. *JVS-Vascular Science* vol. 2 136–148 Preprint at https://doi.org/10.1016/j.jvssci.2021.02.002 (2021).

10. Evrard, S. M. *et al.* Endothelial to mesenchymal transition is common in atherosclerotic lesions and is associated with plaque instability. *Nat. Commun.* **7**, 11853 (2016).

11. Andueza, A. *et al.* Endothelial Reprogramming by Disturbed Flow Revealed by Single-Cell RNA and Chromatin Accessibility Study. *Cell Rep.* **33**, 108491 (2020).

12. Winkels, H. *et al.* Atlas of the Immune Cell Repertoire in Mouse Atherosclerosis Defined by Single-Cell RNA-Sequencing and Mass Cytometry. *Circ. Res.* **122**, 1675–1688 (2018).

13. Cochain, C. *et al.* Single-Cell RNA-Seq Reveals the Transcriptional Landscape and Heterogeneity of Aortic Macrophages in Murine Atherosclerosis. *Circ. Res.* **122**, 1661–1674 (2018).

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

15. Depuydt, M. A. C. *et al.* Microanatomy of the Human Atherosclerotic Plaque by Single-Cell Transcriptomics. *Circ. Res.* **127**, 1437–1455 (2020).

16. Alsaigh, T., Evans, D., Frankel, D. & Torkamani, A. Decoding the transcriptome of

atherosclerotic plaque at single-cell resolution. Preprint at https://doi.org/10.1101/2020.03.03.968123.

17. Hu, Z. *et al.* Single-Cell Transcriptomic Atlas of Different Human Cardiac Arteries Identifies Cell Types Associated With Vascular Physiology. *Arterioscler. Thromb. Vasc. Biol.* **41**, 1408–1427 (2021).
18. Wirka, R. C. *et al.* Atheroprotective roles of smooth muscle cell phenotypic modulation and the TCF21 disease gene as revealed by single-cell analysis. *Nat. Med.* **25**, 1280–1289 (2019).
19. 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* **142**, 2060–2075 (2020).
20. Alencar, G. F. *et al.* Stem Cell Pluripotency Genes Klf4 and Oct4 Regulate Complex SMC Phenotypic Changes Critical in Late-Stage Atherosclerotic Lesion Pathogenesis. *Circulation* **142**, 2045–2059 (2020).
21. Zernecke, A. *et al.* Meta-Analysis of Leukocyte Diversity in Atherosclerotic Mouse Aortas. *Circ. Res.* **127**, 402–426 (2020).
22. Conklin, A. C. *et al.* Meta-Analysis of Smooth Muscle Lineage Transcriptomes in Atherosclerosis and Their Relationships to In Vitro Models. *Immunometabolism* **3**, (2021).
23. Vallejo, J., Cochain, C., Zernecke, A. & Ley, K. Heterogeneity of immune cells in human atherosclerosis revealed by scRNA-Seq. *Cardiovasc. Res.* **117**, 2537–2543 (2021).
24. Germain, P.-L., Lun, A., Garcia Meixide, C., Macnair, W. & Robinson, M. D. Doublet identification in single-cell sequencing data using. *F1000Res.* **10**, 979 (2021).
25. Yang, S. *et al.* Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol.* **21**, 57 (2020).
26. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* vol. 177 1888–1902.e21 Preprint at https://doi.org/10.1016/j.cell.2019.05.031 (2019).

27. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
28. Tran, H. T. N. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21**, 12 (2020).
29. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
30. Meyer, K. B., Wilbrey-Clark, A., Nawijn, M. & Teichmann, S. A. The Human Lung Cell Atlas: a transformational resource for cells of the respiratory system. *Lung Stem Cells in Development, Health and Disease* 158–174 Preprint at https://doi.org/10.1183/2312508x.10010920 (2021).
31. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
32. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
33. Tabula Sapiens Consortium* *et al.* The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
34. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
35. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
36. van der Harst, P. & Verweij, N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circ. Res.* **122**, 433–443 (2018).
37. Hartiala, J. A. *et al.* Genome-wide analysis identifies novel susceptibility loci for myocardial infarction. *Eur. Heart J.* **42**, 919–933 (2021).
38. Franceschini, N. *et al.* GWAS and colocalization analyses implicate carotid intima-media

thickness and carotid plaque loci in cardiovascular outcomes. *Nat. Commun.* **9**, 5141 (2018).

