Uncovering the Pathology of Rheumatoid Arthritis with Single Cell Immunoprofiling

A dissertation presented

by

Chamith Yohan Fonseka

to

the Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biological and Biomedical Sciences

Harvard University

Cambridge, Massachusetts

October 2019

Uncovering the Pathology of Rheumatoid Arthritis with Single Cell Immunoprofiling

Abstract

Rheumatoid arthritis (RA) is a chronic multi-systemic autoimmune disorder affecting nearly 25 million people worldwide, yet its underlying causes remain unclear. Genetic studies of patients with RA have highlighted the role of dysfunction in the adaptive immune system, particularly among CD4+ T cell populations. While defining the precise CD4+ T cell subsets that are dysregulated in RA patients is critical to deciphering pathogenesis, much of the work in the field has relied on animal models or derives from bulk analyses of immune cells, which can lead to overlooking rare or transitional cell types due to the heterogenous nature of immune populations. The recent introduction of high dimensional single-cell analyses have improved the ability to resolve complex mixtures of cells; however, identifying disease-associated cell types or cell states in patient samples remains challenging due to technical and inter-individual variation. In particular, case-control analysis of disease using single cell data requires a quantitative approach to determining which cells provide the most information (and which cells are uninformative) while accounting for confounding effects from batch or technical variation; properly grouping those cells into biologically relevant populations, and then determining whether the abundance of these populations is statistically different between cases and controls. Mixed effects modeling of Associations of Single Cells (MASC) is a novel reverse single cell association strategy to determine if a cellular subpopulation is associated with case-control status while controlling for technical confounders and biological covariates. This method revealed important changes in the abundance of disease-associated immune and stromal

populations – specifically an expansion of cytotoxic CD4+ T cells and *HLA*+ sublining fibroblasts in RA. Compared to peripheral blood, synovial fluid and synovial tissue samples from RA patients were significantly enriched for both of these populations, indicating that these cell types are present in high abundance at the specific locus of RA pathogenesis. The methods developed for the analysis of single cell data are broadly applicable, support performing association testing with high-dimensional single cell data, and can help identify other cellular populations that are critical to rheumatic disease pathogenesis.

Table of Contents

Acknowledgements

I would like to thank my thesis advisor, Soumya Raychaudhuri, for his invaluable mentorship, excellent advice, and unrelenting support over the years. It's not often that a PhD student has the opportunity to work with a principal investigator that they mesh with so well both academically and personally. From Soumya, I have learned how to be a rigorous bioinformatician, carefully developing and applying new and interesting statistical methods while always being willing to consider the results with appropriate skepticism. Not content to simply be a good scientific mentor, Soumya has endured my various non-academic pursuits in science policy and politics with his usual good sense of humor. Although I knew next to nothing about rheumatoid arthritis and nothing about single cell mass cytometry when I started in the lab, Soumya encouraged me to step outside my comfort zone and dive right into the project.

Starting a new project with a new technology in a new field can be intimidating, but the incredibly supportive environment of the Raychaudhuri Lab made the process much easier. I never felt inhibited about asking questions or trying out new ideas with this collaborative and curious group. From going on group ski trips and snow tubing, to retreats on the harbor islands, to the many, many late beer nights, I cannot think of a better place to learn, laugh, and live than the Raychaudhuri lab. In particular, I thank Nicola Teslovich for the mentorship, technical expertise, and great sense of humor that I was privileged to experience while working on the mass cytometry study together. Nick, Susan Hannes, and Jessica Beynor deserve endless credit for their diligent work preparing and analyzing all of the biological samples that I have used in my doctoral work. I thank Jamie Valerus, Xinli Hu, Kamil Slowikowski, and Maria Gutierrez-Arcelus for their enthusiasm and friendship when I first joined the lab; it made all the difference to have people willing to answer any and all of my questions as I found my place. I thank Rachel Kenvel for putting up with all of my long-winded diatribes about international politics; Harm-

Jan Westra for his excellent advice, fashion sense, and mastery of Excel; and Emma Davenport, for both her boundless scientific knowledge and her masterful baking skills. Ilya Korsunsky has been an amazing mentor in the time we've been in the lab together, always willing to cover the wall with equations as we puzzled through the inner workings of one method or another. Despite being stuck in the cubicle next to mine for nearly three years now, Tiffany Amariuta has somehow remained as friendly and sympathetic as she was at the beginning – I will miss dishing and hearing about all of the hot lab goss! To my trivia partners, Aparna Nathan and Joseph Mears – I am thankful for their extensive assistance and advice, cheerfulness and friendship, and for introducing me to 1UP. To all the other members of the Raychaudhuri lab: thank you for making the lab an enjoyable, exciting, and food-filled place to be. I cannot wait to see what all of you do in the future.

The vast majority of scientific research is the product of collaboration, and I have been extraordinarily lucky to have such great collaborators over the years. I want to sincerely thank Michael Brenner for his mentorship and support from the very beginning of my academic career. Deepak Rao, my co-first author on many publications, showed unlimited patience with me as I learned about immunology from the ground up, for which I am supremely grateful. Kevin Wei is a first-class immunologist and first-class co-author as well, and I have greatly benefited from the time I got to spend working with him. I thank Helena Jonssson, who has been a font of knowledge on all things granzymes and T cells. I truly enjoyed working with all of them in the AMP RA/SLE consortium and its many meetings. I also want to thank the members of my dissertation advisory committee, Steve Gygi, Chirag Patel, and Branch Moody for all the helpful scientific advice and thoughtful questions throughout the years as I navigated through these projects and grew as a scientist.

Harvard Medical School can seem like a tough place at first, but the BBS office makes everything go so much easier. I thank Kate Hodgins, Maria Stevens Bollinger, Danny Gonzalez, and Anne O'Shea for being an excellent resource whenever I needed help or advice, and for

planning and supervising our program retreat in Provincetown, which will always be a highlight of my time here. When I started classes first year, I had the fairly normal fear that I would fail to make friends in my program. Little did I know that the classmates who would go on to become my close friends happened be sitting right next to me in Genetics 201! To Adam, Zach, Lindsay, Elizabeth, and Jessalyn – it has been an incomparable experience to go through this process with such great friends. It's not just that I've gotten to experience Halloween parties, corner rooms, formal cruises, Eurovision spectacles, beach adventures, or amazing weddings with all of you, but that you have always been there, whether it was advice I needed to hear, or frustrations I needed to vent, or even last minute questions about dissertation formatting!

Finally, I want to thank my family. My brother, Janaka Fonseka, has been a rock in my life and I'm thankful for his unflagging belief in my abilities and talents. To my parents, Hema and Preethi Fonseka – it sometimes still shocks me how much you sacrificed and endured to give me this opportunity. I am thankful for your unwavering support and unconditional love throughout my entire life. Mom, it was so many years ago that you spent hours learning the taxonomy of snakes and whales just to humor me. I didn't appreciate it at the time, but your devotion to learning about these random subjects in depth and then teaching them to me left a deep impression; it truly was my first scientific training. Dad – thank you for taking care of me through the thick and thin, for always checking in on me when I needed a break, and for encouraging me to pursue my dreams. I could not have asked for a more accepting and supportive family, even if I occasionally show hints of frustration.

Finally, to the love of my life and my partner, Monica. You've known me for so long that you really do know me better than I know myself; moreover, you believed in me when I couldn't. Thank you for everything you've done as we've gone on this journey together. From making sure I remembered to eat to painstakingly editing figures in Adobe when I didn't know how – I am forever grateful. I cannot wait to spend the rest of my life with you.

Chapter 1:

Introduction

Autoimmune disorders result from immunological dysfunction wherein the capacity of an organism's immune system to ignore or tolerate its own tissues is compromised, leading to the targeted destruction of healthy cells and attendant pathological complications. There is a wide spectrum of diseases classified as autoimmune disorders, including both diseases that affect a specific organ or tissue – like type 1 diabetes (pancreas)[1], celiac (intestine)[2], and multiple sclerosis (central nervous system)[3] – as well as diseases that affect a variety of tissues and organ systems, like systematic lupus erythematosus[4]. The class of autoimmune disorders consists of over 100 distinct syndromes and have a global prevalence of 3-5%, although the accuracy of these estimates are complicated by the observation that many autoimmune diseases have varied presentation and share symptoms and subphenotypes, which impedes accurate diagnosis and classification[5,6]. Broadly, autoimmune disorders are well known for demonstrating clear sex differences in prevalence; in general, females are more frequently affected than males, although the exact sex bias differs between specific conditions and does not necessarily indicate sex-specific differences in disease severity[7].

Generally, autoimmune disorders are considered to share at least one fundamental etiology – the breakdown of immunological tolerance. Tolerance is the process of how the immune system normally prevents itself from targeting self-molecules, cells or tissues instead of exogenous pathogens[8]. A critical step in this mechanism is central tolerance, which takes place in the thymus and bone marrow where T and B lymphocytes develop. Lymphocytes that demonstrate the capacity to react to self-antigens are typically negatively selected against and removed, although this is not perfect and some autoreactive lymphocytes are able to escape this process. However, various methods of peripheral tolerance allow for the control of autoreactive lymphocytes that have escaped central tolerance and prevent further damage. Thus, the mere presence of autoreactive lymphocytes is not necessarily pathogenic on its own; rather, the development of autoimmunity likely involves deficiencies in both central and peripheral

tolerance – leading to an increase in escape of autoreactive cells and a reduced ability to control them (Figure 1-1).

Dysfunction in Immune Tolerance Can Lead to Autoimmunity

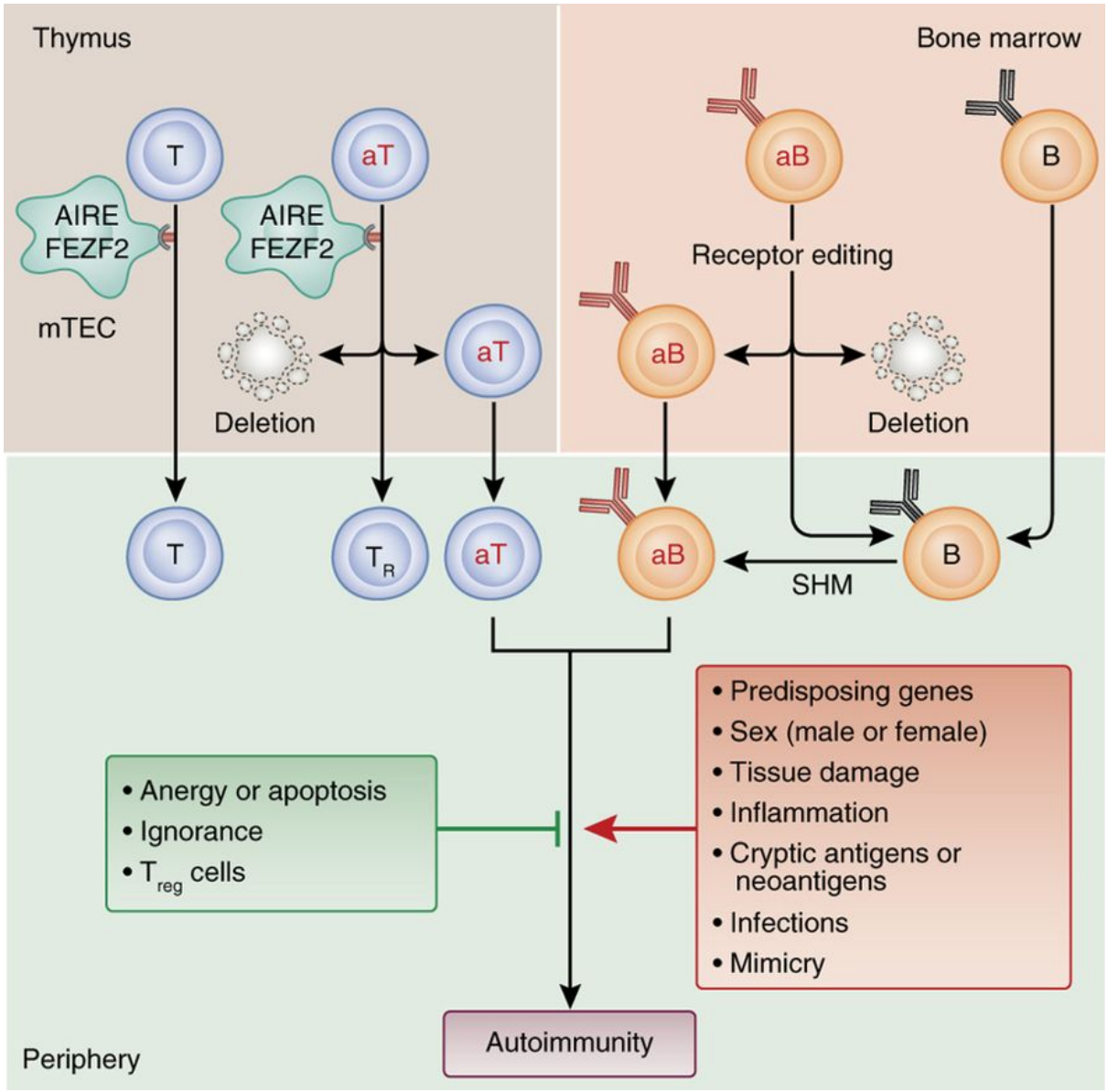


Figure 1-1. A simplified overview of the mechanism of immunological tolerance. In central tolerance, T and B lymphocytes develop in the thymus and bone marrow, respectively, and cells that do not strongly react to the presentation of self-antigens are exported to the periphery. Autoreactive cells (aT and aB) are typically selected against and deleted or induced to become less autoreactive. If autoreactive cells make it to the periphery, they are typically controlled by the mechanisms listed in the green box on the right. However, endogenous and exogenous insults to this system, as shown in red box in the right, can promote the survival and proliferation of autoreactive T and B cells, leading to the development of autoimmunity. (Adapted from Theofilopoulos *et al.*, 2017)[9]

The idea that autoimmunity results from a break in adaptive tolerance is generally accepted, although recent findings suggest that a substantial number of autoimmune diseases involve disruption to the innate immune system as well[10]. Regardless, the immune pathologies underlying specific autoimmune disorders remain unclear; while some diseases share common signatures of autoreactivity and are amenable to treatment with similar therapeutics, the fact that the course of an autoimmune disease often differs from patient to patient and through different phases within the same patient may indicate that different pathogenic mechanisms are at work at any given time or in any given individual. Individuals may suffer from more than one autoimmune disorder simultaneously; however, whether this is the result of a shared immunopathogenic mechanism is complicated by the observation that many autoimmune diseases share symptoms and subphenotypes which complicate accurate classification[5]. Overall, autoimmune disorders represent a set of highly complex and heterogeneous syndromes that require detailed experimentation and analysis to understand. The remainder of this work will focus on the specific immunopathologies that underlie rheumatoid arthritis.

Rheumatoid arthritis (RA) is a chronic, multisystemic autoimmune disorder affecting 0.5-1% of the adult population[11]. The most notable symptom of RA is persistent inflammation of the synovial tissues, leading to swelling at the flexible joints and eventual destruction of the surrounding cartilage and bone. Damage to these tissues causes loss of joint function and severe disability[12] in RA patients, while increased bone fragility and comorbid cardiovascular disease are thought to be major contributors to increased mortality[13-15]. Although various treatments are available for the management of RA[11,16,17], the disease is currently incurable and rarely resolves spontaneously. Overall, RA is a debilitating disease that causes significant burden: it is estimated that RA decreases expected lifespan by up to 10 years[18,19].

RA susceptibility is thought to involve the complex interplay of environmental risk factors, genetic risk factors, and autoimmunity[20]. The best understood environmental trigger for RA is smoking, which nearly doubles disease risk and affects RA severity in a dose-dependent

manner[18]. Estimates of RA heritability from twin studies range from 50-65%, about half of which is explained by known disease-associated variants[21,22]. The strongest genetic associations for RA are observed at the major histocompatibility complex (MHC) locus, driven by variants in the genes *HLA-DRB1*, *HLA-DPB1,* and *HLA-B*[23]. The MHC, located on chromosome 6, is consists of genes that encode molecules involved in antigen presentation; thus, genetic variants in this region are likely to play a critical role in distinguishing self from nonself. Although roughly three times as much phenotypic variance in RA is explained by MHC associations as opposed to non-MHC associations, numerous genome-wide association studies (GWAS) have identified over a hundred other RA risk loci outside of the MHC[24-26].

The genetic signature of RA can potentially highlight the specific immune systems that are affected by disease; in this case, associations both in the MHC and outside of it indicate a significant role for a specific compartment of the adaptive immune system. *HLA-DRB1* and *HLA-DPB1* are components of the MHC class II molecule, which antigen presenting cells use to present antigens to CD4+ T cells. Polymorphisms in this locus affect the range of antigens that MHC class II molecules can bind and present in order to activate CD4+ T cells[23,27]. Genetic risk alleles outside of the MHC locus also point to a role for CD4+ T cells, playing important roles at various points in pathways important for T cell activation, for the differentiation of regulatory (T_{reg}) and effector (T_{eff}) cell subsets, and for maintenance of subset identity[25,28]. In addition, non-MHC RA risk loci are enriched among genes preferentially expressed in effector memory CD4+ T cells over other immune cell types[29,30]. CD4+ T cells are frequently found infiltrating the synovium in RA, often in dense lymphocyte aggregates[31,32]; interfering with T cell activation by blocking costimulatory signals with abatacept (CTLA4-Ig) is an effective therapy for clinical RA[33].

While it is clear that CD4+ T cells play an important role in promoting RA pathology, pinpointing the specific T cell phenotypes or functions that are most relevant in this disease has been challenging. CD4+ T cells are typically categorized by the level of expression of surface and

intracellular proteins that reflect functionally distinct cell types[34,35]. In response to cytokine stimulation, naïve CD4+ T cells are polarized into effector and regulatory subsets that activate specific gene expression programs, produce cytokines and other signaling molecules, and perform different functions[36-38] (Figure 1-2).

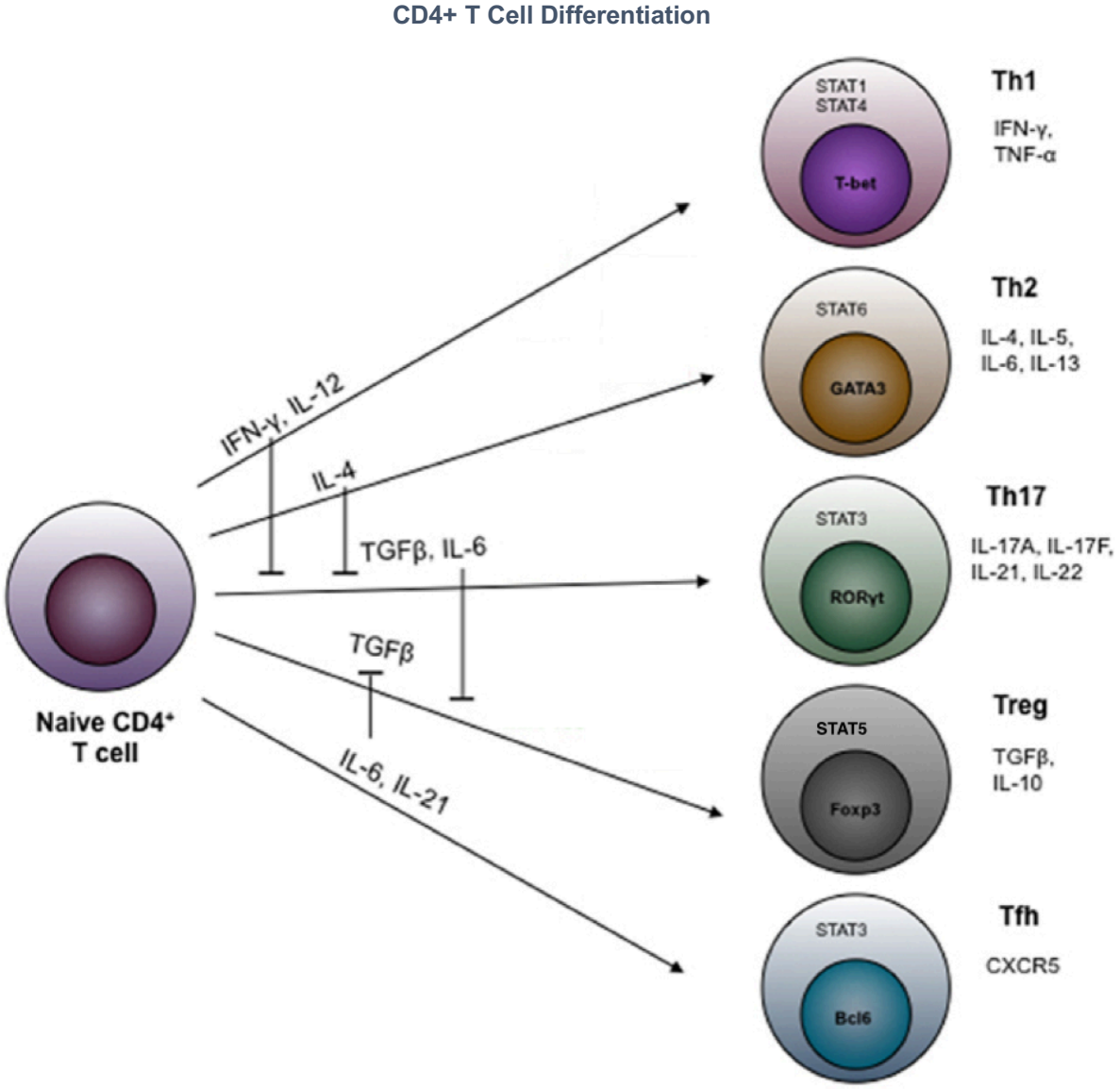


Figure 1-2. An overview of CD4+ T cell differentiation. The addition of cytokines and other molecules labeled along each line promotes the differentiation of naïve CD4+ T cells into a specific subset; as shown by the blocking arrows, some of these molecules can repress differentiation into other subsets as well. The proteins labeled inside each cells' nucleus represent the lineage-defining transcription factor for that subset (along with the corresponding STAT signaling molecule), while the molecules listed to the right are specifically produced by that subset. (Adapted from Sethi *et al.*, 2012)[37].

There is considerable evidence that the activity of pro-inflammatory T_{eff} cells and suppressive T_{reg} cells is dysregulated in rheumatoid arthritis[39-41]. At various times, studies have implicated an imbalance in the abundance and capacity of T helper 1 (T_h1), T helper 2 (T_h2), T helper 17 (T_h17), and T_{reg} subtypes[42-48]. Yet the fundamental cell types and mechanisms that underlie the pathogenesis of RA remain unclear. In part, this is because much of what is known about the role of CD4+ T cells in autoimmunity has been derived from animal models of RA[49]. While these models may adequately replicate the phenotypical characteristics of the disease, the differences between the two species' immune systems mean that cell types or functions that may be directly relevant for human pathology could play a minor role or be completely absent in the mouse[50-52].

Moreover, CD4+ T cells are highly heterogeneous, displaying diverse combinations of surface markers and effector functions. This heterogeneity makes it difficult to describe T cell infiltrates as bulk populations; such analyses are liable to miss potentially relevant cell phenotypes because they are rare or transitory in the sample. This heterogeneity also explains the extent of contradictory literature on the function and role of CD4+ T cells in RA – for example, it may indeed be true that in two separate experiments, T_{reg} cells appear to be compromised or functional, solely due to the proportions of cells sampled in the bulk mixture and the markers used to define the populations. The highly complex and plastic nature of these cells means that pinpointing the specific T cell phenotypes or functions that are most relevant in this disease has been challenging and has highlighted the value of single cell analyses to resolve the diverse D4+ T cell compartment.

The recent rapid expansion of single cell technologies has led to a dramatic advance in the ability to study complex populations in large-scale with high dimensionality (Figure 1-3). This high-dimensional single cell profiling may lead to the identification of specific T cell populations or states that are mechanistically linked to disease and ideal for therapeutic targeting. The following section reviews recent advances in single cell immunoprofiling and

describe their early application in RA in advance of the next chapter describing the necessary methodological and bioinformatic considerations to maximize the potential of single cell technologies in its application to define mechanisms of immune-mediated diseases[53].

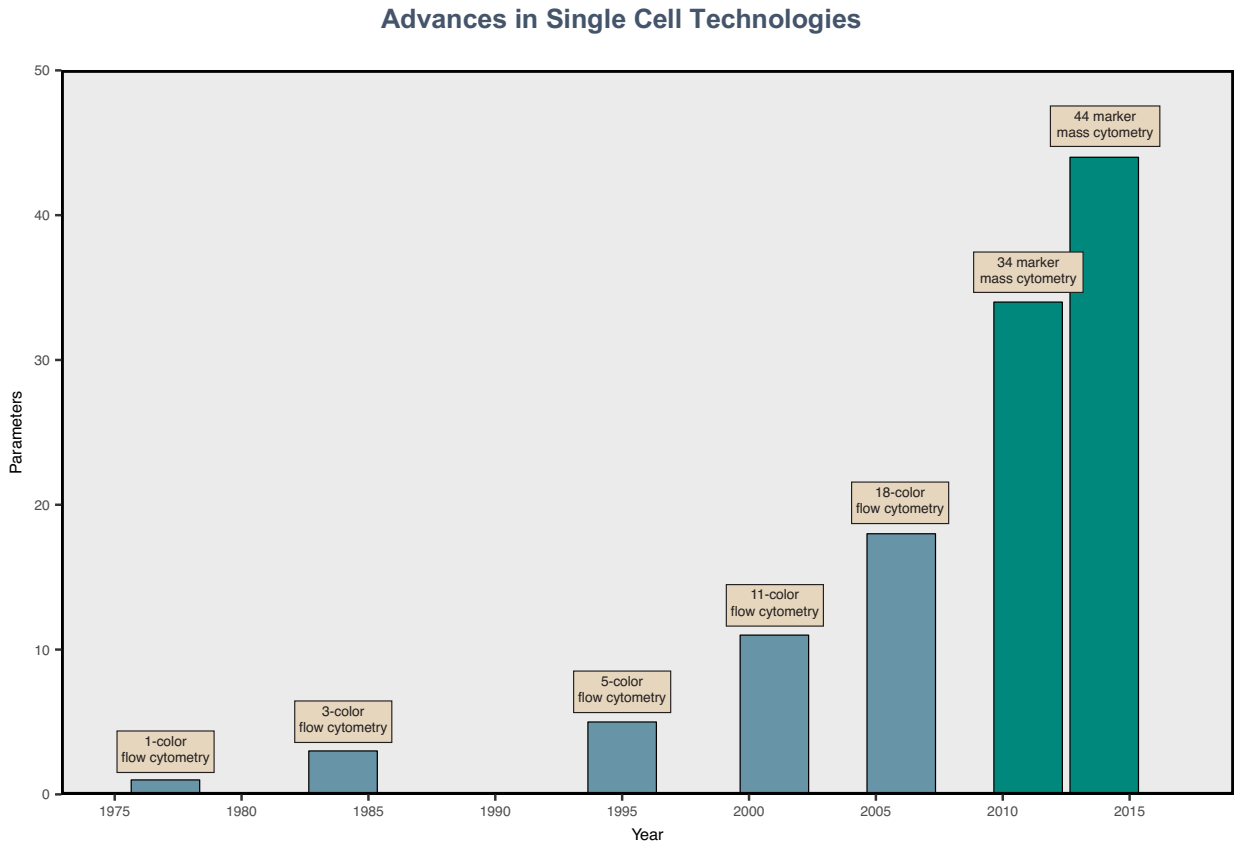


Figure 1-3. The number of unique molecules that can be simultaneously characterized for a single cell has progressively increased. The introduction of new fluorchromes has improved polychromatic flow cytometry and enabled the development of 18-color assays. Mass cytometry, which uses stable isotopes of non-biological rare earth metals linked to antibodies to detect protein epitopes, is currently capable of acquiring 44 markers simultaneously. Current equipment for experiments are limited by the availability of isotopically pure reagents.

Low-dimensional single cell analysis of T cells in RA

Single cell assays have a long history in the field of autoimmunity, beginning in 1969 with the initial use of fluorescent assays to label and sort immune cell populations[54-57]. Cytometry has been thoroughly exploited in the exploration of lymphocyte heterogeneity in

RA[58-62]. Subsequent improvements in flow cytometry technology have steadily increased the number of parameters that can be measured for each cell, provided access to cytoplasmic and nuclear protein expression through intracellular staining, and facilitated measurement of cell signaling using antibodies specific for the phosphorylation state of signaling molecules[63]. Flow cytometric analyses of T cells from RA synovial tissue and fluid have highlighted the dramatic 'activated' phenotype of T cells within the RA joint, consistent with an ongoing autoimmune response directed at the synovium[64,65]. Synovial T cells frequently express CXCR3, suggesting Th1 differentiation, and loss of CD27, suggesting a chronically activated state[66-68].

Immunophenotyping of peripheral blood CD4+ T cells from RA patients has also identified characteristic changes, including expansion of Th17 cells relative to Tregs[40,69], and an expansion of CD28- T cells[61,70]. Unfortunately, studies of peripheral blood T cells in RA have often yielded inconsistent results. For example, the abundance of T_{reg} cells in RA peripheral blood has been observed to be reduced or expanded compared to healthy controls in different studies[71-75]; in addition, conflicting results have been reported concerning the suppressive capability of T_{reg} cells in RA[41,76-79]. While single cell experiments can overcome the limitations of bulk assays of heterogenous populations, some of this inconsistency is rooted in methodological issues that will need to be addressed as investigators begin to apply single cell technologies to autoimmune diseases. Specific issues have included the use of small sample sizes, variability in cohorts, technical noise resulting in batch effects, publication bias, and the lack of principled statistical methodology and criteria.

High-dimensional analyses reveal an expanded view of CD4+ T cell heterogeneity.

The recent development of mass cytometry - a fusion of mass spectrometry and flow cytometry that is capable of the simultaneous acquisition of over 40 parameters on a single cell level – has further extended the dimensionality of single cell cytometric assays[80]. Mass cytometry relies upon staining cells with the same target-specific antibodies that are commonly

used in flow cytometry to tag markers of interest; however, in mass cytometry antibodies are labeled with pure, non-radioactive rare earth isotopes instead of fluorescent proteins. After staining, single cells are analyzed by a time-of-flight mass spectrometer by integrating the detection of heavy metal reporter ions to determine expression levels for each labeled antibody[81-83].

Single cell immunoprofiling by mass cytometry has already been used to reveal remarkable heterogeneity within conventional T cell subsets. Wong et al. used mass cytometry to profile CD4+ T cells across eight human tissue types and described 75 different populations, including multiple T_h1 populations for each T_H subset. Many cell populations were tissue-specific and differed based the expression of trafficking receptors and cytokine production[84]. They observed that certain populations co-expressed "key" cytokines like IFN-γ, IL-4, and IL-17A that are typically restricted to a single CD4+ T_H subset, in line with previous findings highlighting the phenotypic plasticity between CD4+ T_H lineages[85-88], reviewed in[36]. Other studies have taken advantage of high-dimensional single cell mass cytometry analysis to describe multiple populations of T_{REG} and T_{FH} cells[89,90].

While advances in flow cytometry and mass cytometry enable users to define single cells across many parameters, the set of proteins to be measured must be decided *a priori*, limiting the use of these technologies in unbiased discovery studies. In contrast, single cell transcriptomic analysis presents an opportunity to define single cell expression profiles without relying on prior knowledge. Several different single cell RNA-seq (scRNA-seq) methods have been developed over the past decade[91-95] and successfully applied in various immunological studies, such as identifying differentiation pathways in immune cell lineages[96,97], establishing novel transcriptional regulatory networks[98], and revealing functional diversity among lymphoid cell populations[99,100].

Single cell RNA-seq technologies provide an orthogonal approach to cytometry-based methods for establishing CD4+ T cell heterogeneity. As CD4+ T cell subsets are differentiated by

their putative functionality, quantifying of transcript expression on the single cell level can be used to identify gene expression programs that underlie those functional divisions. Single cell sequencing of T cells isolated from patients with liver cancer identified 11 distinct CD4 and CD8 T cell populations, some of which were expanded in hepatocellular carcinoma and marked by specific gene signatures[101]. The functional diversity of natural killer T (NKT) cells is difficult to characterize using cytometry alone; however, single cell RNA-seq analysis revealed differential patterns of gene expression that resolve NKT subsets and indicate potential functions[102]. Single cell transcriptomic profiling is also particularly useful for understanding T cell differentiation and proliferation, as the expression of key transcription factors and other regulatory genes can be easily ascertained and used to assign cells to differentiation trajectories[103,104].

Early high-dimensional analyses of T cells in RA

These same technologies are already being used in RA tissue and blood to define key features of pathogenic CD4+ T cell populations in RA. For example, mass cytometry was applied to evaluate the heterogeneity of CD4+ T cells that infiltrate RA synovium[105]. This high-dimensional analysis identified a T 'peripheral helper' (T_{PH}) cell population that is markedly expanded in RA synovium, constituting ~25% of synovial CD4+ T cells. T_{PH} cells, characterized as PD-1^{hi} $CXCR5^{-}$ CD4+, display a unique capacity to infiltrate inflamed tissues and enhance local B cell antibody production and differentiation into plasma cells. A preliminary single-cell RNA-seq analysis of a single RA synovial sample also demonstrated the presence of multiple T cell subsets, including a population of peripheral helper T cells, in the RA T cell infiltrate[106].

In a distinct approach, Ishigaki and colleagues used parallel single cell transcriptomics and T cell receptor (TCR) sequencing to identify and analyze expanded CD4+ T cell clones in RA patients[107]. Expanded memory CD4+ T cells in both the synovium and periphery are phenotypically similar in expression to senescent T cells, upregulating Granzyme B and downregulating CD28. Intriguingly, the majority of expanded memory T cell clones did not

belong to the well-defined T_H1 or T_H17 subsets despite their established association with RA[40,42,108]. Although the findings are limited by the small number of donors studied, this study suggests that as yet undefined CD4+ T cell populations may undergo expansion in RA and may be relevant to RA pathology.

One potential benefit of characterizing the extent of CD4+ T cell diversity with high-dimensional analyses is that it may provide a means to differentiate between pathogenic and non-pathogenic variants of known T cell subsets. For example, single cell RNA-seq was used to define a spectrum of pathogenicity for T_H17 cells isolated from mice with experimental autoimmune encephalomyelitis (EAE) and identify key genes involved in the process[109]. Similarly, immunoprofiling of T_{reg} cells in RA described the discovery of a novel senescent-like T_{reg} cell population characterized by the loss of CD28 expression and increased numbers of double stranded DNA breaks. Compared to standard T_{REG} cells, $CD28^-$ T_{reg} cells had impaired suppressive function and produced higher amounts of proinflammatory cytokines IFN-γ and TNF[110].

Identifying biomarkers through cell phenotyping

As the diversity, precision, and cost of therapeutics in RA has increased, the importance of being able to determine the option best-suited for a given patient up front has become increasingly clear. There is now a major need for biomarkers to predict response to therapies with distinct mechanisms of action; however, efforts using multiplexed cytokine profiling and genetic variation have not yet led to clinically applicable tools[111,112]. The increased resolution of single cell assays is an asset for revealing disease biomarkers, as the ability to characterize the diversity of lymphocyte populations can be leveraged to monitor the abundances of multiple populations longitudinally or in a case-control context. Changes in the frequency of disease-associated populations that can be easily measured in peripheral blood can be used as a powerful readout of disease state in less accessible compartments.

Several studies have suggested the potential ability to identify specific lymphocyte populations whose peripheral frequencies are predictive of treatment response in order to guide therapeutic decisions. Tracking CD4+ T cell populations by flow cytometry in patients with early RA receiving methotrexate and healthy controls revealed that higher abundances of naïve CD4+ T cells are significantly associated with increased chances of remission[113]. Response to treatment with tocilizumab, an IL-6 receptor inhibitor, is associated with higher baseline frequencies of natural killer ($CD3^{-}CD56^{+}$) cells[114] and higher increases in the frequencies of T_{reg} cells in the periphery[115]. A case-control study of RA patients and healthy controls demonstrated that IL-10+ producing $LAG3^{+}$ T_{reg} cells are specifically increased after treatment with abatacept, and that the magnitude of this increase is correlated with the strength of response[116]. Immunoprofiling studies have also revealed changes in the function of lymphocyte populations in response to therapy: for example, RA patients who respond well to anti-TNF treatment have higher production of GM-CSF from T cells[117]. Response to TNF inhibition therapy is also associated with a higher abundance of CD8+ T cells that are specifically reactive to apoptotic epitopes[118]. Studies such as these fuel hope for the development of predictive cellular biomarkers, though none have been prospectively validated and adopted for use clinically to date.

Recent advances in availability and throughput have made single cell technologies a practical choice for conducting immunoprofiling studies to understand mechanisms of disease and define predictive biomarkers. The application of these methods in RA include the profiling of blood, as many studies referred to above already have done, but also performing immunoprofiling in human tissue. For human immunology to successfully leverage the large quantities of observational data that emerge from single cell queries of the immune system, we will need to develop and reliably apply robust statistical methods and study design principles in single cell studies. Taking full advantage of the power of single cell analysis will require overcoming technical, methodological, and bioinformatic challenges.

Chapter 2:

Methodology of Single Cell Analysis

The analysis of single cell data is complicated by a unique set of factors not typically considered when conducting analyses of bulk transcriptomic or proteomic data. Unlike in bulk analyses, where variation among particular cells is masked within each sample analyzed individually, single cell experiments are sensitive to cell-to-cell differences. This variation represents both variation due to biological differences between cells and variation due to technical effects that affect how well or poorly the cell itself is analyzed. Thus while bulk analyses allow for the assumption that differential informativeness of cells will be averaged within an experiment, the proper analysis of single cell data requires the application of novel pre-processing and analysis methods to maximize the biological information captured in the experiment. For example, given that single cell data is particularly vulnerable to batch effects, a good analysis methodology will rely on both good experimental design and post-assay analytic techniques to maximize the power of the data. Important factors to consider in design include ensuring that samples are collected from the same source, handled in the same fashion, and assayed using the same protocols to the extent that it is possible. Ideally, samples would be prepared using the same lot of reagents; however, this can be difficult to achieve, and steps such as RNA preparation or antibody staining should be performed in a limited number of batches. Given that large-scale association studies typically require performing assays in batches, sample randomization is crucial. Mixing cases and controls within each batch guards against the possibility of discovering biological associations that are perfectly confounded with batch, which can be difficult to account for *post-hoc*; if possible, it is also best to randomize samples in respect to other known factors that may confound analyses, such as samples that can be stratified by sex or medication history. Moreover, sample processing is best conducted in a short window of time and using the same equipment to minimize technical variation to the greatest possible extent.

The choice of tools for computational analysis of high-dimensional data is another important consideration in conducting single cell immunoprofiling studies. Although produced

using very different technologies, both transcriptomic and cytometric single cell data can be analyzed similarly by treating the data as matrices where rows represent single cells and columns represent expression measurements for transcripts or proteins. Beyond just dealing with batch and technical effects, other important issues must be dealt with prior to performing single cell analyses. In the context of studying disease association, analysis of single cell immunoprofiling data can be split into two steps: clustering, where the goal is to identify groups of cells that are related by similarity of expression, and association testing, where the goal is to determine significant changes in the abundance or character of immune cell populations in disease.

Addressing batch and technical effects

Among the many considerations that must be taken into account when designing single cell immunophenotyping experiments, one of the most prominent is determining how to handle batch effects. Here we use the term 'batch' to refer to a set of samples processed together in a single experimental run, and the term 'batch effect' to refer to variation in a dataset caused by technical variation in the processing of different batches of samples. Large-scale microarray assays powerfully illustrated the dramatic effects that differences in machine sensitivity, preparation or handling of samples, or protocol variations can have on the results of transcriptomic analyses[119-122]. Single-cell technologies such as mass cytometry and scRNA-seq are even more vulnerable to confounding from batch effects due to extensive intra-individual and inter-individual heterogeneity of expression among single cells. Application of single cell profiling to human tissues, where cases and controls may respond differently to sample processing and manipulation, could provide an additional source of batch effects.

Indeed, Hicks et al. has demonstrated that variable detection rate and other technical effects account for much of the "biological" variation that was presented in some of the early single cell transcriptomic studies[123]. Careful experimental design can partially alleviate the

influence of batch effects in single cell profiling studies; however, others have shown that common normalization methods for scRNA-seq like spike-in controls and the use of unique molecular identifiers (UMIs) are insufficient for fully removing technical variation[124]. For single cell transcriptomic studies, critical steps include applying quality control methods to remove poorly captured cells and quantifying transcripts to determine cell expression levels. In single cell cytometry studies, quality control is often performed by selecting cells for analysis based upon forward and side scatter parameters (flow cytometry) or DNA content (mass cytometry) and inclusion of a live/dead marker, while marker expression quantification is normally provided by onboard software.

However, since batch variability is difficult to completely eliminate *post hoc*, careful experimental design is essential. First, the importance of minimizing variation in experimental procedure cannot be overstated. Best practices include ensuring that samples are collected from the same source, handled in the same fashion, and assayed using the same protocols to the extent that it is possible. Ideally, samples would be prepared using the same lot of reagents; however, this can be difficult to achieve, and steps such as RNA preparation or antibody staining should be performed in a limited number of batches. Second, as large-scale studies typically require performing assays in batches, sample randomization is crucial. Interspersing cases and controls within each batch guards against the possibility of discovering biological associations that are perfectly confounded with batch. Finally, ensuring that sample processing is done in a short window of time and that samples are assayed using the same equipment also minimizes technical variation. For example, the AMP RA/SLE network significantly reduced batch effects by processing and assaying samples in a single location, as opposed to trying to analyze data obtained at different sites[125].

Applying quality control methods to single cell data

In Fonseka, Rao, *et. al.*, I analyzed single cell mass cytometry data acquired from 26 RA case and osteoarthritis control peripheral blood samples[126]. This study – including the specific settings used in to perform quality control – will be described in further detail in upcoming chapters, but briefly: we obtained peripheral blood samples from cases and controls, purified the samples for CD4+ memory T cells (T_{mem}) using negative selection, and then assayed the samples with a 32 marker mass cytometry panel. We divided each sample and performed non-antigenic stimulation with anti-CD3/anti-CD28 beads, yielding two experimental conditions that we labeled resting (rest) and stimulated (stim). Upon investigating the single cell protein expression data, I found that technical and batch effects confounded my ability to detect biological differences in CD4 T_{mem} populations between cases and controls, despite explicitly balancing the numbers of cases and controls across stimulation conditions in each batch.

To mitigate the influence of batch effects and spurious clusters, I first removed poorly recorded events and low-quality markers before further analysis. I removed those markers (1) that have little expression, as these markers are not informative, (2) that were either uniformly negative or positive across batches, as this indicated that the antibody for that marker was not binding specifically to its target, and (3) with significant batch variability. I concatenated samples by batch and measured the fraction of cells negative and positive for each marker, then calculated the ratio of between-batch variance to total variance for each marker's negative and positive populations, allowing us to rank and retain 20 markers that were the least variable between batches (Figure 2-1). For single-cell transcriptomic data, an analogous step would involve removing genes with low numbers of supporting reads or genes whose expression varies widely between batches.

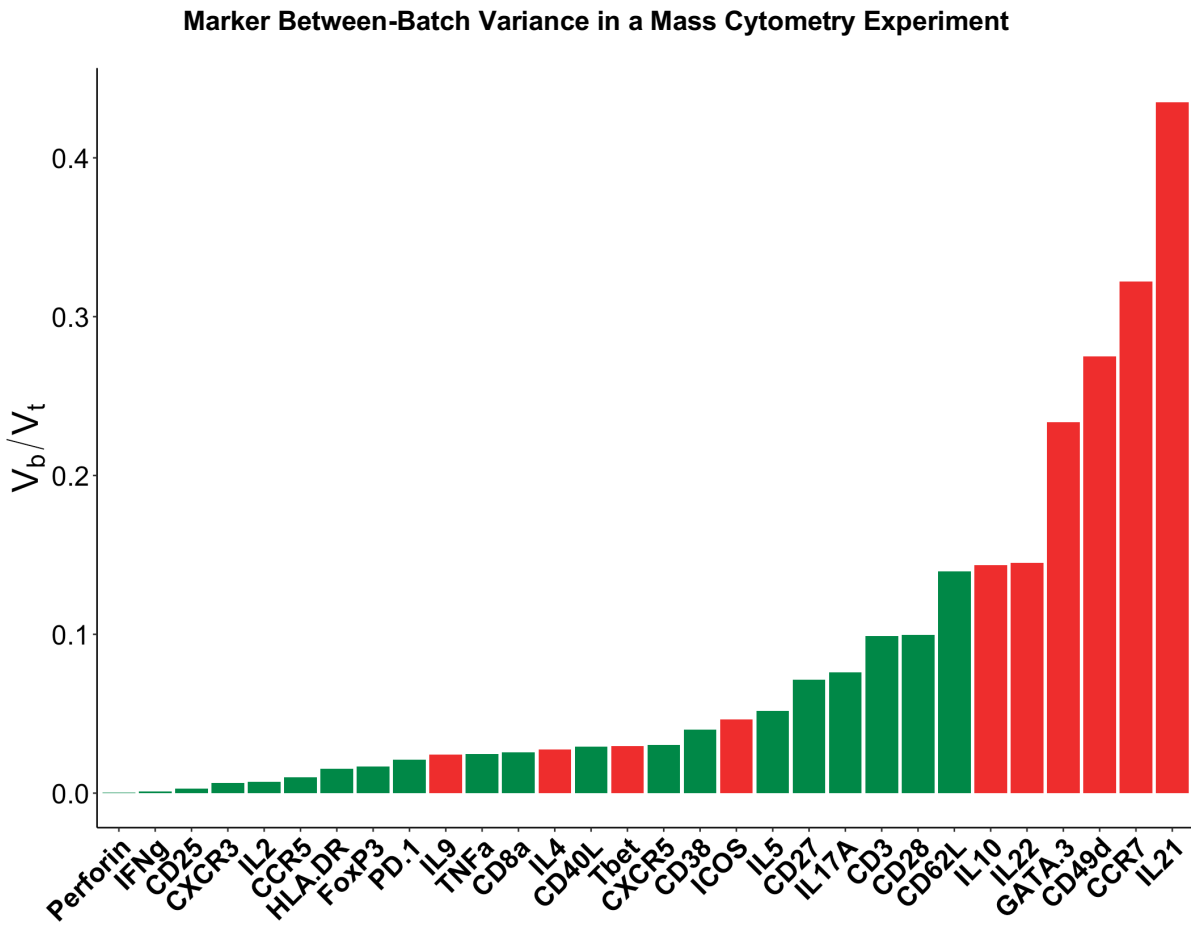


Figure 2-1. The ratio of between-batch variance to total variance for 30 markers used in Fonseka, Rao *et. al.* The top six markers were removed from downstream analysis as they showed a high amount of their variance was driven by batch and confounding biological case-control differences. Four other markers were also removed as they were uniformly negative or positive across batches and samples, indicating that the marker was uninformative or that the antibody was not demonstrating specific binding for its target, respectively.