39. Giri, A. *et al.* Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nat. Genet.* **51**, 51–62 (2019).
40. Jansen, I. E. *et al.* Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).
41. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
42. Timshel, P. N., Thompson, J. J. & Pers, T. H. Genetic mapping of etiologic brain cell types for obesity. *Elife* **9**, (2020).
43. Mammana, S. *et al.* The Role of Macrophages in Neuroinflammatory and Neurodegenerative Pathways of Alzheimer's Disease, Amyotrophic Lateral Sclerosis, and Multiple Sclerosis: Pathogenetic Cellular Effectors and Potential Therapeutic Targets. *International Journal of Molecular Sciences* vol. 19 831 Preprint at https://doi.org/10.3390/ijms19030831 (2018).
44. Dawson, A., Wang, Y., Li, Y., LeMaire, S. A. & Shen, Y. H. New Technologies With Increased Precision Improve Understanding of Endothelial Cell Heterogeneity in Cardiovascular Health and Disease. *Front Cell Dev Biol* **9**, 679995 (2021).
45. Tombor, L. S. *et al.* Single cell sequencing reveals endothelial plasticity with transient mesenchymal activation after myocardial infarction. *Nat. Commun.* **12**, 681 (2021).
46. Kalucka, J. *et al.* Single-Cell Transcriptome Atlas of Murine Endothelial Cells. *Cell* **180**, 764–779.e20 (2020).
47. Willemsen, L. & de Winther, M. P. Macrophage subsets in atherosclerosis as defined by single-cell technologies. *J. Pathol.* **250**, 705–714 (2020).
48. Tillie, R. J. H. A., van Kuijk, K. & Sluimer, J. C. Fibroblasts in atherosclerosis: heterogeneous and plastic participants. *Curr. Opin. Lipidol.* **31**, 273–278 (2020).

49. Chowdhury, R. R. *et al.* Human Coronary Plaque T Cells Are Clonal and Cross-React to Virus and Self. *Circ. Res.* **130**, 1510–1530 (2022).

50. Domínguez Conde, C. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022).

51. Tanaka, M. *et al.* The endothelial adrenomedullin-RAMP2 system regulates vascular integrity and suppresses tumour metastasis. *Cardiovasc. Res.* **111**, 398–409 (2016).

52. Thiriot, A. *et al.* Differential DARC/ACKR1 expression distinguishes venular from non-venular endothelial cells in murine tissues. *BMC Biol.* **15**, 45 (2017).

53. Jiang, Y. *et al.* Endothelial Aquaporin-1 (AQP1) Expression Is Regulated by Transcription Factor Mef2c. *Mol. Cells* **39**, 292–298 (2016).

54. Rutkovskiy, A. *et al.* Aquaporin-1 in cardiac endothelial cells is downregulated in ischemia, hypoxia and cardioplegia. *J. Mol. Cell. Cardiol.* **56**, 22–33 (2013).

55. Elmasri, H. *et al.* Endothelial cell-fatty acid binding protein 4 promotes angiogenesis: role of stem cell factor/c-kit pathway. *Angiogenesis* **15**, 457–468 (2012).

56. Xu, S. *et al.* Endothelial Dysfunction in Atherosclerotic Cardiovascular Diseases and Beyond: From Mechanism to Pharmacotherapies. *Pharmacol. Rev.* **73**, 924–967 (2021).

57. Newman, A. A. C. *et al.* Multiple cell types contribute to the atherosclerotic lesion fibrous cap by PDGFRβ and bioenergetic mechanisms. *Nature Metabolism* vol. 3 166–181 Preprint at https://doi.org/10.1038/s42255-020-00338-8 (2021).

58. Johnson, L. A. & Jackson, D. G. Inflammation-induced secretion of CCL21 in lymphatic endothelium is a key regulator of integrin-mediated dendritic cell transmigration. *Int. Immunol.* **22**, 839–849 (2010).

59. Lim, H. Y. *et al.* Hyaluronan Receptor LYVE-1-Expressing Macrophages Maintain Arterial Tone through Hyaluronan-Mediated Regulation of Smooth Muscle Cell Collagen. *Immunity* **49**, 1191 (2018).

60. Boltjes, A. & van Wijk, F. Human dendritic cell functional specialization in steady-state and

inflammation. *Front. Immunol.* **5**, 131 (2014).

61. Villani, A.-C. *et al.* Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, (2017).
62. Swiecki, M. & Colonna, M. The multifaceted biology of plasmacytoid dendritic cells. *Nat. Rev. Immunol.* **15**, 471–485 (2015).
63. Jahrsdörfer, B. *et al.* Granzyme B produced by human plasmacytoid dendritic cells suppresses T-cell expansion. *Blood* **115**, 1156–1165 (2010).
64. Ziegler, S. F., Ramsdell, F. & Alderson, M. R. The activation antigen CD69. *STEM CELLS* vol. 12 456–465 Preprint at https://doi.org/10.1002/stem.5530120502 (1994).
65. Schluns, K. S., Kieper, W. C., Jameson, S. C. & Lefrançois, L. Interleukin-7 mediates the homeostasis of naïve and memory CD8 T cells in vivo. *Nat. Immunol.* **1**, 426–432 (2000).
66. Li, J., Huston, G. & Swain, S. L. IL-7 promotes the transition of CD4 effectors to persistent memory cells. *J. Exp. Med.* **198**, 1807–1815 (2003).
67. Saigusa, R., Winkels, H. & Ley, K. T cell subsets and functions in atherosclerosis. *Nat. Rev. Cardiol.* **17**, 387–401 (2020).
68. Andreatta, M. & Carmona, S. J. UCell: Robust and scalable single-cell gene signature scoring. *Comput. Struct. Biotechnol. J.* **19**, 3796–3798 (2021).
69. Dobnikar, L. *et al.* Disease-relevant transcriptional signatures identified in individual smooth muscle cells from healthy mouse vessels. *Nat. Commun.* **9**, 4567 (2018).
70. Speer, M. Y. *et al.* Smooth muscle cells give rise to osteochondrogenic precursors and chondrocytes in calcifying arteries. *Circ. Res.* **104**, 733–741 (2009).
71. Cheng, P. *et al.* Smad3 regulates smooth muscle cell fate and mediates adverse remodeling and calcification of the atherosclerotic plaque. *Nature Cardiovascular Research* vol. 1 322–333 Preprint at https://doi.org/10.1038/s44161-022-00042-8 (2022).
72. Kavousi, M. *et al.* Multi-ancestry genome-wide analysis identifies effector genes and druggable pathways for coronary artery calcification. *medRxiv* 2022.05.02.22273844

(2022) doi:10.1101/2022.05.02.22273844.

73. Jin, S. *et al.* Inference and analysis of cell-cell communication using CellChat. *Nat. Commun.* **12**, 1088 (2021).

74. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).

75. Grootaert, M. O. J., Finigan, A., Figg, N. L., Uryga, A. K. & Bennett, M. R. SIRT6 Protects Smooth Muscle Cells From Senescence and Reduces Atherosclerosis. *Circ. Res.* **128**, 474–491 (2021).

76. Steck, E. *et al.* Chondrocyte secreted CRTAC1: a glycosylated extracellular matrix molecule of human articular cartilage. *Matrix Biol.* **26**, 30–41 (2007).

77. Yang, X. *et al.* Overexpression of Spry1 in chondrocytes causes attenuated FGFR ubiquitination and sustained ERK activation resulting in chondrodysplasia. *Dev. Biol.* **321**, 64–76 (2008).

78. Alvarez, M. J. *et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.* **48**, 838–847 (2016).

79. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.* **29**, 1363–1375 (2019).