Once low-quality markers were identified and removed, I removed events that were likely to be artifacts. I first removed events that had extremely high signal for a single marker: events that have recorded expression values at or above the 99.9th percentile for that marker are removed. These events were considered unlikely to be intact, viable cells, given that their measured protein expression was orders of magnitude higher than normal. Next, a composite "information content" score (eq. 1) for each event i was created in the following manner: the expression x for each marker M is rescaled from 0 to 1 across the entire dataset to create

normalized expression values y_i for each event i. The sum of these normalized expression values was used to create the event's information content score.

1) $INFO_i = \sum_{m=1}^{m=M} y_{i,m}$

The information content score reflects that events with little to no expression in every channel are less informative than events that have more recorded expression. Events with low scores ($INFO_i < 0.05$) were considered unlikely to be informative in downstream analysis and were removed. In addition, events that derived more than half of their information content score from expression in a single channel were also removed (eq. 2):

2) $INFO_i * 0.5 < \max_{m \in M}(y_{i,m})$

Potential explanations for these events include poorly stained cells or artifacts caused by the clumping of antibodies with DNA fragments. These antibody-DNA clumps would pass the quality control metrics inherent to the CyTOF 2 platform and preliminary gating as they resembled real cells based on DNA content. A final filtering step retained events that were recorded as having detectable expression in at least M_{min} markers, where M_{min} may vary from experiment to experiment based on the panel design and expected level of co-expression between channels. In this experiment, because we had isolated samples for CD4+ T_{mem} cells, we expected considerable co-expression between different markers in the panel, and could use the lack of co-expression as an indication of which cells were of low complexity and informativeness. The quality control steps described here are specific for mass cytometry analysis and need to be optimized separately for use with transcriptomic data.

Clustering methods for single cell data

As previously stated, the goal of clustering algorithms is to group single cells into biologically meaningful populations using some metric of similarity, principally gene expression for single cell transcriptomic data and protein expression for single cell flow and mass cytometry

data. There are a wide-range of clustering algorithms available with different sets of parameters; however, perhaps one of the most important features of any clustering algorithm is whether eventual cell-cluster assignment relies on "hard" or "soft" clustering. While an algorithm that uses hard clustering will produce a set of one-to-one assignments between cells and clusters, a soft clustering assignment will typically provide the probability of a given cell belonging to all clusters. Therefore, soft clustering algorithms allow for a cell to be assigned to multiple clusters at the same time, albeit with differing levels of confidence. Another important feature of clustering algorithms is the number of clusters detected; while algorithms like *k-means* clustering allow for this number to be explicitly chosen, others use one or more hyperparameters to control the eventual number of clusters in an indirect fashion.

While many different algorithms have been applied to the analysis of single cell data, the following methods represent some of the state-of-the-art tools for performing single cell immunophenotyping studies. Seurat is an R package that contains multiple methods for clustering and visualizing single cell sequencing data, as well as performing differential expression testing between groups and finding associations[127]. It is currently widely used in single cell RNA-sequencing studies Multiple clustering methods have been developed for the analysis of flow cytometry[128,129] and mass cytometry[81,130-134]; a recent comparison of these methods identified FlowSOM[130] and PhenoGraph[132] as the best performers[135]. In Fonseka, Rao *et. al.*, I performed clustering after preprocessing data from each sample using the quality control metrics described previously. After applying quality control measures to each sample, I combined data from cases and controls into a single dataset. It was critical to ensure that each sample contributed equal numbers of cells to this dataset, as otherwise the largest samples would dominate the analysis and confound association testing. After sampling an equal number of cells from each sample, I partitioned these cells into populations using the DensVM algorithm[134]. This clustering algorithm requires as input a dimensionally-reduced version of the

expression data alongside the expression data itself; next, it uses a kernel density estimator to identify peaks of density on the dimensionally-reduced projection over a range of bandwidths (effectively, the width of the peak in two dimensions). This step returns a set of clusterings at each bandwidth tested; the "correct" bandwidth is then determined using the elbow method – that is, by identifying the first place where increasing the bandwidth yields the same number of peaks. After selecting a set of peaks, the algorithm then treats uses the clusters defined by these peaks as training data for a support vector machine model that assigns all cells into clusters based upon the full expression matrix.

In order to compare the robustness of the clustering result found by this algorithm, I performed additional clustering using two other algorithms, FlowSOM and Phenograph. We independently clustered the resting dataset with Phenograph and FlowSOM using the same cells and markers used to cluster the data with DensVM. We set *k* to 19 for FlowSOM clustering to explicitly match the number of clusters found by DensVM; for Phenograph, we used the default setting of k = 30 for the resolution hyperparameter (importantly, this does *not* ask the algorithm to produce 30 clusters). To evaluate quantify the ability of different clustering algorithms to define clusters that was explaining marker fluctuations, I defined an information theory-based metric to evaluate the relative information content captured by each set of clusters in terms of marker intensity, which I named the Cluster Informativeness Metric (CIM). I selected this approach since it is separate from the objective functions that the clustering algorithms were attempting to optimize.

First, for each cell, I normalized marker intensities so that they summed to one. Then I defined a null Q_i representing the average normalized intensity for marker i across all cells. I also defined $P_{i,j}$ which is the mean intensity of marker i of cells from cluster j. Then for each cluster j I calculated their KL divergence for each of the M markers (eq. 3).

3) $D_{KL,j}(P_j \parallel Q) = \sum_i^M P_{i,j} \ln \frac{P_{i,j}}{Q_i}$

A cluster with low divergence from the average expression of markers across the entire dataset will capture less marker intensity information than one with a high divergence, as biologically valid clusters will have unique marker profiles that differ greatly from one another and from the average marker expression profile.

I defined a similar metric to quantify the extent to which individual batches were accounting for differences in cluster composition. In this instance I calculated $P_{i,j}$ which is the proportion of cells from cluster j that batch i contributed. I also calculate Q_i which is the proportion of cells that batch i contributes overall to the dataset. With this definition I can calculate the KL divergence for each of the M batches (eq. 4).

4) $D_{KL,j}(P_{\cdot j} \parallel Q) = \sum_i^M P_{i,j} \ln \frac{P_{i,j}}{Q_i}$

A cluster that contains cells with low divergence from the null distribution of cells across batches is affected less by batch effects than one with a high divergence score, and a cluster completely free of batch effects should have a K-L divergence of zero. Thus, this approach allowed me to evaluate clustering methods agnostically and measure their ability to identify biologically distinct populations that were not driven by confounded batch effects. For this study, I demonstrated that the DensVM method produced biologically informative clusters with less influence from batch effects than the other two algorithms; however, this finding is unique to the dataset in question and not necessarily generalizable (Figure 2-2).

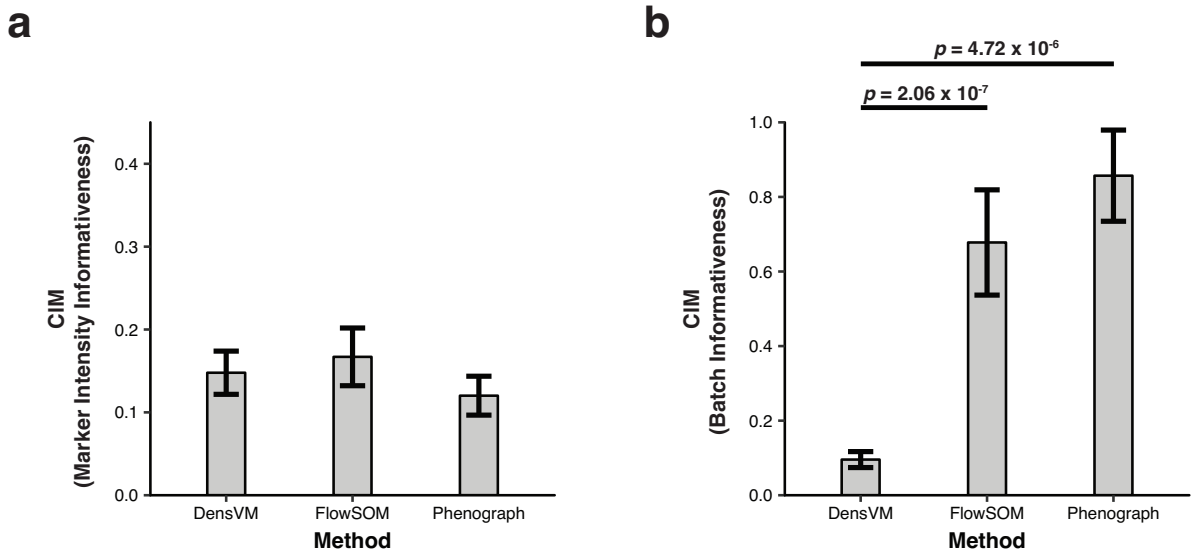


Figure 2-2. We clustered the same dataset using three different clustering algorithms, DensVM, Phenograph, and FlowSOM. These algorithms identified 19 (DensVM and FlowSOM) or 21 (Phenograph) clusters. (a) Clusters found by DensVM, Phenograph, and FlowSOM had similar average CIM scores when considering marker expression, indicating that the clusters found by these algorithms were similarly informative. That is, marker intensities were different from the average marker expression profile across clusters to the same extent. (b) Clusters found by Phenograph and FlowSOM had a significantly higher CIM score when considering batch than those found by DensVM, indicating that the Phenograph and FlowSOM clusters were more affected by batch effects. We assessed significance using a Wilcoxon rank sum test and p-values were Bonferroni adjusted to control for multiple testing.

Performing association testing with single cell data

While the set of algorithms available for clustering single cell data is rapidly expanding, there is a relative paucity of methods designed to perform association testing with single cell data[136,137]. This is a significant issue because inter-individual variation and technical variation can influence cell population frequencies and must be accounted for in an association framework. For example, it is well established that the ratio of naïve to memory T cell proportions shifts with age, with older individuals having a higher frequencies of the latter cell types[138-141]. Consider a hypothetical single cell association study comparing the abundance of T cell populations between two groups that yields a finding that individuals from group 1 have a lower frequency of T_{mem} cells than individuals in group 2. While this differential abundance phenotype appears be associated with group status, it is actually driven by group 2 disproportionally consisting of older individuals and causing an apparent shift in cell

frequencies between groups. Avoiding a false-positive result like this requires a association testing framework capable controlling for both technical effects and the high levels of inter-individual variability in the human immune system.

At first glance, a straightforward approach would be to use a difference-of-means test to determine if the abundance of a given single cell population was associated with a given attribute, like case-control status of the sample. Assume that proper pre-processing and clustering of a single cell dataset has yielded a set of biologically meaningful single cell cluster assignments and that we have data linking each cell to the feature we want to test for association, such as whether the cell comes from a case or control sample. Under this strategy, we would reduce the single cell data of each sample to a set of frequencies for each cluster – that is, we would create a matrix with k samples by n clusters, where each row described the proportion of cells in sample k in each of n clusters. We can then group the samples by the feature we want to test for association and test whether the average proportion of cells in a given cluster is significantly different between cases and controls using a standard parametric or non-parametric difference-of-means test. As an example, suppose we have clustered single cell data from 10 cases and 10 controls into 11 clusters and we wish to determine if the frequency of cluster 9 is different in cases and controls. We can calculate the average proportion of cells in cluster 9 for each sample, then group the sample and perform a t-test (Figure 2-3). Conversely, we could use a binomial testing approach where we measure the number of total cells in the cluster from cases and controls, then compare that to the expected number given the size of the cluster relative to the entire dataset. Under this approach, we determine whether there is a significant association between cluster abundance and case-control status by testing for deviation from a chi-square distribution (Figure 2-4). Note that in the following examples, the Barnes-Hut implementation of the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm has been used to provide a two-dimensional visualization of the simulated data[142].

Performing Single Cell Association Testing Under a Difference-of-Means Framework

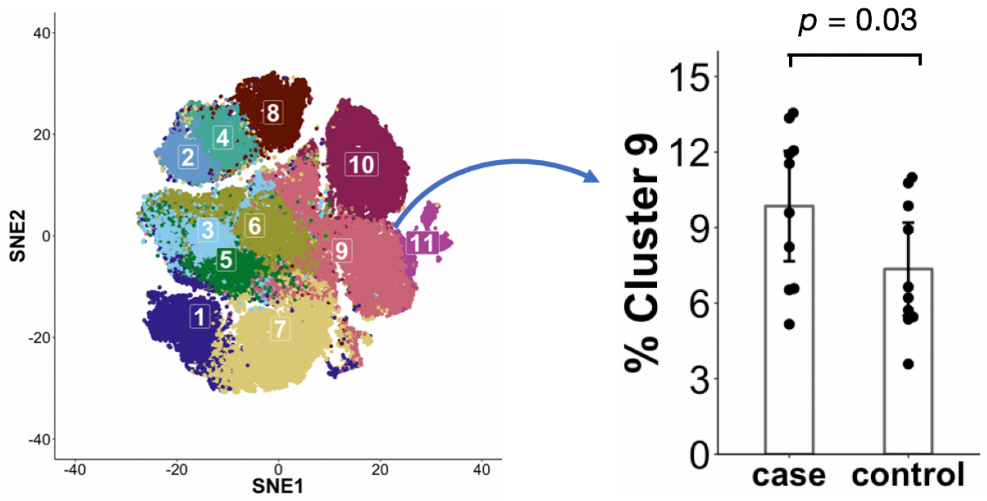


Figure 2-3. In this simulated example of a single cell association study, cells have been combined across 10 case and 10 control samples and clustered into 11 populations. The cells are shown projected into a two-dimensional embedding using the t-SNE algorithm on the left. On the right, each dot represents the percentage of a given sample's cells that were assigned to cluster 9, grouped by case-control status. A two-sample t-test fails to reject that there is no significant difference between the abundance of cluster 9 in cases and controls.

Performing Single Cell Association Testing Under a Binomial Framework

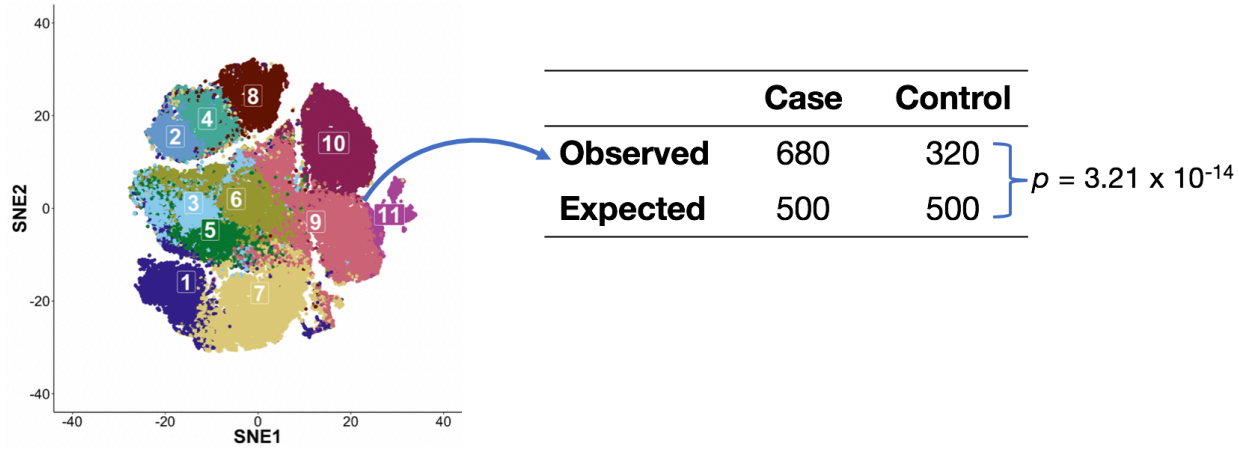


Figure 2-4. Similarly to Figure 2-3, simulated cells have been combined across 10 case and 10 control samples and clustered into 11 populations. The cells are shown projected into a two-dimensional embedding using the t-SNE algorithm on the left. On the right, a two-by-two contingency table shows the number of cells observed and expected to be in cluster 9 from case and control samples. In this simulation, equal numbers of cells were used from each sample so the expectation is that the groups should be balanced. Under a binomial testing framework, the observed number of case cells is significantly more than would be expected.

It turns out that neither of these approaches is particularly appropriate for performing single cell data association testing. Under the difference-of-means testing approach, all single-cell observations are collapsed into per-sample proportions for each cluster. This is problematic for populations that are in low-frequency in the dataset; namely, the error around the point estimate of the cluster's frequency in each sample increases with as the number of observations – here, single cells assigned to the cluster – decreases. Using a binomial framework to perform association testing suffers from inflated Type 1 error; the significant p-values obtained are likely to be false positives because this approach assumes that each cell is an independent measure, which is a poor assumption for single cell data. Given the strong influence of batch effects upon single cell data and high levels of inter-individual variability in the human immune system, it is important that any association testing model is able to take these factors into consideration in a transparent manner. These observations led me to develop a novel statistical method for performing association testing with single cell data, one that would be able to account for technical covariates and inter-individual differences while directly testing for associations between single cells and the outcome of interest.

Mixed-effects modeling of Associations of Single Cells (MASC)

MASC is a 'reverse' association strategy where the case-control status is an independent variable, rather than the dependent variable, and uses mixed-effects logistic regression to test at the single cell level the association between population clusters and disease status. MASC accepts user-identified populations regardless of clustering method, directly reports the significance of case-control associations for each cluster, provides an estimate of the effect size of the association itself, and incorporates both technical covariates (e.g. batch) and clinical covariates when modeling associations, a key feature when analyzing high-dimensional datasets of large disease cohorts. This approach allows for capturing inter-individual differences between donors, as well as modeling the influence of technical and clinical covariates that might

influence a cell to be included as a member of one cluster versus another, allowing the user to directly assess the contribution of these covariates to differential cluster abundance.

Importantly, MASC is not dependent on any particular method of clustering, allowing the user to partition their single cell data using the method of their choice – even by using traditional bivariate gating to define populations using cytometry data. The rest of this chapter will cover the statistical framework underlying MASC, while the following chapters will demonstrate the successes of MASC when applied to single cell association studies of rheumatoid arthritis.

Given a single cell dataset in which all cells have been assigned to a given cluster, the relationship between single cells and clusters can be modeled using mixed-effects logistic regression to account for donor effects and other technical variation (eq. 5). Employing the model used in Fonseka, Rao *et al.* as an example, I was able to model the age and sex of sample k *as* fixed effect covariates, whereas the donor and batch that cell i belongs to were modeled as random effects. The random effects variance-covariance matrix treated each sample and batch as independent gaussians. Each cluster was individually modeled. Note that this baseline model did not explicitly include any single cell expression measures.

$$5) \quad \log\left[\frac{Y_{i,j}}{1-Y_{i,j}}\right] = \theta_j + \beta_{clinical} X_{i,k} + (\phi_i|k) + (\kappa_i|m)$$

where $Y_{i,j}$ is the odds of cell i belonging to cluster j, θ_j is the intercept for cluster j, $\beta_{clinical}$ is a vector of clinical covariates for the kth sample, $(\phi_i|k)$ is the random effect for cell i from kth sample, $(\kappa_i|m)$ is the random effect for cell i from batch m.

To determine if any clusters were associated with case-control status, I included an additional covariate that indicated whether the kth sample is a case or control (eq. 6)

6) $\log\left[\frac{Y_{i,j}}{1-Y_{i,j}}\right] = \theta_j + \beta_{clinical}X_{i,k} + (\phi_i|k) + (\kappa_i|m) + \beta_{case}X_{i,k}$

Here, $Y_{i,j}$ is the odds of cell i belonging to cluster j, θ_j is the intercept for cluster j, $\beta_{clinical}$ is a vector of clinical covariates for the kth sample, $(\phi_i|k)$ is the random effect for cell i from kth sample, $(\kappa_i|m)$ is the random effect for cell i from batch m, β_{case} indicates the effect of kth sample's case-control status.

7) $D = -2 * \ln\left(\frac{\text{likelihood for null model}}{\text{likelihood for full model}}\right)$

8) $p = 1 - \left(\frac{t^{(\nu-2)/2}e^{-t/2}}{2^{\nu/2}\Gamma(\frac{\nu}{2})}\right)$

I compared the two models (baseline and full) using a likelihood ratio test (eq. 7) to find the test statistic *D*, which is the ratio of the likelihoods for the baseline and full models. The term *D* is distributed under the null by a χ^2 distribution with 1 degree of freedom, as there is only one additional parameter in the full model compared to the null (case-control status). I then derived a *p*-value by comparing test statistic *D* of the likelihood ratio test to the value of the χ^2 distribution with 1 degree of freedom (eq. 6), allowing us to find clusters in which case-control status significantly improves model fit. A significant result ($p < 0.05$ after multiple testing correction) indicated that cluster membership for a single cell is influenced by case-status after accounting for technical and clinical covariates. The effect size of the case-control association can be estimated by calculating the odds ratio from β_{case}. Note that if a feature of interest includes multiple groups, then MASC can be used to test for association between *g* groups using *g*-1 indicator variables.

To test the robustness of MASC compared to other association testing approaches, I took the clustered mass cytometry data from Fonseka, Rao *et al.*, and permuted the sample case-control labels 10,000 times to break up any cluster associations with case-control status. I then tested

for case-control associations using MASC and binomial tests for each cluster in each of the permuted datasets and recorded the p-values produced by each method. Since the effect of this permutation strategy is to remove any case-control associations, 5% of trials should obtain p-values below 0.05 purely due to random chance; conversely, higher proportions of trials obtaining p values < 0.05 would indicate that the method has inflated type 1 error. MASC demonstrated only a slight inflation in type 1 error, with 6.5% of trials obtaining $p < 0.05$, while the binomial association testing approach was highly inflated, with 66.1% of trials obtaining $p < 0.05$ (Figure 2-5).

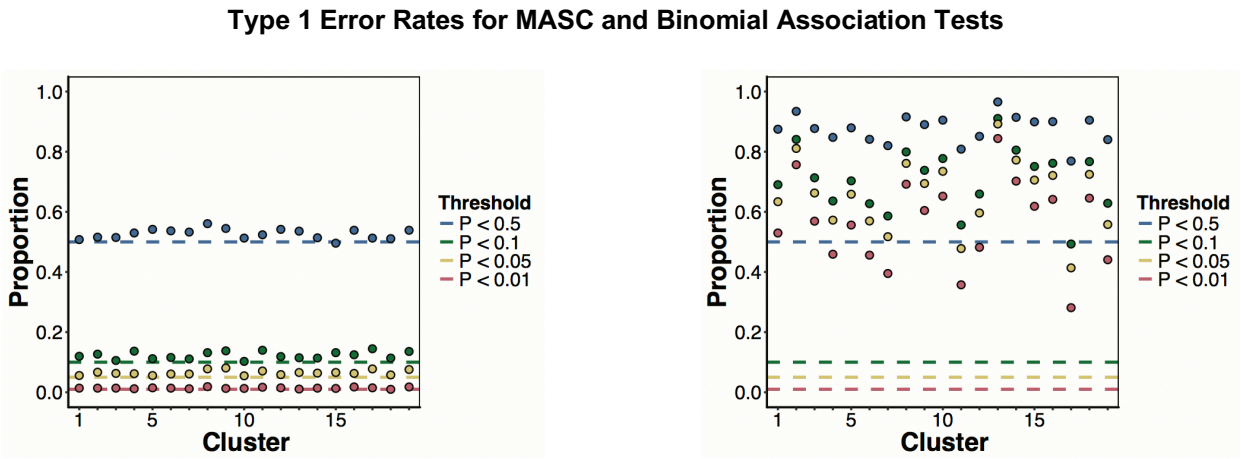


Figure 2-5. MASC demonstrates well-controlled type 1 error rates. (left) MASC was run on the resting dataset after randomizing case-control labels 10000 times to eliminate any case-control associations. The proportion of p-values at different thresholds are plotted for each cluster. (right) P-values obtained in the same manner for binomial association tests on clusters found in the resting dataset.

I also used the same permutation framework (randomizing case-control labels on a sample-by-sample basis) to confirm that the p-values generated by MASC were coherent. In each permutation, we tested each cluster for how often the number of cells from cases was in excess of their observed number and generated explicit permutation p-values for each cluster found in both experimental conditions. This showed that the p-values produced by MASC were concordant with the p-values derived from explicit permutation testing (Figure 2-6).

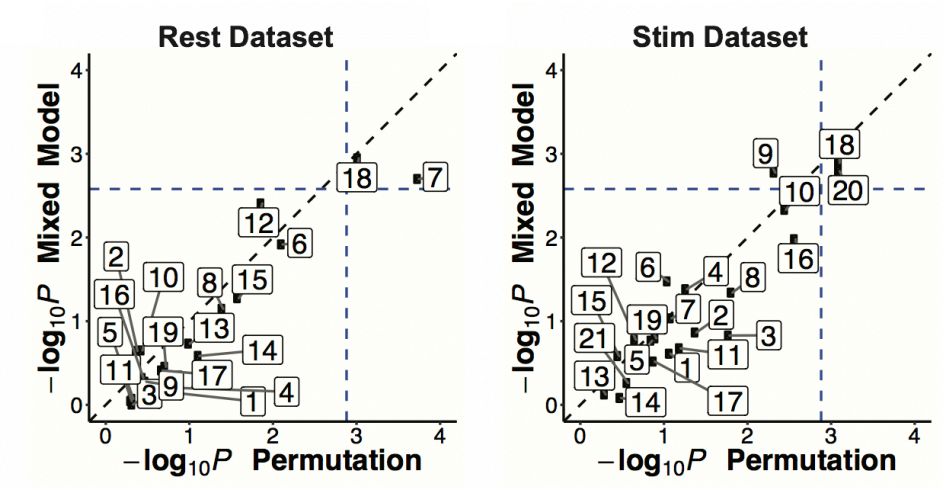


Figure 2-6. Association p-values as calculated by MASC (y-axis) and by explicit permutation (x-axis) correlate in both resting and stimulated datasets. Spearman's correlation coefficients for (left) and (right) were $r_s = 0.82$ and $r_s = 0.86$, respectively.

MASC is a robust framework for performing association testing with any form of single cell data; while it was developed for use with mass cytometry data, it has been successfully applied to single cell transcriptomic datasets as well. Applying MASC to scRNA-seq data requires including a different set of technical covariates in the model – for example, it is useful to include cell-specific metrics of quality, like complexity or read depth. I applied MASC to a single cell transcriptomic dataset from a study that analyzed frozen kidney biopsies of patients with lupus nephritis (LN) and healthy controls[143]. The dataset clustered single cells into 21 leukocyte populations, which I tested for case-control association using MASC. Here, I included the sex of the sample as a technical covariate in the model, alongside two other fixed-effect covariates: the percentage of mitochondrial reads and the number of unique molecular identifiers (UMIs), which serve as measures of complexity. This model also included the sample identifier itself as a random effect to account for interindividual differences. MASC was able to identify six populations that were significantly altered between cases and controls (Table 2-1).

MASC Analysis of Single Cell Transcriptomic Data

Table 1-1. MASC was run on a single cell RNA-seq dataset derived from kidney biopsies of lupus nephritis patients and healthy controls[143]. Out of 21 clusters, MASC identified 6 clusters that were significantly associated with case-control status. The name of the cluster and number of cells are listed, followed by the association p-value generated by MASC, the resulting q values from applying an false discovery rate correction of 5%, and the odds ratio of the case-control status term in the model.

Cluster	Cells	MASC *p* value	MASC *q* value	MASC Odds Ratio
CB0	245	9.06×10^{-6}	1.62×10^{-4}	13515835.1
CB1	81	2.27×10^{-3}	9.09×10^{-3}	49941195.9
CM2	86	1.62×10^{-5}	1.62×10^{-4}	0.1
CT0	220	5.26×10^{-5}	3.51×10^{-4}	0.1
CT2	348	1.30×10^{-4}	6.50×10^{-4}	23.0
CT4	195	5.24×10^{-3}	1.75×10^{-2}	6.8

The results from this MASC analysis indicate that two of the clusters (CM2, CT0) were decreased in lupus nephritis samples compared to healthy controls, while the other four clusters were expanded. Although at first glance, the odds ratios calculated for clusters CB0 and CB1 may seem incorrect given how extreme they are, these results merely reflect the fact that these cell types are effectively exclusive to the LN kidney samples (Figure 2-7). Arazi *et al.* identified these populations as activated B cells (CB0) and plasma cells (CB1), noting that B cells were almost entirely absent from the healthy control samples anyway. MASC also identified two CD8+ T cell populations that were expanded in LN samples: CT2, which resembled cytotoxic T lymphocytes, and CT4, which was described as a second population of CD8+ cells specifically marked by expression of the granzyme *GZMK*. Correspondingly, the T cell population that was depleted from LN samples, CT0, was most similar to naïve CD4+ T cells. Finally, the CM2 cluster that was also identified as depleted from LN samples represents typical kidney macrophages that have relatively similar abundances in cases and controls, but differ in their gene expression programs.

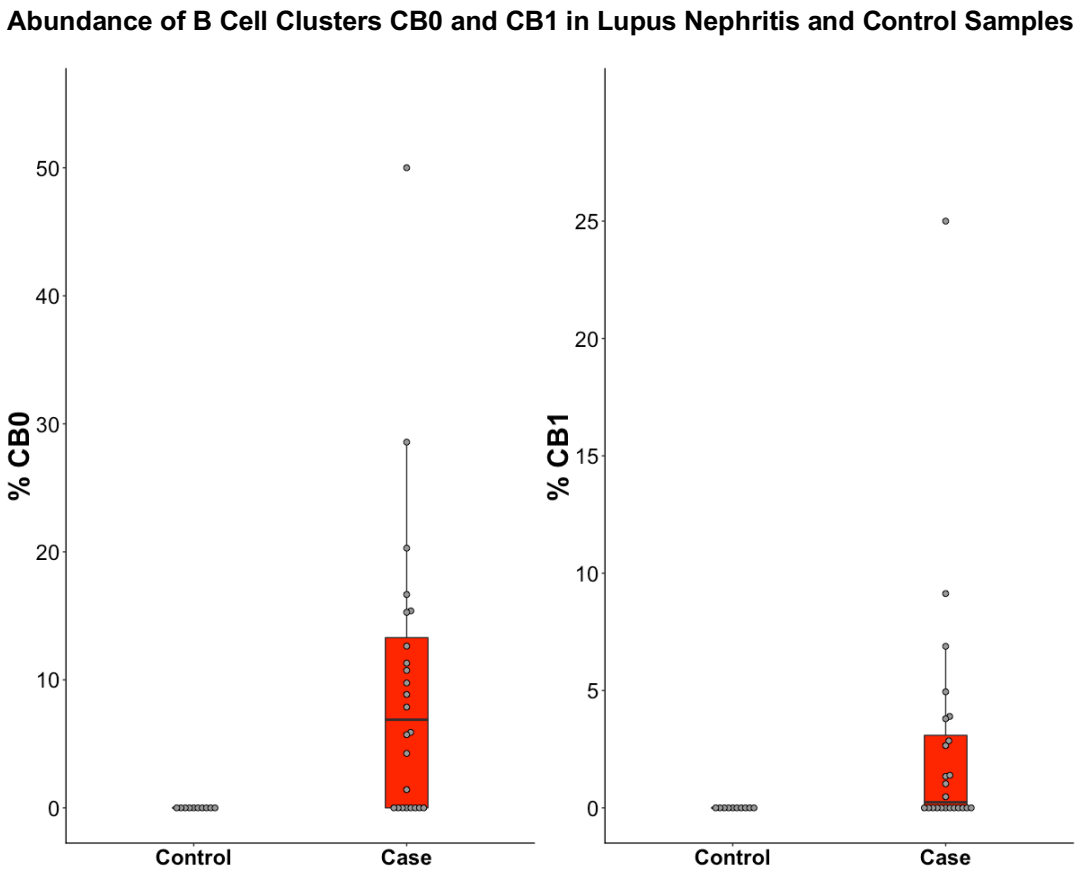


Figure 2-7. Per-sample abundance data calculated for the B cell clusters CB0 and CB1 that were detected as significantly expanded by MASC. Both of these clusters are exclusive to the lupus nephritis kidney samples and were not present in healthy controls. Data is calculated from 24 cases and 10 controls.

In summary, MASC is a single cell association testing method that is adaptable to any case-control experiment in which single cell data are available. Such experimental setups could include cytometry assays, single cell transcriptomics, and even single cell ATAC-seq – all MASC requires is a set of cluster assignments for all the single cells under analysis. Because the framework of MASC is flexible, it supports building models with technical covariates that are important to control for in a specific context or experimental setup. MASC is capable of testing for associations between cluster abundance and any feature of interest, such as disease progression, medication usage, or even patient ancestry. MASC is a useful tool for performing single cell immunophenotyping studies, and its ability to resolve novel subpopulations of T cells and fibroblasts in RA patients will be covered in the following two chapters. However, another

potential application of MASC is using the method on simulated single cell datasets to understand how to best power single cell association testing experiments.

Power analysis of single cell association testing studies

In planning any association testing study, it is useful to have an estimate of the minimum number of samples needed to reliably detect an effect of a given size. Power analyses provide a method of calculating these metrics by simulating data under a varied set of conditions and then measuring how often association tests detect significantly significant results for a given effect size. In the context of single cell association studies, I focus here on calculating the power to detect differential abundance – that is, a change in the frequency of a population or clusters of single cells – and its association with a factor of interest. MASC presented itself as an obvious option to perform these analyses; its flexible framework allows for directly measuring the effects of experimental design choices like the number of donors or batches upon study power.

To test the effectiveness of using MASC in this manner, I started by using the mass cytometry data collected in Fonseka, Rao *et al.* as a model of a single cell study. These data were collected from 26 case and 26 control samples over 10 batches, assayed by mass cytometry, and then analyzed as a concatenated dataset created by randomly sampling 1000 cells from each donor (after pre-processing to remove low-quality and uninformative cells). For this analysis, I created synthetic single cell mass cytometry datasets by randomly downsampling either the number of donors or the number of cells per donor in the dataset. To avoid creating unbalanced replicates when downsampling donors, I ensured that I removed equal numbers of cases and controls each time. I then ran MASC on each synthetic replicate and calculated the proportion of trials in which MASC was able to detect an association between a differentially abundant population, which has an abundance of 2.8% in RA samples and about half as much in controls, and case-control status.

Reducing the number of cells included from each donor while maintaining that the synthetic datasets had 26 case and 26 control donors reduced the power to detect the differentially abundant population, with roughly 300 cells per donor being required to maintain statistical power of 50%. Intriguingly, reducing the number of donors while keeping the number of cells sampled per donor at 1000 cells degraded power much more quickly; in these analyses, removing just two or three donors yielded simulations in which the differentially abundant population could only be detected half of the time (Figure 2-8). While these results could obviously be specific to this dataset and not generalizable, they fit a hypothesis for single cell association testing study design. Especially in the context of human immunology, in which inter-individual differences in population frequencies are fairly large, more donors are needed to detect differential abundance than cells contributed per donor. It is likely that there is a saturation point (which will vary according to experiment) after which additional cells provide minimal information about the true frequencies of cell populations; i.e. estimates of population frequencies within a donor stabilize more quickly than the variance in population frequencies across donors.

To explore this further, I used MASC to conduct power analyses on simulated single cell transcriptomic data, allowing me to directly set the number of donors and cells per donor rather than using downsampling, as well as incorporate varying amounts of batch and donor variance. The simulation strategy I followed relies on an existing single cell transcriptomic dataset – namely, a study of patients in Lima who were either diagnosed as active tuberculosis (TB) progressors or as having latent TB. Performing a power analysis of this TB dataset was particularly of interest because we had received sequencing data generated from 48 donors and wanted to estimate what our power to detect differentially abundant cell populations would be once all 259 donors had been sequenced.

Power to Detect Differential Abundance in a Single Cell Mass Cytometry Dataset

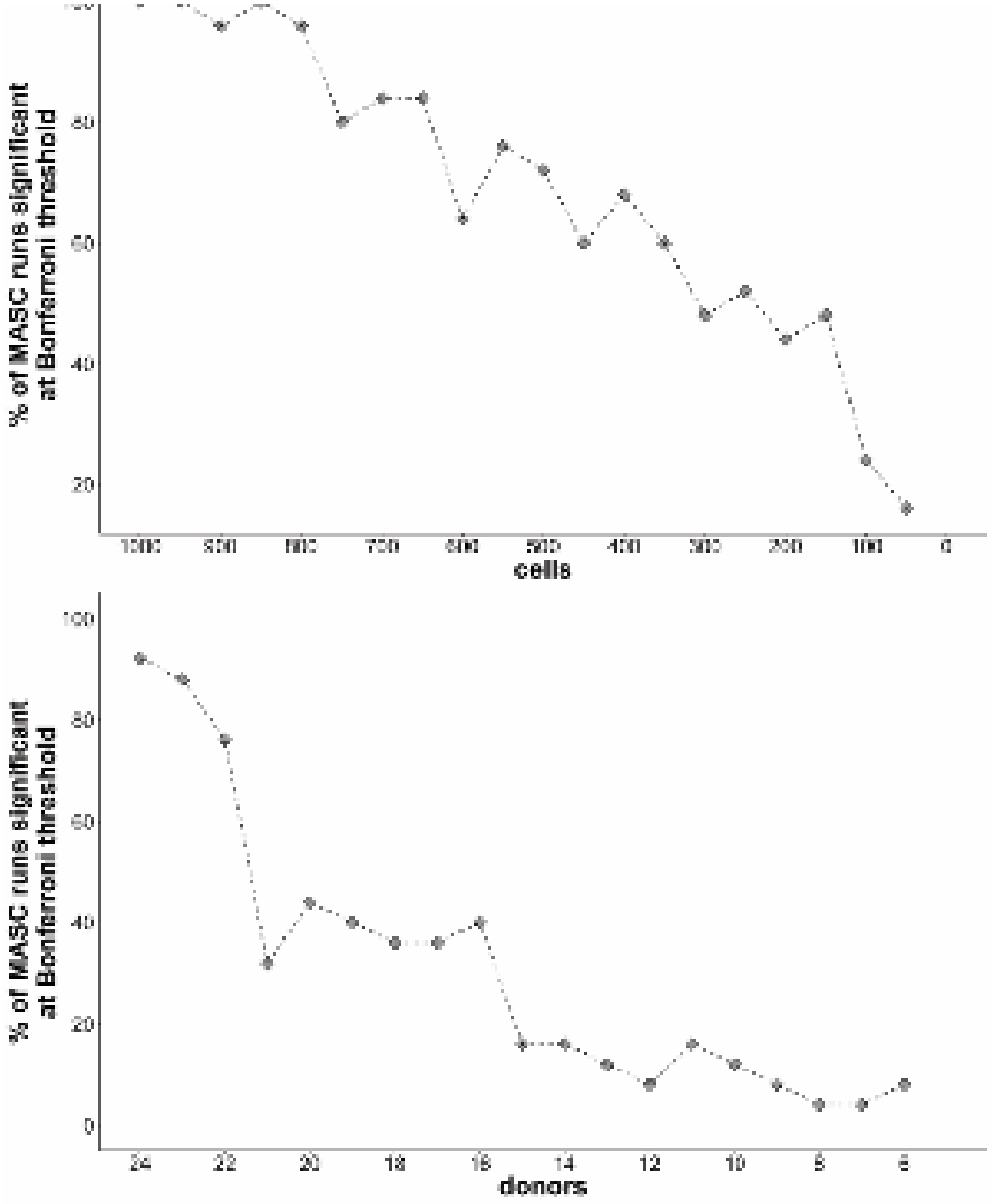


Figure 2-8. The percentage of trials that detected a significant differential abundance case-control association in synthetic replicates of the mass cytometry data used in Fonseka, Rao *et al.* (top) The number of cells contributed by each donor was decreased in synthetic datasets and MASC was used to detect differential abundance for each replicate. (bottom) Same as top, except that the number of donors was decreased in synthetic datasets while the number of cells per donor was fixed. Each datapoint represents a proportion calculated across 25 replicates.

In this experiment, memory T cells were isolated from patient samples by negative selection and subjected to simultaneous mRNA and protein assay following the Total-seq protocol, which uses antibody-tagged oligos to detect protein expression alongside mRNA quantification[144]. The samples were multiplexed prior to sequencing, but because they were genotyped, we were able to use demuxlet to accurately assign cells to samples and remove doublets[145]. After performing standard normalization and QC, the final dataset consisted of 69972 cells, with an average of 1458 cells and 1113 non-zero genes per donor. We then partitioned the dataset into 30 clusters using Seurat[146,147] after integrating the single cell and protein expression data with Harmony[148].

To conduct power analyses with this data, I used the following strategy to create simulated datasets. First, I used principal components analysis to project the data into a low-dimensional representation and then summarized each of the single cell clusters as two dimensional centroids using the first two principal components. All of the cells belonging to a cluster were then used to define their a cluster-specific variance in each dimension, yielding a set of 30 clusters defined by their means and variances in two dimensions. Next, I defined a vector indicating the expected frequency of all thirty clusters for each donor. This vector differed between cases and controls, as it was generated by measuring the observed cluster frequency in cases and controls, as well as the between-donor variance of each clusters' proportion. This ensured that cases and controls did not have the exact same proportions cells in each cluster. Because these simulations were intended to calculate the power to detect differential abundance once the full set of samples had been sequenced, I simulated donors using 1500 cells and 1000 genes. The simulated cells for each donor were probabilistically assigned clusters according to the expected cluster frequency vector for that donor. Then, cells in each cluster were placed in low-dimensional space using a bivariate normal distribution parametrized by the cluster-specific centroids and variance; as these dimensions are derived from a principal components analysis and are orthonormal, the off-diagonal elements of the variance-covariance matrix were

set to 0. To simulate the effect of inter-individual differences, the two-dimensional locations for each cell are perturbed by adding noise generated from independent mean-0, variance-1 gaussians. This process yields a matrix of cells by two dimensions for each donor. I then multiplied that matrix by the transpose of a matrix of with the shape of genes by two-dimensions, where I assumed each gene's expression was modeled by a random gaussian. While a gaussian distribution does not necessarily describe the observed expression patterns of single cell transcriptomics, the types of distributions do best model the single cell expression are not currently well-defined in the field[149-152]. Finally, the simulated expression matrices were concatenated across cases and controls, the resulting data was clustered using the same settings used to generate the original clustering, and the simulated clusters were tested for case-control associations with MASC. I created simulations for two different sample sizes – 48 and 240 donors – using balanced numbers of cases and controls and 100 replicates per simulation; I then aggregated the power results across the replicates. I grouped clusters by their simulated fold-change between cases and controls, where a fold-change of 1 indicates that the cluster was at the same frequency in cases and controls, and a fold-change of 2 indicates that the cluster was twice as abundant in cases than controls. Simulated studies of 48 donors were only able to reliably detect fold-changes of 1.5 or greater, while simulations with 240 donors achieved 75% power for clusters that were between 1.2 and 1.3-fold expanded in cases (Figure 2-9).

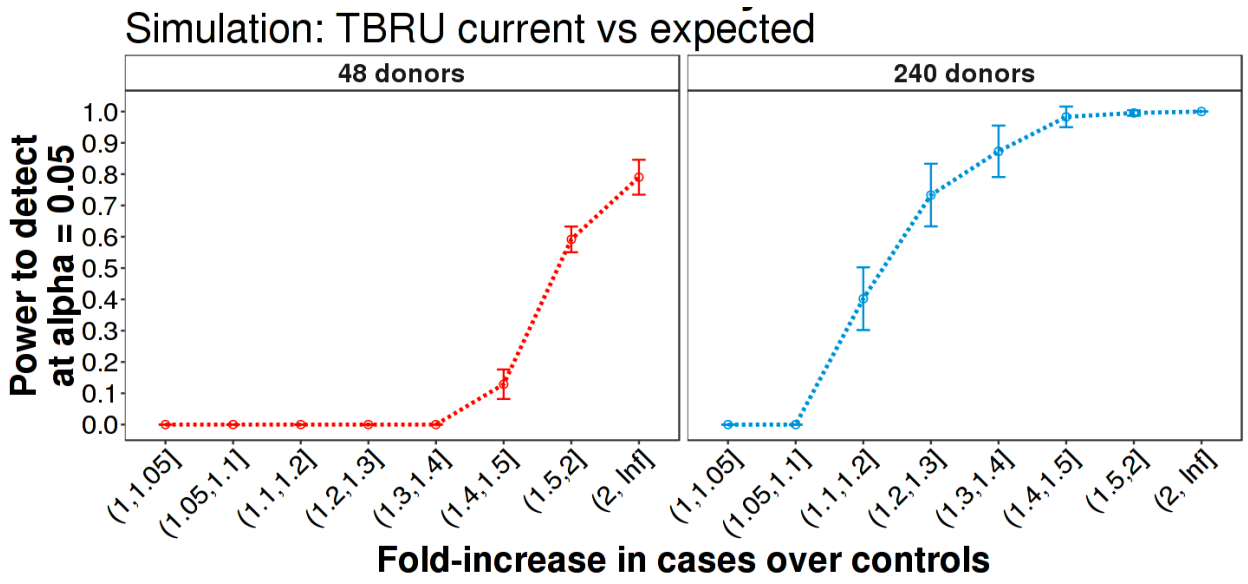


Figure 2-9. The power to detect differentially abundant populations increases with the number of donors included in the study. MASC was used to perform association testing on simulated single cell RNA-seq datasets with either 48 or 240 donors split evenly into cases and controls. One hundred replicates were performed for each simulation; results were grouped by the effect size of the population in question.