80. Turner, A. W. *et al.* Author Correction: Single-nucleus chromatin accessibility profiling highlights regulatory mechanisms of coronary artery disease risk. *Nat. Genet.* **54**, 1259 (2022).

81. Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).

82. Durham, A. L., Speer, M. Y., Scatena, M., Giachelli, C. M. & Shanahan, C. M. Role of smooth muscle cells in vascular calcification: implications in atherosclerosis and arterial stiffness. *Cardiovascular Research* vol. 114 590–600 Preprint at

https://doi.org/10.1093/cvr/cvy010 (2018).

83. Leopold, J. A. Vascular calcification: Mechanisms of vascular smooth muscle cell calcification. *Trends in Cardiovascular Medicine* vol. 25 267–274 Preprint at https://doi.org/10.1016/j.tcm.2014.10.021 (2015).

84. Sanchez, C. *et al.* Chondrocyte secretome: a source of novel insights and exploratory biomarkers of osteoarthritis. *Osteoarthritis Cartilage* **25**, 1199–1209 (2017).

85. Szilagyi, I. *et al.* OP0111 PLASMA PROTEOMICS IDENTIFIES CRTAC1 AS BIOMARKER FOR OSTEOARTHRITIS SEVERITY AND PROGRESSION. *Annals of the Rheumatic Diseases* vol. 80 61.1–62 Preprint at https://doi.org/10.1136/annrheumdis-2021-eular.1888 (2021).

86. Pustlauk, W. *et al.* Induced osteogenic differentiation of human smooth muscle cells as a model of vascular calcification. *Scientific Reports* vol. 10 Preprint at https://doi.org/10.1038/s41598-020-62568-w (2020).

87. Kim, J.-S., Ryoo, Z. Y. & Chun, J.-S. Cytokine-like 1 (Cytl1) regulates the chondrogenesis of mesenchymal cells. *J. Biol. Chem.* **282**, 29359–29367 (2007).

88. Mahmoud, A. D. *et al.* The Human-Specific and Smooth Muscle Cell-Enriched LncRNA SMILR Promotes Proliferation by Regulating Mitotic CENPF mRNA and Drives Cell-Cycle Progression Which Can Be Targeted to Limit Vascular Remodeling. *Circ. Res.* **125**, 535–551 (2019).

89. Koplev, S. *et al.* A mechanistic framework for cardiometabolic and coronary artery diseases. *Nature Cardiovascular Research* **1**, 85–100 (2022).

90. Mietus-Snyder, M., Gowri, M. S. & Pitas, R. E. Class A Scavenger Receptor Up-regulation in Smooth Muscle Cells by Oxidized Low Density Lipoprotein. *Journal of Biological Chemistry* vol. 275 17661–17670 Preprint at https://doi.org/10.1074/jbc.275.23.17661 (2000).

91. Beyea, M. M. *et al.* The oxysterol 24(s),25-epoxycholesterol attenuates human smooth

muscle-derived foam cell formation via reduced low-density lipoprotein uptake and enhanced cholesterol efflux. *J. Am. Heart Assoc.* **1**, e000810 (2012).