Although this "quick and dirty" simulation strategy was effective for conducting power analyses in this study, the method has many potential areas of improvement. First, the underlying distributions used to model gene expression could be changed from random gaussians to a more appropriate distribution for simulating count data, such as the Poisson or negative binomial. Second, while the two principal components captured a significant amount of variability in this dataset, a more robust method could be expanded to use any number of dimensions to annotate cluster centroids. Finally, a different method of simulation could rely on the fact that the transform matrix of a PCA already provides a link between the low-dimensional representation of cells and their projection into an expression matrix. Here, batch, donor, and other technical effects would be modeled as shifts in low-dimensional space, pushing cells away from their cluster centroids in a consistent manner. That is to say that cells from the same donor and batch would have the same batch and donor-effect vectors added to them, while cells from the same batch but different donors would have a different donor-effect vector to represent

inter-individual differences. After perturbation, the modified representation can then be multiplied by the transpose of the PCA transform matrix to return a single cell expression matrix. It is currently unclear what the best strategy for simulating single cell transcriptomic data is, especially in the context of performing association testing for differential abundance; the majority of tools in the field have been designed for calculating power for differential expression analyses[153-157]. While there is certainly more to be done, MASC can play an important role as a state-of-the-art method for conducting differential abundance testing with single cell data. The following two chapters of this work will demonstrate the ability of MASC to identify disease relevant changes in immune cell populations when applied to single cell studies.

Chapter 3:

Mixed-Effects Association of Single Cells Identifies an Expanded Effector CD4+ T Cell Subset in Rheumatoid Arthritis

Authors:
Chamith Y. Fonseka[1,2,3,5†], Deepak A. Rao[2†], Nikola C. Teslovich[2], Ilya Korsunsky[1], Susan K. Hannes[2], Kamil Slowikowsi[1,2,3,4], Michael F. Gurish[2], Laura T. Donlin[6,7,8], James A. Lederer[1], Michael E. Weinblatt[2], Elena M. Massarotti[2], Jonathan S. Coblyn[2], Simon M. Helfgott[2], Derrick J. Todd[2], Vivian P. Bykerk[8,9], Elizabeth W. Karlson[2], Joerg Ermann[2], Yvonne C. Lee[2,11], Michael B. Brenner[2], and Soumya Raychaudhuri[1,2,3,4,10 *]
[†]Co-first authors
*Corresponding author. Email: soumya@broadinstitute.org

Affiliations:
1) Center for Data Sciences, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115 USA.
2) Division of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115 USA.
3) Division of Genetics, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115 USA.
4) Program in Medical and Population Genetics, Broad Institute of Massachusetts Technical Institute and Harvard University, Cambridge, MA 02138, USA.
5) Department of Biomedical Informatics, Harvard University, Cambridge, MA 02138, USA.
6) Arthritis and Tissue Degeneration Program, Hospital for Special Surgery, New York, New York 10021, USA.
7) David Z. Rosensweig Genomics Research Center, Hospital for Special Surgery, New York, New York 10021, USA.
8) Department of Medicine, Weill Cornell Medical College, Cornell University, New York, NY 10021, USA.
9) Division of Rheumatology, Hospital for Special Surgery, 535 E 70th Street, New York, NY 10021 USA.
10) Institute of Inflammation and Repair, University of Manchester, Manchester, UK.
11) Division of Rheumatology, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA

Attributions

C.Y.F., S.R. conceived the MASC association testing method. C.Y.F., S.R., and K.F.S. conducted all statistical analyses. D.A.R., N.C.T., S.K.H., M.F.G., J.E., M.B.B., and S.R. designed and conducted all mass cytometry, flow cytometry, and functional immunology assays. D.A.R., L.T.D., M.E.W., E.M.M., J.S.C., S.M.H., D.J.T., V.P.B., E.W.K., and Y.C.L. recruited patients and obtained samples for this study. C.Y.F, D.A.R. and S.R. wrote the initial manuscript. All authors edited and revised the final manuscript.

Abstract

High dimensional single-cell analyses have improved the ability to resolve complex mixtures of cells from human disease samples; however, identifying disease-associated cell types or cell states in patient samples remains challenging due to technical and inter-individual variation. Here we present Mixed-effects modeling of Associations of Single Cells (MASC), a reverse single cell association strategy for testing whether case-control status influences the membership of single cells in any of multiple cellular subsets while accounting for technical confounders and biological variation. Applying MASC to mass cytometry analyses of CD4+ T cells from the blood of rheumatoid arthritis (RA) patients and controls revealed a significantly expanded population of CD4+ T cells, identified as CD27- HLA-DR+ effector memory cells, in RA patients (OR = 1.7; $p = 1.1 \times 10^{-3}$). The frequency of CD27- HLA-DR+ cells was similarly elevated in blood samples from a second RA patient cohort, and CD27- HLA-DR+ cell frequency decreased in RA patients who responded to immunosuppressive therapy. Mass cytometry and flow cytometry analyses indicated that CD27- HLA-DR+ cells were associated with RA (meta-analysis $p = 2.3 \times 10^{-4}$). Compared to peripheral blood, synovial fluid and synovial tissue samples from RA patients contained ~5-fold higher frequencies of CD27- HLA-DR+ cells, which comprised ~10% of synovial CD4+ T cells. CD27- HLA-DR+ cells expressed a distinctive effector memory transcriptomic program with Th1- and cytotoxicity-associated features, and produced abundant IFN-γ and granzyme A protein upon stimulation. We propose that MASC is a broadly applicable method to identify disease-associated cell populations in high-dimensional single cell data.

Introduction

The advance of single cell technologies has enabled investigators to resolve cellular heterogeneity with unprecedented resolution. Single cell assays have been particularly useful in the study of the immune system, in which diverse cell populations often consisting of rare and transitional cell states may play an important role[158]. Application of single cell transcriptomic and cytometric assays in a case-control study has the potential to reveal expanded pathogenic cell populations in immune-mediated diseases.

Rheumatoid arthritis (RA) is a chronic, systemic disease affecting 0.5-1% of the adult population, making it one of the most common autoimmune disorders worldwide[19]. RA is triggered by environmental and genetic risk factors, leading to activation of autoreactive T cells and B cells that mediate an autoimmune response directed at the joints[20,22]. CD4+ T cells have been strongly implicated in RA pathogenesis[39,40]. For one, the strongest genetic association to RA is with the *HLA-DRB1* gene within the MHC; these polymorphisms affect the range of antigens that MHCII molecules can bind and present in order to activate CD4+ T cells[22,23,27]. Furthermore, many RA risk alleles outside of the MHC locus also lie in pathways important for CD4+ T cell activation, differentiation into effector (T_{eff}) and regulatory (T_{reg}) subsets, and maintenance of subset identity[20,24,25,27-29]. Defining the precise CD4+ T cell subsets that are expanded or dysregulated in RA patients is critical to deciphering pathogenesis. Such cell populations may be enriched in antigen-specific T cells and may aid in discovery of dominant disease-associated autoantigens. In addition, these populations may directly carry out pathologic effector functions that can be targeted therapeutically[105].

For many autoimmune diseases, directly assaying affected tissues is difficult because samples are only available through invasive procedures. Instead, querying peripheral blood for altered immune cell populations is a rapidly scalable strategy that achieves larger sample sizes and allows for serial monitoring. Flow cytometric studies have identified alterations in specific circulating T cell subsets in RA patients, including an increased frequency of CD28- CD4+ T

cells[61,159,160]; however, the expansion of CD28- T cells represents one of the relatively few T cell alterations that has been reproducibly detected by multiple groups. Limited reproducibility may be the consequence of differences in clinical cohorts, small sample sizes, methodologic variability, and use of limited, idiosyncratic combinations of phenotypic markers[53].

The advent of mass cytometry now allows for relatively broad assessment of circulating immune cell populations with >30 markers[81], enabling detailed, multiparametric characterization of lymphocyte subsets. This technology provides the potential to define and quantify lymphocyte subsets at high resolution using multiple markers. While the discovery of novel cellular populations has been enabled by rapid progress in developing sensitive clustering methods[91,132,161-163], a key challenge that remains is establishing methods to identify cell populations associated with a disease. In particular, inter-individual variation and technical variation can influence cell population frequencies and need to be accounted for in an association framework. Single cell association studies, with either mass cytometry or single-cell RNA-seq, require statistical strategies robust to inter-individual donor variability and technical effects that can skew cell subset estimates. For example, mass cytometry studies need to control for variability in machine sensitivity, reagent staining, and sample handling that can lead to batch effects. Inter-individual differences can lead to real shifts in cell population frequencies at baseline, while technical effects can lead to apparent shifts in cell population frequencies.

Here we describe a robust statistical method to test for disease associations with single cell data called MASC (Mixed-effects modeling of Associations of Single Cells), which tests at the single cell level the association between population clusters and disease status. It is a 'reverse' association strategy where the case-control status is an independent variable, rather than the dependent variable. We applied MASC to identify T cell subsets associated with RA in a mass cytometry case-control immunophenotyping dataset that we generated focused on CD4+ T cells. This high-dimensional analysis enabled us to identify disease-specific changes in canonical as well as non-canonical CD4+ T cell populations using a panel of 32 markers to reveal cell lineage,

activation, and function[82,164]. Using MASC, we identified a population of memory CD4+ T cells, characterized as CD27- HLA-DR+, which was expanded in the circulation of RA patients. Further, we found that CD27- HLA-DR+ T cells were enriched within inflamed RA joints, rapidly produced IFN-γ and cytolytic factors, and contracted with successful treatment of RA.

Results

Statistical and computational strategy

We acquired single cell mass cytometry data from RA case and osteoarthritis (OA) control peripheral blood samples (**Figure 3-1**), and then applied MASC after stringent quality controls to remove technical artifacts and poorly stained cells that are typically observed in mass cytometry data (**Materials and Methods**). First, we objectively defined subsets of cells using an equal number of random cells from each sample so that they contributed equally to the subsequent analyses. We then used DensVM[134] to cluster the mass cytometry data. Finally, we applied MASC to identify differentially abundant cellular populations associated with disease. MASC is a reverse association strategy that uses single cell logistic mixed-effect modeling to test individual cellular populations for association by predicting the subset membership of each cell based upon fixed effects (e.g. sex) and random effects (e.g. batch, donor). It assumes a null model where the subset membership of each single cell is estimated by fixed and random effects without considering the case-control status of the samples. Thus, under this null framework, we assume that variation in cluster frequencies are not associated with case-control status. We then measured the improvement in model fit when a fixed effect term for the case-control status of the sample was included with a likelihood ratio test. This framework allowed us to evaluate the significance and effect size of the case-control association for each subset while controlling for inter-individual and technical variability.

Mixed-effects modeling of Associations of Single Cells (MASC) overview

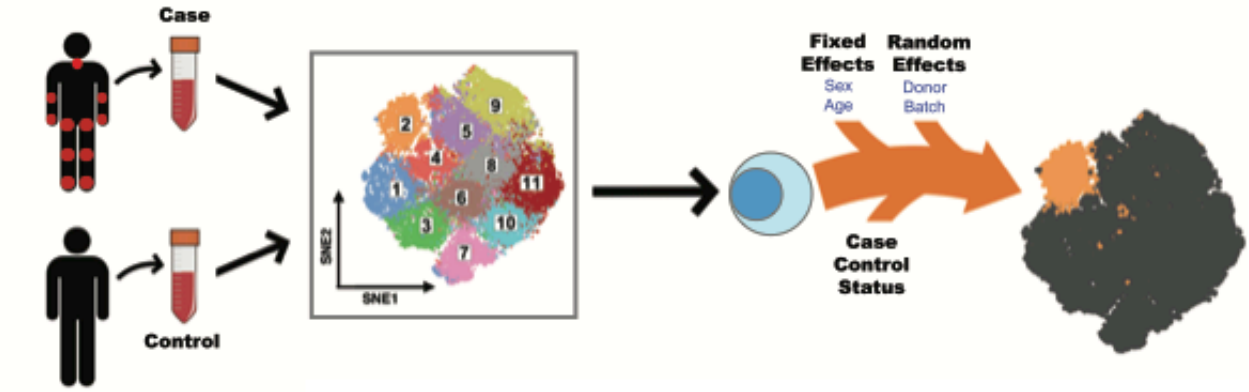


Figure 3-1. Single cell transcriptomics or proteomics are used to assay samples from cases and controls, such as immunoprofiling of peripheral blood. The data is then clustered to define populations of similar cells. Mixed-effects logistic regression is used to predict individual cell membership in previously defined populations. The addition of a case-control term to the regression model allows the user to identify populations for which case-control status is significantly associated.

To ensure that MASC had appropriately calibrated type 1 error, we ran the analysis 10,000 times after permuting case-control labels using the dataset described below to obviate almost any case-control association that might be present. We then recorded the reported p-values for each cluster. Under ideal circumstances, 5% of these 10000 trials would achieve a p-value < 0.05 purely by chance in this random a data set; conversely, an inflated method would have much greater than 5% of trials obtaining a p-value < 0.05. This approach demonstrated that MASC has only a modestly inflated type I error rate for the 19 clusters in the dataset, with 6.5% of trials obtaining p<0.05 (**Figure S1A**). We found that including both donor and batch random effects was critical; eliminating random effects in the model led to highly inflated p-values with 66.1% of trials obtaining p<0.05 (**Figure S1B**). As an alternative and frequently used strategy, we also tested a simple binomial test for case-control association. This approach is limited in our view as it fails to model donor specific and technical effects. Unsurprisingly, we found that this commonly used approach produced highly inflated results in comparison to MASC, with 65.7% of trials obtaining p<0.05 (**Figure S1C**).

Experimental strategy

We applied MASC to mass cytometric analysis of memory CD4+ T cells that were magnetically isolated from peripheral blood mononuclear cells from patients with established RA (cases) and non-inflammatory OA controls (Table 3-1). We either (1) rested the cells for 24 hours or (2) stimulated the cells with anti-CD3/anti-CD28 beads for 24 hours before analyzing cells with a 32-marker mass cytometry panel that included 22 markers of lineage, activation, and function (**Table S1**). This experimental design allowed us to interrogate immune states across cases and controls and also capture stimulation-dependent changes. After stringent quality control measures, we analyzed a total of 26 RA cases and 26 controls.

Clinical characteristics of patient cohorts

Table 3-1. Mean ± SD is shown. Parentheses indicate percentages. CRP: C-reactive protein. CDAI: clinical disease activity index. "Other biologics" includes rituximab, tofacitinib, and abatacept.

		Mass cytometry cohort		Flow cytometry cohort	
		Controls	**Cases**	**Controls**	**Cases**
Blood Cross-sectional cohorts	**Number**	26	26	27	39
	Age	57 ± 15	66 ± 9	61 ± 14	58 ± 14
	Female	19 (73)	20 (77)	18 (64)	30 (77)
	ACPA- or RF-positive	N/A	22 (85)	N/A	39 (100)
	CRP (mg/L)	N/A	8.6 ± 16.9	N/A	9.8 ± 17.9
	CDAI	N/A	9.3 ± 4.4	N/A	13.7 ± 7.4
	Methotrexate	0	18 (69)	0	18 (46)
	Anti-TNF	0	10 (38)	0	16 (41)
	Other biologics	0	5 (19)	0	9 (23)

		Longitudinal cohort
Blood Longitudinal cohort	**Number**	18
	Age	49 ± 17
	Female	17 (94)
	ACPA- or RF-positive	18 (100)
	CDAI Before	17.6 ± 9.3
	CDAI After	6.3 ± 4.2
	Started methotrexate	7
	Started anti-TNF	4
	Started other biologic	7

	Patient	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**	**#8**	**#9**
Synovial Tissue Donors	**Age**	57	54	76	46	46	79	62	63	52
	Sex	F	F	F	F	F	F	M	M	F
	CRP (mg/L)	25	8	8	11	17	19	13	66	76
	CDAI	14	9	17	15	21	25	5	9	N/A
	Methotrexate	No	Yes	No	No	No	No	No	Yes	No
	Biologic therapy	Yes	Yes	Yes	Yes	Yes	No	No	No	No

We randomly sampled 1000 cells from each of the 52 samples so that each sample contributed equally to the analysis, preventing samples that happened to have more cells captured by the mass cytometry assay from being overrepresented. We projected resting and stimulated T cell data separately using the Barnes-Hut modification to the t-SNE (*t*-distributed stochastic neighbor embedding) algorithm[142] so that all cells from all samples were projected into the same two dimensions using all markers in the panel with the exception of CD4 and CD45RO, which we used to gate CD4+ memory T cells for analysis.

Clustering Approach

Projecting the data into t-SNE space revealed areas of local density that consisted predominantly of cells from RA or OA samples (**Figure S2**). We wanted to identify CD4+ memory T cell populations in an unsupervised manner. Currently, clustering high-dimensional single cell data (such as mass cytometry or single cell RNA-seq data) is an active area of research, and there is no consensus on the best clustering strategy. Hence we objectively considered multiple clustering algorithm options. We evaluated DensVM[134], FlowSOM[130], and Phenograph[132], which were identified as among the best-performing in a recent benchmarking comparison of clustering methods for high dimensional cytometry data[135]. The DensVM method uses an t-SNE projection of the dataset to first estimate the number of clusters by searching for local densities on the projection with varying bandwidths before classifying cells based on the similarity of expression. Phenograph works by creating a graph representing phenotypic similarities between cells and identifying clusters using Louvain community detection. The FlowSOM algorithm builds a self-organizing map with a minimum spanning tree to detect populations, then classifies cells in a meta-clustering step using consensus hierarchical clustering. We clustered cells with all three methods using the same selection of markers that was used to create t-SNE projections.

After running all three algorithms on the dataset, we identified 19, 19, and 21 clusters for DensVM, FlowSOM, and Phenograph respectively. An ideal clustering algorithm would define clusters that are distinct from each other in terms of marker intensities. However, cluster intensity differences should not be driven by batch effects; that is, clusters should not be disproportionately constituted by cells from any individual batch.

In order to quantify the extent to which clustering approaches defined clusters with distinct marker intensities, we utilized a Marker Informativeness Metric (MIM) (**Materials and Methods**). All three methods were similar in their ability to generate clusters with distinct marker intensities (**Figure S3A**). Then, to determine whether those intensity differences were dependent on batch differences, we used a Cluster Informativeness Metric (CIM)-based metric to assess whether clusters were disproportionately represented by individual batches (**Materials and Methods**). Here, we observed that Phenograph and FlowSOM were much more sensitive to batch effects than clusters identified by DensVM ($p < 2 \times 10^{-3}$, Wilcoxon rank sum test, **Figure S3B**). Consistent with this quantitative assessment, we observed that in our data FlowSOM and Phenograph produced clusters that were constituted exclusively of or dominated by cells from a specific batch. Thus, we chose to analyze clusters identified by the DensVM algorithm going forward.

Landscape of CD4+ Memory T Cell Subsets

We observed substantial diversity among resting CD4+ memory T cells in both cases and controls, consistent with previous reports demonstrating a breadth of phenotypes in CD8+ T cells and CD4+ T cells[84,89,165]. We identified 19 distinct subsets in resting (*R*) memory CD4+ T cells (**Figure 3-2A, S4A, S5**). Central memory T cells (T_{CM}) segregated from effector memory T cells (T_{EM}) by the expression of CD62L (**Figure 3-2B**). Five subsets (subsets 1 – 5) of T_{CM} cells, all expressing CD62L, varied in expression of CD27 and CD38, highlighting the heterogeneity within the T_{CM} compartment (**Figure 3-2E**). We identified two T_h1 subsets (subsets 8 and 12) as

well as two T_{reg} subsets (subsets 7 and 11). Both T_{reg} subsets expressed high levels of CD25 and FoxP3, and subset 11 also expressed HLA-DR, reflecting a known diversity among T_{reg} populations in humans *(32)*.

When applied to the stimulated CD4+ T cell data in a separate analysis, DensVM identified 21 subsets amongst all case and control samples (**Figure 3-2C, S4B, S5**). As expected, certain activation markers, such as CD25 and CD40L, were broadly expressed across most subsets after stimulation (**Figure 3-2D**). Mass cytometry robustly detected cytokine production from stimulated CD4+ memory T cells (**Figure 3-2D, 3-2F**). Activated effector cells identified by the production of IL-2 were separated into three groups (subsets 1 – 3) according to relative expression of TNF. Cytokine expression after stimulation also improved the ability to resolve certain CD4+ T effector subsets, such as T_h17 cells (subset 15) and T_h1 cells (subset 12). However, we were unable to resolve the T_{reg} subtypes that were observed in the resting T cells; after stimulation, all CD25+ FoxP3+ cells were grouped together (subset 21). The set of cells that did not activate after stimulation (subsets 16, 17, and 19) were easily identified by the lack of expression of activation markers such as CD25, CD40L, and cytokines, and the retention of high CD3 expression.

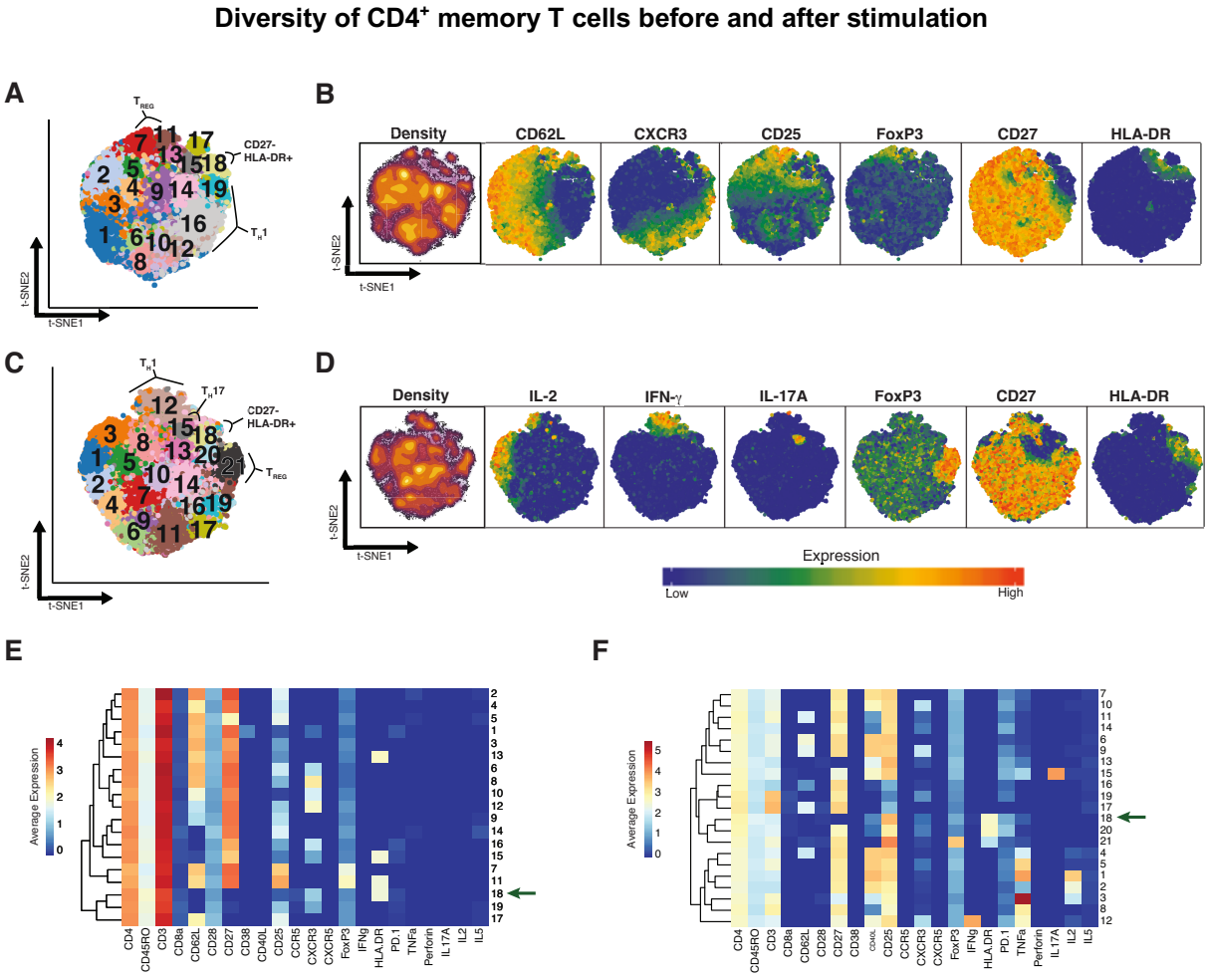


Figure 3-2. (A) t-SNE projection of 50,000 resting CD4+ memory T cells sampled equally from RA patients (n=24) and controls (n=26). DensVM identified 19 populations in this dataset. (B) Same t-SNE projections as in (A) colored by the density of cells on the SNE plot or the expression of the markers labeled above each panel. (C) t-SNE projection of 52,000 CD4+ stimulated memory T cells sampled equally from RA patients (n=26) and controls (n=26). Cells were stimulated for 24 hours with anti-CD3/anti-CD28 beads. (D) Same t-SNE projections as in (C) colored by the density of cells on the SNE plot or the expression of the markers labeled above each panel. (E) Heatmap showing mean expression of indicated markers across the 19 populations found in resting cells. (F) Heatmap showing mean expression of indicated markers across the 21 populations found after stimulation. Protein expression data are shown after arcsinh transformation. All markers but CD4 and CD45RO were used to create t-SNE projections and perform clustering.

Identifying Populations that are Enriched or Depleted in RA Samples

We next sought to identify subsets that were significantly overrepresented or underrepresented in patient cells. The frequency of RA cells in each subset ranged from 36.7% to 63.6% in the resting state, and 38.9% to 65.7% in the stimulated state (**Table S2**, **Figure S6A**). Visualizing the density of the t-SNE projections revealed that related cells clustered into dense groups both at rest and after stimulation (**Figure 3-2B, 3-2D**), and plotting t-SNE projections of cells from cases and controls separately while coloring by density clearly suggested differential abundance of RA cells among clusters (**Figure S2**). Accounting for subject-specific and batch-specific random effects with MASC (**Materials and Methods**), we observed three populations with significantly altered proportions of cells from cases in the resting T cell data. Most notably, we observed enrichment in subset 18 ($p = 5.9 \times 10^{-4}$, **Table 3-2**, **Figure 3-3A**); this subset consisted 3.1% of total cells from cases compared to 1.7% of cells from controls and achieved an odds ratio (OR) of 1.9 (95% confidence interval [CI] = 1.3 – 2.7). Conversely, subset 7 ($p = 8.8 \times 10^{-4}$, OR = 0.6, 95% CI = 0.5 – 0.8) and subset 12 ($p = 2.0 \times 10^{-3}$, OR = 0.5, 95% CI = 0.4 – 0.8) were underrepresented for RA cells.

Overview of subsets found to be significantly expanded in RA

Table 3-2. RA proportion reflects the fraction of cells in the subset that were from RA donors. The 95% confidence interval is shown next to the odds ratio.

Condition	Description	Subset	RA Proportion	*P* value	Odds Ratio	Test
Resting	HLA-DR+,CD27-	18	0.636	$\mathbf{5.9 \times 10^{-4}}$	1.9 (1.3 – 2.7)	MASC
Stimulated	HLA-DR+, CD27-	18	0.619	$\mathbf{1.3 \times 10^{-3}}$	1.7 (1.2 – 2.2)	MASC
Flow cytometry replication	Gated HLA-DR+, CD27-	NA	NA	$\mathbf{4.4 \times 10^{-2}}$	NA	One-tailed *t* test
Meta-analysis	HLA-DR+, CD27-	NA	NA	$\mathbf{2.3 \times 10^{-4}}$	NA	Stouffer's Z-score method

MASC identifies a population that is expanded in RA

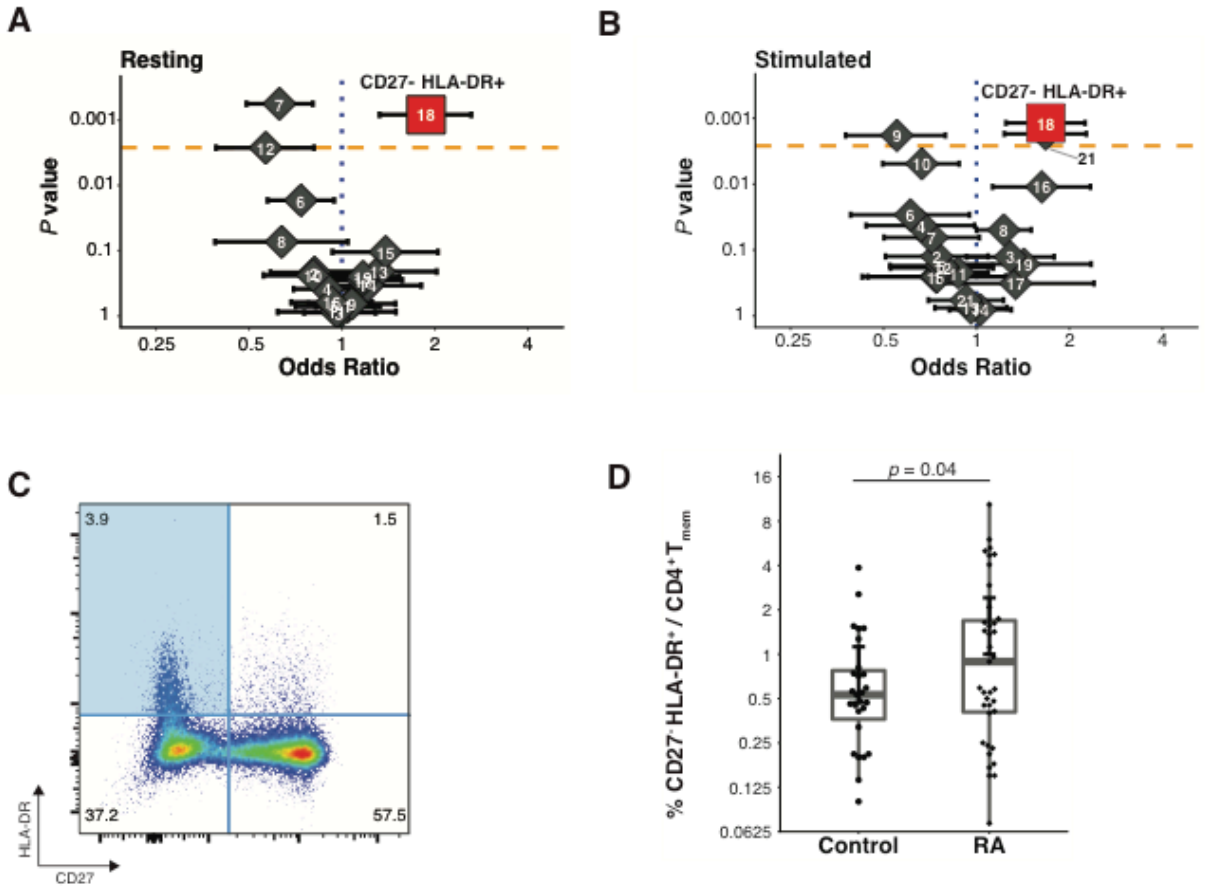


Figure 3-3. (A, B) Odds ratios and association p-values were calculated by MASC for each population identified the resting (A) and stimulated (B) datasets. The yellow line indicates the significance threshold after applying the Bonferroni correction for multiple testing. (C) Flow cytometry dot plot of gated memory CD4+ T cells from a single RA donor shows the gates used to identify $CD27^-$ $HLA\text{-}DR^+$ memory CD4+ T cells (blue quadrant). (D) Flow cytometric quantification of the percentage of $CD27^-$, $HLA\text{-}DR^+$ cells among blood memory CD4+ T cells in an independent cohort of seropositive RA patients (n = 39) and controls (n = 27). Statistical significance was assessed using a one-tailed *t*-test after assessing normality with a Shapiro-Wilk test ($p > 0.52$).

To confirm the robustness of this analytical model, we conducted a stringent permutation test that was robust to potential batch effects. To control for batch, we reassigned case-control labels randomly to each sample, retaining the same number of cases and controls in each batch. For each T cell subset we recorded the number of case cells after randomization to define a null distribution. In the real data, we observed that 753 out of the 1184 cells in subset 18 were from case samples. In contrast, when we permuted the data we observed a mean of 550 cells and a standard deviation of 63 cells from case samples. In only 93 out of 100,000 permutations did we observe more than 753 cells from case samples in the randomized data (permutation $p = 9.4 \times 10^{-4}$, **Table S2**). We applied permutation testing to every subset and found that the *p*-values produced by permutation were similar to *p*-values derived from the mixed-effects model framework (Spearman's r = 0.86, **Figure S6B-C**). When we considered the subsets identified as significant by MASC, both subsets 18 (permutation $p = 9.4 \times 10^{-4}$) and 7 (permutation $p = 1.8 \times 10^{-4}$) retained significance whereas subset 12 demonstrated nominal evidence of case-control association (permutation $p = 1.3 \times 10^{-2}$).

Next, for each resting T cell subset, we identified corresponding subsets in the stimulated T cell dataset using a cluster centroid alignment strategy to calculate the distance between subsets across datasets (**Materials and Methods**). Subset 18 in the resting dataset was most similar to subset 18 in the stimulated dataset, while subset 7 in the resting dataset was most similar to subset 21 in the stimulated dataset (**Figure S6D-E**). Applying MASC, we observed case-control association for subset 18 in the stimulated data ($p = 1.2 \times 10^{-3}$, OR = 1.7, 95% CI = 1.2 – 2.2), while subset 21 (p = 0.55) was not significant in the stimulated data (**Table 3-2, Figure 3-3B**). We identified one additional population subset that was significant in the stimulated dataset: subset 20 ($p = 1.7 \times 10^{-3}$, OR = 1.7, 95% CI = 1.2 – 2.3) (**Table S3**).
We wanted to ensure that the results we observed were not an artifact of using DensVM to cluster the cytometry data. When we used Phenograph and FlowSOM to cluster the same mass cytometry dataset, we observed a CD27-, HLA-DR+ cluster with either method (**Figure S7A,**

S7C) with association to RA by MASC (**Figure S7B, S7D**). We also assessed how analysis by MASC compared to Citrus[137], an algorithm that uses hierarchical clustering to define cellular subsets and then builds a set of models using the clusters to stratify cases and controls. When applied to the same case-control resting dataset, Citrus was unable to produce models with an acceptable cross-validation error rate, regardless of the method used (**Figure S8**).

CD27- HLA-DR+ CD4+ Effector Memory T Cell Expansion in RA

After noting that subset 18 demonstrated robust association with RA, we interrogated the key features of this population. The lack of CD62L expression in subset 18 indicated that this subset was an effector memory T cell population. To define the markers that best differentiated this subset from other cells, we calculated the MIM for the expression of each marker in the subset for both the resting and stimulated datasets (**Materials and Methods**, **Figure 3-4A**). The expression of HLA-DR and perforin were notably increased in subset 18 compared to all other cells, while the expression of CD27 was decreased in this subset (**Figure 3-4B**). We observed that gating on CD27 and HLA-DR largely recapitulated subset 18 in both the resting cells and the stimulated cells (**Figure 3-4C and 3-4D**), with an F measure of 0.8 and 0.7 respectively (**Materials and Methods**). Although HLA-DR is a known to be expressed on T cells in response to activation, it takes several days to induce strong HLA-DR expression[166]. Thus, it is likely that cells in subset 18 expressed HLA-DR prior to stimulation, such that analyses of both resting and stimulated cells identified the same HLA-DR+ CD27- effector memory T cell population.

CD27 and HLA-DR expression specifically mark the expanded population

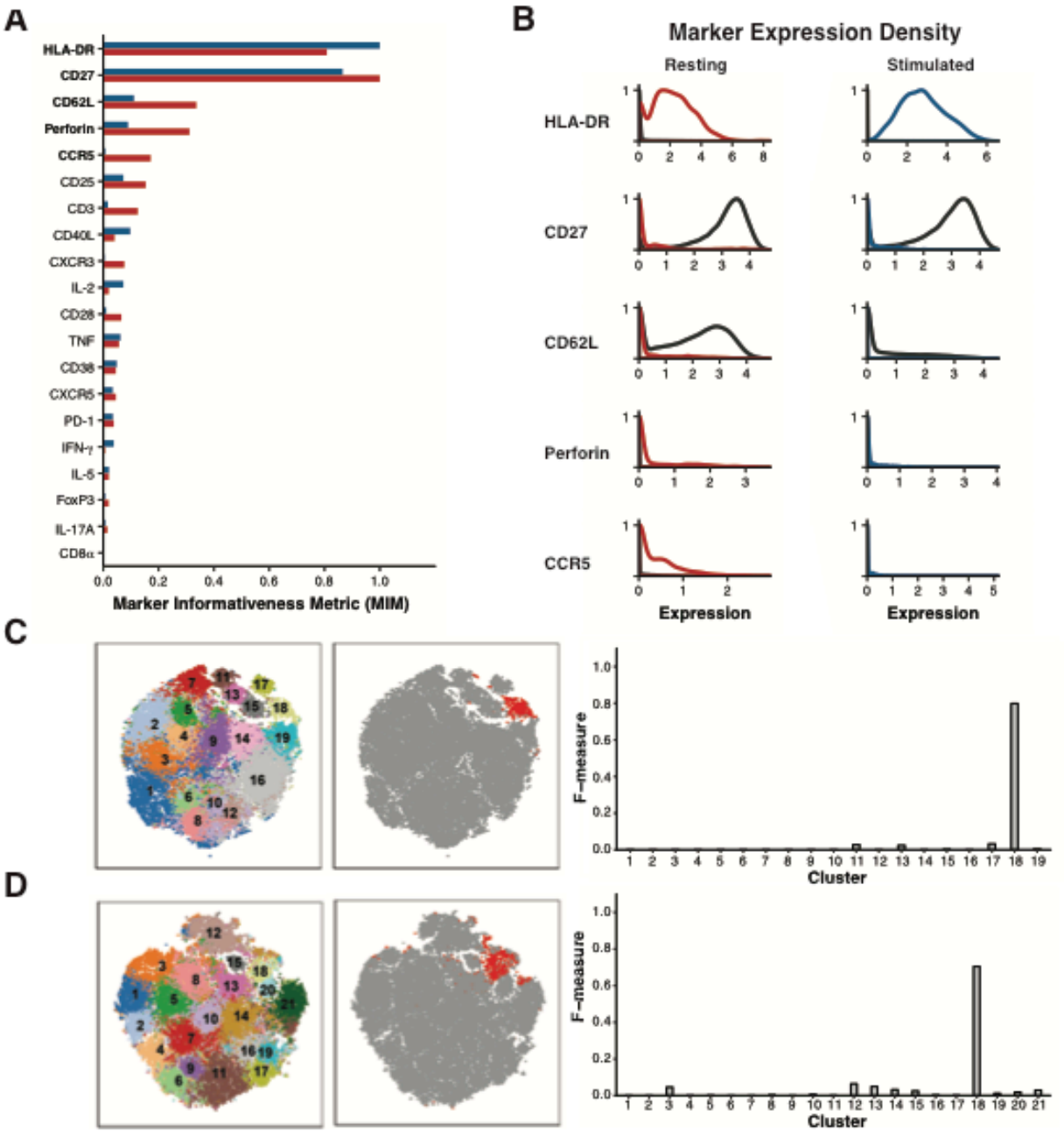


Figure 3-4. (A) Plot of the Kullback-Liebler divergence for each marker comparing cluster 18 to all other cells (grey) in both the resting dataset (red) and the stimulated dataset (blue). (B) Density plots showing expression of the five markers most different between cluster 18 cells (resting = red, stimulated = blue) and all other cells in the same dataset (black line). (C) Left: t-SNE projection of clusters identified in resting dataset; Middle: Same t-SNE projection, with cells gated as CD27- HLA-DR+ colored in red; Right: F-measure scores were calculated for the overlap between gated cells and each cluster in the resting dataset. (D) Left: t-SNE projection of clusters identified in stimulated dataset; Middle: Same t-SNE projection, with cells gated as CD27- HLA-DR+ colored in red; Right: F-measure scores were calculated for the overlap between gated cells and each cluster in the stimulated dataset.

In order to confirm the expansion of the CD27- HLA-DR+ T cell population in RA patients that we observed by mass cytometry, we evaluated the frequency of CD27- HLA-DR+ T cells in an independent cohort of 39 seropositive RA patients and 27 non-inflammatory OA controls using conventional flow cytometry (**Table 3-1**). We determined the percentage of memory CD4+ T cells with a CD27- HLA-DR+ phenotype by gating individual samples from each group (**Figure 3-3C, Figure S9A, Figure S10**). Consistent with the mass cytometry analysis, CD27- HLA-DR+ cells were significantly expanded in the RA patient samples ($p = 0.044$, one-tailed *t* test, Shapiro Wilk normality test $p > 0.52$, **Figure 3-3D**). The frequency of this subset was 0.8% in controls and 1.7% in RA samples, which was similar to the two-fold enrichment we observed in the mass cytometry data. We then considered the mass cytometry and flow cytometry association results together in a meta-analysis, confirming that CD27- HLA-DR+ cells significantly associated with RA ($p = 2.3 \times 10^{-4}$, Stouffers Z-score method, **Table 3-2**).

To assess the effect of RA treatment on CD27- HLA-DR+ cell frequency, we quantified CD27- HLA-DR+ cell frequencies in 23 RA patients before and 3 months after initiation of a new medication for RA. We dichotomized patients as those who experienced a clinical response, defined as a reduction in CDAI (Clinical Disease Activity Index) *(61)* scores (**Δ**CDAI-), versus those that did not, defined as an increase in CDAI scores (**Δ**CDAI+). We observed that changes in CD27- HLA-DR+ cell frequency tracked with clinical response to treatment initiation ($p = 3.49 \times 10^{-4}$, Wilcoxon signed-rank test, **Figure 3-5A**). Specifically, in the 18 **Δ**CDAI- patients, the frequency of CD27- HLA-DR+ cells significantly reduced by 0.7-fold ($p = 0.006$, Wilcoxon signed-rank test, **Figure S11B**); in contrast, the 5 **Δ**CDAI+ patients in this same trial had an 1.8-fold increase in CD27- HLA-DR+ cells, although the increase was not statistically significant on its own. We did observe a significant difference in the CD27- HLA-DR+ frequency fold-change between patients experiencing a CDAI reduction versus not ($p = 0.02$, Wilcoxon rank sum test, **Figure S11A**).

CD27$^-$ HLA-DR$^+$ memory CD4+ T cells are expanded in the blood and joints of patients with active RA

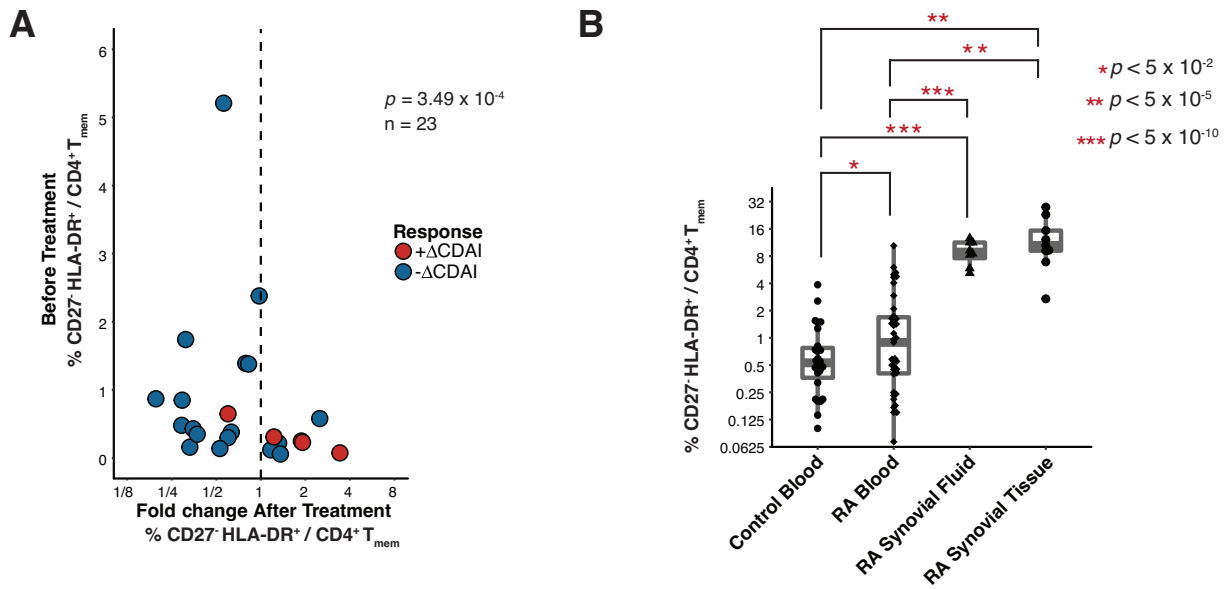


Figure 3-5. (A) Flow cytometric quantification of the frequency of CD27- HLA-DR+ memory CD4+ T cells in 18 RA patients prior to starting a new medication, plotted against change in cell frequency after 3 months of new therapy. Treatment significantly reduced CD27- HLA-DR+ cell frequency as determined by a Wilcoxon signed-rank test. (B) Flow cytometric quantification of the percentage of memory CD4+ T cells with a CD27- HLA-DR+ phenotype in cells from seropositive RA synovial fluid (n=8) and synovial tissue (n=9), compared to blood samples from RA patients and controls. Blood sample data are the same as shown in Fig. 3D. Significance was assessed using one-tailed t-test after determining normality with a Shapiro-Wilk test ($p > 0.52$) and applying a Bonferroni correction for multiple testing.

The decrease in frequency of CD27- HLA-DR+ cells in patients responding to treatment escalation was accompanied by a slight increase in the frequency of CD27+ HLA-DR- cells: CD27+ HLA-DR- cells represented an average of 86.2% of CD4+ memory T cells before treatment, and an average of 87.9% of CD4+ memory T cells afterwards ($p = 0.009$, Wilcoxon signed-rank test, **Figure S11C**). These results confirm that CD27- HLA-DR+ CD4+ T cells were expanded in the circulation of RA patients and decreased with effective disease treatment. We also examined the relationship between the frequency of CD27- HLA-DR+ CD4+ T cells and disease activity or therapeutic use, but we found no significant associations (**Figure S12**).
To determine whether CD27- HLA-DR+ T cells were further enriched at the sites of inflammation in seropositive RA patients, we evaluated T cells in inflamed synovial tissue samples obtained at the time of arthroplasty (**Table 3-1**) and in inflammatory synovial fluid samples from RA patients. In a set of 9 synovial tissue samples with lymphocytic infiltrates observed by histology, the frequency of CD27- HLA-DR+ cells was significantly increased 5-fold ($p < 0.01$, Wilcoxon rank-sum, median 10.5% of memory CD4+ T cells) compared to blood (**Figure 3-5B**). Notably, in 2 of the tissue samples, >20% of the memory CD4+ T cells displayed this phenotype. CD27- HLA-DR+ cells were similarly expanded in synovial fluid samples from seropositive RA patients ($p < 0.001$, Wilcoxon rank-sum, median 8.9% of memory CD4+ T cells, n=8). Thus, CD27- HLA-DR+ T cells were enriched at the primary sites of inflammation in RA patients.

Functional and Transcriptional Features of CD27- HLA-DR+ CD4+ Effector Memory T Cells

To evaluate the potential function of the CD27- HLA-DR+ cell subset, we performed RNA-seq on CD27- HLA-DR+ CD4+ T cells with other effector populations to identify transcriptomic signatures for this subset (**Figure S9B-D**). We sorted and sequenced the following CD4+ T cell populations: naïve CD4+ T cells (TN), central memory CD4+ T cells

(TCM), regulatory CD4+ T cells (TReg), and all four CD27-/+ HLA-DR-/+ subsets – defined as DR+27+ Effector T cells (DR+27+), DR+27- Effector T Cells (DR+27-), DR-27+ Effector T Cells (DR-27+), and DR-27- Effector T Cells (DR-27-) – from peripheral blood mononuclear cells (PBMCs) in 7 RA cases and 6 OA controls. In total we generated 1.1 billion reads for 90 samples. We aligned reads with Kallisto[167] and applied stringent quality controls to remove genes that lacked sufficient expression (**Materials and Methods**), ultimately resulting in a set of 15,234 genes for analysis.

We performed principal component analysis (PCA) on the expression data (**Figure 3-6A**). The first PC, capturing 4% of the total variation, separated cell types along a naïve to effector axis, with the CD27- HLA-DR+ subset representing the extreme case with the highest PC1 values, and naïve T cells with the lowest PC1 values. We examined the genes with the highest PC1 loadings and found that PC1 was associated negatively with *CCR7* and positively with *CXCR3* and *CCR5*. This finding was consistent with the elevated expression of CXCR3 and CCR5 we observed in the mass cytometry analyses of CD27- HLA-DR+ cells. We note that *CXCR3* and *CCR5* are both chemokine receptors associated with a Th1 phenotype. In addition, *PRF1* (perforin), a cytotoxic factor, was also strongly associated with PC1. The extreme position of CD27- HLA-DR+ cells along the continuum of CD4+ T cells suggested a possible late or terminal effector memory phenotype. To further explore this naïve to effector gradient, we used the gene loadings along PC1 to perform gene set enrichment analysis (GSEA) and identify the pathways that were most associated. Intriguingly, the naïve to CD27- HLA-DR+ axis was strongly correlated with naïve vs effector and natural killer (NK) vs CD4+ T cell gene signatures (**Figure 3-6B**).

Transcriptomic characterization of CD27- HLA-DR+ memory CD4+ T cells identified a Th1-skewed cytotoxic phenotype

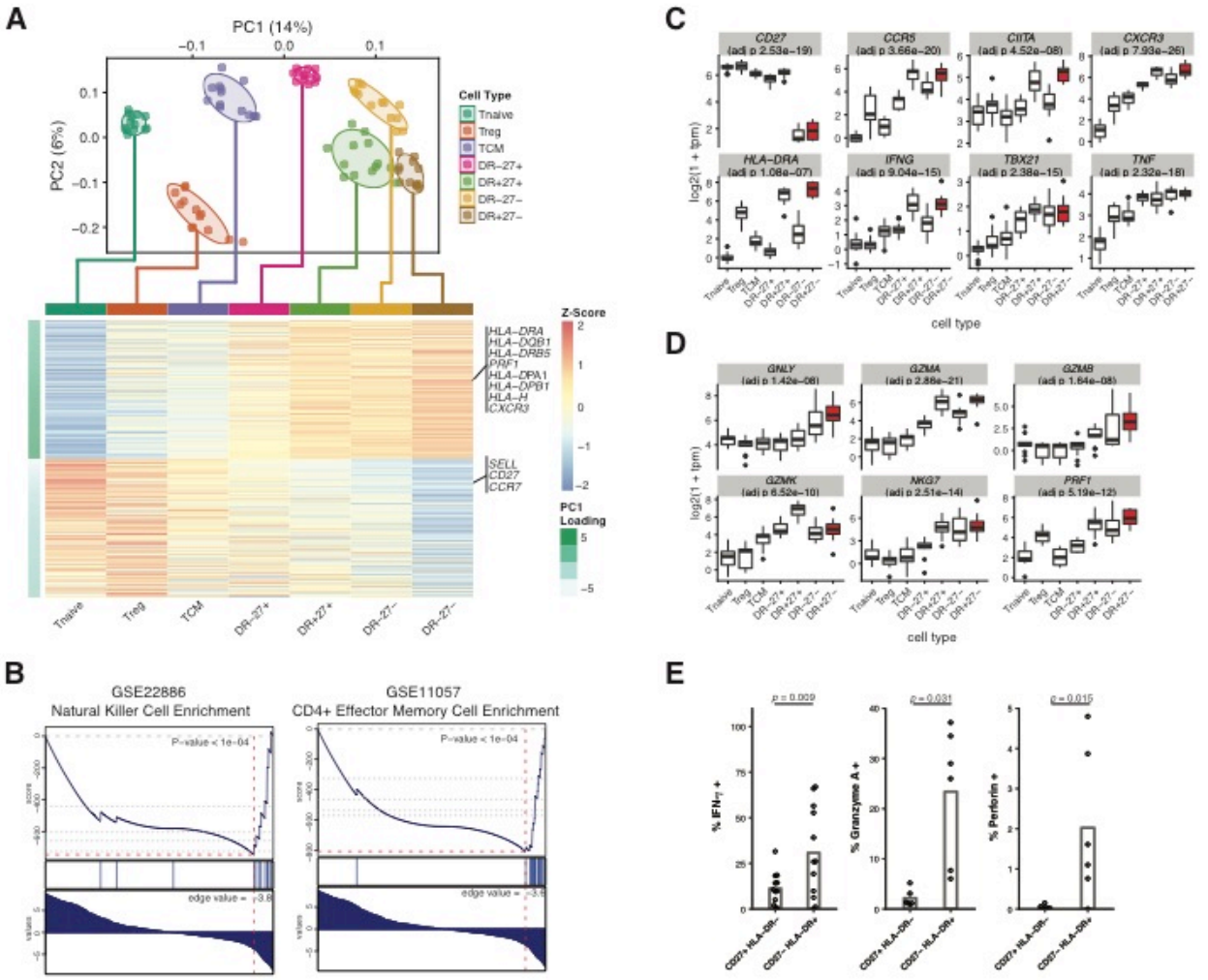


Figure 3-2. RNA-seq characterization of CD27- HLA-DR+ (DR+27-) cells and 6 related CD4+ T cell populations: naive T cells (Tnaive), regulatory T cells (Treg), central memory t cells (TCM), and three populations of effector memory T cells, CD27+ HLA-DR- (DR-27+), CD27+ HLA-DR+ (DR+27+), and CD27- HLA-DR- (DR-27-). (A) PCA plot (top) and PC1 gene loadings (bottom) of 90 samples from the 7 CD4+ T cell populations. Cells were colored on the PCA plot according to known cell type. Normal confidence ellipses at 1 standard deviation were plotted for each cell type. The 300 most positive and 300 most negative PC1 gene loadings for each cell type were averaged and plotted in the heatmap. Genes relevant to the CD27- HLA-DR+ population were labeled. (B) Gene set enrichment analysis was performed on all genes, ranked on their PC1 loadings. Two significantly enriched gene sets: NK signature (GSE22886 NAIVE CD4 T CELL VS NK CELL DN) and effector memory t cell signature (GSE11057 NAIVE VS EFF MEMORY CD4 T CELL) are shown. (C) Distribution of log-scaled expression of six canonical Th1 genes: CCR5, CIITA, CXCR3, IFNG, TBX21 (Tbet), and TNF. Populations are ordered by PC1 loading values, with CD27- HLA-DR+ population highlighted in red. (D) Distribution of log-scaled gene expression of six canonical cytotoxic genes: GNLY, GZMA, GZMB, GMZK, NKG7, and PRF1. Populations are ordered by PC1 loading values, with the CD27- HLA-DR+ population highlighted in red. Reported p-values in (C) and (D) correspond to a linear model of gene expression against ordered cell type (as an ordinal variable), with p-values adjusted for multiple testing by the Benjamini Hochberg procedure. (E) Cytokine expression determined by intracellular cytokine staining of peripheral effector memory CD4+ T cells after in vitro stimulation with PMA/ionomycin. The percentage of cells positive for

Figure 3-6 (continued). each stain is plotted for CD27+ HLA-DR- and CD27- HLA-DR+ subsets. Each dot represents a separate donor (n = 12; 6 RA patients and 6 controls, except for the quantification of Granyzme A and perforin where n = 6; 3 RA patients and 3 controls). Statistical significance was assessed using a Wilcoxon signed-rank test.

The association of CXCR3 and CCR5 with PC1 prompted us to examine the expression of Th1-associated genes in these cells. The effector memory cell populations showed increased expression of multiple Th1-associated genes, including *IFNG* (IFN-g) and *TBX21* (Tbet), compared to naïve, Treg, and central memory cells. In addition, expression of these Th1-associated genes was higher in CD27- HLADR+ compared to CD27+ HLADR- effector memory cells, which constituted the majority of the effector memory T cell pool (**Figure 3-6C**). In contrast, cytokines characteristic of other polarized Th subsets (*IL17A*, *IL4*, *IL21*, *TGFB1*) and transcription factors associated with other Th subsets (*RORC*, *GATA3*, *BCL6*, *FOXP3*) were not elevated in CD27- HLA-DR+ cells compared to other effector populations (**Figure S13A**). We noted that the transcription factor *CIITA* was also increased in CD27- HLA-DR+ cells (**Figure 3-6C**). *CIITA* is well-known for its role as a regulator of MHC class II; indeed, we observed that targets of *CIITA* such as HLA genes were significantly overexpressed in the CD27- HLA-DR+ population compared to other populations (**Figure S13B**).

As the expression of *PRF1* was positively associated with PC1, we wanted to assess the relationship between the expression of cytotoxic gene programs and the naïve to effector PC1 gradient. We used gene set enrichment analysis (GSEA) to query whether NK cell-associated genes were upregulated in the CD27-HLA-DR+ population. Specifically, we ranked genes by their degree of differential expression, comparing expression in CD27-HLA-DR+ cells versus that seen in all other cell types. For NK cell-associated genes, we used the previously defined geneset from MSigDB that defines genes upregulated in NK cells versus CD4+ T cells. Indeed, genes that were highly expressed in NK cells similarly showed high expression in CD27- HLA-DR+ cells, while genes with low expression in NK cells showed similarly low expression in CD27- HLADR+ cells (**Table S4**). Consistent with the enrichment results, a set of genes

characteristically associated with cytolytic cells, including *PRF1*, *GZMB*, *GZMA*, *GNLY*, and *NKG7* were all increased in expression in CD27- HLADR+ cells compared to most other T cell populations analyzed (**Figure 3-6D**). These results indicate that the CD27- HLA-DR+ cells expressed a transcriptomic signature characteristic of cytotoxic cells.

We also evaluated the production and expression of cytokines and cytolytic factors at the protein level by intracellular flow cytometry. We assessed production of effector molecules by cells after in vitro stimulation with PMA + ionomycin, a stimulation method that readily reveals T cell capacity for cytokine production. Consistent with RNA-seq analyses, CD27- HLA-DR+ cells also more frequently expressed perforin ($p = 0.031$, one-tailed Wilcoxon signed-rank test) and granzyme A ($p = 0.015$, one-tailed Wilcoxon signed-rank test) than did CD27+ HLA-DR- cells, which constituted the majority of memory CD4+ T cells. In addition, CD27- HLA-DR+ cells produced IFN-γ at a much higher frequency ($p = 0.009$, one-tailed Wilcoxon signed-rank test, **Figure 3-6E**). Percent positivity for each marker was determined by selecting an expression level threshold to determine positive staining (**Figure S14**). Taken together, broad analyses of gene expression and targeted measures of effector molecule production at the protein level indicated that CD27 HLA-DR+ T cells are a Th1-skewed T cell population capable of producing cytotoxic molecules.

Discussion

Mass cytometry has been successfully applied to decipher the heterogeneity of human T cells in multiple settings, including for the identification of disease-specific changes to circulating immune cell populations and mapping of developmental pathways[84,90,133]. It and other emerging single cell strategies offer a promising avenue to characterize the T cell and other immunological features of a wide-range of diseases in humans. However, successful application of mass cytometry on a larger scale (>20 samples) to conduct association studies on clinical phenotypes in humans has been limited thus far[168-171], in part due to the limited availability of effective association testing strategies.

Although there are many approaches to cluster single cell data[91,130,132,134,161-163], extension of these methods to perform case-control association testing is not straightforward. The simplest strategy would be to define subsets of interest by clustering the cells and then comparing frequencies of these subsets in cases and controls with a univariate test such as a t-test or a non-parametric Mann-Whitney test. While commonly used in flow cytometry analysis, this approach is dramatically underpowered, as it relies upon reducing single-cell data to potentially inaccurate per-sample subset frequencies.

In contrast, our methodology takes full advantage of single cell measurements by using a "reverse-association" framework to test whether case-control status influences the membership of a given cell in a population - that is, each single cell was treated as a single event. However, these cells are not entirely statistically independent. Failing to account for dependencies, for example by using a binomial test or not accounting for random effects, can result in considerably inflated statistics and irreproducible associations (**Figure S1**). Using a mixed-effects logistic regression model allowed us to account for covariance in single cell data induced by technical and biological factors that could confound association signals (**Figure 3-1, Figure S1B**), without inflated association tests. MASC also allows users to utilize technical covariates that might be relevant to a single cell measurement (e.g. read depth per cell for single cell RNA-

seq or signal quality for mass cytometry) that might influence cluster membership for a single cell.

We note that Citrus, an association strategy to automatically highlight statistical differences between experimental groups and identify predictive populations in mass cytometry data[137], was unable to identify the expanded CD27- HLA-DR+ population. We believe that our methodology compared favorably to Citrus because it incorporated both technical covariates (e.g. batch) and clinical covariates when modeling associations, a key feature when analyzing high-dimensional datasets of large disease cohorts. Additionally, the agglomerative hierarchical clustering framework in Citrus substantially increased the testing burden and limited power. Although we found that clustering with DensVM outperformed FlowSOM and Phenograph on our data (**Figure S3**), we want to emphasize that many clustering methods have emerged to analyze both cytometry and single cell RNA-seq data; clustering single cell data remains an active field of research[172,173]. There continues to be controversy as to the best strategies, most appropriate metrics, and the optimal parameter choices[173-176]. Available clustering approaches utilize a variety of methods, such as graph-based community detection, self-organizing maps[130], and density-based clustering[132,162,177]. As neural networks and as a subclass, autoencoders, have been proven to be powerful general nonlinear models, we implemented an autoencoder-based clustering approach that we tested on our datasets. We found this clustering method also had potential and should be further investigated in the future (**Figure S14**). As different clustering strategies may be more appropriate for different datasets, we have implemented MASC specifically to allow for different clustering options based on user preference.

The CD27- HLA-DR+ T cells identified by MASC in blood samples from RA patients were further enriched in both synovial tissue and synovial fluid of RA patients. The accumulation of these cells in chronically inflamed RA joints, combined with lack of CD27, suggests that these cells have been chronically activated. Loss of CD27 is characteristic of T cells that have been repeatedly activated, for example T cells that recognize common restimulation

antigens[178-180]. Importantly, the broader CD27- CD4+ T cell population did not itself differ in frequency between RA patients and controls, in contrast to the expanded population of CD27- HLA-DR+ cells identified by MASC. We hypothesize the CD27- HLA-DR+ T cell population in RA patients may be enriched in RA antigen-specific T cells, offering a potential tool to identify relevant antigens in RA.

Expression of HLA-DR suggests recent or continued activation of these cells in vivo. The well-known HLA class II association to RA may also suggest an important role in disease susceptibility for this subset. Despite the suspected chronic activation of these cells, CD27- HLA-DR+ cells did not appear functionally exhausted. CD27- HLA-DR+ cells rapidly produced multiple effector cytokines upon stimulation in vitro, with an increased predisposition to IFN-☐ production. These cells also showed increased expression of cytolytic molecules such as granzyme A and perforin on both the transcriptomic and proteomic level.
CD4+ cytotoxic T cells expressing granzyme A and perforin, also with a CD27- phenotype, are reported to be expanded in patients with chronic viral infections[181] *(53)*. Of note, CD4+ cells that are perforin+ and granzyme A+ have been observed in RA synovial samples[182,183] and other chronic inflammatory conditions[66,184]. Interestingly, cytolytic CD4+ T cells lack the capacity to provide B cell help, suggesting that this is a distinct population from the expanded T peripheral helper cell population in seropositive RA[105,182]. These findings nominate CD27- HLA-DR+ T cells as a potential pathogenic T cell population that may participate in the chronic autoimmune response in RA.

Although we were able to detect significant associations between the frequency of CD27- HLA-DR+ cells and clinical response (**Figure 3-5A**), we recognize the need for a larger study to detect other subtle subphenotypic associations, such as association with specific therapeutics or disease activity (**Figure S15**). To this end, larger prospective cohorts with deep immunophenotyping of CD4+ T cells in blood and tissue will be critical.

We note that this study has certain limitations. First, the degree of expansion of CD27- HLA-DR+ T cells in the periphery is modest and varies between individuals. While we were able to recover the majority of cells identified as expanded by MASC in our mass cytometry experiments using CD27 and HLA-DR as markers; we note that the more fine-grained immunophenotyping may help to refine phenotypes that even more specifically pinpoint the disease-associated T cell population. While we have characterized these cells as effector memory CD4+ T cells that produce molecules associated with cytotoxicity and Th1 phenotype, further characterization is needed in future studies. For example, we assigned an effector memory phenotype to the CD27- HLA-DR+ cell subset based on upon mass cytometry data; however, the mass cytometry panel did not specifically measure the expression of CD45RA. While this would appear to leave open the possibility that these cells might belong to a T_{EMRA} (CD45RO+ CD45RA+) subset, we note that our sorting strategy for cells prior to assay by mass cytometry specifically excluded CD45RA+ cells and consider this possibility unlikely. In addition, our study focused exclusively on CD45RO+ memory CD4+ T cells, thus we have not assessed other CD4+ T cells populations that may also have cytotoxic features, including T effector memory CD45RA+ (T_{EMRA}) cells[185]. A broader cytometric assessment of T cell and other immune cell phenotypes in RA will be of interest in subsequent studies. Also, we did observe that the mRNA and protein expression of specific markers was not always concordant. Applying more recently developed single cell techniques that ascertain protein and mRNA expression simultaneously would better describe the dynamics of CD27- HLA-DR+ T cells in response to stimulation[144,186]. We anticipate that MASC will be applied test for case-control associated differential abundance across multiple cell types in the future.

In summary, the MASC single-cell association modeling framework identified a Th1-skewed cytotoxic effector memory CD4+ T cell population expanded in RA using a case-control mass cytometry dataset. The MASC method is adaptable to any case-control experiment in which single cell data are available, including flow cytometry, mass cytometry, and single cell

RNA-seq datasets. Although current single cell RNA-seq studies are not yet large scale, ongoing projects may benefit from using the MASC framework in case-control testing, or testing for other clinical subphenotypes such as specific treatment response or disease progression.

Materials and Methods

Study Design

The objective of this research study was to profile immune subsets in peripheral blood samples from RA patients and osteoarthriris (OA) controls to identify disease-related alterations or changes in frequency among CD4+ T cells. Human subject research was performed in accordance with the Institutional Review Boards at Partners HealthCare and Hospital for Special Surgery via approved protocols (Partners HealthCare Protocol 2014P00255) with appropriate informed consent as required. Patients with RA fulfilled the ACR 2010 Rheumatoid Arthritis classification criteria, and electronic medical records were used to ascertain patients' rheumatoid factor and anti-CCP antibody status, C-reactive protein level, and medication use. Synovial tissue samples for mass and flow cytometry were collected from seropositive RA patients undergoing arthroplasty at the Hospital for Special Surgery, New York or at Brigham and Women's Hospital, Boston. Samples with lymphocytic infiltrates on histology were selected for analysis. Sample inclusion criteria were established prospectively, with the exception of samples that were excluded from the study due to poor acquisition by the mass cytometer. The study sample size was a result of including all samples that passed quality control and not set prospectively.

Synovial fluid samples were obtained as excess material from a separate cohort of patients undergoing diagnostic or therapeutic arthrocentesis of an inflammatory knee effusion as directed by the treating rheumatologist. These samples were de-identified; therefore, additional clinical information was not available.

Blood samples for clinical phenotyping were obtained from consented patients seen at Brigham and Women's Hospital. We performed medical record review for ACPA (anti-citrullinated protein antibody) positivity according to CCP2 (cyclic citrullinated protein) assays, C-reactive protein (CRP), and RA-specific medications including methotrexate and biologic DMARDS. For blood cell analyses in the cross-sectional cohort, the treating physician measured the clinical disease activity index (CDAI) on the day of sample acquisition. For RA patients followed longitudinally, a new disease-modifying antirheumatic drug (DMARD) was initiated at the discretion of the treating rheumatologist, and CDAIs were determined at each visit by trained research study staff. Blood samples were acquired before initiation of a new biologic DMARD or within 1 week of starting methotrexate and 3 months after initiating DMARD therapy[187]. Concurrent prednisone at doses ≤10mg/day were permitted. All synovial fluid and blood samples were subjected to density centrifugation using Ficoll-Hypaque to isolate mononuclear cells, which were cryopreserved for batched analyses.

Sample Preparation for Mass Cytometry

We rapidly thawed cryopreserved PBMCs and isolated total CD4+ Memory T cells by negative selection using MACS magnetic bead separation technology (Miltenyi). Subsequently, we rested the CD4+ Memory T cells for 24 hours in complete RPMI (Gibco) sterile-filtered and supplemented with 15% FBS, 1% Pen/Strep (Gibco), 0.5% Essential and Non-Essential Amino Acids (Gibco), 1%Sodium Pyruvate (Gibco), 1% HEPES (Gibco), and 55μM 2-mercaptoethanol (Gibco). We activated the cells using Human T-Activator CD3/CD28 Dynabeads (ThermoFisher) at a density of 1 bead:2 cells. At 6 hours prior to harvesting (t=18 hours of stimulation), we added Monensin and Brefeldin A 1:1000 (BD GolgiPlug and BD GolgiStop). After 24 hours of stimulation, we incubated the cells with a rhodium metallointercalator (Fluidigm) in culture at a final dilution of 1:500 for 15 minutes as a viability measure. We then harvested cells into FACS tubes and washed with CyTOF Staining Buffer (CSB) composed of PBS with 0.5% BSA (Sigma-

Aldrich), 0.02% sodium azide (Sigma-Aldrich), and 2μM EDTA (Ambion). We spun the cells at 500xg for 7 minutes at room temperature. We incubated the resulting cell pellets in 10ul Fc Receptor Binding Inhibitor Polyclonal Antibody (eBioscience) and 40ul of CSB for 10 minutes at 4C. The samples were then incubated for 30 minutes at 4C on a shaker rack with 1ul of the following eighteen CyTOF surface antibodies in a cocktail brought to a volume of 50ul/sample in CSB: Anti-Human CD49D (9F10)-141Pr (Fluidigm), Anti-Human CCR5 (CD195)(–P-6G4) -144Nd (Fluidigm), Anti-Human CD4 (RPA-T4) -145Nd (Fluidigm), Anti-Human CD8a (RPA-T8) -146Nd (Fluidigm), Anti-Human CD45RO-147Sm (Brigham and Women's Hospital CyTOF Core), Anti-Human CD28-148Nd (BWH CyTOF Core), Anti-Human CD25 (IL-2R– (2A3) -149Sm (Fluidigm), Anti-Human PD1-151Eu (BWH CyTOF Core), Anti-Human CD62L (DREG-56)-153Eu (Fluidigm), Anti-Human CD3 (UCHT1) -154Sm (Fluidigm), Anti-Human CD27 (L128) -155Gd (Fluidigm), Anti-Human CD183[CXCR3](G025H7)-156Gd (Fluidigm), Anti-Human CCR7-170Er (BWH CyTOF Core), Anti-Human ICOS-160Gd (BWH CyTOF Core), Anti-Human CD38 (HIT2)-167Er (Fluidigm), Anti-Human CD154 (CD40L) (24-31)-168Er (Fluidigm), Anti-Human CXCR5[CD185](51505)-171Yb (Fluidigm), and Anti-Human HLA-DR (L243) -174Yb (Fluidigm). We washed the cells with 1 ml of CSB and spun at 700xg for 5 minutes at RT. Post spin, we aspirated the buffer from pellet and added 1 ml 1:4 ratio of concentrate to diluent of a Foxp3 / Transcription Factor Staining Buffer Set (eBioscience) supplemented with formaldehyde solution (Sigma-Aldrich #F1268) to a final concentration of 1.6%. We incubated the cells at room temperature on a gentle shaker in the dark for 45 minutes. We washed the cells with two ml of CSB + 0.3% saponin (CSB-S), and spun at 800xg for 5 minutes. We incubated the cell pellet with 1ul of the following fourteen intracellular antibodies and 10ul of a solution of Iridium (1:25) in CSB-S brought to a total volume of 100ul for 35 minutes at r.t. on a gentle shaker: Anti-Human IL-4 (MP4-25D2)-142 (Fluidigm), Anti-Mouse/Human IL-5 (TRFK5) -143Nd (Fluidigm), Anti-Human IL-22 (22URTI) -150Nd (Fluidigm), Anti-Human TNFα (Mab11) -152Sm (Fluidigm), Anti-Human IL-2 (MQ1-17H12)-

158Gd (Fluidigm), Anti-Human IL21-159Tb (BWH CyTOF Core), Anti-Human IFNg (B27) - 165Ho (Fluidigm), Anti-Human GATA3-166Er (BWH CyTOF Core), Anti-Human IL9-172Yb (BWH CyTOF Core), Anti-Human Perforin (B-D48)-175Lu (Fluidigm), Anti-Human IL10-176Yb (BWH CyTOF Core), Anti-Human IL17A-169Tm (BWH CyTOF Core), Anti-Human Foxp3 (PCH101)-162Dy (Fluidigm), and Anti-Human Tbet-164Dy (BWH CyTOF Core). Post-incubation, we washed the cells in PBS and spun them at 800xg for 5 minutes. We resuspended the pellet in 1 ml of 4% formaldehyde prepared in CSB and incubated the cells for 10 minutes at r.t. on a gentle shaker, followed by another PBS wash and spin at 800xg for 5 minutes. We washed the pellet in 1ml of MilliQ deionized water, spun at 800xg for 6 minutes, and subsequently resuspended the resulting pellet in deionized water at a concentration of 700,000 cells per ml for analysis via the CyTOF 2. We transferred the suspensions to new FACS tubes through a 70µm cell strainer and added MaxPar EQ Four Element Calibration Beads (Fluidigm #201078) at a ratio of 1:10 by volume prior to acquisition.

Mass Cytometry Panel Design

We designed an antibody panel for mass cytometry with the goal of both accurately identifying CD4+ effector memory T cell populations and measuring cellular heterogeneity within these populations. We chose markers that fell into one of five categories to generate a broadly informative panel: chemokine receptors, transcription factors, lineage markers, effector molecules, and markers of cellular activation and exhaustion (**Table S1**).

Mass Cytometry Data Acquisition

We analyzed samples at a concentration of 700,000 cells/ml on a Fluidigm-DVS CyTOF 2 mass cytometer. We added Max Par 4-Element EQ calibration beads to every sample that was run on the CyTOF 2, which allowed us to normalize variability in detector sensitivity for samples run in different batches using previously described methods[188]. We used staining for iridium and

rhodium metallointercalators to identify viable singlet events. We excluded samples where acquisition failed or only yielded a fraction of the input cells (< 5%). The criteria for sample exclusion were not set prospectively but were maintained during all data collection runs.
As samples were processed and analyzed on different dates, we ran equal numbers of cases and controls each time to guard against batch effects. However, as CyTOF data are very sensitive to day-to-day variability, we took extra steps to pre-process and normalize data across the entire study.

Flow Cytometry Sample Preparation

For flow cytometry analysis of the validation and longitudinal blood cohorts and synovial samples, cryopreserved cells were thawed into warm RPMI/10% FBS, washed once in cold PBS, and stained in PBS/1% BSA with the following antibodies for 45 minutes: anti-CD27-FITC (TB01), anti-CXCR3-PE (CEW33D), anti-CD4-PE-Cy7 (RPA-T4), anti-ICOS-PerCP-Cy5.5 (ISA-3), anti-CXCR5-BV421 (J252D4), anti-CD45RA-BV510 (HI100), anti-HLA-DR-BV605 (G46-6), anti-CD49d-BV711 (9F10), anti-PD-1-APC (EH12.2H7), anti-CD3-AlexaFluor700 (HIT3A), anti-CD29-APC-Cy7 (TS2/16), propidium iodide. Cells were washed in cold PBS, passed through a 70-micron filter, and data acquired on a BD FACSAria Fusion or BD Fortessa using FACSDiva software. Samples were analyzed in uniformly processed batches containing both cases and controls.

Flow Cytometry Intracellular Cytokine Staining

Effector memory $CD4^+$ T cells were purified from cryopreserved PBMCs by magnetic negative selection (Miltenyi) and rested overnight in RPMI/10%FBS media. The following day, cells were stimulated with PMA (50ng/mL) and ionomycin (1µg/mL) for 6 hours. Brefeldin A and monensin (both 1:1000, eBioscience) were added for the last 5 hours. Cells were washed twice in cold PBS, incubated for 30 minutes with Fixable Viability Dye eFluor 780 (eBioscience), washed

in PBS/1%BSA, and stained with anti-CD4-BV650 (RPA-T4), anti-CD27-BV510 (TB01), anti-HLA-DR-BV605 (G46-6), anti-CD20-APC-Cy7 (2H7), and anti-CD14-APC-Cy7 (M5E2). Cells were then washed and fixed and permeabilized using the eBioscience Transcription Factor Fix/Perm Buffer. Cells were then washed in PBS/1%BSA/0.3% saponin and incubated with anti-IFN-γ-FITC (B27), anti-TNF-PerCp/Cy5.5 (mAb11), anti-IL-10-PE (JES3-9D7) and anti-IL-2-PE/Cy7 (MQ1-17H12) or anti-granzyme A-AF647 (CB9) and anti-perforin-PE/Cy7 (B-D48) for 30 minutes, washed once, filtered, and data acquired on a BD Fortessa analyzer. Gates were drawn to identify singlet T cells by FSC/SSC characteristics, and dead cells and any contaminating monocytes and B cells were excluded by gating out eFluor 780-positive, CD20+, and CD14+ events.

Synovial Tissue Processing

Synovial samples were acquired after removal as part of standard of care during arthroplasty surgery. Synovial tissue was isolated by careful dissection, minced, and digested with 100μg/mL LiberaseTL and 100μg/mL DNaseI (both Roche) in RPMI (Life Technologies) for 15 minutes, inverting every 5 minutes. Cells were passed through a 70μm cell strainer, washed, subjected to red blood cell lysis, and cryopreserved in Cryostor CS10 (BioLife Solutions) for batched analyses.

RNA Library Preparation and Sequencing

Seven RA case samples and six OA control samples were flow sorted into the following seven cellular subsets for low-input RNA-Sequencing: TCM, TN, TReg, DR+27+, DR+27-, DR-27+, DR-27-. We rapidly thawed the case and control samples of cryopreserved peripheral blood mononuclear cells, and MACS enriched the samples for CD4+ T cells (Miltenyi). We rested the samples overnight in complete RPMI/10% FBS. Following the rest, we prepared the samples for fluorescence activated cell sorting (FACS). We washed the cells once in cold PBS, and incubated them with eBioscience human FC Receptor Binding Inhibitor (Thermo). Subsequently, we

stained the samples in PBS/5% FBS for 45 minutes with the following antibodies: FITC CD27 (Biolegend), PE CD25 (Biolegend), Pe/Cy7 CD127(IL-7Ra) (Biolegend), Brilliant Violet 510 HLA-DR (Biolegend), Brilliant Violet 605 CD45RA (Biolegend), APC CD62L (Biolegend), Alexa Fluor 700 CD4 (Biolegend), APC/Cy7 CD14 (Biolegend), and APC/Cy7 CD19 (Biolegend). Post-stain, we washed the cells in cold PBS, passed them through a 70-micron filter, and acquired the samples on a BD FACSAria Fusion cytometer. 1000 cells from each subset were sorted and collected into 5ul of TCL Lysis Buffer (Qiagen) with 1% b-me. We processed and collected the samples uniformly in batches containing both cases and controls, and randomized the samples within the plate. We prepared sequencing libraries using the Smart-Seq2 protocol. Sequenced libraries were pooled and sequenced with the Illumina HiSeq 2500 using 25bp paired-end reads. We removed one outlier with low read depth. The remaining libraries were sequenced to a depth of 6-19M reads.

Flow Cytometry Data Analysis

Flow cytometry data were analyzed using FlowJo 10.0.7 (TreeStar Inc.), with serial gates drawn to identify singlet lymphocytes by FSC/SSC characteristics. Viable memory CD4+ T cells were identified as propidium iodide-negative CD3+ CD4+ CD45RA- cells. We then calculated the frequency of various populations from the pool of memory CD4+ T cells.

Gene Expression Quantification

We quantified cDNAs on canonical chromosomes (autosomal, X, Y, and mitochondrial) in Ensembl release 83 with *Kallisto* v0.43.1 in transcripts per million (TPM). This analysis quantified inferred counts and length-normalized expression (TPM) of transcripts. To quantify gene expression, we collapsed transcripts mapping to the same HGNC genes symbol by summing the TPM values over. We filtered transcripts that were not sufficiently well expressed, omitting those that did not have at least 5 counts in at least 10 samples. This resulted in 15,234

well expressed genes out of 27,717 total genes. For differential expression analysis and principal components analysis (PCA), we used log (base 2) transformed TPM values.

Principal Component Analysis

Using the gene-level log2 TPM expression metric described above, we selected the top 500 most variably expressed genes for PCA, excluding genes with lower than 1 mean or 0.75 standard deviation expression. We used the *removeBatchEffect* function in the R *Limma* package, with default parameters, to regress out donor-specific contributions to gene expression. To prevent the absolute range of a strongly expressed gene from dominating the signal in the PCA, we scaled gene expression using the base R scale function on the rows of the expression matrix. This function centers and scales a vector by subtracting the mean and dividing by standard deviation. We then re-normalized the samples by centering and scaling the columns, with the same R function. Finally, we used *prcomp* in R to perform PCA on the resulting gene expression matrix.

Correlation Analysis

For individual genes, we computed association of their expression with the proposed ordering of cell types, along the naïve to effector gradient. In this association, we modeled expression as a linear function of cell type, encoded as an ordinal variable, and donor, one-hot encoded as a categorical variable. The statistical significance of the association was estimated with the *lm* function in R, followed by adjustment using the Benjamini-Hochberg procedure.

Gene Set Enrichment Analysis

We performed gene set enrichment analysis using the R package *gage* directly on orderings defined by previous analyses. To enrich pathways from the PCA results, we used gene loadings for each principle component. For differential expression, we used the estimated *t* statistics. In

order to produce barcode plots for select pathways, we reanalyzed these pathways using the R *liger* package.

Mixed-effects modeling of Associations of Single Cells (MASC)

Here, we present a flexible method of finding significant associations between subset abundance and case-control status that we have named MASC (Mixed-effects modeling of Associations of Single Cells). The MASC framework has three steps: (1) stringent quality control, (2) definition of population clusters, and (3) association testing. Here we assume that we have single cell assays each quantifying M possible markers where markers can be genes (RNA-seq) or proteins (cytometry).

Quality control. To mitigate the influence of batch effects and spurious clusters, we first removed poorly recorded events and low-quality markers before further analysis. We removed those markers (1) that have little expression, as these markers are not informative, and (2) with significant batch variability. First, we concatenated samples by batch and measured the fraction of cells negative and positive for each marker. We then calculated the ratio of between-batch variance to total variance for each marker's negative and positive populations, allowing us to rank and retain 20 markers that were the least variable between batches. We also removed markers that were either uniformly negative or positive across batches, as this indicated that the antibody for that marker was not binding specifically to its target. For single-cell transcriptomic data, an analogous step would involve removing genes with low numbers of supporting reads or genes whose expression varies widely between batches.

Once low-quality markers were identified and removed, we removed events that were likely to be artifacts. We first removed events that had extremely high signal for a single marker: events that have recorded expression values at or above the 99.9^{th} percentile for that marker are removed. These events were considered unlikely to be intact, viable cells. Next, a composite

"information content" score (eq. 1) for each event i was created in the following manner: the expression x for each marker M is rescaled from 0 to 1 across the entire dataset to create normalized expression values y_i for each event i. The sum of these normalized expression values was used to create the event's information content score.

1) $INFO_i = \sum_{m=1}^{m=M} y_{i,m}$

The information content score reflects that events with little to no expression in every channel are less informative than events that have more recorded expression. Events with low scores ($INFO_i < 0.05$) were considered unlikely to be informative in downstream analysis and were removed. In addition, events that derived more than half of their information content score from expression in a single channel were also removed (eq. 2):

2) $INFO_i * 0.5 < \max_{m \in M}(y_{i,m})$

Potential explanations for these events include poorly stained cells or artifacts caused by the clumping of antibodies with DNA fragments. A final filtering step retained events that were recorded as having detectable expression in at least M_{min} markers, where M_{min} may vary from experiment to experiment based on the panel design and expected level of co-expression between channels. The quality control steps described here are specific for mass cytometry analysis and will need to be optimized for use with transcriptomic data.

Clustering. After applying quality control measures to each sample, we combined data from cases and controls into a single dataset. It was critical to ensure that each sample contributed equal numbers of cells to this dataset, as otherwise the largest samples would dominate the analysis and confound association testing. After sampling an equal number of cells from each sample, we partitioned these cells into populations using DensVM *(26)*, which performs unsupervised clustering based on marker expression. We note that partitioning the data can be accomplished with different clustering approaches – such as SPADE or PhenoGraph for mass

cytometry data – or even by using traditional bivariate gating, as MASC is not dependent on any particular method of clustering (**Figure S5**)

Association testing. Once all cells were assigned to a given cluster, the relationship between single cells and clusters was modeled using mixed-effects logistic regression to account for donor or technical variation (eq. 3). We modeled the age and sex of sample *k as* fixed effect covariates, whereas the donor and batch that cell *i* belongs to were modeled as random effects. The random effects variance-covariance matrix treated each sample and batch as independent gaussians. Each cluster was individually modeled. Note that this baseline model did not explicitly include any single cell expression measures.

3) $\log\left[\frac{Y_{i,j}}{1-Y_{i,j}}\right] = \theta_j + \beta_{clinical}X_{i,k} + (\phi_i|k) + (\kappa_i|m)$

where $Y_{i,j}$ is the odds of cell i belonging to cluster j, θ_j is the intercept for cluster j, $\beta_{clinical}$ is a vector of clinical covariates for the kth sample, $(\phi_i|k)$ is the random effect for cell i from kth sample, $(\kappa_i|m)$ is the random effect for cell i from batch m.

To determine if any clusters were associated with case-control status, we included an additional covariate that indicated whether the kth sample is a case or control (eq. 4)

4) $\log\left[\frac{Y_{i,j}}{1-Y_{i,j}}\right] = \theta_j + \beta_{clinical}X_{i,k} + (\phi_i|k) + (\kappa_i|m) + \beta_{case}X_{i,k}$

Here, $Y_{i,j}$ is the odds of cell i belonging to cluster j, θ_j is the intercept for cluster j, $\beta_{clinical}$ is a vector of clinical covariates for the kth sample, $(\phi_i|k)$ is the random effect for cell i from kth sample, $(\kappa_i|m)$ is the random effect for cell i from batch m, β_{case} indicates the effect of kth sample's case-control status.

5) $D = -2 * \ln\left(\frac{\text{likelihood for null model}}{\text{likelihood for full model}}\right)$

6) $p = 1 - \left(\frac{t^{(v-2)/2}e^{-t/2}}{2^{v/2}\Gamma(\frac{v}{2})}\right)$

We compared the two models using a likelihood ratio test (eq. 5) to find the test statistic *D*, which is the ratio of the likelihoods for the baseline and full models. The term *D* is distributed under the null by a χ^2 distribution with 1 degree of freedom, as there is only one additional parameter in the full model compared to the null (case-control status). We derived a *p*-value by comparing test statistic *D* of the likelihood ratio test to the value of the χ^2 distribution with 1 degree of freedom (eq. 6), allowing us to find clusters in which case-control status significantly improves model fit. A significant result ($p < 0.05$ after multiple testing correction) indicated that cluster membership for a single cell is influenced by case-status after accounting for technical and clinical covariates. The effect size of the case-control association can be estimated by calculating the odds ratio from β_{case}. If a dataset includes multiple groups, then we can test for association between *g* groups using *g*-1 indicator variables. This approach allowed us to capture inter-individual differences between donors, as well as model the influence of technical and clinical covariates that might influence a cell to be included as a member of one cluster versus another.

Mass Cytometry Data Analysis

We analyzed 50 magnetically sorted peripheral blood samples (26 cases, 24 controls) in the resting condition and 52 samples (26 cases, 26 controls) in the stimulated condition by CyTOF. Here, we stimulated samples by incubating them with Human T-Activator CD3/CD28 Dynabeads (ThermoFisher) at a density of 1 bead:2 cells for 24 hours. Two samples were only analyzed after stimulation due to low numbers of available PBMCs. We ran aliquots of a standard PBMC sample alongside cases and controls with each CyTOF run to allow us to measure batch variability directly, as these aliquots should not be biologically dissimilar. We then used these data to find markers that had stained poorly or varied significantly between batches and removed them analysis. After acquisition, each sample was gated to a $CD4^+$,

CD45RO$^+$ population using FlowJo 10.1 (TreeStar, USA) and combined into a single dataset before analyzing the data using MASC as previously described. We performed the data filtration steps requiring cells to demonstrate measurable expression (arcsinh-transformed expression > 0) in at least 5 markers (M_{min} = 5). This removed 3.0-6.0% of all events captured in each sample. We first removed the initial noise factor applied to all zero expression values in mass cytometry by subtracting 1 from expression values and setting any negative values to 0, then applied the inverse hyperbolic sine transform with a cofactor of 5 to the raw expression data, using the following equation: $y = \sinh^{-1}\frac{\max(x-1,0)}{5}$

To partition the data, we first randomly selected 1000 cells from each sample and applied the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm (Barnes-Hut implementation)[142] to the reduced dataset with the following parameters: perplexity = 30 and theta = 0.5. We did not include channels for CD4 or CD45RO in the t-SNE clustering as these markers were only used in gating samples to confirm the purity of CD4 memory T cell selection. We performed separate t-SNE projections for resting and stimulated cells. To identify high-dimensional populations, we used a modified version of DensVM[134]. DensVM performs kernel density estimation across the dimensionally reduced t-SNE map to build a training set, then assigns cells to clusters by their expression of all markers using an SVM classifier. We modified the DensVM code to increase the range of potential bandwidths searched during the density estimation step and to return the SVM model generated from the t-SNE projection.

To create elbow plots, we ran DensVM using 25 bandwidth settings evenly spaced along the interval [0.61, 5] for the resting data and [0.66, 5] for the stimulated data. We normalized marker expression to mean 0 variance 1 in each dataset before calculating the fraction of total variance explained by between-cluster variance.

Association testing for each cluster was performed using mixed-effects logistic regression according to the MASC method. Donor and batch were included as random-effect covariates, donor sex and age was included as a fixed-effect covariate, and donors were labeled as either

cases (RA) or controls (OA). To confirm the associations found by MASC, we conducted exact permutation testing in which we permuted the association between case-control status and samples within batches 10,000 times, measuring the fraction of cells from RA samples that contributed to each cluster in each permutation. This allowed us to build an empirical null for the case-control skew of each cluster, and we could then determine for each cluster how often a skew equal or greater to the observed skew occurred. We adjusted the p-values for the number of tests we performed (the number of clusters analyzed in each condition) using the Bonferroni correction.

Cluster Alignment

We aligned subsets between experiments using the following strategy: In each experiment, expression data was first scaled to mean zero, variance one to account for differences in sensitivity. We used the mean expression value for marker column to define a centroid for each cluster in both datasets. For a given cluster in the first dataset (query dataset), we calculated the Euclidean distance (eq. 7) between that cluster and cluster centroids in the second dataset (target dataset) across all shared markers. The cluster that is most similar to the cluster in query dataset is the cluster with the lowest distance in the target dataset, relative to all other clusters in that dataset.

7) $\text{dist}(q, t) = \sum_{k=1}^{K} \sqrt{(q_k - t_k)^2}$

Here, q refers to the query cluster and t to the target cluster, while q_k and t_k indicate the normalized mean expression of marker k (out of K total markers) in clusters q and t respectively.