92. Allahverdian, S., Chehroudi, A. C., McManus, B. M., Abraham, T. & Francis, G. A. Contribution of intimal smooth muscle cells to cholesterol accumulation and macrophage-like cells in human atherosclerosis. *Circulation* **129**, 1551–1559 (2014).
93. Liu, B. *et al.* Genetic Regulatory Mechanisms of Smooth Muscle Cells Map to Coronary Artery Disease Risk Loci. *Am. J. Hum. Genet.* **103**, 377–388 (2018).
94. Turner, A. W. *et al.* Single-nucleus chromatin accessibility profiling highlights regulatory mechanisms of coronary artery disease risk. *Nat. Genet.* **54**, 804–816 (2022).
95. Slenders, L. *et al.* Intersecting single-cell transcriptomics and genome-wide association studies identifies crucial cell populations and candidate genes for atherosclerosis. *Eur Heart J Open* **2**, oeab043 (2022).
96. Jinnouchi, H. *et al.* Calcium deposition within coronary atherosclerotic lesion: Implications for plaque stability. *Atherosclerosis* **306**, 85–95 (2020).
97. Naik, V. *et al.* Sources of cells that contribute to atherosclerotic intimal calcification: an in vivo genetic fate mapping study. *Cardiovasc. Res.* **94**, 545–554 (2012).
98. Iyer, D. *et al.* Coronary artery disease genes SMAD3 and TCF21 promote opposing interactive genetic programs that regulate smooth muscle cell differentiation and disease risk. *PLoS Genet.* **14**, e1007681 (2018).
99. Pardali, E. TGFβ Signaling and Cardiovascular Diseases. *International Journal of Biological Sciences* 195–213 Preprint at https://doi.org/10.7150/ijbs.3805 (2012).
100. Furumatsu, T., Tsuda, M., Taniguchi, N., Tajima, Y. & Asahara, H. Smad3 induces chondrogenesis through the activation of SOX9 via CREB-binding protein/p300 recruitment. *J. Biol. Chem.* **280**, 8343–8350 (2005).
101. Furumatsu, T., Ozaki, T. & Asahara, H. Smad3 activates the Sox9-dependent transcription on chromatin. *Int. J. Biochem. Cell Biol.* **41**, 1198–1204 (2009).

102. Otsuka, F., Finn, A. V. & Virmani, R. Do vulnerable and ruptured plaques hide in heavily calcified arteries? *Atherosclerosis* vol. 229 34–37 (2013).

103. Mauriello, A. *et al.* Coronary calcification identifies the vulnerable patient rather than the vulnerable Plaque. *Atherosclerosis* vol. 229 124–129 Preprint at https://doi.org/10.1016/j.atherosclerosis.2013.03.010 (2013).

104. Hutcheson, J. D. *et al.* Genesis and growth of extracellular-vesicle-derived microcalcification in atherosclerotic plaques. *Nat. Mater.* **15**, 335–343 (2016).

105. Germain, P.-L., Lun, A., Macnair, W. & Robinson, M. D. Doublet identification in single-cell sequencing data using scDblFinder. *F1000Research* vol. 10 979 Preprint at https://doi.org/10.12688/f1000research.73600.1 (2021).

106. Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J. & Peterson, H. gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. *F1000Res.* **9**, 709 (2020).

107. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).

108. Xi, N. M. & Li, J. J. Protocol for executing and benchmarking eight computational doublet-detection methods in single-cell RNA sequencing data analysis. *STAR Protoc* **2**, 100699 (2021).

109. Sikkema, L. *et al.* An integrated cell atlas of the human lung in health and disease. *bioRxiv* 2022.03.10.483747 (2022) doi:10.1101/2022.03.10.483747.

110. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols* vol. 4 1184–1191 Preprint at https://doi.org/10.1038/nprot.2009.97 (2009).

111. Rainer, J., Gatto, L. & Weichenberger, C. X. ensembldb: an R package to create and use Ensembl-based annotation resources. *Bioinformatics* vol. 35 3151–3153 Preprint at https://doi.org/10.1093/bioinformatics/btz031 (2019).

112. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* vol. 31 2382–2383 Preprint at https://doi.org/10.1093/bioinformatics/btv145 (2015).

113. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).

114. Yu, G. *et al.* GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* vol. 26 976–978 Preprint at https://doi.org/10.1093/bioinformatics/btq064 (2010).

115. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

116. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 1–16 (2011).

117. Didangelos, A. *et al.* Proteomics Characterization of Extracellular Space Components in the Human Aorta *. *Mol. Cell. Proteomics* **9**, 2048–2062 (2010).

118. Comparative analysis of statistical methods used for detecting differential expression in label-free mass spectrometry proteomics. *J. Proteomics* **129**, 83–92 (2015).

119. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

120. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring regulatory networks from expression data using tree-based methods. *PLoS One* **5**, (2010).

121. Shu, L. *et al.* Mergeomics: multidimensional data integration to identify pathogenic perturbations to biological systems. *BMC Genomics* **17**, 874 (2016).