Marker Informativeness Metric (MIM)

We wanted to determine which markers best separated a given population from the rest of the data in a quantitative manner, as finding a set of population-specific markers is crucial for isolating the population *in vivo*. In order to do this, we examined the distribution of expression

for each marker individually in the entire dataset (*Q*) and the population of interest (*P*). We then grouped expression values for *P* and *Q* into 100 bins, normalized the binned vector to 1, and calculated the Kullback-Leibler divergence from *Q* to *P* for that marker with the following equation:

$$D_{KL}(P \parallel Q) = \sum_{i=1}^{i=100} P_i \log \frac{P_i}{Q_i}$$

The divergence score can be interpreted as a measure of how much the distribution of expression for a given marker in the entire dataset resembles the distribution of expression for that marker in the population of interest. Higher scores represent lower similarity of the marker's expression distributions for *P and Q*, indicating that the expression profile of that marker is more specific for that population. By calculating this score for every marker, we can rank and identify markers that best differentiate the population of interest from the dataset.

Biaxial Gating and Cluster Overlap

To determine the concordance between biaxial gating of CD27- HLA-DR+ cells and cluster 18, we first selected cells that had normalized expression values of CD27 < 1 and HLA-DR >= 1. We then calculated an F-measure statistic between the cells selected using the CD27 and HLA-DR gates and each cluster identified in the resting and stimulated datasets (eq. 8). Here, precision is defined as the number of cells in each cluster tested that fall into the CD27- HLA-DR+ gate, while recall is defined as the number of cells gated as CD27- HLA-DR+ that are in each cluster.

8) $$F_{measure} = 2 \times \frac{precision \times recall}{precision + recall}$$

Clustering Informativeness Metric (CIM)

We compared clustering sets that contained either 19 (FlowSOM and DensVM) or 21 (Phenograph) clusters. We independently clustered the resting dataset with Phenograph and FlowSOM using the same cells and markers used to cluster the data with DensVM. We set *k* to

19 for FlowSOM clustering to match the number of clusters found by DensVM; for Phenograph, we used the default setting of k = 30.

To evaluate quantify the ability of different clustering algorithms to define clusters that was explaining marker fluctuations, we defined an information theory-based metric to evaluate the relative information content captured by each set of clusters in terms of marker intensity. We selected this approach since it is separate from the objective functions that the clustering algorithms were attempting to optimize.

First, for each cell, we normalized marker intensities so that they summed to one. Then we defined a null Q_i representing the average normalized intensity for marker i across all cells. We also defined $P_{i,j}$ which is the mean intensity of marker i of cells from cluster j. Then for each cluster j we calculate their KL divergence for each of the M markers (eq. 9).

9) $D_{KL,j}(P_j \parallel Q) = \sum_i^M P_{i,j} \ln \frac{P_{i,j}}{Q_i}$

A cluster with low divergence from the average expression of markers across the entire dataset will capture less marker intensity information than one with a high divergence, as biologically valid clusters will have unique marker profiles that differ greatly from one another and from the average marker expression profile.

We defined a similar metric to quantify the extent to which individual batches were accounting for differences in cluster composition. In this instance we calculated $P_{i,j}$ which is the proportion of cells from cluster j that batch i contributed. We also calculate Q_i which is the proportion of cells that batch i contributes overall to the dataset. With this definition we calculate the KL divergence for each of the M batches (eq. 10).

10) $D_{KL,j}(P_{\cdot j} \parallel Q) = \sum_i^M P_{i,j} \ln \frac{P_{i,j}}{Q_i}$

A cluster that contains cells with low divergence from the null distribution of cells across batches is affected less by batch effects than one with a high divergence score, and a cluster completely free of batch effects should have a K-L divergence of zero.

Autoencoder Clustering

We performed clustering analysis using a deep autoencoder and a Gaussian Mixture Model (GMM). The autoencoder was designed with an architecture of 3 hidden layers, with depths of 8, 2, and 8 nodes, respectively. We trained the model with no regularization and 300 epochs. The two modes from the middle hidden layer were then used as features to learn a GMM. The number of clusters was chosen a priori to match the number discovered in the DensVM analysis. All parameters not specific in this section were set to default values.

Meta-Analysis

We used Stouffer's Z-score method to define a meta-analysis p-value for the significant expansion of CD27- HLA-DR+ cells in RA. We converted p-values from the resting mass cytometry and flow cytometry analyses to Z-scores, and found a meta-analysis Z-score by taking the sum of these scores divided by the square-root of the number of scores – which was two, in our case. We then derived a meta-analysis p-value from the Z-score using the standard normal distribution.

All analyses were performed using custom scripts for R 3.4.0. We used the following packages: *flowCore*[189] to read and process FCS files for further analysis, *lme4*[190] to apply mixed-effects logistic regression, *ggplot2*[191], *pheatmap*[192] for data visualization, and *cytofkit*[193] for the implementation of FlowSOM and Phenograph clustering algorithms. RNA-seq analyses were conducted using *Kallisto*[167] to align reads and *gage*[194] and *liger*[195] to perform gene set

enrichment analysis. The *h2o*[196] package and *mclust*[197] packages were used to implement the autoencoder clustering method.

Acknowledgements

This work was supported in part by funding from the National Institutes of Health: UH2AR067677 (S.R.), U19AI111224 (S.R.), and 1R01AR063759 (S.R.), R01 AR064850-03 (Y.C.L.), the Doris Duke Charitable Foundation Grant #2013097 (S.R.), T32 AR007530-31 (M.B.B., D. A. R.), the William Docken Inflammatory Autoimmune Disease Fund (M.B.B., S. R.), the Ruth L. Kirschstein National Research Service Award F31-AR070582 (K. S.), and the Rheumatology Research Foundation Tobe and Stephen Malawista, MD Endowment in Academic Rheumatology (D.A.R).

Competing Interests

The authors declare that they have no competing interests. I.K. has been a paid bioinformatics consultant for Outlier Bio LLC since November 2017.

Data and Materials Availability

All data associated with this study can be found in the paper or supplementary materials. The data that support the findings of this study has been deposited in the database GEO (GSE118209). RNA sequencing reads have been deposited in the SRA database (SRP156530). MASC and all other custom scripts used in this analysis are available at https://www.github.com/immunogenomics.

Chapter 4:

Defining Inflammatory Cell States in Rheumatoid Arthritis Joint Synovial Tissues by Integrating Single-cell Transcriptomics and Mass Cytometry

Authors:

Fan Zhang[1,2,3,4,5,^], Kevin Wei[5,^], Kamil Slowikowski[1,2,3,4,5,^], Chamith Y. Fonseka[1,2,3,4,5,^], Deepak A. Rao[5,^], Stephen Kelly[6], Susan M. Goodman[7,8], Darren Tabechian[9], Laura B. Hughes[10], Karen Salomon-Escoto[11], Gerald F. M. Watts[5], Anna H. Jonsson[5], Javier Rangel-Moreno[9], Nida M. Pellett[9], Cristian Rozo[8], William Apruzzese[5], Thomas M. Eisenhaure[4], David J. Lieb[4], David L. Boyle[12], Arthur M. Mandelin II[13], Accelerating Medicines Partnership: RA Phase 1[14], AMP RA/SLE, Brendan F. Boyce[15], Edward DiCarlo[8,16], Ellen M. Gravallese[11], Peter K. Gregersen[17], Larry Moreland[18], Gary S. Firestein[12], Nir Hacohen[4], Chad Nusbaum[4], James A. Lederer[19], Harris Perlman[13], Costantino Pitzalis[20], Andrew Filer[21,22], V. Michael Holers[23], Vivian P. Bykerk[7,8], Laura T. Donlin[8,24,*], Jennifer H. Anolik[9,25,*], Michael B. Brenner[5,*], Soumya Raychaudhuri[1,2,3,4,5,26,*+]

^Co-first authors
*Co-senior authors
+Corresponding author. Email: soumya@broadinstitute.org

Affiliations:

[1]Center for Data Sciences, Brigham and Women's Hospital, Boston, MA 02115, USA
[2]Division of Rheumatology and Genetics, Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA
[3]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115 USA
[4]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[5]Division of Rheumatology, Immunology, Allergy, Brigham and Women's Hospital and Harvard Medical School, MA 02115, USA
[6]Department of Rheumatology, Barts Health NHS Trust, London, E1 1BB, UK
[7]Division of Rheumatology, Hospital for Special Surgery, New York, NY 10021, USA
[8]Department of Medicine, Weill Cornell Medical College, New York, NY 10065, USA
[9]Division of Allergy, Immunology and Rheumatology, Department of Medicine, University of Rochester Medical Center, Rochester, NY 14642, USA
[10]Division of Clinical Immunology and Rheumatology, Department of Medicine, University of Alabama at Birmingham, Birmingham, AL 35294-2182, USA
[11]Division of Rheumatology, Department of Medicine, University of Massachusetts Medical School, Worcester, MA 01605, USA
[12]Department of Medicine, Division of Rheumatology, Allergy and Immunology, University of California, San Diego, La Jolla, CA 92093 USA
[13]Division of Rheumatology, Department of Medicine, Northwestern University Feinberg School of Medicine. Chicago, IL 60611, USA
[14]AMP RA Phase 1: full list of members and affiliations appears in the end of the paper

[15]Department of Pathology and Laboratory Medicine, University of Rochester Medical Center, Rochester, NY 14642, USA
[16]Department of Pathology and Laboratory Medicine, Hospital for Special Surgery, New York, NY 10021, USA
[17]Feinstein Institute for Medical Research, Northwell Health, Manhasset, New York, NY 11030, USA
[18]Division of Rheumatology and Clinical Immunology, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261 USA
[19]Department of Surgery, Brigham and Women's Hospital and Harvard Medical School, MA 02115, USA
[20]Centre for Experimental Medicine & Rheumatology, William Harvey Research Institute, Queen Mary University of London, E1 4NS, UK
[21]Rheumatology Research Group, Institute of Inflammation and Ageing, The University of Birmingham, Birmingham, B15 2WB, UK
[22]University Hospitals Birmingham NHS Foundation Trust, Birmingham, B15 2TH, UK
[23]Division of Rheumatology, University of Colorado School of Medicine, Aurora, CO 80220, USA
[24]Arthritis and Tissue Degeneration, Hospital for Special Surgery, New York, NY 10021, USA
[25]Center for Musculoskeletal Research, University of Rochester Medical Center, Rochester, NY 14642, USA
[26]Arthritis Research UK Centre for Genetics and Genomics, Centre for Musculoskeletal Research, The University of Manchester, Oxford Road, Manchester, M13 9PT, UK

Attributions

S.K., S.M.G., D.T., L.B.H., K.S.-E., A.M.M., D.L.B., J.H.A., V.P.B., V.M.H., A.F., C.P., H.P., G.S.F., L.M., P.K.G., W.A. and L.T.D. recruited patients and obtained synovial tissues. B.F.B., E.D. and E.M.G. performed histological assessment of tissues. K.W., D.A.R., G.F.M.W., and M.B.B. designed and implemented tissue processing and cell sorting pipeline. J.A.L. obtained mass cytometry data from samples. N.H., C.N., and T.M.E. obtained single cell RNA-seq data from samples. F.Z., K.S., C.Y.F., D.J.L. and S.R. conducted computational and statistical analysis. A.H.J., J.R.-M., N.M.P., and C.R., designed and performed validation experiments. K.S., F.Z., and J.R.M. implemented the website. J.A., S.L.B., C.D.B., J.H.B., J.D., J.M.G., M.G., L.B.I., E.A.J., J.A.J., J.K., Y.C.L., M.J.M., M.M., F.M., J.N., A.N., D.E.O., M.P., C.R., W.H.R., A.S., D.S., J.S., J.D.T., and P.J.U. contributed to the procurement and processing of samples, design of the AMP study. S.R., M.B.B., J.H.A., and L.T.D. supervised the research. F.Z., K.W., K.S, and S.R. generated figures and wrote the initial draft. K.S, C.Y.F. D.A.R, L.T.D., J.H.A, M.B.B. edited the draft, and all the authors participated in writing the final manuscript.

Abstract

To define the cell populations that drive joint inflammation in rheumatoid arthritis (RA), we applied single-cell RNA sequencing (scRNA-seq), mass cytometry, bulk RNA-seq and flow cytometry to T cells, B cells, monocytes and fibroblasts from 51 samples of synovial tissue from patients with RA or osteoarthritis. Utilizing an integrated strategy based on canonical correlation analysis of 5,265 scRNA-seq profiles, we identified 18 unique cell populations. Combining mass cytometry and transcriptomics together revealed cell states expanded in RA synovia: *THY1(CD90)*$^{+}$*HLA-DRA*^{hi} sublining fibroblasts, *IL1B*$^{+}$ pro-inflammatory monocytes, *ITGAX*$^{+}$*TBX21*$^{+}$ autoimmune-associated B cells and *PDCD1*$^{+}$ T peripheral helper (Tph) and T follicular helper (Tfh). We defined distinct subsets of CD8^{+} T cells characterized by a *GZMK*$^{+}$, *GZMB*$^{+}$ and *GNLY*$^{+}$ phenotype. We mapped inflammatory mediators to their source cell populations; for example, we attributed *IL6* expression to *THY1*$^{+}$*HLA-DRA*^{hi} fibroblasts, and *IL1B* production to pro-inflammatory monocytes. These populations are potentially key mediators of RA pathogenesis.

Introduction

Rheumatoid arthritis (RA) is an autoimmune disease affecting up to 1% of the population where a complex interplay between many different cell types drives chronic inflammation in the synovium of the joint tissue[33,198,199]. This inflammation leads to joint destruction, disability and shortened life span[200]. Defining key cellular subsets and their activation states in RA is a critical step to define new therapeutic targets for RA. $CD4^+$ T cell subsets[39,182], B cells[201], monocytes[202,203], and fibroblasts[204-206] have established relevance to RA pathogenesis. Here, we use single cell technologies to view all of these cell types simultaneously across a large collection of samples from inflamed joints. We believe a global single-cell portrait of how different cell types work together would advance our understanding of therapeutics.

Application of transcriptomic and cellular profiling technologies to whole synovial tissue has already identified specific cell populations associated with RA[198,207-209]. However, most studies have focused on a pre-selected cell type, surveyed whole tissues rather than disaggregated cells, or used only a single technology platform. The latest advances in single-cell technologies offer an opportunity to identify disease-associated cell subsets in human tissues at high resolution in an unbiased fashion[84,106,210,211]. These technologies have already been used to discover roles for T peripheral helper (Tph) cells[105] and HLA-DR^+CD27^- cytotoxic T cells[126] in RA pathogenesis. Studies using scRNA-seq has defined myeloid cell heterogeneity in human blood[212] and identified a distinct subset of $PDPN^+CD34^-THY1^+$ (THY1, also known as CD90) fibroblasts enriched in RA synovial tissue[106,213].

To generate high-dimensional multi-modal single-cell data from synovial tissue samples collected across a collaborative network of research sites, we developed a robust pipeline[214] in the Accelerating Medicines Partnership Rheumatoid Arthritis and Lupus (AMP RA/SLE) consortium. We collected and disaggregated tissue samples from patients with RA and osteoarthritis (OA), and then subjected constituent cells to scRNA-seq, sorted-population bulk RNA-seq, mass cytometry, and flow cytometry. We developed a unique computational strategy

based on canonical correlation analysis (CCA) to integrate multi-modal transcriptomic and proteomic profiles at a single cell level. A unified analysis of single cells across data modalities can precisely define contributions of specific cell subsets to pathways relevant to RA and chronic inflammation.

Results

Generation of parallel mass cytometric and transcriptomic data from synovial tissue

In phase 1 of AMP RA/SLE, we recruited 36 patients with RA that met the 1987 American College of Rheumatology (ACR) classification criteria and 15 patients with OA from 10 clinical sites over 16 months (**Table S1**) and obtained synovial tissues from ultrasound-guided biopsies or joint replacements (**Methods**, **Figure 4-1a**). We required that all tissue samples included had synovial lining documented by histology. Synovial tissue disaggregation yielded an abundance of viable cells for downstream analyses (362,190 +/- 7,687 (mean +/- SEM) cells per tissue). We used our validated strategy for cell sorting[214] (**Figure 4-1a**) to isolate B cells ($CD45^+CD3^-CD19^+$), T cells ($CD45^+CD3^+$), monocytes ($CD45^+CD14^+$), and stromal fibroblasts ($CD45^-CD31^-PDPN^+$) (**Figure S1a**). We applied bulk RNA-seq to all four sorted subsets for all 51 samples. For samples with sufficient cell yield (**Methods**), we also measured single-cell protein expression using a 34-marker mass cytometry panel (n=26, **Table S2)**, and single-cell RNA expression in sorted cell populations (n=21, **Figure 4-1b**).

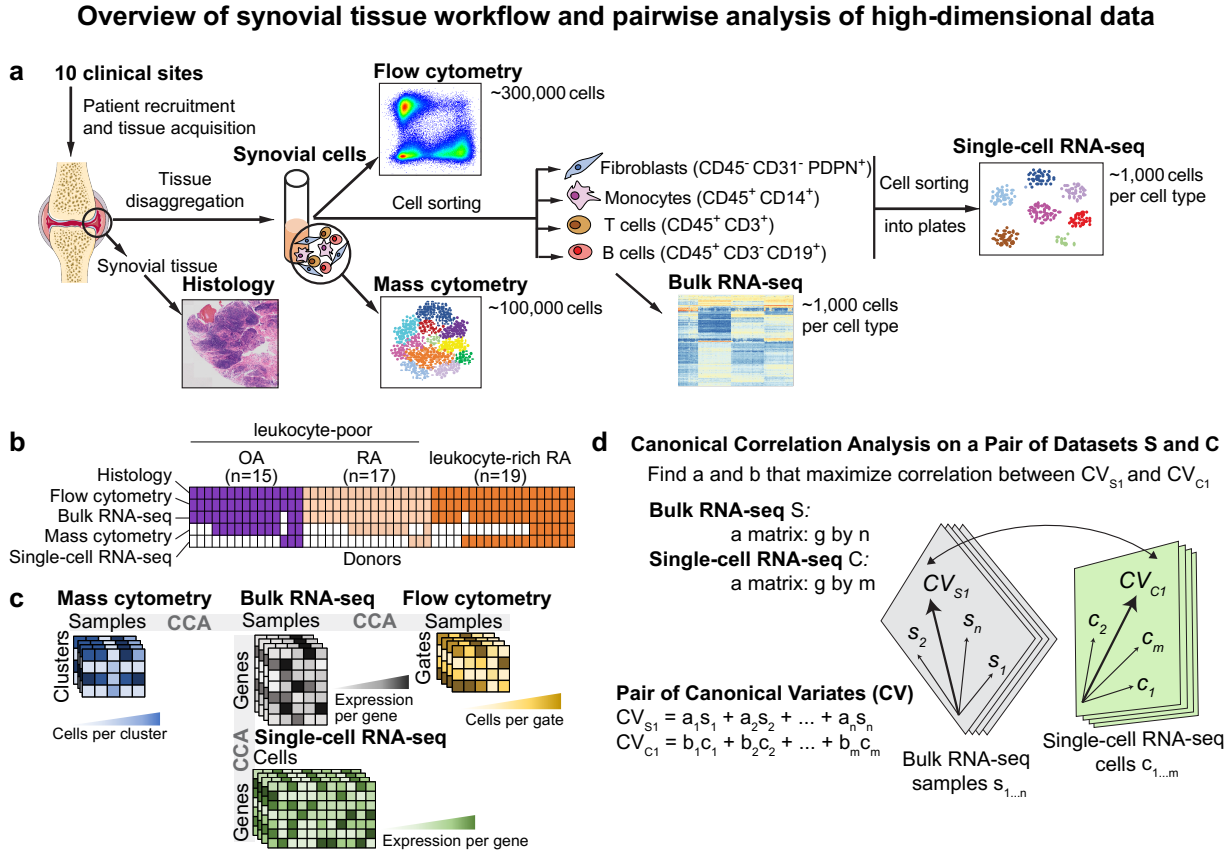


Figure 4-1. **a**. We acquired synovial tissue, disaggregated the cells, sorted them into four gates representing fibroblasts ($CD45^-CD31^-PDPN^+$), monocytes ($CD45^+CD14^+$), T cells ($CD45^+CD3^+$), and B cells ($CD45^+CD3^-CD19^+$). We profiled these cells with mass cytometry, flow cytometry, sorted low-input bulk RNA-seq, and single-cell RNA-seq. Here, we use Servier Medical Art by Servier for the joint picture. **b**. Presence and absence of five different data types for each tissue sample. **c**. Schematic of each dataset and the shared dimensions used to analyze each of the three pairs of datasets with canonical correlation analysis (CCA). **d**. CCA finds a common mapping for two datasets. For bulk RNA-seq and single-cell RNA-seq, we first find a common set of g genes present in both datasets. Each bulk sample s_i gets a coefficient a_i and each cell c_i gets a coefficient b_i. The linear combination of all samples $s_{1...n}$ arranges bulk genes along the canonical variate CV_{s1} and the linear combination of all cells $c_{1...m}$ arranges single-cell genes along CV_{c1}. CCA finds the coefficients $a_{1...n}$ and $b_{1...m}$ that arrange the genes from the two datasets in such a way that the correlation between CV_{s1} and CV_{c1} is maximized. After CCA finds the first pair of canonical variates, the next pair is computed on the residuals, and so on.

Summary of computational data integration strategy to define cell populations

To confidently define RA-associated cell populations, we integrated multiple data modalities (**Figure 4-1b, c**). We use bulk RNA-seq data as the reference point because it was available for all of the donors and most of the cell types, it had the highest dimensionality and least sensitive to technical artifacts (**Figure 4-1b**).

Integrating scRNA-seq with bulk RNA-seq data ensures robust discovery of cell populations. Here, we used CCA to find linear combinations of bulk RNA-seq samples and scRNA-seq cells (**Figure 4-1c, d**) to create gene expression profiles that were maximally correlated. These linear combinations captured sources of shared variation between the two datasets and allowed us to identify individual cell populations that drive variation in the bulk RNA-seq data. We analyzed the scRNA-seq data by using the canonical variate coefficients for each cell to compute a nearest neighbor network, identifying clusters with a community detection algorithm, and evaluating the separation between clusters with Silhouette analysis (**Methods**, **Figure S2b**).

We identified cell clusters in mass cytometry data with density-based clustering[134]. Next, we used CCA to identify linear combinations of bulk RNA-seq genes and mass cytometry cluster abundances that maximize correlation across patients. These canonical variates offer a way to visualize genes and mass cytometry clusters together. We then queried this CCA result with the best marker genes from scRNA-seq to establish a relationship between each scRNA-seq cluster and each mass cytometry cluster (**Methods**). We also used CCA to associate bulk gene expression in each sample with proportions of cells in different flow cytometry gates.

Flow cytometry features define a set of RA synovia that are leukocyte-rich

Histology of RA synovial tissues revealed heterogeneous tissue composition with variable lymphocyte and monocyte infiltration (**Figure 4-2a**, **b, Figure S2c, d**).

Distinct cellular composition in synovial tissue from OA, leukocyte-poor RA, and leukocyte-rich RA patients

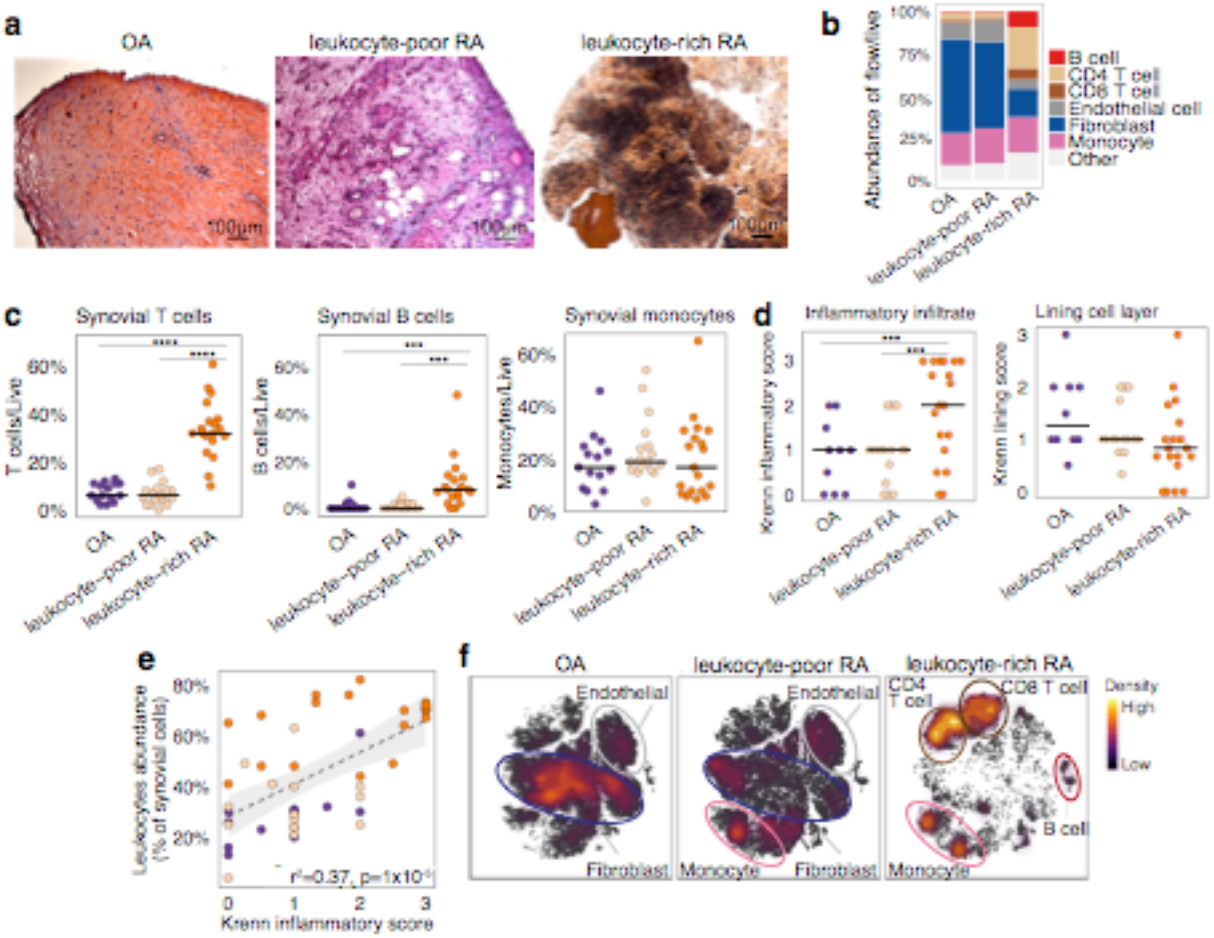


Figure 4-2. **a**. Histological assessment of synovial tissue derived from OA (*n* = 15 independent tissue samples), leukocyte-poor RA (*n* = 17 independent tissue samples), and leukocyte-rich RA (*n* = 19 independent tissue samples). **b**. Cellular composition of major synovial cell types by flow cytometry. **c**. Synovial T cells, B cells, and monocytes by flow cytometry in samples from OA (*n* = 15), leukocyte-poor RA (*n* = 17), and leukocyte-rich RA (*n* = 19). Leukocyte-rich RA tissues were significantly higher infiltrated in synovial T cells (Student's one-sided t-test $P = 4x10^{-9}$, t-value = 8.92, df = 22.27) compared to leukocyte-poor RA and OA. Leukocyte-rich RA tissues were significantly higher infiltrated in synovial B cells (Student's one-sided t-test $P = 1x10^{-3}$, t-value = 3.50, df = 20.56) compared to leukocyte-poor RA and OA. Center value is mean. Statistical significance levels: ****$P<1x10^{-4}$ and ***$P<1x10^{-3}$. **d**. Quantitative histologic inflammatory scoring of both sublining cell layer and lining layer. Leukocyte-rich RA samples (*n* = 19) exhibited higher (Student's one-sided t-test $P = 1x10^{-3}$, t-value = 3.21, df = 30.66) Krenn inflammation scores than leukocyte-poor RA (n=15) and OA tissues (*n* = 10) samples. Center value is mean. **e**. Correlation between leukocyte infiltration assessed by cytometry with histologic inflammation score (*n* = 44 biologically independent samples). Student's one-sided t-test $P = 3x10^{-09}$, t-value = 7.15, df = 46.51. **f**. tSNE visualization of synovial cell types in OA, leukocyte-poor RA, and leukocyte-rich RA by mass cytometry density plot.

This heterogeneity was expected, because variation in tissue immune cell infiltration reflects local disease activity in the source joint. Consequently, we employed a data-driven approach to separate samples based on flow cytometry of lymphocyte and monocyte infiltration in each tissue sample (**Figure S1b,c**). We calculated a multivariate normal distribution of these parameters based on OA samples as a reference, and for each RA sample we calculated the Mahalanobis distance from OA[215]. We defined the maximum OA distance (4.5) as the threshold for defining leukocyte-rich RA (>4.5, n=19) or leukocyte-poor RA (<4.5, n=17) samples (**Methods**, **Figure S1d**). Whereas leukocyte-rich RA tissues had significant infiltration of synovial T cells and B cells, leukocyte-poor RA tissues had cellular compositions more similar to OA (**Figure 4-2c**). Synovial monocyte abundances were similar between RA and OA (**Figure 4-2c**).

To test if our classification indicates inflammation, we assessed tissue histology and assigned each sample a Krenn inflammation score[216]. Samples we classified as leukocyte-rich RA had a significantly higher Krenn inflammation score than leukocyte-poor RA or OA (**Figure 4-2d**). In contrast, synovial lining membrane hyperplasia was not significantly different between leukocyte-rich RA, leukocyte-poor RA, and OA samples (**Figure 4-2d**). We observed significant correlation between synovial leukocyte infiltration measured by flow cytometry and the histological Krenn inflammation score (**Figure 4-2e**). Mass cytometry in 26 synovial tissues was consistent with flow cytometry and histology. OA and leukocyte-poor RA samples were characterized by high abundance of fibroblasts and endothelial cells; while leukocyte-rich RA tissues were characterized by high abundance of CD4 T, CD8 T, and B cells (**Figure 4-2f**, **Figure S3a**).

Single-cell RNA-seq analysis reveals distinct cell subpopulations

Next, we analyzed 5,265 scRNA-seq profiles passing quality control (**Methods**), including 1,142 B cells, 1,844 fibroblasts, 750 monocytes, and 1,529 T cells. We used canonical variates (from CCA with bulk RNA-seq) to define 18 cell clusters that were independent of donor (n=21) and technical plate (n=24) effects (**Figure 4-3a, b**, **Figure S2c, Figure S4a**). In contrast, conventional PCA-based clustering led to clusters that were confounded by batch effects (**Figure S4b**). All of the clusters in the PCA-based clustering, excluding clusters confounded by batch, were identified in CCA-based clustering. Next, we compared expression values between cells in the cluster and all other cells to select cluster marker genes (**Methods**, **Table S4**). For selected genes, we show expression values in each cell positioned in a t-distributed Stochastic Neighbor Embedding (tSNE) (**Figure 4-3c-f**). Among fibroblasts, we identified four putative subpopulations (**Figure 4-3c**): *CD34*$^{+}$ sublining fibroblasts (SC-F1), *HLA-DRA*^{hi} sublining fibroblasts (SC-F2), *DKK3*$^{+}$ sublining fibroblasts (SC-F3), and *CD55*$^{+}$ lining fibroblasts (SC-F4). In monocytes (**Figure 4-3d**), we identified *IL1B*$^{+}$ pro-inflammatory monocytes (SC-M1), *NUPR1*$^{+}$ monocytes (SC-M2), *C1QA*$^{+}$ monocytes (SC-M3), and interferon (IFN) activated monocytes (SC-M4). In T cells (**Figure 4-3e**), we identified three CD4^{+} clusters: *CCR7*$^{+}$ T cells (SC-T1), *FOXP3*$^{+}$ regulatory T cells (T_{reg} cells) (SC-T2), and *PDCD1*$^{+}$ Tph and T follicular helper (Tfh) (SC-T3); and three CD8^{+} clusters: *GZMK*$^{+}$ T cells (SC-T4), *GNLY*$^{+}$*GZMB*$^{+}$ cytotoxic lymphocytes (CTLs) (SC-T5), and *GZMK*$^{+}$*GZMB*$^{+}$ T cells (SC-T6). Within B cells (**Figure 4-3f**), we identified four cell clusters, including naive *IGHD*$^{+}$*CD27*$^{-}$ (SC-B1) and *IGHG3*$^{+}$*CD27*$^{+}$ memory B cells (SC-B2). We identified an autoimmune-associated B cells (ABCs) cluster (SC-B3) with high expression of *ITGAX* (also known as *CD11c*) and a plasmablast cluster (SC-B4) with high expression of immunoglobulin genes *and XBP1*, a transcription factor for plasma cell differentiation[217].

High-dimensional transcriptomic scRNA-seq clustering reveals distinct cell type subpopulations.

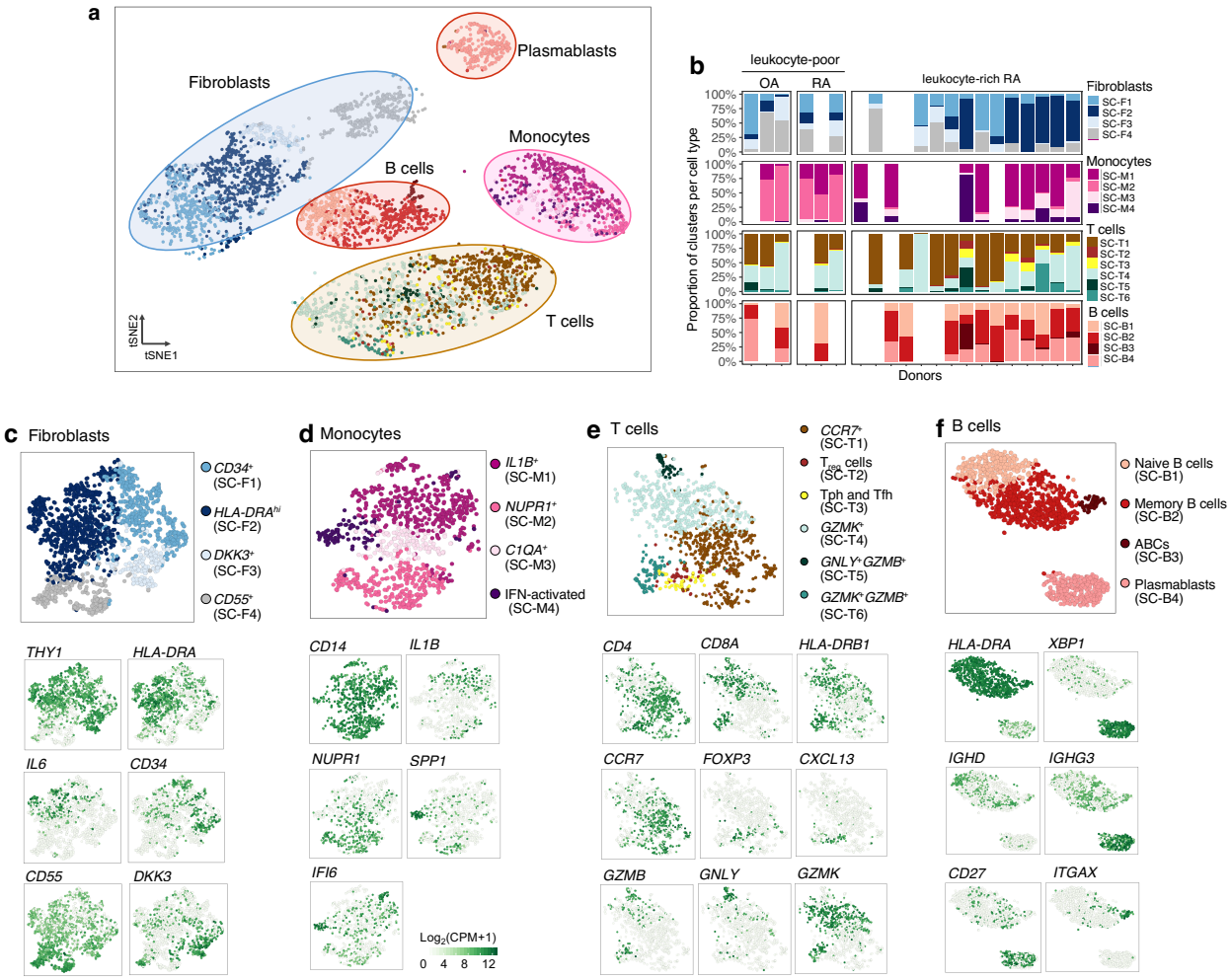


Figure 4-3. **a**. 18 clusters across 5,265 cells from all cell types on a tSNE visualization. **b**. Cluster abundances across donors. **c**. Fibroblasts: three types of THY1$^+$ sublining fibroblasts (SC-F1, SC-F2, and SC-F3) and CD55$^+$ lining fibroblasts (SC-F4). **d**. Monocytes: two activated cell states of IL1B$^+$ pro-inflammatory (SC-M1) and IFN-activated (SC-M4) monocytes. **e**. T cells: CD4$^+$ subsets: SC-T1, SC-T2, SC-T3, and CD8$^+$ subsets: SC-T4, SC-T5, and SC-T6. **f**. B cells: HLA$^+$ (SC-B1, SC-B2, and SC-B3) and plasmablasts (SC-B4). The cluster colors in **c-f** are consistent with (**a**).

We assessed protein fluorescence measurements of typical cell type markers, which were consistent with our identified scRNA-seq clusters (**Figure S2e)**. Cell density quantified from 10 histology samples was correlated with the lymphocyte flow cytometric cell yields, suggesting that samples with the most single cell measurements are those with the best yields and the most inflammation (**Figure S5**).

Distinct synovial fibroblasts defined by cytokine activation and MHC II expression

To identify the fibroblast subpopulations overabundant in leukocyte-rich RA synovia, we selected marker genes for each cluster and assessed their expression levels in bulk RNA-seq from sorted fibroblasts ($CD45^{-}PDPN^{+}$) from RA and OA patients. For example, genes associated with *HLA-DRA*^{hi} (SC-F2) fibroblasts were more highly expressed in bulk RNA-seq samples from leukocyte-rich RA than OA (*t*-test $p<1x10^{-3}$ for *HLA-DRA*, *IFI30*, and *IL6*) (**Figure 4-4a**). Since the expression profile of a bulk tissue sample is an aggregate of the expression profiles of its constituent cell populations, this result suggests expansion of *HLA-DRA*^{hi} (SC-F2) fibroblasts in RA tissues. Genes associated with *CD55*$^{+}$ fibroblasts (SC-F4) were significantly more highly expressed in bulk RNA-seq samples from OA than leukocyte-rich RA (*t*-test $p<1x10^{-3}$ for *HBEGF*, *CLIC5*, *HTRA4*, and *DNASE1L3*) (**Figure 4-4a**). *CD55*$^{+}$ fibroblasts (SC-F4) were the most transcriptionally distinct subset from the three *THY1*$^{+}$ clusters (SC-F1-3), including the highest expression of lubricin (*PRG4*), suggesting that these cells represent synovial lining fibroblasts and *THY1*$^{+}$ fibroblasts (SC-F1-3) represent sublining (**Figure 4-4a**). Next, we use the averaged expression level of the best marker genes for each scRNA-seq cluster (AUC>0.7) and tested for differential expression in bulk RNA-seq fibroblast samples from leukocyte-rich RA and OA synovia. The gene averages for *HLA-DRA*^{hi} sublining fibroblasts (SC-F2) and *CD34*$^{+}$ sublining fibroblasts (SC-F1) were higher in leukocyte-rich RA compared to OA (*t*-test $p=2x10^{-6}$

and $p=2x10^{-3}$, respectively), while the gene averages for *CD55*$^{+}$ lining fibroblasts (SC-F4) were higher in OA than leukocyte-rich RA (*t*-test $p=5x10^{-7}$) (**Figure 4-4b**).

Distinct synovial fibroblast subsets defined by cytokine activation and MHC II expression

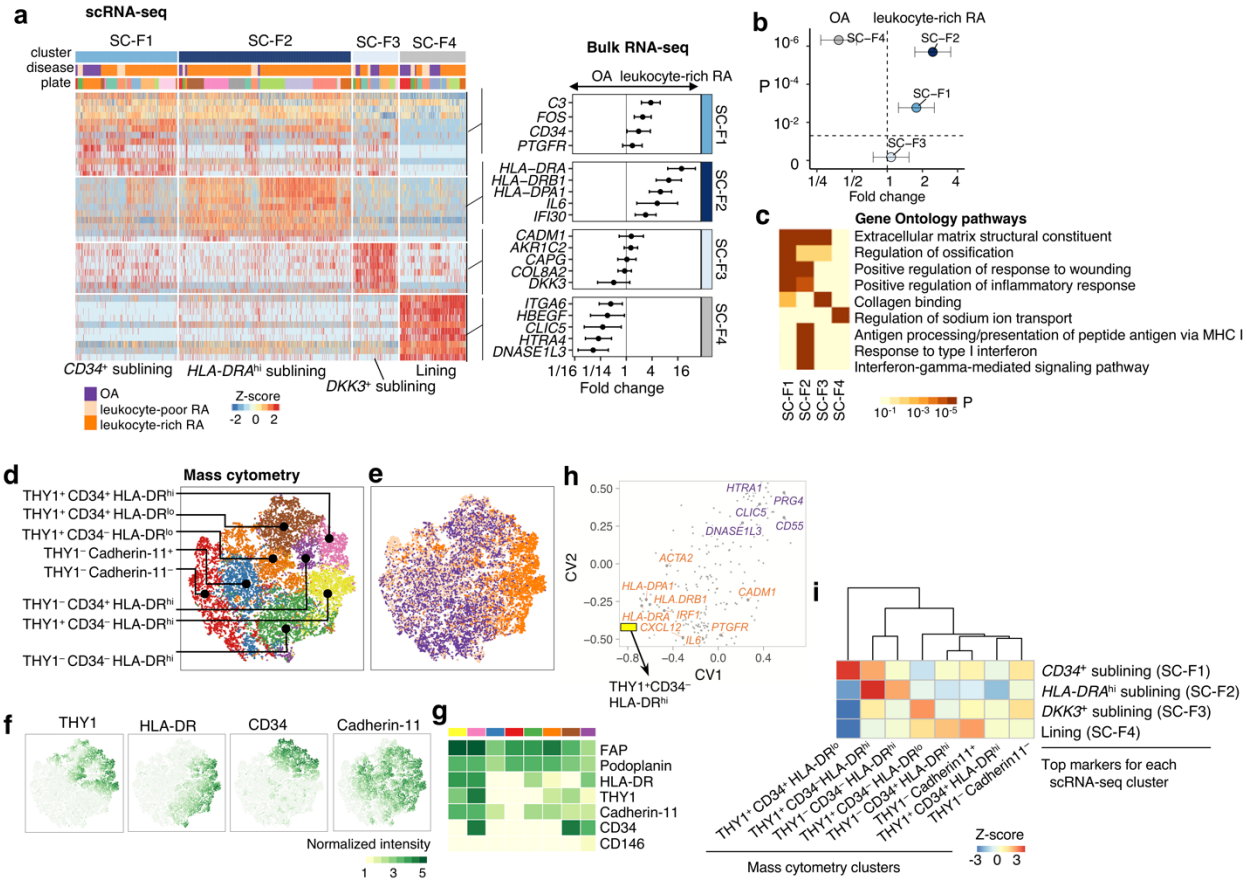


Figure 4-4. **a**. scRNA-seq analysis identified three sublining subsets, *CD34*$^{+}$ (SC-F1), *HLA*^{hi} (SC-F2), and *DKK3*$^{+}$ (SC-F3) and one lining subset (SC-F4). Differential analysis between leukocyte-rich RA (n = 16) and OA (n = 12) bulk RNA-seq fibroblast samples shows marker genes upregulated or downregulated in leukocyte-rich RA. Fold changes with 95% confidence interval (CI). **b**. By querying the leukocyte-rich RA (n = 16) and OA (n = 12) fibroblast bulk RNA-seq samples, scRNA-seq cluster *HLA-DRA*^{hi} (SC-F2) and *CD34*$^{+}$ (SC-F1) fibroblasts are significantly overabundant (two-sided Student's t-test P=2x10^{-6}, t-value=6.2, df = 23.91 and P=2x10^{-3}, t-value = 3.20, df = 25.41, respectively) in leukocyte-rich RA relative to OA. Lining fibroblasts (SC-F4) are overabundant (two-sided Student's t-test P=5x10^{-7}, t-value=-5.31, df =21.97) in OA samples. Fold changes with 95% CI. **c**. Pathway enrichment analysis for each cluster. Two-sided Kolmogorov-Smirnov test with 10^{5} permutations; Benjamini-Hochberg FDR is shown. **d-e**. Identified subpopulations from fibroblasts (n = 25,161) and disease status from 6 leukocyte-rich RA, 9 leukocyte-poor RA, and 8 OA by mass cytometry on the same gating with scRNA-seq. **f-g**. Normalized intensity of distinct protein markers shown in tSNE visualization and averaged for each cluster heatmap. **h.** CCA projections of mass cytometry clusters and bulk RNA-seq genes. First two canonical variates (CVs) separated genes upregulated in leukocyte-rich RA from genes upregulated in OA. *HLA*^{hi} genes are highly associated with THY1^{+}CD34^{-}HLA-DR^{hi} by mass cytometry. **i.** Integration of mass cytometry clusters with scRNA-seq clusters based on the top markers (AUC > 0.7) for each scRNA-seq cluster using top 10 canonical variates in the low-dimensional CCA space. We computed the spearman correlation between each pair of scRNA-seq cluster and mass cytometry cluster in the CCA space and performed permutation test 10^{4} times. Z-score is calculated based on permutation p-value. We observed HLAhigh sublining fibroblasts by scRNA-seq are strongly correlated with THY1^{+}CD34^{-}HLA-DR^{hi} fibroblasts by mass cytometry.

Consistent with the role of synovial fibroblasts in matrix remodeling, the sublining fibroblast subsets (SC-F1-3) expressed genes encoding extracellular matrix constituents (**Figure 4-4c**). *HLA-DRA*^{hi} sublining fibroblasts (SC-F2) expressed genes related to MHC class II presentation and the interferon gamma-mediated signaling pathway (*IFI30*) (**Figure 4-4a,c**), suggesting upregulation of MHC class II in response to interferon-gamma signaling in these cells. We identified a novel sublining fibroblast subtype (SC-F3) that is characterized by high expression of *DKK3*, *CADM1* and *COL8A2* (**Figure 4-4a**).

To independently confirm the presence of four fibroblast subpopulations discovered by scRNA-seq, we analyzed CD45^{-}PDPN^{+} cells in mass cytometry data, and found eight putative cell clusters with differential protein levels of THY1, HLA-DR, CD34, and Cadherin-11 without obvious batch effects (**Figure 4-4d-g**, **Figure S3b**). CCA revealed that greater abundance of THY1^{+}CD34^{-}HLA-DR^{hi} fibroblasts measured by mass cytometry is associated with higher expression of *IL6*, *CXCL12*, and *HLA-DRA* in bulk RNA-seq of the same samples, suggesting these cells are in an active cytokine-producing state (**Figure 4-4h**). CCA allowed us to place mass cytometry clusters in the same space as bulk RNA-seq genes, so we could query the positions of scRNA-seq genes within this space to find the correspondence between scRNA-seq clusters and mass cytometry clusters (**Figure 4-4i, Methods**). We found *HLA-DRA*^{hi} sublining fibroblasts (SC-F2) correspond to THY1^{+}CD34^{-}HLA-DR^{hi} fibroblasts (z-score=2.8), and *CD34*$^{+}$ sublining fibroblasts (SC-F1) correspond to THY1^{+}CD34^{+}HLA-DR^{lo} fibroblasts (z-score=2.7) (**Table 4-1**). Consistent with differential expression analysis of bulk RNA-seq, we found that THY1^{+}CD34^{-}HLA-DR^{hi} cells in the mass cytometry data were overabundant in leukocyte-rich RA relative to leukocyte-poor RA and OA controls (36% versus 2% of fibroblasts, MASC OR = 33.8 (95% CI: 11.7-113.1), one-sided MASC p=1.9x10^{-5}) (**Table 4-1**).

Connection between cell populations determined by mass cytometry and scRNA-seq clusters and disease associations

Table 4-1. Bold mass cytometry clusters are significantly enriched in leukocyte-rich RA (one-sided Benjamini-Hochberg FDR q value < 0.05). Two significant digits are given to the one-sided F-tests conducted on nested models with MASC. 95% confidence interval (CI) for the odds ratio (OR) is given for each mass cytometry cluster. Where possible, we have identified the most similar scRNA-seq clusters for each cluster found by mass cytometry. The mass cytometry analysis is performed on downsampled datasets of 25,161 fibroblasts from 23 patients, 15,298 monocytes from 26 patients, 19,985 T cells from 26 patients, and 8,179 B cells from 23 patients.

scRNA-seq cluster	mass cytometry cluster	leukocyte-poor RA and OA	leukocyte-rich RA	One-sided MASC p value	leukocyte-rich OR (CI)
	THY1$^-$ Cadherin-11$^-$	21%	4%	1.00	0.04 (0-0.2)
Lining (SC-F4)	THY1$^-$ Cadherin-11$^+$	18%	2%	1.00	0.1 (0-0.3)
	THY1$^-$ CD34$^+$ HLA-DR^{hi}	7%	3%	0.87	0.5 (0.3-1.2)
	THY1$^-$ CD34$^-$ HLA-DR^{hi}	17%	15%	0.48	1.2 (0.3-4.4)
HLA^{hi} sublining (SC-F2)	**THY1$^+$ CD34$^-$ HLA-DR^{hi}**	**2%**	**36%**	**1.9x10^{-5}**	**33.8 (11.7-113.1)**
DKK3$^+$ sublining (SC-F3)	THY1$^+$ CD34$^-$ HLA-DR^{low}	16%	15%	0.66	0.8 (0.3-1.8)
CD34$^+$ sublining (SC-F1)	THY1$^+$ CD34$^+$ HLA-DR^{low}	18%	4%	1.00	0.2 (0.1-0.4)
	THY1$^+$ CD34$^+$ HLA-DR^{hi}	**2%**	**21%**	**1.6x10^{-4}**	**25.5 (7.5-101.8)**
NUPR1$^+$ (SC-M2)	CD11c$^-$	30%	4%	1.00	0.1 (0-0.4)
IL1B$^+$ (SC-M1), IFN-activated (SC-M4)	CD11c$^+$ CCR2$^+$	34%	40%	0.23	1.6 (0.7-3.6)
	CD11c$^+$ CD38$^-$	13%	2%	1.00	0.1 (0-0.3)
	CD11c$^+$ CD38$^-$ CD64$^+$	13%	3%	0.93	0.3 (0.1-1)
IL1B$^+$ (SC-M1), IFN-activated (SC-M4), C1QA$^+$ (SC-M3)	**CD11c$^+$ CD38$^+$**	**15%**	**51%**	**6.7x10^{-5}**	**7.8 (3.6-17.2)**
	CD4$^-$ CD8$^-$	15%	9%	0.95	0.6 (0.3-1)
CCR7$^+$(SC-T1)	CD4$^+$ CCR2$^+$	26%	13%	1.00	0.4 (0.2-0.7)
	CD4$^+$ HLA-DR$^+$	6%	2%	0.83	0.7 (0.2-4.1)
	CD4$^+$ PD-1$^+$ ICOS$^-$	13%	12%	0.81	0.9 (0.5-1.6)
Tph and Tfh (SC-T3)	**CD4$^+$ PD-1$^+$ ICOS$^+$**	**11%**	**25%**	**2.7x10^{-4}**	**3.0 (1.7-5.2)**
	CD8$^+$ PD-1$^-$ HLA-DR$^-$	14%	9%	0.76	0.7 (0.3-1.5)
GZMK$^+$*GZMB*$^+$(SC-T6), *GZMK*$^+$ (SC-T4), CTLs (SC-T5)	CD8$^+$ PD-1$^-$ HLA-DR$^+$	2%	1%	0.64	0.9 (0.4-2.2)
	CD8$^+$ PD-1$^+$ HLA-DR$^-$	13%	14%	0.40	1.1 (0.6-1.9)
Tph and Tfh (SC-T3)	**CD8$^+$ PD-1$^+$ HLA-DR$^+$**	**1%**	**15%**	**9.2x10^{-5}**	**11.8 (4.9-34.2)**
plasmablasts (SC-B4)	**CD38^{++} CD20$^-$ IgM$^-$ IgD$^-$**	**6%**	**12%**	**0.01**	**3.3 (1.2-10.5)**
	CD38^{++} CD20$^-$ IgM$^+$ HLA-DR$^+$	**1%**	**3%**	**0.01**	**6.9 (1.3-83.1)**
Memory B cells (SC-B2)	IgM$^-$ IgD$^-$ HLA-DR$^-$	27%	2%	1.00	0.1 (0-0.3)
	CD38$^+$ HLA-DR^{++} CD20$^-$ CD11c$^+$	19%	6%	0.56	0.9 (0.1-6.7)
ABCs (SC-B3)	**IgM$^-$ IgD$^-$ HLA-DR^{++} CD20$^+$ CD11c$^+$**	**4%**	**12%**	**2.7x10^{-3}**	**5.7 (1.8-22.3)**
	IgM$^-$ IgD$^-$ HLA-DR$^+$	32%	20%	0.98	0.4 (0.2-1)
	IgA$^+$ IgM$^-$ IgD$^-$	5%	4%	0.68	0.9 (0.5-1.6)
Naïve B cells (SC-B1)	IgM$^+$ IgD$^-$	22%	11%	0.97	0.5 (0.2-1)
	IgM$^+$ IgD$^+$ CD11c$^-$	12%	26%	0.02	4.0 (1.3-12.0)
	IgM$^+$ IgD$^+$ CD11c$^+$	4%	7%	0.14	2.2 (0.74 - 7.7)

To validate that the protein surface markers from mass cytometry were capturing the same transcriptional populations from scRNA-seq, we isolated fibroblasts from 10 synovial tissue samples based on surface protein levels of THY1 and HLA-DR and applied bulk RNA-seq (**Figure S6a**). We trained a linear discriminant analysis (LDA) classifier on fibroblast scRNA-seq data and used it to determine the most similar scRNA-seq cluster for each bulk RNA-seq sample. The sorted THY1^{+}HLA-DR^{+} fibroblast population was similar to *THY1*$^{+}$*HLA-DRA*^{hi} (SC-F2) and the THY1^{-} HLA-DR^{-} population was similar to *THY1*$^{-}$ (SC-F4) (**Figure S7a-d**). Genes upregulated in the sorted THY1^{+}HLA-DR^{+} fibroblasts included the interleukin *IL6* and the chemokine *CXCL12*, consistent with the scRNA-seq data.

Activation states define heterogeneity among synovial monocytes

We identified four transcriptionally distinct monocyte subsets in the scRNA-seq data: *IL1B*$^{+}$ pro-inflammatory monocytes (SC-M1), *NUPR1*$^{+}$ monocytes (SC-M2), *C1QA*$^{+}$ monocytes (SC-M3) and IFN-activated *SPP1*$^{+}$ monocytes (SC-M4) (**Figure 4-5a**). In bulk RNA-seq monocyte samples from leukocyte-rich RA and OA donors, we found that genes associated with *IL1B*$^{+}$ monocytes (SC-M1), including *NR4A2*, *HBEGF*, *PLAUR* and the IFN-activated gene *IFITM3* were significantly upregulated in leukocyte-rich RA samples (*t*-test $p<1x10^{-4}$). In contrast, marker genes associated with *NUPR1*$^{+}$ monocytes (SC-M2) were downregulated in leukocyte-rich RA relative to OA (**Figure 4-5a**). Next, we took the average of the top marker genes (AUC>0.7) for each monocyte scRNA-seq subset and tested for differential expression of these averages in the bulk RA versus OA RNA-seq data. This analysis suggests that leukocyte-rich RA synovia have a greater abundance of *IL1B*$^{+}$ monocytes (*t*-test $p=6x10^{-5}$) and IFN-activated monocytes (*t*-test $p=6x10^{-3}$), but lower abundance of *NUPR1*$^{+}$ monocytes (*t*-test $p=2x10^{-5}$) (**Figure 4-5b**). These data suggest that cytokine activation drives expansion of unique monocyte populations in active RA synovia.

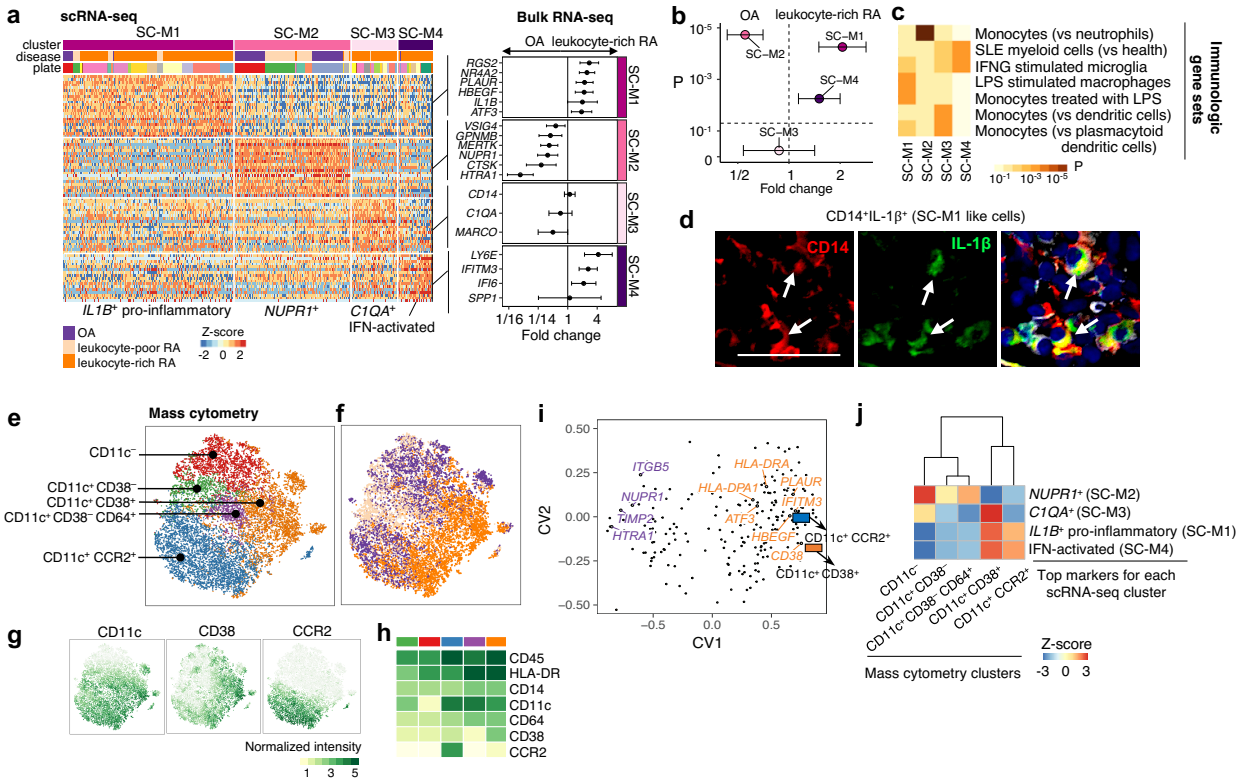


Figure 4-5. **a**. scRNA-seq analysis identified four subsets: *IL1B*$^{+}$ pro-inflammatory monocytes (SC-M1), *NUPR1*$^{+}$ monocytes (SC-M2) with a mixture of leukocyte-poor RA and OA cells, *C1QA*$^{+}$ (SC-M3), and IFN-activated monocytes (SC-M4). Differential analysis by bulk RNA-seq on leukocyte-rich RA samples (n = 17) and OA samples (n = 13) revealed upregulation/downregulation of cluster marker genes. Effect sizes with 95% CI are given. **b**. By querying the bulk RNA-seq, we found scRNA-seq cluster *IL1B*$^{+}$ pro-inflammatory monocytes (two-sided Student's t-test P=6x10^{-5}, t-value=4.56, df =26.33) and IFN-activated monocytes (two-sided Student's t-test P=6x10^{-3}, t-value=3.28, df =23.68) are upregulated in leukocyte-rich RA (n = 17) compared to OA (n = 13), while SC-M2 is depleted (two-sided Student's t-test P=2x10^{-5}, t-value=-5.62, df=26.81) in leukocyte-rich RA. Error bars indicate mean and 95% CI. **c**. Pathway enrichment analysis indicates the potential pathways for each subset. Two-sided Kolmogorov-Smirnov test with 10^{5} times permutation was performed; Benjamini-Hochberg was used to control the FDR of multiple tests. The standard names for the immunological gene sets from up to bottom are: Genes down-regulated in neutrophils versus monocytes (GSE22886); Genes down-regulated in healthy myeloid cells versus SLE myeloid cells (GSE10325); Genes down-regulated in control microglia cells versus those 24 h after stimulation with IFNG (GSE1432); Genes down-regulated in unstimulated macrophage cells versus macrophage cells stimulated with LPS (GSE14769); Genes up-regulated monocytes treated with LPS versus monocytes treated with control IgG (GSE9988); Genes up-regulated in monocytes versus myeloid dendritic cells (mDC) (GSE29618); Genes up-regulated in monocytes versus plasmacytoid dendritic cells (pDC) (GSE29618). **d**. Detection of pro-inflammatory IL-1β in inflamed synovium by multicolor immunofluorescent staining with antibodies CD14 (red), IL-1β (green), and counterstained with DAPI (blue) identified CD14^{+}IL-1β^{+} cells (white arrow). The experiment was repeated > 5 times with staining of 6 independent leukocyte-rich RA samples with similar results. Image was acquired at 200 magnification. Scale bar is 50 μm. **e-f.** Identified subpopulations from monocytes (n = 15,298) and disease status from 6 leukocyte-rich RA, 9 leukocyte-poor RA, and 11 OA by mass cytometry on the same gating with scRNA-seq. **g-h**. Normalized intensity of distinct protein markers by tSNE visualization and averaged for each cluster in heatmap. **i**. Integration of identified mass cytometry clusters with bulk RNA-seq reveals genes that are associated with CD11c+CD38+ and CD11c+CCR2+, like IFITM3, CD38, HBEGF, ATF3, and HLA+ genes. **j**. Integration of mass cytometry clusters and scRNA-seq clusters revealed that

Figure 4-5 (continued). CD11c+CD38+ by mass cytometry are significantly associated with *IL1B*$^+$ pro-inflammatory (SC-M1) monocytes.

With GSEA, we tested MSigDB immunologic gene sets and found *IL1B*$^+$ monocytes (SC-M1) have relatively high expression levels of genes defining the LPS response in monocytes and macrophages (**Figure 4-5b**). This suggests *IL1B*$^+$ monocytes (SC-M1) are similar to TLR-activated IL-1-producing pro-inflammatory monocytes. Among Gene Ontology gene sets, we found *SPP1*$^+$ monocytes (SC-M4) express genes induced by type I and II IFN (**Figure S8a**), including *IFITM3* and *IFI6* (**Figure 4-5a**). The transcriptional profiles of monocytes in SC-M2 and SC-M3 do not align with known activation states, possibly indicating that these clusters represent cell phenotypes tailored to the unique homeostatic needs of the synovium. Immunofluorescence staining confirmed the presence of CD14 and IL-1β positive cells in 6 tissue samples, consistent with an enrichment of the *IL1B*$^+$ pro-inflammatory monocytes (SC-M1) phenotype in RA synovium (**Figure 4-5d, Figure S9a,b**).

In the mass cytometry data, we identified five CD14$^+$ monocyte clusters (**Figure 4-5e-h**, **Figure S3c**)**.** Using CCA to integrate mass cytometry and bulk RNA-seq data, we found that samples with a greater abundance of CD11c$^+$CCR2$^+$ and CD11c$^+$CD38$^+$ using mass cytometry also had a higher expression of *IFITM3*, *PLAUR*, *CD38,* and *HLA* genes (**Figure 5i**). This was consistent with a correspondence between the CD11c$^+$CD38$^+$ mass cytometry cluster and the activated monocyte scRNA-seq cluster *IL1B*$^+$ (SC-M1) and *SPP1*$^+$ (SC-M4) (z-score=2.3 and 2.3, respectively) (**Figure 4-5j**, **Table 4-1**). Supporting this finding, we confirmed that CD11c$^+$CD38$^+$ monocytes are significantly expanded in leukocyte-rich RA (OR = 7.8 (95% CI: 3.6-17.2), one-sided MASC p=6.7x10^{-5}) (**Table 4-1**). Conversely, *NUPR1*$^+$ monocytes (SC-M2) correspond to CD11c$^-$ monocytes in mass cytometry and are inversely correlated with inflammatory monocyte populations (z-score=2.7) (**Figure 4-5j**, **Table 4-1**).

To confirm that putative populations from mass cytometry correspond to those identified by scRNA-seq clusters, we sorted CD14$^+$ monocytes from 4 synovial tissue samples

using CD11c and CD38 protein markers and assayed them with RNA-seq (**Figure S6c**). Importantly, we found that $CD14^+$ synovial cells had high expression of both CD11c and CD38 particularly in the RA samples. The $CD14^+CD11c^{+++}CD38^{+++}$ and $CD14^+CD11c^+CD38^-$ sorted cells were consistent with *IL1B*$^+$ pro-inflammatory (SC-M1) and *NUPR1*$^+$ (SC-M2) cells, respectively (**Figure S7e-h)**. These data, alongside the mass cytometry data, support the findings of greater abundance of *IL1B*$^+$ pro-inflammatory (SC-M1) monocytes and lower abundance of *NUPR1*$^+$ (SC-M2) monocytes in leukocyte-rich RA samples.

Heterogeneity in synovial CD4 and CD8 T cells defined by effector functions

We found three $CD4^+$ and three $CD8^+$ T cell subsets in the scRNA-seq data (**Figure 4-6a**). *CCR7*$^+$ T cells (SC-T1) expressed genes in the MSigDB immunologic gene set for central memory T cells (**Figure 4-6a, c**). The two other $CD4^+$ populations, *FOXP3*$^+$ T_{reg} cells and *PDCD1*$^+$ Tph and Tfh cells, were marked by high expression of *FOXP3* (SC-T2) and C*XCL13* (SC-T3) by examining differentially expressed genes between these two clusters[18] (**Figure S8c**). *CXCL13*, a chemokine expressed by Tph cells, was upregulated in bulk-sorted T cells ($CD45^+CD14^-CD3^+$) from leukocyte-rich RA compared to OA (*t*-test $p=1.2x10^{-4}$) (**Figure 4-6a**). We found that the average of marker genes for Tph and Tfh cells (SC-T3) (AUC>0.7) was higher in leukocyte-rich RA than OA samples (*t*-test p=0.01) (**Figure 4-6b**), suggesting greater abundance of Tph and activated T cells in RA than OA. We identified three CD8 T cell subsets characterized by distinct expression patterns of effector molecules *GZMK, GZMB, GZMA* and *GNLY* (**Figure 4-6a**). We defined these populations as *GZMK*$^+$ (SC-T4), *GNLY*$^+$*GZMB*$^+$ cytotoxic T lymphocytes (CTLs) (SC-T5), and *GZMK*$^+$*GZMB*$^+$ T cells (SC-T6). *GZMK*$^+$*GZMB*$^+$ T cells (SC-T6) also expressed *HLA-DPA1* and *HLA-DRB1*, and other genes suggestive of an effector phenotype (**Figure 4-6a, c**).

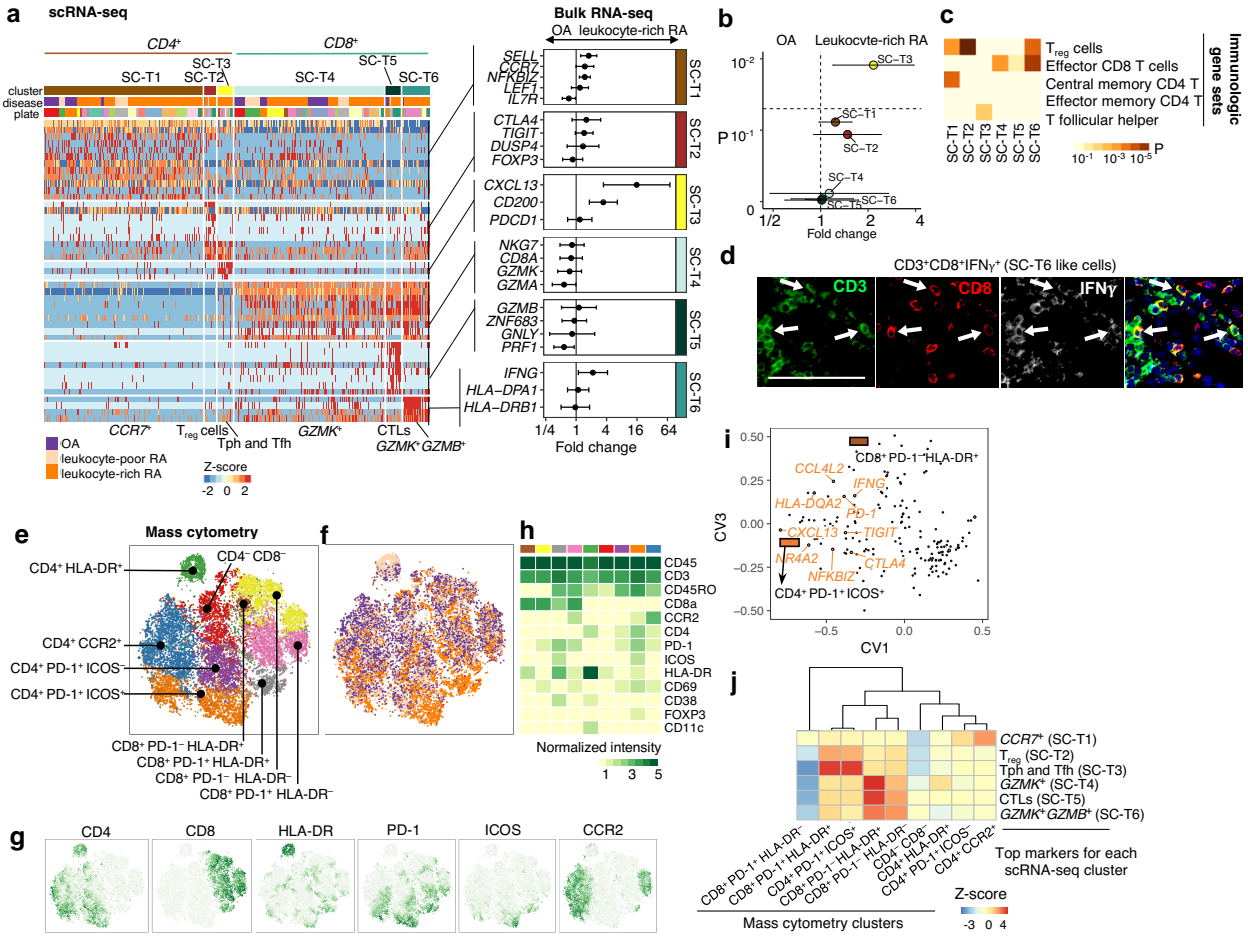


Figure 4-6. **a**. scRNA-seq analysis identified three CD4$^+$ subsets: CCR7$^+$ (SC-T1), T$_{reg}$ cells (SC-T2), and Tph and Tfh (SC-T3); and three CD8$^+$ subsets: GZMK$^+$ (SC-T4), CTLs (SC-T5), and GZMK$^+$GZMB$^+$ (SC-T6). Differential expression analysis on leukocyte-rich RA (n = 18) comparing with OA (n = 13) on sorted T cell bulk RNA-seq samples revealed that CXCL13 is most significantly enriched in leukocyte-rich RA compared to OA. Effect sizes with 95% CI are given. **b**. Disease association of scRNA-seq clusters by aggregating top markers (AUC>0.7) by comparing leukocyte-rich RA (n = 18) with OA (n = 13) using bulk RNA-seq. Tph and Tfh cells (SC-T4) are upregulated (two-sided Student's t-test p=0.01, t-value=2.73, df =29.00) in leukocyte-rich RA. Error bars indicate mean and 95% CI. **c**. Pathway analysis based on immunologic gene set enrichment indicates the potential enriched T cell states pathways. Two-sided Kolmogorov-Smirnov test with 10^5 times permutation was performed; Benjamini-Hochberg was used to control the FDR of multiple tests. The brief description of the standard names from up to bottom are: Genes up-regulated in CD4 high cells from thymus: Treg versus T conv (GSE42021); Genes up-regulated in comparison of effector CD8 T cells versus memory CD8 T cells (GOLDRATH); Genes down-regulated in comparison of effector memory T cells versus central memory T cells from peripheral blood mononuclear cells (PBMC) (GSE11057); Genes up-regulated in comparison of effective memory CD4 T cells versus Th1 cells (GSE3982); Genes up-regulated in comparison of T follicular helper (Tfh) cells versus Th17 cells (GSE11924). **d**. Detection of CD3$^+$CD8$^+$IFNγ$^+$ (white arrow) in inflamed RA synovium by multicolor immunofluorescent staining with antibodies CD3 (green), CD8 (red), IFNγ (white), and counterstained with DAPI (blue). The experiment was repeated > 5 times with staining of 6 independent leukocyte-rich RA samples with similar results. Image was acquired at 200 magnification. Scale bar is 50 μm. **e-f**. Identified subpopulations from T cells (n = 19,985) and disease status from 6 leukocyte-rich RA, 9 leukocyte-poor RA, and 11 OA by mass cytometry. **g-h**. Distinct patterns of protein markers by tSNE

Figure 4-6 (continued). and heatmap that define these clusters. **i**. Integration of identified mass cytometry clusters with bulk RNA-seq using CCA reveals bulk genes that are associated with $CD4^+PD\text{-}1^+ICOS^+$ and $CD8^+PD\text{-}1^-HLA\text{-}DR^+$ by mass cytometry. **j**. Integration of mass cytometry clusters with scRNA-seq clusters on the top markers (AUC>0.7) for each scRNA-seq cluster in the top 10 canonical variates. Z-score based on permutation test reveals that $CD4^+PD\text{-}1^+ICOS^+$ and $CD8^+PD\text{-}1^+HLA\text{-}DR^+$ by mass cytometry are highly associated with Tph and Tfh (SC-T3) by scRNA-seq; $CD8^+PD\text{-}1^-HLA\text{-}DR^+$ T cells by mass cytometry are highly associated with $CD8^+$ T cells (SC-T4, SC-T5, and SC-T6).

To confirm these findings, we applied intracellular staining to tissues from RA samples and RNA-seq to sorted CD8 T cells. Intracellular staining of GZMK and GZMB proteins in disaggregated tissue samples from patients with RA revealed that the majority of CD8 T cells in synovial tissue express GZMK (**Figure S10a**). Furthermore, we found that most $HLA\text{-}DR^+$ CD8 T cells express both GZMB and GZMK by intracellular protein staining (**Figure S10b**). In a comparison of 7 synovial tissue samples, CD8 T cells had higher proportion of $IFN\gamma^+$ cells than CD4 T cells from the same sample (**Figure S10c,d**). We also applied immunofluorescence to 6 synovial tissue samples and found that $IFN\gamma^+CD3^+CD8^+$ T cells were more frequent in RA than OA (**Figure 4-6d, Figure S9c,d**). Overall, these results closely mirror the findings from the scRNA-seq clusters.

Using mass cytometry, we identified nine putative T cell clusters among the synovial T cells ($CD45^+CD14^-CD3^+$) (**Figure 4-6e-h**, **Figure S3d**). By integrating bulk RNA-seq with mass cytometry cluster abundances, we found that higher gene expression of *CXCL13* and inhibitory receptors *TIGIT* and *CTLA4* was associated with greater abundance of the $CD4^+PD\text{-}1^+ICOS^+$ mass cytometry cluster. Greater abundance of $CD8^+$ $PD\text{-}1^-HLA\text{-}DR^+$ cells was associated with greater expression of *IFNG* (**Figure 4-6i**). We found correspondence between Tph and Tfh cells (SC-T3) and $CD4^+PD\text{-}1^+ICOS^+$ T cells (z-score = 3.4). $CD8^+$ subsets including $GZMK^+GZMB^+$ (SC-T6), CTLs (SC-T5), and $GZMK^+$ (SC-T4) tracked with $CD8^+PD\text{-}1^-HLA\text{-}DR^+$ T cells by mass cytometry (**Figure 4-6j**, **Table 4-1**). In addition, $CD4^+PD\text{-}1^+ICOS^+$ cells were significantly overabundant in leukocyte-rich RA (MASC OR = 3 (95% CI: 1.7-5.2), one-sided MASC $p=2.7x10^{-4}$) (**Table 4-1**).

Autoimmune-associated B cells expanded in RA synovium by single-cell RNA-seq

We identified four synovial B cell clusters with scRNA-seq: naive B cells (SC-B1), memory B cells (SC-B2), *ITGAX*$^{+}$ ABC cells (SC-B3), and plasmablasts (SC-B4) (**Figure 4-7a**). GSEA with Gene Ontology pathways suggested that SC-B1, SC-B2, and SC-B3 clusters represent activated B cells (**Figure S8b**). GSEA with MSigDB immunological gene sets revealed that SC-B1 cells express naive B cell genes, while SC-B2 and SC-B3 cells express IgM and IgG memory B cell genes (**Figure 4-7b**). SC-B3 cells express high levels of *ITGAX* and *TBX21 (T-bet)*, which are markers of autoimmunity-associated B cells **(Figure 4-3f** and **Figure 4-7a**)[218,219], as well as markers of recently activated B cells including *ACTB*[220]. High expression of *AICDA* is consistent with the recently reported transcriptomic analysis of CD11c^{+} B cells from SLE peripheral blood[221]. Interferon stimulated genes (*GBP1* and *ISG15*) are also expressed in ABCs (SC-B3) and upregulated in leukocyte-rich RA (**Figure 4-7a**). While ABCs (SC-B3) constitute a relatively small proportion of all B cells, they are almost exclusively derived from two patients with leukocyte-rich RA (**Figure 4-3b**). To confirm the presence of ABCs in human tissues, we applied immunofluorescence staining to 6 synovial tissue samples. RA synovium had increased numbers of CD20^{+}T-bet^{+} CD11c^{+} B cells compared to OA synovium. Specifically, we observed ABC cells in tissue sections from the same inflamed tissue samples that had a high proportion of ABCs by scRNA-seq analysis (**Figure 4-7c, Figure S9e, f)**.

Synovial B cells display heterogeneous subpopulations in RA synovium

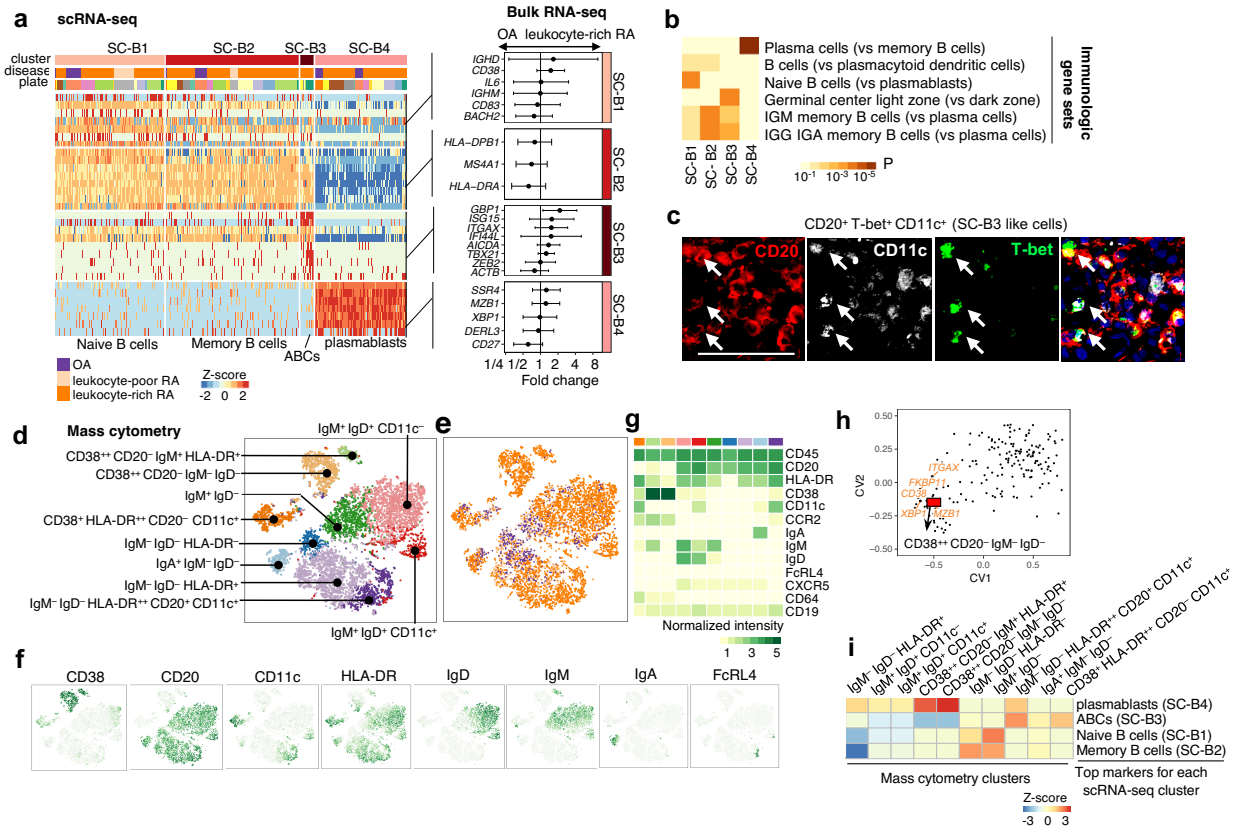


Figure 4-7. **a**. scRNA-seq analysis identified naive B cells (SC-B1), memory B cells (SC-B2), autoimmune-associated B cells (ABCs) (SC-B3), and plasmablasts (SC-B4). Differential expression analysis is given by comparing leukocyte-rich RA (n = 16) with OA (n = 7) using bulk RNA-seq B cell samples. Effect size with 95% CI are given. **b**. Pathway enrichment analysis using immunologic gene sets indicates the distinct enriched pathways for each scRNA-seq cluster. Two-sided Kolmogorov-Smirnov test with 10^5 times permutation was performed; Benjamini-Hochberg was used to control the FDR of multiple tests. The standard names for the immunological gene sets from up to bottom are: Genes up-regulated in plasma cells versus memory B cells (GSE12366); Genes up-regulated in comparison of B cells versus plasmacytoid dendritic cells (pDC) (GSE29618); Genes up-regulated in B lymphocytes: naive versus plasmablasts (GSE42724); Genes up-regulated in B lymphocytes: human germinal center light zone versus dark zone (GSE38697); Genes up-regulated in comparison of memory IgM B cells versus plasma cells from bone marrow and blood (GSE22886); Genes up-regulated in comparison of memory IGG and IGA B cells versus plasma cells from bone marrow and blood (GSE22886). **c**. Detection of $CD20^+$T-bet^+CD11c^+ (white arrow) in inflamed synovium by multicolor immunofluorescence. Immunofluorescent staining with antibodies CD20 (red), CD11c (white), T-bet (green), and counterstained with DAPI (blue). The experiment was repeated > 5 times with staining of 6 independent leukocyte-rich RA samples with similar results. Image was acquired at 200 magnification. Scale bar is 50 μm. **d-e**. Identified subpopulations of B cells (n = 8,179) and disease status from 6 leukocyte-rich RA, 9 leukocyte-poor RA, and 8 OA by mass cytometry. **f-g**. Distinct expression patterns of protein markers by tSNE and averaged for each cluster in heatmap. **h**. Integrating mass cytometry clusters with bulk RNA-seq data using CCA shows that $CD38^+CD20^-Ig^-$ (plasmablasts) population is highly associated with gene expression of plasma cells makers, like XBP1. **i**. Integration of mass cytometry clusters with scRNA-seq clusters suggested that $CD38^{++}CD20^-IgM^+HLA\text{-}DR^+$ and $CD38^{++}CD20^-IgM^-IgD^-$ are significantly associated with plasmablast (SC-B4); $IgM^-IgD^-HLA\text{-}DR^{++}CD20^+CD11c^+$ B cells are associated with ABCs (SC-B3).

We identified 10 putative B cell clusters in the mass cytometry data ($CD45^{+}CD3^{-}CD14^{-}CD19^{+}$) (**Figure 4-7d-g**, **Figure S3e**). CCA analysis showed that samples with higher gene expression of *CD38*, *MZB1*, and plasma cell differentiation factor *XBP1* had greater abundance of $CD38^{++}CD20^{-}IgM^{-}IgD^{-}$ plasmablasts (**Figure 4-7h**). Plasmablasts (SC-B4) corresponded with $CD38^{++}CD20^{-}IgM^{-}IgD^{-}$ B cells (z-score=2.7) (**Figure 4-7i**, **Table 4-1**). ABCs (SC-B3) corresponded with the IgM^{-} IgD^{-} HLA-DR^{++} $CD20^{+}$ $CD11c^{+}$ mass cytometry cluster (z-score=1.6), which is significantly overabundant in leukocyte-rich RA (OR = 5.7 (95% CI: 1.8-22.3), one-sided MASC $p=2.7x10^{-3}$) (**Figure 4-7i**, **Table 4-1**). Mass cytometry analysis further identified three putative subsets within $CD11c^{+}$ cells: $IgM^{-}IgD^{-}$HLA-$DR^{++}CD20^{+}CD11c^{+}$, $CD38^{+}$HLA-$DR^{++}CD20^{-}CD11c^{+}$, and $IgM^{+}IgD^{+}CD11c^{+}$, which is suggestive of additional heterogeneity within ABCs.

To demonstrate that $CD19^{+}CD11c^{+}$ cells by surface protein markers correspond to SC-B3 (ABCs), we flow-sorted $CD19^{+}CD11c^{+}$ cells from an independent cohort of 6 RA synovial samples and applied RNA-seq **(Figure S6b**). We show that these RNA-seq profiles are most consistent with ABC cells (**Figure S7i-k)**. In these sorted samples, we found more putative marker genes (e.g. *ZEB2* and *CIITA*) and interferon-induced genes (*IFITM3* and *IFI27*) for the ABC population (**Figure S7l**).

Inflammatory pathways and effector modules revealed by global single cell profiling

We used bulk and single cell transcriptomes of sorted synovial cells to examine pathologic molecular signal pathways. First, principal component analysis (PCA) on post-QC OA and RA bulk RNA-seq samples (**Figure S11a,b**) showed that cell type accounted for most of the data variance. Each cell type expressed specific marker genes, *PDGFRA* for fibroblasts, *C1QA* for monocytes, *CD3D* for T cells, and *CD19* for B cells (**Figure S11c**). Within each cell type, PCA

showed that leukocyte-rich RA samples separated from OA and leukocyte-poor RA samples (**Figure S11d-g**). Differential gene expression analysis between leukocyte-rich RA and OA (FC>2 and FDR<0.01) revealed genes upregulated in leukocyte-rich RA tissues: 173 in fibroblasts, 159 in monocytes, 10 in T cells, and 5 in B cells. To define the pathways relevant to leukocyte-rich RA, we used GSEA weighted by gene effect sizes on Gene Ontology pathways and identified type I interferon response and inflammatory response (monocytes and fibroblasts) (**Figure S11h-i**), Fc receptor signaling (monocytes), NF-kappa B signaling (fibroblasts), and interferon gamma (T cells) (**Figure 4-8a**). Leukocyte-rich RA samples had significantly higher expression of some genes in fibroblasts and monocytes: inflammatory response genes (*PTGS2, PTGER3,* and *ICAM1*), interferon response genes (*IFIT2, RSAD2, STAT1,* and *XAF1*), and chemokine or cytokine genes (*CCL2* and *CXCL9*) (**Figure 4-8b**), consistent with a coordinated chemotactic response to interferon activation. T cells had upregulation of interferon regulatory factors (IRFs), including *IRF7* and *IRF9*, and monocytes had upregulation of *IRF7*, *IRF8* and *IRF9*. Taken together, pathway analysis suggests crosstalk between immune and stromal cells in leukocyte-rich RA synovia. Inflammatory response genes upregulated in leukocyte-rich RA had comparable expression levels between leukocyte-poor RA and OA synovial cells (**Figure 4-8b**)

Transcriptomic profiling of synovial cells reveals upregulation of inflammatory pathways in RA synovium

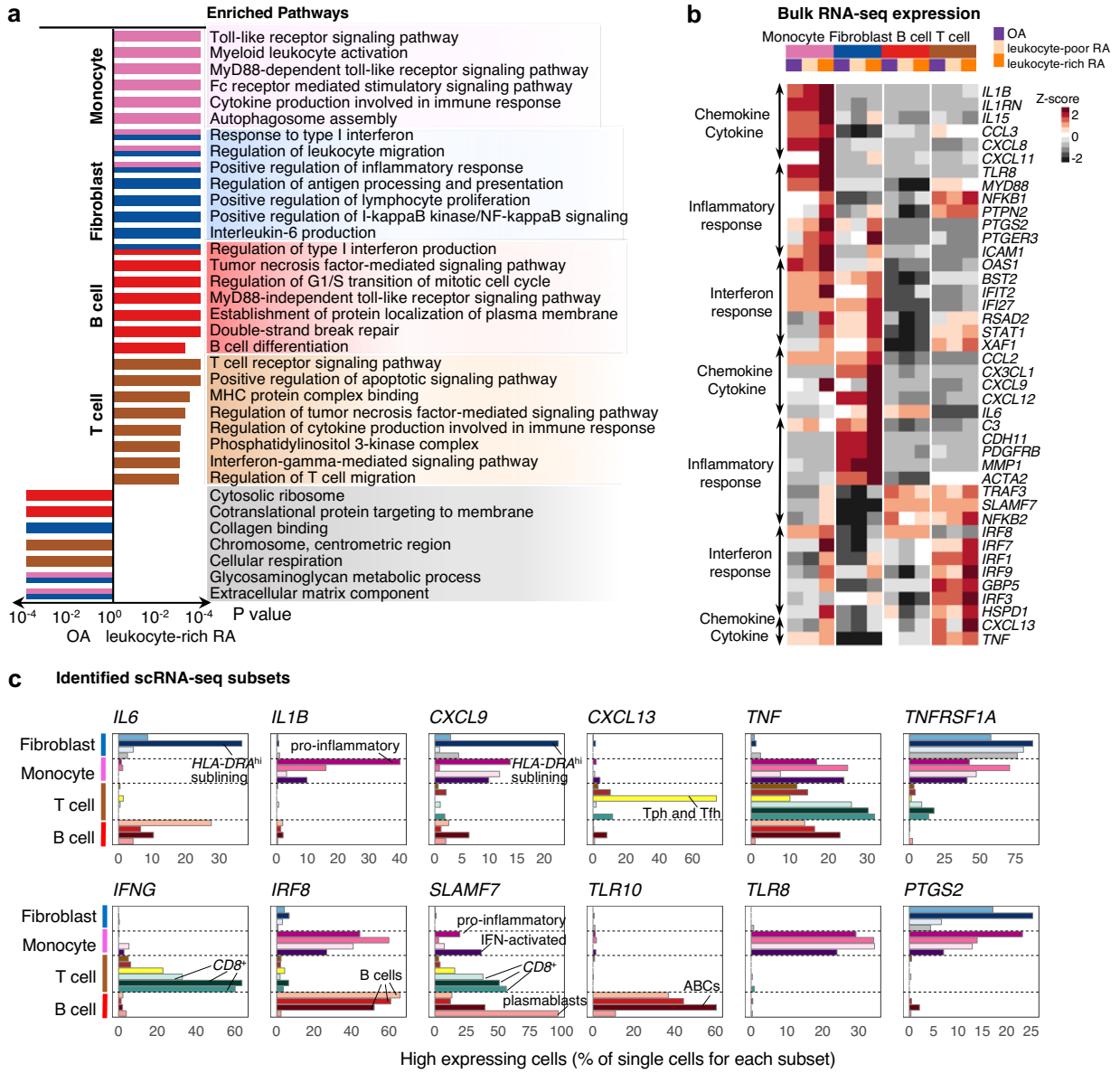


Figure 4-8. **a**. Pathway enrichment using bulk RNA-seq identified shared and unique inflammatory response pathways for each cell type. Two-sided Kolmogorov-Smirnov test with 10^5 permutations was performed on 18 leukocyte-rich RA, 17 leukocyte-poor RA, and 14 OA. **b**. Bulk RNA-seq profiling of genes obtained from the significantly enriched pathways from (**a**) shows the averaged gene expression from each group (18 leukocyte-rich RA, 17 leukocyte-poor RA, and 14 OA) normalized across all cell type samples. **c**. scRNA-seq profiling resolved that inflammatory cytokines/chemokines, interferon responsive, and inflammatory responsive genes were driven by a global upregulation within a synovial cell type or discrete cell states.

Next, we asked whether inflammatory cytokines upregulated in leukocyte-rich RA are driven by global upregulation within a single synovial cell type, or specific upregulation within a discrete cell subset defined by scRNA-seq. Whereas *TNF* was produced at a high level by multiple monocyte, B cell and T cell populations; *IL6* expression was restricted to *HLA-DRA*^{hi} sublining fibroblasts (SC-F2) and a subset of B cells (SC-B1) (**Figure 4-8c**); CD8 T cells, rather than CD4 T cells, were the dominant source of *IFNG* transcription in leukocyte-rich synovia.

We also observed cell subset-specific responses to inflammatory pathways. Toll-like receptor signaling pathway was enriched in B cells and monocytes in leukocyte-rich RA tissues (**Figure 4-8a**). At the single cell level, we observed that *TLR10* was only expressed by activated B cells, indicating that *TLR10* has a functional role within the B cell lineage. In contrast, *TLR8* was elevated in all RA monocyte subsets. The hematopoietic cell-specific transcription factor *IRF8* was expressed in a significant fraction of monocytes and B cells that cooperatively regulate differentiation of monocytes and activated B cells in RA synovium. *SLAMF7* is highly expressed by pro-inflammatory monocytes (SC-M1), IFN-activated monocytes (SC-M4), CD8 T cells, and plasmablasts (SC-B4).

Furthermore, mass cytometry analysis across all identified cell clusters revealed that leukocyte-rich RA patients show high cell abundances of HLA-DR^{hi} fibroblast populations, Tph cells, CD11c^{+}CD14^{+} monocytes, and CD11c^{+} B cell populations (**Figure S3f**).

Discussion

Using multi-model, high-dimensional synovial tissue data we defined stromal and immune cell populations overabundant in RA and described their transcriptional contributions to essential inflammatory pathways. Recognizing the considerable variation in disease duration and activity, treatment types, and joint histology scores[222], we elected to use a molecular parameter, based on percent leukocytes of the total cellularity, to classify our samples at the

local tissue level. We note that differences in leukocyte enrichment of joint replacement samples and biopsy samples were best explained by leukocyte infiltration and not by the histological scores (**Figure S1**, **Figure S11d-g**).

This study and a previous study[223] have highlighted sublining fibroblasts as a potential therapeutic target in RA. Sublining fibroblasts are a major source of pro-inflammatory cytokines such as *IL6* (**Figure 4-4**), and a specific subset of sublining fibroblasts expressing MHC II (SC-F2, *THY1*$^{+}$*CD34*$^{-}$*HLA-DR*^{hi}) was >15 fold expanded in RA tissues. Further studies are needed to define molecular mechanisms that regulate sublining fibroblast expansion in RA. T cells, B cells, and monocyte proportions track with expression of individual fibroblast genes (**Figure S11j**). We found *DNASE1L3,* a gene whose loss of function is associated with RA[224] and systemic lupus erythematosus[225] to be highly expressed in *CD55*$^{+}$ lining fibroblasts (SC-F4) (**Figure 4-4a**). We identified a novel fibroblast subset (SC-F3) with high expression of *DKK3*+ (**Figure 4-4**), encoding Dickkopf3, a protein upregulated in OA that prevents cartilage degradation in vitro[226].

Transcriptional heterogeneity in the synovial monocytes indicated that distinct RA-enriched subsets are driven by inflammatory cytokines and interferons (**Figure 4-5**). This suggests monocytes may be differentially polarized by unique cytokine combinations in local microenvironments. These newly identified inflammatory phenotypes align with RA therapeutic targets, including anti-TNF therapies and interferon pathway JAK kinase inhibitors[227]. The *NUPR1*$^{+}$ (SC-M2) monocytes were inversely correlated with tissue inflammation, and expressed high levels of monocyte tissue remodeling factors such as *MERTK* (**Figure 4-5**)[228]. Alternatively, *NUPR1*$^{+}$ markers such as osteoactivin (*GPNMB*) and cathepsin K (*CTSK*) may indicate a subset of osteoclast progenitors that control bone remodeling (**Figure 4-5**)[227,229]. Furthermore, spatial studies—particularly focused on lining versus sublining, perivascular and lymphocyte aggregate-associated monocytes—will help understand the functional roles of these subsets.

Single cell classification of T cell subsets in RA synovium demonstrated $CD4^+$ T cell heterogeneity that is consistent with distinction between the homing capacity and effector functions of these subsets. Consistent with previous studies, we observed expansion of *PDCD1*$^+$*CD4*$^+$ Tph cells (SC-T3) within leukocyte-rich RA. We also found CD8 T cell subsets (SC-T4-6) characterized by a distinct granzyme expression pattern (**Figure 4-6a**). A larger study may be better powered to differentiate the relative expansion of individual subpopulations.

This study is the first to report the presence of autoimmune-associated B cells (SC-B3) by transcriptomic sequencing in human leukocyte-rich synovial RA and, in fact, in any human autoimmune target tissue. This B cell population was first reported in aging mice and subsequently seen in autoimmune mice and SLE patient peripheral blood[221,230]. We observed a heterogeneity of $CD11c^+$ B cells detectable in both IgD^+ and switched B cell populations by mass cytometry. The gene expression of other ABCs markers suggests a balance between germinal center (*IRF8* and *AID*) and plasma cell (*SLAMF7*) differentiation within the RA synovium. We have few B cells from OA synovia (**Figure 4-2b**), which limited our ability to identify RA-associated B cell subsets through case-control comparisons (**Figure 4-7g**).

A critical unmet need in RA is identifying therapeutic targets for patients failing to respond to disease-modifying antirheumatic drugs (DMARDs)[231]. We observed upregulation of chemokines (*CXCL8*, *CXCL9*, and *CXCL13*), cytokines (*IFNG* and *IL15*[232,233]), and surface receptors (*PDGFRB* and *SLAMF7*) in distinct immune and stromal cell populations, suggesting potential novel targets. This study was enabled by advances in the statistical integration of single-cell data and our recent work optimizing robust methodologies for disaggregation of synovial tissue[214].

We developed advanced strategies to integrate multiple molecular datasets by modulating technical artifact from single cell technologies[123], while emphasizing biological signals. CCA has been successfully employed in other contexts to integrate high-dimensional

biological data[234,235]. Our CCA-based strategy analyzed scRNA-seq data using canonical variates that capture variance that are present in both single-cell and bulk RNA-seq data. The shared variances likely represent biological trends, and not technical factors that would likely be uncorrelated in these two independent datasets. We further confirmed that the identified scRNA-seq clusters are well correlated with the bulk RNA-seq data and also the mass cytometry data (**Figure S12, S13**).

The two single cell modalities used in this study, mass cytometry and scRNA-seq, complement each other. Single-cell RNA-seq captures expression of thousands of genes, but at the cost of sparse data[236]. Mass cytometry captures hundreds of thousands of individual cells, but measures a limited number (~40) of pre-selected markers[81]. However, since markers are backed with decades of experimental experience they can be effective at defining cellular heterogeneity[237]. To make the analysis consistent, we gated mass cytometry cells on the same markers upon which the scRNA-seq was gated. Combining mass cytometry with the extended dimensionality of scRNA-seq enables quantification of well-established cell populations and discovery of novel cell states, such as the CD8 T cell states noted here. As an ongoing AMP phase 2 study, we are examining larger numbers of ungated cell populations from ~100 synovial tissue patients with RA by capturing mRNA and protein expression simultaneously[186] with detailed clinical data and ultrasound score evaluation of synovitis. We anticipate that this larger study will enable us to not only discover additional subpopulations, but to better define their link to clinical sub-phenotypes.

It is essential to interrogate the tissue infiltration of diseases other than RA, including SLE, type I diabetes, psoriasis, multiple sclerosis and other organ targeting conditions. Application of multiple single-cell technologies together can help define key novel populations, thereby providing new insights about etiology and potential therapies.

Materials and Methods

Study design and patient recruitment

The study was performed in accordance with protocols approved by the institutional review board. A multicenter, cross-sectional study of individuals undergoing elective surgical procedures and a prospective observational study of synovial biopsy specimens from patients with RA ≥ age 18, with at least one inflamed joint, recruited from 10 contributing sites in the network. Synovial tissues were obtained from joint replacement procedures or ultrasound-guided biopsies, followed by cryopreservation in cryopreservation media Cryostor CS10 (Sigma-Aldrich) and transit to a central technology site.

Histological assessment of synovial tissue and quality control

Synovial tissue quality and grading of synovitis were evaluated in formalin-fixed, paraffin-embedded sections by histologic analysis (H and E staining). Specimens were identified as synovium by the presence of a lining layer or by characteristic histologic features of synovium, including the presence of loose fibrovascular or fatty tissue lacking a lining layer. Samples consisting of dense fibrous tissue, joint capsule or other tissues were determined not to be synovium. For each histological and molecular analysis, we generated pooled data from 6-8 separate fragments from different sites in the same joint. Thus, this should be representative of the whole tissue and mitigate much of the biopsy site-to-site variability. Krenn lining scores (0-3) and inflammation scores (0-3) for each tissue sample were determined independently by three pathologists[216].

Tissue disaggregation for mass cytometry and RNA-sequencing

For pipeline analysis, synovial tissue samples stored in cryovials were disaggregated into single cell suspension as describe. Briefly, synovial tissue fragments were separated mechanically and

enzymatically in digestion buffer (Liberase TL (Sigma-Aldrich) 100 ug/mL and DNase I (New England Biolabs) 100 ug/ml in RPMI) in a 37°C water bath for 30 minutes. Single cell suspensions from disaggregated synovial tissues were assessed for cell quantity and cell viability by trypan Blue. For samples with more than 200,000 viable synovial cells, 50% of all synovial cells were allocated for analysis by mass cytometry and the remaining cells were allocated for RNA-seq. For samples with less than 200,000 viable synovial cells, all synovial cells were utilized for RNA-seq analysis.

Synovial cell sorting strategy for RNA sequencing

Synovial T cells, B cells, monocytes, and fibroblasts were isolated from disaggregated synovial tissue, as described[214]. Briefly, disaggregated synovial cells were stained with antibodies against CD45 (HI30), CD90 (5E10), podoplanin (NZ1.3), CD3 (UCHT1), CD19 (HIB19), CD14 (M5E2), CD34 (4H11), CD4 (RPA-T4), CD8 (SK1), CD31 (WM59), CD27 (M-T271), CD235a (KC16), using human TruStain FcX in 1% BSA in Hepes-Buffered Saline (HBS,20 mM HEPES, 137 mM NaCl, 3mM Kcl, 1mM CaCl2) for 30 minutes. 1000 viable (PI-) T cells ($CD45^+$, $CD3^+$, $CD14^-$), monocytes ($CD45^+$, $CD3^-$, $CD14^+$), B cells ($CD45^+$, $CD3^-$, $CD14^-$, $CD19^+$), and synovial fibroblasts ($CD45^-$, $CD31^-$, $PDPN^+$) were collected by fluorescence-activated cell sorting (BD FACSAria Fusion) directly in buffer RLT (Qiagen) for bulk RNA-seq. For single cell RNA-seq, live cells of each population were re-sorted into 384-well plates single cells with a maximum of 144 cells for each cell type, per patient sample.

Flow sorting strategy for bulk RNA-seq experimental validation

For bulk RNA-seq validation experiments, RA and OA synovial tissue were disaggregated and synovial cells were stained with cell-type specific antibody panels. For each cell subset, up to 1000 cells were collected directly into buffer TCL (Qiagen). Antibody panels used to define cell

subsets are fibroblasts: CD90 (5E10), podoplanin (NZ1.3), HLA-DR (G46-6); B cell subsets: HLA-DR (G46-6), CD11c (3.9), CD19 (SJ25C1), CD27 (M-T271), IgD (IA6-2), CD3 (UCHT1), CD14 (M5E2), CD38 (HIT2); Monocyte subsets: CD14-BV421 (M5E2), CD38-APC (HB-7), and CD11c-PECy7 (B-ly6). Immediately prior to sorting, DAPI or LIVE/DEAD viability dye was added to cell suspensions and cells were passed through a 100μm filter. Synovial cell subsets were sorted based on flow cytometry gating schema shown in **Supplementary Fig. 6**. In all, we sorted $THY1^{-}$ DR^{-} populations from 4 OA samples, $THY1^{+}DR^{-}$ population from 4 OA and 6 RA samples, and $THY1^{+}$ DR^{+} population from 6 RA samples. For monocytes, we sorted $CD14^{+}CD11c^{+++}CD38^{+++}$ population from 2 RA samples and $CD14^{+}CD11c^{+}$ $CD38^{-}$ population from 2 OA samples. For B cells, we sorted $CD11c^{-}IgD^{-}CD27^{+}$ population from 6 RA samples, $CD11c^{-}IgD^{+}CD27^{-}$ population from 3 RA samples, $CD19^{+}CD11c^{+}$ population from 3 RA samples, and plasma cells from 3 RA samples.

To validate the identified single-cell populations using bulk RNA-seq, we fit an LDA (Linear Discriminant Analysis) classifier on the scRNA-seq cell clusters and then classified each flow sorted bulk RNA-seq sample. For each cell type, 1) we trained an LDA model on the scRNA-seq clusters with the top 500 marker genes for each cluster; 2) Next, we applied this LDA model to classify each sample of bulk sorted cells and estimated the maximum posterior probability for each sample. In summary, we tested if we could sort new cells from new, independent samples and see the same gene expression profiles in the new bulk samples as the original scRNA-seq samples.

Multicolor immunofluorescent staining of paraffin synovial tissue

Briefly, 5 mm thick formalin fixed paraffin sections were incubated in a 60°C oven to melt paraffin. Slides were quickly transferred to xylenes to completely dissolve the paraffin and after 5 minutes transferred to absolute ethanol. Slides were left in absolute ethanol for 5 minutes and

then transferred to 95% ethanol. At the end of the 5 minutes immersion in 95% ethanol, slides were rinsed several times with distilled water and transfer to a plastic coplin jar filled with 1X DAKO retrieval solution (S1699, Dakocytomation). Antigens were unmasked by immersing of plastic coplin jar in boiling water for 30 minutes. Slides were let cool down for 10 minutes at room temperature and washed several times with distilled water. Non-specific binding was blocked with 5% normal donkey serum (017-000-121, Jackson ImmunoResearch Laboratories,) dissolved in PBS containing 0.1% Tween 20 and 0.1% Triton X-100. Without washing, blocking solution was removed from slides and combinations of primary antibodies were added to PBS containing 0.1% Tween 20 and 0.1% Triton X-100. Primary antibodies to detect $IFNg^+$ T cells include goat anti-CD3 epsilon (clone M-20, Santa Cruz Biotechnology), mouse anti-human CD8 (clone 144B, GeneTex), and rabbit anti-human IFNg (Biorbyt, orb214082). To visualize ABC, we incubated slides with goat anti-human CD20 (LifeSpan Biosciences, LS-B11144), rabbit anti-Tbet (H-210, Santa Cruz Biotechnology) and biotinylated mouse anti-human CD11c (clone 118/A5, Thermo Fisher Scientific). To identify $IL1B^+$ monocytes, we used a mixture of goat anti human CD14 (119-13402, RayBiotech) biotinylated rabbit anti-human IL1b (OABF00305-Biotin, Aviva Systems Biology) and mouse anti-human CD16 (clone DJ130c, LifeSpan Biosciences). Finally, slides were probed with rabbit monoclonal anti-human CD90 (2694-1, Epitomics), rat anti-human HLADR (cloneYE2/36 HLK, LifeSpan Biosciences) and mouse anti-human CD45 (clone F10-89-4, abcam) to detect fibroblasts, Class II expressing cells and hematopoietic cells, respectively. Slides with primary antibodies were incubated in a humid chamber at room temperature, overnight. Next morning, primary antibodies for triple T cell stain and for detecting ABC's were revealed with Alexa Fluor 568 donkey anti-goat IgG (A-11057, Thermo Fisher Scientific), Alexa Fluor 488 donkey anti-rabbit (771-546-152, Jackson ImmunoResearch Laboratories) and Alexa fluor 647 donkey anti-mouse (715-606-151, Jackson ImmunoResearch Laboratories) . Primary antibodies in the stain for monocytes were revealed with Alexa Fluor 568 donkey anti-goat Ig G, Alexa fluor 488 streptavidin (S11223, Thermo

Fisher Scientific) and Alexa Fluor 647 donkey anti-mouse Ig G. Primary antibodies in the stain for fibroblasts and hematopoietic cells were detected with Cy3 donkey anti-rabbit (711-166-152, Jackson ImmunoResearch Laboratories), Alexa Fluor 488 donkey anti-rat Ig G (A-21208, Thermo Fisher Scientific) and Alexa Fluor 647 donkey anti-mouse Ig G. After 2 hours of incubation, slides were washed and mounted with Vectashield mounting media with DAPI (H-1200, Vector Laboratories). Pictures were taken with an Axioplan Zeiss microscope and recorded with a Hamamatsu camera. Double immunofluorescence pictures were obtained by merging individual channels in NIH Image J software.

Estimation of number of cells by counting nuclei

To estimate number of cells, we counted number of nuclei in 5 random 200x fields that show synovial lining with Image J NIH software. Briefly, original color TIFF files were first transformed into 8-bit grayscale images. We use similar settings to adjust threshold in 8-bit images (Lower threshold level: 0, Upper threshold level: 60). Next, we used process: binary: watershed to separate nuclei. In the analyze icon, we select analyze particles and we use equal settings to count particles in our images (Size (pixel^2): 50-infinity, circularity 0.00-1.00, Show: outlines) and we selected to display results. We visually confirmed that individual nuclei were outlined in the final image and calculate the average number of cells/200x field in individual samples.

Tissue samples classification based on leukocyte infiltration

We classified RA tissue samples into leukocyte-poor RA and leukocyte-rich RA based on Mahalanobis distance from OA samples computed on leukocyte abundance measured by flow cytometry. We first took OA samples as a reference, and calculated a multivariate normal distribution of the percentages of live T cells, B cells, and monocytes. Here we used the

mahalanobis function in R: data x = a matrix of all 51 samples by flow gates of T cells, B cells, and monocytes; center = mean of T cells, B cells, and monocytes for all OA samples; covariance = covariance of T cells, B cells, and monocytes for all OA samples. We calculated the square root to get Mahalanobis distance for each sample,

$$mah = \sqrt{(x-\mu)'\Sigma^{-1}(x-\mu)}.$$

We then defined the maximum value of all OA samples (4.5) as a threshold to define 19 leukocyte-rich RA (>4.5) and 17 leukocyte-poor RA (<4.5) samples in our cohort (**Supplementary Fig. 1d**).

Bulk RNA-seq gene expression quantification

We sorted cells into the major immune and stromal cell populations: T cells, B cells, monocytes and synovial fibroblasts. We then performed RNA sequencing. Full-length cDNA and sequencing libraries were performed using Illumina Smart-eq2 protocol[238]. Libraries were sequenced on MiSeq from Illumina to generate 35 base paired-end reads. Reads were mapped to Ensembl version 83 transcripts using kallisto 0.42.4 and summed expression of all transcripts for each gene to get transcripts per million (TPM) for each gene[167].

Bulk RNA-seq quality control

For quality control of bulk RNA-seq data, we began by defining common genes as the set of genes detected with at least 1 mapped fragment in 95% of the samples. Then, for each sample, we computed the percent of common genes detected in that sample. Low quality samples are those that have less than 99% of common genes detected, and these were discarded. We found that the low-quality samples also had low cell counts (**Supplementary Fig. 11a**). After discarding 25 low quality samples, we used 167 good

quality samples, including 45 fibroblast samples, 46 monocyte samples, 47 T cell samples, and 29 B cell samples in all bulk RNA-seq analyses. Cell lineage markers, *PDGFRA*, *C1QA*, *CD3D*, and *CD19*, are expressed selectively by fibroblasts, monocytes, t cells, and b cells, respectively (**Supplementary Fig. 11c**).

Single-cell RNA-seq gene expression quantification

Single-cell RNA-seq was performed using the CEL-Seq2 method[47] with the following modifications. Single cells were sorted into 384-well plates containing 0.6 µL 1% NP-40 buffer in each well. Then, 0.6 µL dNTPs (10mM each; NEB) and 5 nl of barcoded reverse transcription primer (1 µg/µL) were added to each well along with 20 nL of ERCC spike-in (diluted 1:800,000). Reactions were incubated at 65°C for 5 min, and then moved immediately to ice. Reverse transcription reactions were carried out, as previously described (Hashimshony *et al.*, 2016), and cDNA was purified using 0.8X volumes of Agencourt RNAClean XP beads (Beckman Coulter). *In vitro* transcription reactions (IVT) were performed, as described followed by EXO-SAP treatment. Amplified RNA (aRNA) was fragmented at 80°C for 3 min and purified using Agencourt RNAClean XP beads (Beckman Coulter). The purified aRNA was converted to cDNA using an anchored random primer and Illumina adaptor sequences were added by PCR. The final cDNA library was purified using Agencourt RNAClean XP beads (Beckman Coulter). Paired-end sequencing was performed on the HiSeq 2500 in High Output Run Mode with a 5% PhiX spike-in using 15 bases for Read 1, 6 bases for the Illumina barcode and 36 bases for Read 2. We mapped Read2 to human reference genome hg19 using STAR 2.5.2b, and removed samples with outlier performance using Picard. We quantified gene levels by counting UMIs (Unique Molecular Identifiers) and transforming the counts to Log_2(CPM+1) (Counts Per Million).

Single-cell RNA-seq quality control

For quality control of single-cell RNA-seq data, we filtered out molecules that are likely to be contamination between cells, and we used several metrics to exclude poor quality cells. We identified molecules that are likely to represent cell-to-cell cross-contamination as follows. Many single-cell RNA-seq library preparation protocols include pooling and amplification of cDNA molecules from a large number of cells. This can introduce cell-to-cell contamination. We found that molecules represented by a small number of reads are more likely to be contaminant molecules derived from other cells. We developed a simple algorithm to set a threshold for the minimum number of reads per molecule, and we ran it separately for each quadrant of 96 wells in each 384-well plate. We used 2 marker genes expected to be exclusively expressed in each of the 4 cell types: *PDGFRA* and *ISLR* for fibroblasts, *CD2* and *CD3D* for T cells, *CD79A* and *RALGPS2* for B cells, and *CD14* and *C1QA* for monocytes. We counted nonzero expression of these genes in the correct cell type as a true positive and nonzero expression in the incorrect cell type as a false positive. Then we tried each threshold for reads per molecule from 1-20 and chose the threshold that maximizes the ratio of true positive to false positive (**Supplementary Fig. 14**). This left us with 7,127 cells and 32,391 genes. Next, we discarded cells with fewer than 1,000 genes detected with at least one fragment. We also discarded cells that had more than 25% of molecules coming from mitochondrial genes. This left us with 5,265 cells. We discarded genes that had nonzero expression in fewer than 10 cells. We show all post-QC single cells based on the number of genes detected and percent of molecules from mitochondrial genes for each identified cluster (**Supplementary Fig. 15**).

Mass cytometry sample processing and quality control

We collected 6 leukocyte-rich, 9 leukocyte-poor RA, and 11 OA samples for mass cytometry analysis, and processed the samples, as described previously[214]. Briefly, we analyzed samples on

a Helios instrument (Fluidigm) after antibody staining and fixation (**Supplementary Table 2**). Mass cytometry data were normalized using EQ™ Four Element Calibration Beads (Fluidigm), as previously described[188]. Cells were first gated to live DNA+ cells prior to gating for specific cell populations using the following scheme: B cells ($CD3^-CD14^-CD19^+$), fibroblasts ($CD45^-PDPN^+$), monocytes ($CD3^-CD14^+$), and T cells ($CD3^+CD14^-$). All biaxial gating was performed using FlowJo 10.0.7.

Integrative computational pipeline for scRNA-seq clustering

We developed a graph-based unbiased clustering pipeline based on canonical correlation analysis to take advantage of the shared variation between single-cell RNA-seq and bulk RNA-seq. We used this computational pipeline to analyze single cells from each cell type. The overall flowchart is shown in **Supplementary Fig. 2**. We describe the details of each step as follows:

1) We first selected the highly variable genes such that the mean and standard deviation are in the top 80% of the density distributions from the single-cell RNA-seq matrix C(g genes by m cells, $c_{1,\dots,m}$) and bulk RNA-seq matrix (g genes by n samples, $s_{1,\dots,n}$), respectively. We focused on the highly variable genes detected in both scRNA-seq and bulk RNA-seq datasets.

2) Based on the shared highly variable genes, we integrated single-cell RNA-seq with bulk RNA-seq by finding a linear projection of bulk samples and single cells such that the correlation between the genes are maximized using the CCA method[239]. CCA finds two vectors a and b that maximize the linear correlations $cor(CV_{s1}, CV_{c1})$, where $CV_{s1} = a_1s_1 + a_2s_2 + \dots + a_ns_n$ and $CV_{c1} = b_1c_1 + b_2c_2 + \dots + b_mc_m$. Each bulk sample s_i gets a coefficient a_i and each cell c_i gets a coefficient b_i. The linear combination of all samples $s_{1,\dots,n}$ arranges bulk genes along the canonical variate CV_{s1} and the linear combination of all cells $c_{1,\dots,m}$ arranges single-cell genes along CV_{c1}. CCA defines the coefficients $a_{1,\dots,n}$ and $b_{1,\dots,n}$ that arrange the genes from the two

datasets in such a way that the correlation between CV_{s1} and CV_{c1} is maximized. After CCA finds the first pair of canonical variates, the next pair is computed on the residuals, and so on.

3) We calculated the cell-to-cell similarity matrix using Euclidean distance on the top ten CCA canonical variates.

4) We built up a K-nearest neighbors (KNN) graph based on the cell-to-cell similarity matrix (Euclidean distance) based on local ordinal embedding (LOE), a graph embedding method. We then converted the KNN neighbor relation matrix into an adjacency matrix using the graph.adjacency function from igraph R package;

5) We clustered the cells using the Infomap algorithm for community detection by applying a cluster_infomap function from igraph R package to decompose the cell-to-cell adjacency matrix into major modules by minimizing a description of the information flow;

6) We then constructed a low dimensional embedding using tSNE based on the cell-to-cell distance matrix using the following parameters: perplexity = 50 and theta = 0.5;

7) We identified and prioritized significantly differentially expressed genes for each distinct cluster based on percent of non-zero expressing cells, AUC score[240], and fold-change;

8) For pathway analysis, we downloaded gene sets from Gene Ontology (GO) terms on April 2017. This included 9,797 GO terms and 15,693 genes. We also used the immunological signatures from 4872 hallmark gene sets from MSigDB[241] to test enrichment of all the tested genes sorted by decreased AUC scores for each cluster by 10^5 permutation tests[242]. We used the liger R package (https://github.com/JEFworks/liger) to do gene set enrichment analysis (GSEA).

To identify the most reasonable and stable clusters, we ran this pipeline repeatedly while tuning the number of top canonical variates (4, 8, 12, 16, and 20) that were incorporated for the cell-to-cell similarity matrix, and the number of k (50, 100, 150, 200, 250, and 300) to build up the K-

nearest neighbors' graph. We chose the clusters that yielded the greatest number of differentially expressed genes. We used Silhouette analysis[243,244] on the cell-to-cell Euclidean distance matrix to evaluate our clustering results (**Supplementary Fig. 2b**). For each cell, the silhouette width $s(i)$ is defined as follows:

$$s(\mathrm{i}) = \frac{b(i)-a(i)}{\max(a(i),b(i))},$$

where $a(i)$ is the average dissimilarity between a cell and all the other cells in the same cluster and $b(i)$ is the average distance between a cell and all cells in the nearest cluster to which the cell does not belong. The measure range is [-1, 1], where a value near 1 indicates a cell is far from neighboring clusters, a value near 0 indicates a cell is near a decision boundary, and a value near -1 indicates the cell is closer to a neighboring cluster than its own cluster.

Thus, for each pair of single-cell RNA-seq and bulk RNA-seq, we ran our pipeline on the shared samples that have both datasets for each cell type **(Figure 4-1b)**. For integrating fibroblast data, we used 45 bulk RNA-seq samples, 1,844 single cells and 7,016 shared highly variable genes; for integrating monocyte data, we used 47 bulk RNA-seq samples, 750 single cells and 7,016 shared highly variable genes; for integrating T cell data, we used 47 bulk RNA-seq samples, 1,716 single cells and 7,003 shared highly variable genes; for integrating B cell data, we used 29 bulk RNA-seq samples, 1,142 single cells and 7,023 shared highly variable genes.

Mass cytometry clustering

We created mass cytometry datasets for analysis by concatenating cells from all individuals for each cell type. For donors with more than 1,000 cells, we randomly selected 1,000 cells to ensure that samples were equally represented. In this way, we created downsampled datasets of 25,161 fibroblasts from 23 patients, 15,298 monocytes from 26 patients, 19,985 T cells from 26 patients, and 8,179 B cells from 23 patients for analysis. We then applied the tSNE algorithm

(Barnes-Hut implementation)[142] to each dataset using the following parameters: perplexity = 30 and theta = 0.5. We used all markers except those used to gate each population in the SNE clustering. To identify high-dimensional populations, we used a version of DensVM[134], modified as in[126]. DensVM performs kernel density estimation across the dimensionally reduced SNE map to build a training set, then assigns cells to clusters by their expression of all markers using an SVM classifier. We modified the DensVM code to increase the range of potential bandwidths searched during the density estimation step and to return the SVM model generated from the tSNE projection. We summarized the details of the clusters with proportion of cells from each disease cohort in **Supplementary Table 3**.

Disease association test of cell populations

We tested whether abundances of individual populations were altered in RA case samples compared to OA controls using two ways. First, we assessed whether marker genes (AUC > 0.7, 20 < n < 100) characteristic of each scRNA-seq cluster were differentially expressed in the same direction in scRNA-seq and bulk RNA-seq datasets. Second, we applied MASC[126], a single cell association method for testing whether case-control status influences the membership of single cells in any of multiple cellular subsets while accounting for technical confounds and biological variation. We specified donor identity and batch as random-effect covariates.

Integration of bulk RNA-seq with mass cytometry

We used CCA to associate the abundances of mass cytometry clusters with gene expression in bulk RNA-seq. We started by selecting the samples that had both data types. The mass cytometry data matrix has samples and clusters, where the values represent proportions of cells from each sample in each cluster. The bulk RNA-seq data matrix has samples and genes, where the values represent proportions of gene abundance from each sample in each gene. CCA

identifies canonical variates (a linear combination of bulk RNA-seq genes and a linear combination of mass cytometry cluster proportions) that maximize correlation of samples along each canonical variate. In other words, it tries to arrange samples from each dataset in a similar order along each canonical variate. We ran CCA separately for fibroblasts, monocytes, T cells, and B cells. For fibroblasts, we associated 2,299 genes with 8 mass cytometry clusters on 22 samples. For monocytes, we associated 2,161 genes with 5 mass cytometry clusters on 25 samples. For T cells, we associated 2,255 genes with 9 mass cytometry clusters on 26 samples. For B cells, we associated 22,95 genes with 10 mass cytometry clusters on 17 samples.

Finding correspondence between scRNA-seq clusters and mass cytometry clusters

1) For each cell type, we ran CCA with mass cytometry clusters with bulk RNA-seq. Each gene is correlated with each canonical variate (CV). Also, each mass cytometry cluster is correlated with each CV. By visualizing these correlations, we can see the positions of bulk RNA-seq genes and mass cytometry clusters in the same space (**Figure 4-4h**).

2) We then associated single-cell RNA-seq clusters with mass cytometry clusters by projecting cluster markers (AUC > 0.7) for each single-cell RNA-seq cluster in the CCA space acquired from step 1).

3) We took the average across the cluster marker genes for each single-cell RNA-seq cluster for each CV and obtained an "average CV" matrix.

4) Based on the "average CV" matrix, we computed Spearman correlation between the scRNA-seq average CV and the CV for mass cytometry clusters.

5) Next, we generated a null distribution for the Spearman correlations by shuffling the scRNA-seq gene names and then repeating steps 2-4 10,000 times.

6) For the 10,000 replicates of CCA matrix, we repeated from step 2 to step 5. Then, we counted how many times the correlation of each pair was greater than the observed value from step 4).

$$permutation\ p\ = \frac{1 + sum(cor_{perm} > cor)}{1 + 1e^4}.$$

7) Finally, we converted the to a $permutation\ p$ to a $z - score$.

Differential expression analysis with bulk RNA-seq

We classified all the samples into OA, leukocyte-poor RA, and leukocyte-rich RA synovial tissues based on the quantitative analysis of T cells, B cells, and monocytes by flow cytometry. PCA on bulk RNA-seq samples showed separation of leukocyte-rich and leukocyte-poor RA on the first or second principal components. For differential analysis, we used the limma R package to identify significantly differentially expressed genes. We used the Benjamini-Hochberg method to estimate false discovery rate (FDR).

Identification of markers for distinct scRNA-seq clusters

Based on the single-cell RNA-seq clusters, we identified cluster marker genes by comparing the cells in one cluster with all other clusters from the same cell type, based on Log_2(CPM+1). We prioritized cluster marker genes using three criteria: 1) percent of non-zero expressing cells > 60%; 2) are under the receiver-operator curve (AUC)[240] > 0.7; and 3) fold-change (FC) > 2.

Intracellular flow cytometry of synovial tissue T cell stimulation

Disaggregated synovial tissue cells were incubated with Fixable Viability Dye (eBioscience) and Fc blocking antibodies (eBioscience) followed by staining for surface markers in Brilliant Stain Buffer (BD Bioscience). Cell were then fixed and permeabilized using an intracellular staining

kit (eBioscience), followed by intracellular staining for granzymes or cytokines. Antibodies used in this study include anti-CD45 (clone HI30) from BD Biosciences; anti-CD3 (clone UCHT1), anti-CD8 (clone SK1), anti-CD14 (clone M5E2), anti-CD4 (clone RPA-T4), anti-HLA-DR (clone L243), anti-granzyme B (clone GB11), and anti-granzyme K (clone GM26E7) from Biolegend; and anti-IFNG (clone 4S.B3) and anti-TNF (clone MAb11) from eBioscience. Data were collected on a BD Fortessa flow cytometer and analyzed using FlowJo 10.5 software. Disaggregated synovial tissue cells were incubated with a cell stimulation cocktail containing PMA and ionomycin (eBioscience) in RPMI with 10% fetal calf serum (Gemini). After 15 minutes, brefeldin A (eBioscience) was added. The cells were incubated at 37C 5% CO_2 for an additional 2 hours. The cells were then collected and stained for intracellular cytokines following the protocol above and the data was shown in **Supplementary Fig. 10**.

Statistics

Results are shown as mean with 95% confidence intervals. The statistics tests used were *t*-test and Kolmogorov-Smirnov test, unless otherwise stated, as described with one-sided or two-sided in the figure legends. Benjamini-Hochberg FDR < 0.01 and Fold-change > 2 were considered to be statistically significant when appropriate.

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data Availability

The single-cell RNA-seq data, bulk RNA-seq data, mass cytometry data, flow cytometry data, and the clinical and histological data this study are available at ImmPort (https://www.immport.org/shared/study/SDY998 and https://www.immport.org/shared/study/SDY999, study accession codes SDY998 and SDY999). The raw single-cell RNA-seq and mass cytometry data are deposited in dbGAP (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001457.v1.p1. The source code repository of the computational and statistical analysis is located at https://github.com/immunogenomics/amp_phase1_ra. Data can also be viewed on 3 different websites at https://immunogenomics.io/ampra, https://immunogenomics.io/cellbrowser/, and https://portals.broadinstitute.org/single_cell/study/amp-phase-1.

Competing Financial Interests

The authors declare no competing financial interests.

AMP RA Phase 1 consists of all named authors, and also includes the following authors:
Jennifer Albrecht[9], S. Louis Bridges, Jr.[10], Christopher D. Buckley[20], Jane H. Buckner[27], James Dolan[18], Joel M. Guthridge[28], Maria Gutierrez-Arcelus[1,2,3,4,5], Lionel B. Ivashkiv[8,29,30], Eddie A. James[27], Judith A. James[27], Josh Keegan[18], Yvonne C. Lee[13], Mandy J. McGeachy[17], Michael McNamara[7,8], Joseph R. Mears[1,2,3,4,5], Fumitaka Mizoguchi[5,31], Jennifer Nguyen[18], Akiko Noma[4], Dana E. Orange[7,8,32], Mina Pichavant[33,34], Christopher Ritchlin[9], William H. Robinson[33,34], Anupamaa Seshadri[18], Danielle Sutherby[4], Jennifer Seifert[22], Jason D. Turner[20], Paul J. Utz[33,34]

[27]Translational Research Program, Benaroya Research Institute at Virginia Mason, Seattle, WA 98101, USA
[28]Department of Arthritis & Clinical Immunology, Oklahoma Medical Research Foundation, Oklahoma City, OK 73104, USA
[29]Graduate Program in Immunology and Microbial Pathogenesis, Weill Cornell Graduate School of Medical Sciences, New York, NY 10065, USA
[30]David Z. Rosensweig Genomics Research Center, Hospital for Special Surgery, New York, NY 10021, USA
[31]Department of Rheumatology, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo 113-8519, Japan
[32]The Rockefeller University, New York, NY 10065, USA

[33]Division of Immunology and Rheumatology, Department of Medicine, Stanford University School of Medicine, Palo Alto, CA 94305, USA
[34]The Institute for Immunity, Transplantation, and Infection, Stanford University School of Medicine, CA 94305, USA

Chapter 5:

Discussion

The advent of single cell technologies has the potential to revolutionize the study of RA by offering an unbiased approach to detecting and characterizing cell heterogeneity in blood and tissue. The work in this dissertation serves to advance the field of single cell disease-association studies, providing both an effective method for performing association tests with single cell data while controlling for confounding effects and evidence of multiple populations that are potentially linked to the pathogenesis of rheumatoid arthritis. This section will cover the strengths of MASC and its future utility; the characterization of novel disease-linked immune cell populations; and the future of single cell association studies more broadly.

Single cell disease association studies

Traditionally, the most common way to understand which cell populations are linked to disease has been to study them in isolation. Multi-step experiments in which populations of interest are first isolated based upon phenotypic markers and then characterized by secondary assays have taken advantage of fluorescence-activated cell sorting (FACS) to provide a non-destructive method of partitioning single cells into discrete populations. This approach can be quite useful, especially when the disease-associated populations are known and express markers that allow them to be easily separated by FACS. For example, celiac disease is an autoimmune disorder known to be triggered the presentation of ingested gluten to CD4+ T cells in the small intestine[245,246]. As the disease-relevant population is already known in celiac, researchers have been able to use HLA-bound tetramers that have gluten loaded to specifically isolate gluten-reactive T cells and perform further characterization experiments[247-249]. Conversely, in diseases like RA where the pathogenic immune populations are not obvious, the experimental strategy of sorting and characterization is considerably less effective and such have yielded discordant results. Inconsistencies can be partially attributed to variation in markers used across different studies or the difficulty of resolving highly heterogeneous populations with bulk cell assays – which may be overcome with advancing single cell technologies.

This work details critical elements to consider when conducting single cell disease association studies: first, the importance of proper experimental design to minimize batch and technical variation from the outset; second, the varied approaches to identifying populations in single cell datasets and how to comparatively evaluate single cell clustering methods with quantitative metrics; and finally a robust statistical strategy for performing association testing, MASC (Mixed-effect modeling of Associations of Single Cells). This method accepts user-identified populations regardless of clustering method, directly reports the significance of case-control associations for each cluster, provides an estimate of the effect size of the association itself, and incorporates both technical covariates (e.g. batch) and clinical covariates when modeling associations, a key feature when analyzing high-dimensional datasets of large disease cohorts. MASC outperforms naïve association-testing strategies – like difference-of-means and binomial tests – and demonstrates better controlled error rates. MASC compares favorably to other single cell association methods like Citrus[137], which uses nested hierarchical clustering and penalized regression models to identify features (defined here as clusters of single cells or median expression levels of markers within a cluster) that are predictive of clinical endpoints. In contrast to MASC, which is able to operate on any clustered single cell dataset, Citrus requires down-sampling cells from each sample and does not retain single cell resolution, which impedes the interpretation of clusters found to be predictive. As demonstrated in Chapter 2, MASC is an effective method for performing association testing with single cell transcriptomic data and can easily include method-specific technical covariates, like the sequencing lane samples were processed in. Indeed, we are currently employing MASC to identify differentially abundant T cell populations in a multimodal transcriptomic and proteomic dataset; similarly, MASC could be used to identify associations of clusters defined by epigenomic features in single cell ATAC-seq datasets[250].

Single cell experiments reveal immune populations associated with RA

In two separate single cell association studies, we were able to describe novel immune populations and robustly identify which populations were statistically associated with disease status. As shown in Chapter 3, single cell mass cytometry resolved a diverse set of CD4+ T cell populations, many of which had not been well characterized in the past. While established CD4+ populations like T_h1 and T_{reg} cells were not significantly more abundant in RA patients compared to controls, MASC identified an RA-associated population of CD4+ T effector memory cells defined by the expression of HLA-DR and the lack of CD27. Analyzing the transcriptome of these cells and closely related populations allowed us to define a gradient from naïve CD4+ cells to the expanded CD27- HLA-DR+ population, demonstrating that CD4+ populations could be organized by increased expression of gene sets linked to cytotoxicity and effectorness. While the CD27- HLA-DR+ population was rare in peripheral blood, it was highly expanded in tissue and produced cytotoxic molecules upon stimulation, suggesting that population may represent an attractive target as a biomarker or as a therapeutic target.

Cytotoxic T cells have been traditionally defined as the subset of mature CD8+ T lymphocytes that play a major role in destroying cells targeted by the immune system, such as virally-infected and tumoral cells[251,252]. In cytotoxic CD8+ T cells, cell-killing activity is dependent on interactions between the T cell receptor (TCR) and MHCI-bound peptides[253,254]. Termed cytotoxic T lymphocytes or CTLs, these cells are capable of causing cell death through multiple, well-described mechanisms: primarily the export of perforin through granule exocytosis and the induction of the Fas/Fas-ligand apoptosis pathway, although the release of cytotoxic cytokine molecules like TNF and IFN-γ are also thought to play a role[255]. Although the majority of literature describing cytotoxicity in T cells refers to CTLs, there has been evidence of CD4+ T cells with cytotoxic properties for a long time[256-259]. A recent single cell study in mice demonstrated that age polarizes CD4+ T cells towards extreme effector states, including a

cytotoxic subset characterized by the production of TNF, IFN-γ, perforin, and Granzyme B upon stimulation[260]. In humans, cytotoxic CD4+ T cells have been strongly associated with chronic viral infections, which, like autoimmune disorders, share the feature of cells being exposed to antigens for long periods of time[181,261-263]. Thus, the chronic inflammatory context of rheumatoid arthritis may promote the differentiation of CD4+ effector cells into the CD27- HLA-DR+ phenotype, which resemble CD4+ CTLs and express molecules like perforin. CD4+ CTLs have been observed in other autoimmune disorders[264-266], such as the recent observation of cells with similar surface marker phenotypes and transcriptomic profiles in IgG4-related disease[184] and celiac disease[267].

While the loss of expression of CD27 on CD4+ T cells is well-characterized as a marker of T cells that are antigen-experienced and have reached a terminal differentiation state[260], the functional role of the expression of HLA-DR in CD27- HLA-DR+ cytotoxic CD4+ T cells is less clear. It has been previously observed that a subset of regulatory CD4+ T cells expresses HLA-DR on their surface; these cells are considered to be highly-suppressive compared to typical T_{reg} cells and play an important role in mediating transplant rejection[268-270]. The CD27- HLA-DR+ population observed as expanded in RA were distinctly not similar to T_{reg} cells, lacking expression of the lineage-defining transcription factor *FoxP3*. However, it is intriguing that HLA-DR+ T_{reg} cells are known to be particularly vulnerable to Granzyme B-induced cell death expressed from other CD4+ T cells[271]. One hypothesis for the functional role of CD27- HLA-DR+ cytotoxic CD4+ T cells in rheumatoid arthritis could be that they are particularly effective at nullifying the anti-inflammatory effects of regulatory T cells at peripheral sites of disease where they are most enriched, like the synovium. In keeping with this theory, peripheral CD4+ HLA-DR+ T cells have been shown to compromise the functional capacity as well as resist suppression by T_{reg} cells in tuberculosis, a disease characterized by chronic T cell activation[272]. Finally, it is notable that a recent single cell study of innate and adaptive T cell populations identified an innateness gradient with adaptive cells on one end and natural killer cells on the

other[273]. When characterized transcriptomically, CD27- HLA-DR+ cells were the most innate-like compared to both naïve and other memory CD4+ populations. As the more innate-like cells contained significant amounts of pre-formed mRNA for cytokines like IFN-γ, this finding may suggest that CD27- HLA-DR+ cells are poised for rapid, pro-inflammatory effector functions. Although further characterization experiments are clearly required, there are many lines of evidence suggesting that CD27- HLA-DR+ CD4+ T cells are functionally relevant to RA pathology – as well as to autoimmunity more broadly – and represent a high-priority target for development as biomarkers or therapeutic targets.

Chapter 4 describes a large-scale case-control study of RA synovial tissue samples, from the Accelerating Medical Partnerships Rheumatoid Arthritis/Systemic Lupus Erythematosus (AMP RA/SLE) network, which involves obtaining, disaggregating, and performing single cell profiling on synovial tissue from cases and controls to query both immune infiltration and stromal adaptions. In this study, we were able to align multiple modes of single cell transcriptomic, bulk transcriptomic, and proteomic data to more robustly detect disease-linked populations. In this study, the limited sample size prevented us from performing differential abundance analyses using the single cell transcriptomic data alone. However, by defining cell populations with mass cytometry and using MASC to identify differential abundance, we were able to annotate disease-relevant populations transcriptomically by aligning mRNA-based clusters and protein-based clusters with CCA. We identified sublining fibroblasts, pro-inflammatory monocytes, autoimmune-associated B cells, and PD-1^{hi} T cells as expanded in RA patients compared to osteoarthritis (OA) controls. While sublining fibroblasts had previously been linked to RA[223], the combination of single cell modalities allowed us to resolve a population of THY1+CD34-HLA-DR^{hi} fibroblasts that was nearly 15-fold expanded and produced massive amounts of the pro-inflammatory cytokine *IL6*, highlighting this specific subset as a therapeutic target. We saw that PD-1^{hi} T cells were expanded in the RA synovium, partially replicating previous observations that peripheral helper T cells were expanded in RA. The observation of an

expansion of autoimmune-associated B cells was particularly striking, as this had not previously been demonstrated in RA. An advantage of the AMP RA/SLE network is that these discoveries can be directly followed up on in a larger cohort of about 100 synovial tissue samples from RA patients, which is currently planned for the next phase of the project. The greatly increased sample size of this study should improve our ability to detect more subtle case-control associated shifts in abundance, as the preliminary data in Chapter 2 suggests.

Looking forward with single cell immunoprofiling

The rapid expansion of high-dimensional single-cell technologies in the past decade has revolutionized the study of systems characterized by a diversity of cell types, like the immune system. By capturing cell-to-cell heterogeneity that is obscured by bulk analyses, methods such as mass cytometry and single-cell RNA-seq (scRNA-seq) provide the opportunity to measure proteins and genes that reflect each T cell's functional program. New technologies like single-cell ATAC-seq (scATAC-seq) identify accessible regions of DNA across the genome and support clustering of single cells by their active regulatory elements, which may represent a more functionally-accurate way to define populations[250]. Single cell repertoire sequencing provides the opportunity to uniquely identify individual T and B cell clones as well as trace their expansion in a disease setting[274,275]. Beyond integrating data across studies and across assays, the next stage of advancement for single cell technologies is be the simultaneous acquisition of transcriptomic and proteomic data from a single cell. Multiple methods for conducting such analyses have been described[144,186,276,277] but have yet to be applied in any large-scale immunoprofiling efforts. The ability to obtain this type of data would allow research into the temporal dynamics of transcription and protein expression as well as provide higher-resolution definition of single cells. Alongside the development of combined single-cell transcriptomic and proteomic assays, work is currently ongoing on optimizing methods that can perform both

single cell RNA-seq and repertoire sequencing, or single cell ATAC-seq and repertoire sequencing, simultaneously[278,279].

Given the high levels of inter-individual variability in the human immune system, the ability to aggregate data across multiple studies is an attractive goal for conducting well-powered analyses. Currently, data aggregation is challenging due to the high dimensionality of single cell data and the difficulty of overcoming different datasets for analysis which include differences in the use of specific sequencing protocols, technical batch effects, and differences in sample handling. Standardization of normalization and quality control methods will be key, as small differences in data processing can overpower biological signals in the noisy context of immunoprofiling; for example, the use of different software pipelines for processing single cell RNA-seq data will impede combined analysis. We have demonstrated one method of aligning single cell data across modalities with canonical correlation analysis; another recently developed method called Harmony involves finding a joint projection of different datasets optimized such that the influence of batch and sample effects is minimzed[148].

For immunological applications, a key initial step should be to better characterize human lymphocytes using single cell data. Building a reference map of the human immune system is a difficult and complicated task; however, the dendritic cell atlas or the work of Wong et al. characterizing T cells across tissues provide examples of the power of this approach[84,212]. Incorporating data on from multiple assays to define lymphocyte profiles will be essential for understanding their functional impact, as shown by multiple studies that utilize repertoire sequences or expression data in combination with single cell cytometry to identify disease-relevant populations[168,280,281]. The development of new peptide-MHC multimeric complexes supports the detection and isolation of antigen-specific lymphocytes at much lower frequencies than was previously feasible[282], while new methods have been recently developed to provide high-throughput single cell repertoire sequencing of B and T lymphocytes[283,284].

High-dimensional single cell analyses of RA synovium have revealed novel lymphocyte and stromal cell populations that are pathologically expanded in the joints of RA patients. These cell populations may now be evaluated as potential therapeutic targets. New single cell technologies will enable detailed characterization of the specific clones of CD4+ T cells that are expanded in RA and help highlight new cell phenotypes to pursue as therapeutic targets or biomarkers. However, the increased resolution of single cell analyses will be wasted without defining a set of statistically-sound standards for experiments that enable combining experimental data across batches, assays, and studies. As the magnitude of data that is produced by single cell immunoprofiling increases and reveals unprecedented levels of diversity among immune cell, methodological rigor will be critical for properly deciphering mechanisms of disease.

References

1 Kahaly, G. J. & Hansen, M. P. Type 1 diabetes associated autoimmunity. *Autoimmun Rev* **15**, 644-648, doi:10.1016/j.autrev.2016.02.017 (2016).

2 Schuppan, D., Junker, Y. & Barisani, D. Celiac disease: from pathogenesis to novel therapies. *Gastroenterology* **137**, 1912-1933, doi:10.1053/j.gastro.2009.09.008 (2009).

3 Dendrou, C. A., Fugger, L. & Friese, M. A. Immunopathology of multiple sclerosis. *Nat Rev Immunol* **15**, 545-558, doi:10.1038/nri3871 (2015).

4 Thong, B. & Olsen, N. J. Systemic lupus erythematosus diagnosis and management. *Rheumatology (Oxford)* **56**, i3-i13, doi:10.1093/rheumatology/kew401 (2017).

5 Anaya, J. M. The diagnosis and clinical significance of polyautoimmunity. *Autoimmun Rev* **13**, 423-426, doi:10.1016/j.autrev.2014.01.049 (2014).

6 Wang, L., Wang, F. S. & Gershwin, M. E. Human autoimmune diseases: a comprehensive update. *J Intern Med* **278**, 369-395, doi:10.1111/joim.12395 (2015).

7 Ngo, S. T., Steyn, F. J. & McCombe, P. A. Gender differences in autoimmune disease. *Front Neuroendocrinol* **35**, 347-369, doi:10.1016/j.yfrne.2014.04.004 (2014).

8 Mackay, I. R. Science, medicine, and the future: Tolerance and autoimmunity. *BMJ* **321**, 93-96, doi:10.1136/bmj.321.7253.93 (2000).

9 Theofilopoulos, A. N., Kono, D. H. & Baccala, R. The multiple pathways to autoimmunity. *Nat Immunol* **18**, 716-724, doi:10.1038/ni.3731 (2017).

10 Park, H., Bourla, A. B., Kastner, D. L., Colbert, R. A. & Siegel, R. M. Lighting the fires within: the cell biology of autoinflammatory diseases. *Nat Rev Immunol* **12**, 570-580, doi:10.1038/nri3261 (2012).

11 Scott, D. L., Wolfe, F. & Huizinga, T. W. Rheumatoid arthritis. *Lancet* **376**, 1094-1108, doi:10.1016/S0140-6736(10)60826-4 (2010).

12 Bombardier, C. *et al.* The relationship between joint damage and functional disability in rheumatoid arthritis: a systematic review. *Ann Rheum Dis* **71**, 836-844, doi:10.1136/annrheumdis-2011-200343 (2012).

13 Michaud, K. & Wolfe, F. Comorbidities in rheumatoid arthritis. *Best Pract Res Clin Rheumatol* **21**, 885-906, doi:10.1016/j.berh.2007.06.002 (2007).

14 Ambrosino, P. *et al.* Subclinical atherosclerosis in patients with rheumatoid arthritis. A meta-analysis of literature studies. *Thromb Haemost* **113**, 916-930, doi:10.1160/TH14-11-0921 (2015).

15 Blum, A. & Adawi, M. Rheumatoid arthritis (RA) and cardiovascular disease. *Autoimmun Rev* **18**, 679-690, doi:10.1016/j.autrev.2019.05.005 (2019).

16 Vivar, N. & Van Vollenhoven, R. F. Advances in the treatment of rheumatoid arthritis. *F1000Prime Rep* **6**, 31, doi:10.12703/P6-31 (2014).

17 Rossi, D., Modena, V., Sciascia, S. & Roccatello, D. Rheumatoid arthritis: Biological therapy other than anti-TNF. *Int Immunopharmacol* **27**, 185-188, doi:10.1016/j.intimp.2015.03.019 (2015).

18 Alamanos, Y. & Drosos, A. A. Epidemiology of adult rheumatoid arthritis. *Autoimmun Rev* **4**, 130-136, doi:10.1016/j.autrev.2004.09.002 (2005).

19 Cross, M. *et al.* The global burden of rheumatoid arthritis: estimates from the global burden of disease 2010 study. *Ann Rheum Dis* **73**, 1316-1322, doi:10.1136/annrheumdis-2013-204627 (2014).

20 Kurko, J. *et al.* Genetics of rheumatoid arthritis - a comprehensive review. *Clin Rev Allergy Immunol* **45**, 170-179, doi:10.1007/s12016-012-8346-7 (2013).

21 Smolen, J. S., Aletaha, D. & McInnes, I. B. Rheumatoid arthritis. *Lancet* **388**, 2023-2038, doi:10.1016/S0140-6736(16)30173-8 (2016).

22 Viatte, S., Plant, D. & Raychaudhuri, S. Genetics and epigenetics of rheumatoid arthritis. *Nat Rev Rheumatol* **9**, 141-153, doi:10.1038/nrrheum.2012.237 (2013).

23 Raychaudhuri, S. *et al.* Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat Genet* **44**, 291-296, doi:10.1038/ng.1076 (2012).

24 Eyre, S. *et al.* High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat Genet* **44**, 1336-1340, doi:10.1038/ng.2462 (2012).

25 Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376-381, doi:10.1038/nature12873 (2014).

26 Yamamoto, K., Okada, Y., Suzuki, A. & Kochi, Y. Genetics of rheumatoid arthritis in Asia--present and future. *Nat Rev Rheumatol* **11**, 375-379, doi:10.1038/nrrheum.2015.7 (2015).

27 Perricone, C., Ceccarelli, F. & Valesini, G. An overview on the genetic of rheumatoid arthritis: a never-ending story. *Autoimmun Rev* **10**, 599-608, doi:10.1016/j.autrev.2011.04.021 (2011).

28 Diogo, D., Okada, Y. & Plenge, R. M. Genome-wide association studies to advance our understanding of critical cell types and pathways in rheumatoid arthritis: recent findings and challenges. *Curr Opin Rheumatol* **26**, 85-92, doi:10.1097/BOR.0000000000000012 (2014).

29 Hu, X. *et al.* Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. *Am J Hum Genet* **89**, 496-506, doi:10.1016/j.ajhg.2011.09.002 (2011).

30 Vahedi, G. *et al.* Super-enhancers delineate disease-associated regulatory nodes in T cells. *Nature* **520**, 558-562, doi:10.1038/nature14154 (2015).

31 Young, C. L., Adamson, T. C., 3rd, Vaughan, J. H. & Fox, R. I. Immunohistologic characterization of synovial membrane lymphocytes in rheumatoid arthritis. *Arthritis Rheum* **27**, 32-39, doi:10.1002/art.1780270106 (1984).

32 Takemura, S. *et al.* Lymphoid neogenesis in rheumatoid synovitis. *J Immunol* **167**, 1072-1080, doi:10.4049/jimmunol.167.2.1072 (2001).

33 McInnes, I. B. & Schett, G. The pathogenesis of rheumatoid arthritis. *N Engl J Med* **365**, 2205-2219, doi:10.1056/NEJMra1004965 (2011).

34 Geginat, J. *et al.* The CD4-centered universe of human T cell subsets. *Semin Immunol* **25**, 252-262, doi:10.1016/j.smim.2013.10.012 (2013).

35 Maecker, H. T., McCoy, J. P. & Nussenblatt, R. Standardizing immunophenotyping for the Human Immunology Project. *Nat Rev Immunol* **12**, 191-200, doi:10.1038/nri3158 (2012).

36 Zhu, J., Yamane, H. & Paul, W. E. Differentiation of effector CD4 T cell populations (*). *Annu Rev Immunol* **28**, 445-489, doi:10.1146/annurev-immunol-030409-101212 (2010).

37 Sethi, A., Kulkarni, N., Sonar, S. & Lal, G. Role of miRNAs in CD4 T cell plasticity during inflammation and tolerance. *Front Genet* **4**, 8, doi:10.3389/fgene.2013.00008 (2013).

38 Kondo, Y. *et al.* Review: Transcriptional Regulation of CD4+ T Cell Differentiation in Experimentally Induced Arthritis and Rheumatoid Arthritis. *Arthritis Rheumatol* **70**, 653-661, doi:10.1002/art.40398 (2018).

39 Gizinski, A. M. & Fox, D. A. T cell subsets and their role in the pathogenesis of rheumatic disease. *Curr Opin Rheumatol* **26**, 204-210, doi:10.1097/BOR.0000000000000036 (2014).

40 Wang, W. *et al.* The Th17/Treg imbalance and cytokine environment in peripheral blood of patients with rheumatoid arthritis. *Rheumatol Int* **32**, 887-893, doi:10.1007/s00296-010-1710-0 (2012).

41 Alunno, A. *et al.* Altered immunoregulation in rheumatoid arthritis: the role of regulatory T cells and proinflammatory Th17 cells and therapeutic implications. *Mediators Inflamm* **2015**, 751793, doi:10.1155/2015/751793 (2015).

42 Noack, M. & Miossec, P. Th17 and regulatory T cell balance in autoimmune and inflammatory diseases. *Autoimmun Rev* **13**, 668-677, doi:10.1016/j.autrev.2013.12.004 (2014).

43 Bedoya, S. K., Lam, B., Lau, K. & Larkin, J., 3rd. Th17 cells in immunity and autoimmunity. *Clin Dev Immunol* **2013**, 986789, doi:10.1155/2013/986789 (2013).

44 Zambrano-Zaragoza, J. F., Romo-Martinez, E. J., Duran-Avelar Mde, J., Garcia-Magallanes, N. & Vibanco-Perez, N. Th17 cells in autoimmune and infectious diseases. *Int J Inflam* **2014**, 651503, doi:10.1155/2014/651503 (2014).

45 Astry, B., Venkatesha, S. H. & Moudgil, K. D. Involvement of the IL-23/IL-17 axis and the Th17/Treg balance in the pathogenesis and control of autoimmune arthritis. *Cytokine* **74**, 54-61, doi:10.1016/j.cyto.2014.11.020 (2015).

46 Yang, P. *et al*. Th17 cell pathogenicity and plasticity in rheumatoid arthritis. *J Leukoc Biol*, doi:10.1002/JLB.4RU0619-197R (2019).

47 Schulze-Koops, H. & Kalden, J. R. The balance of Th1/Th2 cytokines in rheumatoid arthritis. *Best Pract Res Clin Rheumatol* **15**, 677-691, doi:10.1053/berh.2001.0187 (2001).

48 Herman, S., Zurgil, N., Langevitz, P., Ehrenfeld, M. & Deutsch, M. Methotrexate selectively modulates TH1/TH2 balance in active rheumatoid arthritis patients. *Clin Exp Rheumatol* **26**, 317-323 (2008).

49 Alzabin, S. & Williams, R. O. Effector T cells in rheumatoid arthritis: lessons from animal models. *FEBS Lett* **585**, 3649-3659, doi:10.1016/j.febslet.2011.04.034 (2011).

50 Zschaler, J., Schlorke, D. & Arnhold, J. Differences in innate immune response between man and mouse. *Crit Rev Immunol* **34**, 433-454 (2014).

51 Gibbons, D. L. & Spencer, J. Mouse and human intestinal immunity: same ballpark, different players; different rules, same score. *Mucosal Immunol* **4**, 148-157, doi:10.1038/mi.2010.85 (2011).

52 Mestas, J. & Hughes, C. C. Of mice and not men: differences between mouse and human immunology. *J Immunol* **172**, 2731-2738, doi:10.4049/jimmunol.172.5.2731 (2004).

53 Fonseka, C. Y., Rao, D. A. & Raychaudhuri, S. Leveraging blood and tissue CD4+ T cell heterogeneity at the single cell level to identify mechanisms of disease in rheumatoid arthritis. *Curr Opin Immunol* **49**, 27-36, doi:10.1016/j.coi.2017.08.005 (2017).

54 Hulett, H. R., Bonner, W. A., Barrett, J. & Herzenberg, L. A. Cell sorting: automated separation of mammalian cells as a function of intracellular fluorescence. *Science* **166**, 747-749, doi:10.1126/science.166.3906.747 (1969).

55 Loken, M. R. & Herzenber, L. A. Analysis of cell populations with a fluorescence-activated cell sorter. *Ann N Y Acad Sci* **254**, 163-171, doi:10.1111/j.1749-6632.1975.tb29166.x (1975).

56 Dean, P. N. & Pinkel, D. High resolution dual laser flow cytometry. *J Histochem Cytochem* **26**, 622-627, doi:10.1177/26.8.357646 (1978).

57 Wilder, M. E. & Cram, L. S. Differential fluorochromasia of human lymphocytes as measured by flow cytometry. *J Histochem Cytochem* **25**, 888-891, doi:10.1177/25.7.70458 (1977).

58 Fox, R. I. *et al*. Synovial fluid lymphocytes differ from peripheral blood lymphocytes in patients with rheumatoid arthritis. *J Immunol* **128**, 351-354 (1982).

59 Pitzalis, C., Kingsley, G., Murphy, J. & Panayi, G. Abnormal distribution of the helper-inducer and suppressor-inducer T-lymphocyte subsets in the rheumatoid joint. *Clin Immunol Immunopathol* **45**, 252-258, doi:10.1016/0090-1229(87)90040-7 (1987).

60 Schmidt, D., Goronzy, J. J. & Weyand, C. M. CD4+ CD7- CD28- T cells are expanded in rheumatoid arthritis and are characterized by autoreactivity. *J Clin Invest* **97**, 2027-2037, doi:10.1172/JCI118638 (1996).

61 Martens, P. B., Goronzy, J. J., Schaid, D. & Weyand, C. M. Expansion of unusual CD4+ T cells in severe rheumatoid arthritis. *Arthritis Rheum* **40**, 1106-1114, doi:10.1002/art.1780400615 (1997).

62 Berner, B., Wolf, G., Hummel, K. M., Muller, G. A. & Reuss-Borst, M. A. Increased expression of CD40 ligand (CD154) on CD4+ T cells as a marker of disease activity in rheumatoid arthritis. *Ann Rheum Dis* **59**, 190-195, doi:10.1136/ard.59.3.190 (2000).

63 Bendall, S. C., Nolan, G. P., Roederer, M. & Chattopadhyay, P. K. A deep profiler's guide to cytometry. *Trends Immunol* **33**, 323-332, doi:10.1016/j.it.2012.02.010 (2012).

64 Steiner, G. *et al.* Cytokine production by synovial T cells in rheumatoid arthritis. *Rheumatology (Oxford)* **38**, 202-213, doi:10.1093/rheumatology/38.3.202 (1999).

65 Brennan, F. M. *et al.* Evidence that rheumatoid arthritis synovial T cells are similar to cytokine-activated T cells: involvement of phosphatidylinositol 3-kinase and nuclear factor kappaB pathways in tumor necrosis factor alpha production in rheumatoid arthritis. *Arthritis Rheum* **46**, 31-41, doi:10.1002/1529-0131(200201)46:1<31::AID-ART10029>3.0.CO;2-5 (2002).

66 Kohem, C. L. *et al.* Enrichment of differentiated CD45RBdim,CD27- memory T cells in the peripheral blood, synovial fluid, and synovial tissue of patients with rheumatoid arthritis. *Arthritis Rheum* **39**, 844-854, doi:10.1002/art.1780390518 (1996).

67 Isomaki, P., Luukkainen, R., Lassila, O., Toivanen, P. & Punnonen, J. Synovial fluid T cells from patients with rheumatoid arthritis are refractory to the T helper type 2 differentiation-inducing effects of interleukin-4. *Immunology* **96**, 358-364, doi:10.1046/j.1365-2567.1999.00712.x (1999).

68 Qin, S. *et al.* The chemokine receptors CXCR3 and CCR5 mark subsets of T cells associated with certain inflammatory reactions. *J Clin Invest* **101**, 746-754, doi:10.1172/JCI1422 (1998).

69 Niu, Q., Cai, B., Huang, Z. C., Shi, Y. Y. & Wang, L. L. Disturbed Th17/Treg balance in patients with rheumatoid arthritis. *Rheumatol Int* **32**, 2731-2736, doi:10.1007/s00296-011-1984-x (2012).

70 Pawlik, A. *et al.* The expansion of CD4+CD28- T cells in patients with rheumatoid arthritis. *Arthritis Res Ther* **5**, R210-213, doi:10.1186/ar766 (2003).

71 Walter, G. J. *et al.* Phenotypic, Functional, and Gene Expression Profiling of Peripheral CD45RA+ and CD45RO+ CD4+CD25+CD127(low) Treg Cells in Patients With Chronic Rheumatoid Arthritis. *Arthritis Rheumatol* **68**, 103-116, doi:10.1002/art.39408 (2016).

72 Matsuki, F. *et al.* CD45RA-Foxp3(high) activated/effector regulatory T cells in the CCR7 + CD45RA-CD27 + CD28+central memory subset are decreased in peripheral blood from patients with rheumatoid arthritis. *Biochem Biophys Res Commun* **438**, 778-783, doi:10.1016/j.bbrc.2013.05.120 (2013).

73 Moradi, B. *et al.* CD4(+)CD25(+)/highCD127low/(-) regulatory T cells are enriched in rheumatoid arthritis and osteoarthritis joints--analysis of frequency and phenotype in synovial membrane, synovial fluid and peripheral blood. *Arthritis Res Ther* **16**, R97, doi:10.1186/ar4545 (2014).

74 Han, G. M., O'Neil-Andersen, N. J., Zurier, R. B. & Lawrence, D. A. CD4+CD25high T cell numbers are enriched in the peripheral blood of patients with rheumatoid arthritis. *Cell Immunol* **253**, 92-101, doi:10.1016/j.cellimm.2008.05.007 (2008).

75 Lawson, C. A. *et al.* Early rheumatoid arthritis is associated with a deficit in the CD4+CD25high regulatory T cell population in peripheral blood. *Rheumatology (Oxford)* **45**, 1210-1217, doi:10.1093/rheumatology/kel089 (2006).

76 Ehrenstein, M. R. *et al.* Compromised function of regulatory T cells in rheumatoid arthritis and reversal by anti-TNFalpha therapy. *J Exp Med* **200**, 277-285, doi:10.1084/jem.20040165 (2004).

77 Mottonen, M. *et al.* CD4+ CD25+ T cells with the phenotypic and functional characteristics of regulatory T cells are enriched in the synovial fluid of patients with rheumatoid arthritis. *Clin Exp Immunol* **140**, 360-367, doi:10.1111/j.1365-2249.2005.02754.x (2005).

78 Flores-Borja, F., Jury, E. C., Mauri, C. & Ehrenstein, M. R. Defects in CTLA-4 are associated with abnormal regulatory T cell function in rheumatoid arthritis. *Proc Natl Acad Sci U S A* **105**, 19396-19401, doi:10.1073/pnas.0806855105 (2008).

79 Cao, D. *et al.* Isolation and functional characterization of regulatory CD25brightCD4+ T cells from the target organ of patients with rheumatoid arthritis. *Eur J Immunol* **33**, 215-223, doi:10.1002/immu.200390024 (2003).

80 Spitzer, M. H. & Nolan, G. P. Mass Cytometry: Single Cells, Many Features. *Cell* **165**, 780-791, doi:10.1016/j.cell.2016.04.019 (2016).

81 Bendall, S. C. *et al.* Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**, 687-696, doi:10.1126/science.1198704 (2011).

82 Ornatsky, O. *et al.* Highly multiparametric analysis by mass cytometry. *J Immunol Methods* **361**, 1-20, doi:10.1016/j.jim.2010.07.002 (2010).

83 Bandura, D. R. *et al.* Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal Chem* **81**, 6813-6822, doi:10.1021/ac901049w (2009).

84 Wong, M. T. *et al.* A High-Dimensional Atlas of Human T Cell Diversity Reveals Tissue-Specific Trafficking and Cytokine Signatures. *Immunity* **45**, 442-456, doi:10.1016/j.immuni.2016.07.007 (2016).

85 Nakayamada, S., Takahashi, H., Kanno, Y. & O'Shea, J. J. Helper T cell diversity and plasticity. *Curr Opin Immunol* **24**, 297-302, doi:10.1016/j.coi.2012.01.014 (2012).

86 Reiner, S. L. & Adams, W. C. Lymphocyte fate specification as a deterministic but highly plastic process. *Nat Rev Immunol* **14**, 699-704, doi:10.1038/nri3734 (2014).

87 Komatsu, N. *et al.* Pathogenic conversion of Foxp3+ T cells into TH17 cells in autoimmune arthritis. *Nat Med* **20**, 62-68, doi:10.1038/nm.3432 (2014).

88 Gagliani, N. *et al.* Th17 cells transdifferentiate into regulatory T cells during resolution of inflammation. *Nature* **523**, 221-225, doi:10.1038/nature14452 (2015).

89 Wong, M. T. *et al.* Mapping the Diversity of Follicular Helper T Cells in Human Blood and Tonsils Using High-Dimensional Mass Cytometry Analysis. *Cell Rep* **11**, 1822-1833, doi:10.1016/j.celrep.2015.05.022 (2015).

90 Mason, G. M. *et al.* Phenotypic Complexity of the Human Regulatory T Cell Compartment Revealed by Mass Cytometry. *J Immunol* **195**, 2030-2037, doi:10.4049/jimmunol.1500703 (2015).

91 Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214, doi:10.1016/j.cell.2015.05.002 (2015).

92 Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* **10**, 1096-1098, doi:10.1038/nmeth.2639 (2013).

93 Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep* **2**, 666-673, doi:10.1016/j.celrep.2012.08.003 (2012).

94 Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res* **21**, 1160-1167, doi:10.1101/gr.110882.110 (2011).

95 Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**, 377-382, doi:10.1038/nmeth.1315 (2009).

96 Ishizuka, I. E. *et al.* Single-cell analysis defines the divergence between the innate lymphoid cell lineage and lymphoid tissue-inducer cell lineage. *Nat Immunol* **17**, 269-276, doi:10.1038/ni.3344 (2016).

97 Olsson, A. *et al.* Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature* **537**, 698-702, doi:10.1038/nature19348 (2016).

98 Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236-240, doi:10.1038/nature12172 (2013).

99 Mahata, B. *et al.* Single-cell RNA sequencing reveals T helper cells synthesizing steroids de novo to contribute to immune homeostasis. *Cell Rep* **7**, 1130-1142, doi:10.1016/j.celrep.2014.04.011 (2014).

100 Bjorklund, A. K. *et al.* The heterogeneity of human CD127(+) innate lymphoid cells revealed by single-cell RNA sequencing. *Nat Immunol* **17**, 451-460, doi:10.1038/ni.3368 (2016).

101 Zheng, C. *et al.* Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing. *Cell* **169**, 1342-1356 e1316, doi:10.1016/j.cell.2017.05.035 (2017).

102 Engel, I. *et al.* Innate-like functions of natural killer T cell subsets result from highly divergent gene programs. *Nat Immunol* **17**, 728-739, doi:10.1038/ni.3437 (2016).

103 Kakaradov, B. *et al.* Early transcriptional and epigenetic regulation of CD8(+) T cell differentiation revealed by single-cell RNA sequencing. *Nat Immunol* **18**, 422-432, doi:10.1038/ni.3688 (2017).

104 Proserpio, V. *et al.* Single-cell analysis of CD4+ T-cell differentiation reveals three major cell states and progressive acceleration of proliferation. *Genome Biol* **17**, 103, doi:10.1186/s13059-016-0957-5 (2016).

105 Rao, D. A. *et al.* Pathologically expanded peripheral T helper cell subset drives B cells in rheumatoid arthritis. *Nature* **542**, 110-114, doi:10.1038/nature20810 (2017).

106 Stephenson, W. *et al.* Single-cell RNA-seq of rheumatoid arthritis synovial tissue using low-cost microfluidic instrumentation. *Nat Commun* **9**, 791, doi:10.1038/s41467-017-02659-x (2018).

107 Ishigaki, K. *et al.* Quantitative and qualitative characterization of expanded CD4+ T cell clones in rheumatoid arthritis patients. *Sci Rep* **5**, 12937, doi:10.1038/srep12937 (2015).

108 Mellado, M. *et al.* T Cell Migration in Rheumatoid Arthritis. *Front Immunol* **6**, 384, doi:10.3389/fimmu.2015.00384 (2015).

109 Gaublomme, J. T. *et al.* Single-Cell Genomics Unveils Critical Regulators of Th17 Cell Pathogenicity. *Cell* **163**, 1400-1412, doi:10.1016/j.cell.2015.11.009 (2015).

110 Fessler, J. *et al.* Novel Senescent Regulatory T-Cell Subset with Impaired Suppressive Function in Rheumatoid Arthritis. *Front Immunol* **8**, 300, doi:10.3389/fimmu.2017.00300 (2017).

111 Cui, J. *et al.* Genome-wide association study and gene expression analysis identifies CD84 as a predictor of response to etanercept therapy in rheumatoid arthritis. *PLoS Genet* **9**, e1003394, doi:10.1371/journal.pgen.1003394 (2013).

112 Cui, J. *et al.* Rheumatoid arthritis risk allele PTPRC is also associated with response to anti-tumor necrosis factor alpha therapy. *Arthritis Rheum* **62**, 1849-1861, doi:10.1002/art.27457 (2010).

113 Ponchel, F. *et al.* An immunological biomarker to predict MTX response in early RA. *Ann Rheum Dis* **73**, 2047-2053, doi:10.1136/annrheumdis-2013-203566 (2014).

114 Daien, C. I. *et al.* High levels of natural killer cells are associated with response to tocilizumab in patients with severe rheumatoid arthritis. *Rheumatology (Oxford)* **54**, 601-608, doi:10.1093/rheumatology/keu363 (2015).

115 Kikuchi, J. *et al.* Peripheral blood CD4(+)CD25(+)CD127(low) regulatory T cells are significantly increased by tocilizumab treatment in patients with rheumatoid arthritis: increase in regulatory T cells correlates with clinical response. *Arthritis Res Ther* **17**, 10, doi:10.1186/s13075-015-0526-4 (2015).

116 Nakachi, S. *et al.* Interleukin-10-producing LAG3(+) regulatory T cells are associated with disease activity and abatacept treatment in rheumatoid arthritis. *Arthritis Res Ther* **19**, 97, doi:10.1186/s13075-017-1309-x (2017).

117 Bystrom, J. *et al.* Response to Treatment with TNFalpha Inhibitors in Rheumatoid Arthritis Is Associated with High Levels of GM-CSF and GM-CSF(+) T Lymphocytes. *Clin Rev Allergy Immunol* **53**, 265-276, doi:10.1007/s12016-017-8610-y (2017).

118 Citro, A. *et al.* CD8+ T Cells Specific to Apoptosis-Associated Antigens Predict the Response to Tumor Necrosis Factor Inhibitor Therapy in Rheumatoid Arthritis. *PLoS One* **10**, e0128607, doi:10.1371/journal.pone.0128607 (2015).

119 Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* **11**, 733-739, doi:10.1038/nrg2825 (2010).

120 Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-127, doi:10.1093/biostatistics/kxj037 (2007).

121 Chen, C. *et al.* Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS One* **6**, e17238, doi:10.1371/journal.pone.0017238 (2011).

122 Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* **3**, 1724-1735, doi:10.1371/journal.pgen.0030161 (2007).

123 Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* **19**, 562-578, doi:10.1093/biostatistics/kxx053 (2018).

124 Tung, P. Y. *et al.* Batch effects and the effective design of single-cell gene expression studies. *Sci Rep* **7**, 39921, doi:10.1038/srep39921 (2017).

125 Zhang, F. *et al.* Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. *Nat Immunol* **20**, 928-942, doi:10.1038/s41590-019-0378-1 (2019).

126 Fonseka, C. Y. *et al.* Mixed-effects association of single cells identifies an expanded effector CD4(+) T cell subset in rheumatoid arthritis. *Sci Transl Med* **10**, doi:10.1126/scitranslmed.aaq0305 (2018).

127 Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495-502, doi:10.1038/nbt.3192 (2015).

128 Hu, X. *et al.* Application of user-guided automated cytometric data analysis to large-scale immunoprofiling of invariant natural killer T cells. *Proc Natl Acad Sci U S A* **110**, 19030-19035, doi:10.1073/pnas.1318322110 (2013).

129 Aghaeepour, N. *et al.* Critical assessment of automated flow cytometry data analysis techniques. *Nat Methods* **10**, 228-238, doi:10.1038/nmeth.2365 (2013).

130 Van Gassen, S. *et al.* FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry A* **87**, 636-645, doi:10.1002/cyto.a.22625 (2015).

131 Shekhar, K., Brodin, P., Davis, M. M. & Chakraborty, A. K. Automatic Classification of Cellular Expression by Nonlinear Stochastic Embedding (ACCENSE). *Proc Natl Acad Sci U S A* **111**, 202-207, doi:10.1073/pnas.1321405111 (2014).

132 Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162**, 184-197, doi:10.1016/j.cell.2015.05.047 (2015).

133 Bendall, S. C. *et al.* Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714-725, doi:10.1016/j.cell.2014.04.005 (2014).

134 Becher, B. *et al.* High-dimensional analysis of the murine myeloid cell system. *Nat Immunol* **15**, 1181-1189, doi:10.1038/ni.3006 (2014).

135 Weber, L. M. & Robinson, M. D. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry A* **89**, 1084-1096, doi:10.1002/cyto.a.23030 (2016).

136 Lun, A. T. L., Richard, A. C. & Marioni, J. C. Testing for differential abundance in mass cytometry data. *Nat Methods* **14**, 707-709, doi:10.1038/nmeth.4295 (2017).

137 Bruggner, R. V., Bodenmiller, B., Dill, D. L., Tibshirani, R. J. & Nolan, G. P. Automated identification of stratifying signatures in cellular subpopulations. *Proc Natl Acad Sci U S A* **111**, E2770-2777, doi:10.1073/pnas.1408792111 (2014).

138 Lefebvre, J. S. & Haynes, L. Aging of the CD4 T Cell Compartment. *Open Longev Sci* **6**, 83-91, doi:10.2174/1876326X01206010083 (2012).

139 Kovaiou, R. D. *et al.* Age-related differences in phenotype and function of CD4+ T cells are due to a phenotypic shift from naive to memory effector CD4+ T cells. *Int Immunol* **17**, 1359-1366, doi:10.1093/intimm/dxh314 (2005).

140 Jackola, D. R., Ruger, J. K. & Miller, R. A. Age-associated changes in human T cell phenotype and function. *Aging (Milano)* **6**, 25-34 (1994).

141 Hong, M. S., Dan, J. M., Choi, J. Y. & Kang, I. Age-associated changes in the frequency of naive, memory and effector CD8+ T cells. *Mech Ageing Dev* **125**, 615-618, doi:10.1016/j.mad.2004.07.001 (2004).

142 Maaten, L. V. D. Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* **15**, 3221-3245 (2014).

143 Arazi, A. *et al.* The immune cell landscape in kidneys of patients with lupus nephritis. *Nat Immunol* **20**, 902-914, doi:10.1038/s41590-019-0398-x (2019).

144 Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* **14**, 865-868, doi:10.1038/nmeth.4380 (2017).

145 Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol* **36**, 89-94, doi:10.1038/nbt.4042 (2018).

146 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821, doi:10.1016/j.cell.2019.05.031 (2019).

147 Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420, doi:10.1038/nbt.4096 (2018).

148 Korsunsky I, M. N., Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner MB, Loh P-R, Raychaudhuri S. Fast, sensitive, and accurate integration of single cell data with Harmony. *Nature Methods* (In Press).

149 Vu, T. N. *et al.* Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics* **32**, 2128-2135, doi:10.1093/bioinformatics/btw202 (2016).

150 Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J. P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat Commun* **9**, 284, doi:10.1038/s41467-017-02554-5 (2018).

151 Chen, W. *et al.* UMI-count modeling and differential expression analysis for single-cell RNA sequencing. *Genome Biol* **19**, 70, doi:10.1186/s13059-018-1438-9 (2018).

152 Chen, H. I., Jin, Y., Huang, Y. & Chen, Y. Detection of high variability in gene expression from single-cell RNA-seq profiling. *BMC Genomics* **17 Suppl 7**, 508, doi:10.1186/s12864-016-2897-6 (2016).

153 Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol* **18**, 174, doi:10.1186/s13059-017-1305-0 (2017).

154 Vieth, B., Ziegenhain, C., Parekh, S., Enard, W. & Hellmann, I. powsimR: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics* **33**, 3486-3488, doi:10.1093/bioinformatics/btx435 (2017).

155 Li, W. V. & Li, J. J. A statistical simulator scDesign for rational scRNA-seq experimental design. *Bioinformatics* **35**, i41-i50, doi:10.1093/bioinformatics/btz321 (2019).

156 Grun, D. & van Oudenaarden, A. Design and Analysis of Single-Cell Sequencing Experiments. *Cell* **163**, 799-810, doi:10.1016/j.cell.2015.10.039 (2015).

157 Abrams, D., Kumar, P., Karuturi, R. K. M. & George, J. A computational method to aid the design and analysis of single cell RNA-seq experiments for cell type identification. *BMC Bioinformatics* **20**, 275, doi:10.1186/s12859-019-2817-2 (2019).

158 Neu, K. E., Tang, Q., Wilson, P. C. & Khan, A. A. Single-Cell Genomics: Approaches and Utility in Immunology. *Trends Immunol* **38**, 140-149, doi:10.1016/j.it.2016.12.001 (2017).

159 Warrington, K. J., Takemura, S., Goronzy, J. J. & Weyand, C. M. CD4+,CD28- T cells in rheumatoid arthritis patients combine features of the innate and adaptive immune systems. *Arthritis Rheum* **44**, 13-20, doi:10.1002/1529-0131(200101)44:1<13::AID-ANR3>3.0.CO;2-6 (2001).

160 Scarsi, M., Ziglioli, T. & Airo, P. Decreased circulating CD28-negative T cells in patients with rheumatoid arthritis treated with abatacept are correlated with clinical response. *J Rheumatol* **37**, 911-916, doi:10.3899/jrheum.091176 (2010).

161 Sorensen, T., Baumgart, S., Durek, P., Grutzkau, A. & Haupl, T. immunoClust--An automated analysis pipeline for the identification of immunophenotypic signatures in high-dimensional cytometric datasets. *Cytometry A* **87**, 603-615, doi:10.1002/cyto.a.22626 (2015).

162 Samusik, N., Good, Z., Spitzer, M. H., Davis, K. L. & Nolan, G. P. Automated mapping of phenotype space with single-cell data. *Nat Methods* **13**, 493-496, doi:10.1038/nmeth.3863 (2016).

163 Lin, P., Troup, M. & Ho, J. W. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol* **18**, 59, doi:10.1186/s13059-017-1188-0 (2017).

164 Ermann, J., Rao, D. A., Teslovich, N. C., Brenner, M. B. & Raychaudhuri, S. Immune cell profiling to guide therapeutic decisions in rheumatic diseases. *Nat Rev Rheumatol* **11**, 541-551, doi:10.1038/nrrheum.2015.71 (2015).

165 Cheng, Y., Wong, M. T., van der Maaten, L. & Newell, E. W. Categorical Analysis of Human T Cell Heterogeneity with One-Dimensional Soli-Expression by Nonlinear Stochastic Embedding. *J Immunol* **196**, 924-932, doi:10.4049/jimmunol.1501928 (2016).

166 Oshima, S. & Eckels, D. D. Selective expression of class II MHC isotypes by MLC-activated human T lymphocytes. *Hum Immunol* **27**, 208-219, doi:10.1016/0198-8859(90)90051-p (1990).

167 Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525-527, doi:10.1038/nbt.3519 (2016).

168 Strauss-Albee, D. M. *et al.* Human NK cell repertoire diversity reflects immune experience and correlates with viral susceptibility. *Sci Transl Med* **7**, 297ra115, doi:10.1126/scitranslmed.aac5722 (2015).

169 Mingueneau, M. *et al.* Cytometry by time-of-flight immunophenotyping identifies a blood Sjogren's signature correlating with disease activity and glandular inflammation. *J Allergy Clin Immunol* **137**, 1809-1821 e1812, doi:10.1016/j.jaci.2016.01.024 (2016).

170 Hansmann, L. *et al.* Mass cytometry analysis shows that a novel memory phenotype B cell is expanded in multiple myeloma. *Cancer Immunol Res* **3**, 650-660, doi:10.1158/2326-6066.CIR-14-0236-T (2015).

171 Gaudilliere, B. *et al.* Clinical recovery from surgery correlates with single-cell immune signatures. *Sci Transl Med* **6**, 255ra131, doi:10.1126/scitranslmed.3009701 (2014).

172 Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat Biotechnol* **34**, 1145-1160, doi:10.1038/nbt.3711 (2016).

173 Chester, C. & Maecker, H. T. Algorithmic Tools for Mining High-Dimensional Cytometry Data. *J Immunol* **195**, 773-779, doi:10.4049/jimmunol.1500633 (2015).

174 Saeys, Y., Van Gassen, S. & Lambrecht, B. N. Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nat Rev Immunol* **16**, 449-462, doi:10.1038/nri.2016.56 (2016).

175 Saeys, Y., Van Gassen, S. & Lambrecht, B. Response to Orlova et al. "Science not art: statistically sound methods for identifying subsets in multi-dimensional flow and mass cytometry data sets". *Nat Rev Immunol* **18**, 78, doi:10.1038/nri.2017.151 (2017).

176 Orlova, D. Y., Herzenberg, L. A. & Walther, G. Science not art: statistically sound methods for identifying subsets in multi-dimensional flow and mass cytometry data sets. *Nat Rev Immunol* **18**, 77, doi:10.1038/nri.2017.150 (2017).

177 Qiu, X. *et al.* Single-cell mRNA quantification and differential analysis with Census. *Nat Methods* **14**, 309-315, doi:10.1038/nmeth.4150 (2017).

178 Schiott, A., Lindstedt, M., Johansson-Lindbom, B., Roggen, E. & Borrebaeck, C. A. CD27- CD4+ memory T cells define a differentiated memory population at both the functional and transcriptional levels. *Immunology* **113**, 363-370, doi:10.1111/j.1365-2567.2004.01974.x (2004).

179 Larbi, A. & Fulop, T. From "truly naive" to "exhausted senescent" T cells: when markers predict functionality. *Cytometry A* **85**, 25-35, doi:10.1002/cyto.a.22351 (2014).

180 De Jong, R. *et al.* The CD27- subset of peripheral blood memory CD4+ lymphocytes contains functionally differentiated T lymphocytes that develop by persistent antigenic stimulation in vivo. *Eur J Immunol* **22**, 993-999, doi:10.1002/eji.1830220418 (1992).

181 Appay, V. *et al.* Characterization of CD4(+) CTLs ex vivo. *J Immunol* **168**, 5954-5958, doi:10.4049/jimmunol.168.11.5954 (2002).

182 Namekawa, T., Wagner, U. G., Goronzy, J. J. & Weyand, C. M. Functional subsets of CD4 T cells in rheumatoid synovitis. *Arthritis Rheum* **41**, 2108-2116, doi:10.1002/1529-0131(199812)41:12<2108::AID-ART5>3.0.CO;2-Q (1998).

183 Griffiths, G. M., Alpert, S., Lambert, E., McGuire, J. & Weissman, I. L. Perforin and granzyme A expression identifying cytolytic lymphocytes in rheumatoid arthritis. *Proc Natl Acad Sci U S A* **89**, 549-553, doi:10.1073/pnas.89.2.549 (1992).

184 Mattoo, H. *et al.* Clonal expansion of CD4(+) cytotoxic T lymphocytes in patients with IgG4-related disease. *J Allergy Clin Immunol* **138**, 825-838, doi:10.1016/j.jaci.2015.12.1330 (2016).

185 Patil, V. S. *et al.* Precursors of human CD4(+) cytotoxic T lymphocytes identified by single-cell transcriptome analysis. *Sci Immunol* **3**, doi:10.1126/sciimmunol.aan8664 (2018).

186 Peterson, V. M. *et al.* Multiplexed quantification of proteins and transcripts in single cells. *Nat Biotechnol* **35**, 936-939, doi:10.1038/nbt.3973 (2017).

187 Lee, Y. C. *et al.* Association Between Pain Sensitization and Disease Activity in Patients With Rheumatoid Arthritis: A Cross-Sectional Study. *Arthritis Care Res (Hoboken)* **70**, 197-204, doi:10.1002/acr.23266 (2018).

188 Finck, R. *et al.* Normalization of mass cytometry data with bead standards. *Cytometry A* **83**, 483-494, doi:10.1002/cyto.a.22271 (2013).

189 B. Ellis, P. H., F. Hahne, N. Le Meur, N. Gopalakrishnan, J. Spidlen, M. Jiang. flowCore: Basic structures for flow cytometry data. *R package version 1.46.1* (2018).

190 Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *2015* **67**, 48, doi:10.18637/jss.v067.i01 (2015).

191 Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag, 2016).

192 Kolde, R. pheatmap: Pretty Heatmaps. *R package version 1.0.10* (2018).

193 Chen, H. *et al.* Cytofkit: A Bioconductor Package for an Integrated Mass Cytometry Data Analysis Pipeline. *PLoS Comput Biol* **12**, e1005112, doi:10.1371/journal.pcbi.1005112 (2016).

194 Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D. & Woolf, P. J. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* **10**, 161, doi:10.1186/1471-2105-10-161 (2009).

195 J. Fan, P. K. liger: Lightweight Iterative Geneset Enrichment. *R package version 0.1* (2018).

196 E. LeDell, N. G., S. Aiello, A. Fu, A. Candel, C. Click, T. Kraljevic, T. Nykodym, P. Aboyoun, M. Kurka, M. Malohlava. h2o: R Interface for 'H2O'. *R package version 3.20.0.2* (2018).

197 Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J* **8**, 289-317 (2016).

198 Orr, C. *et al.* Synovial tissue research: a state-of-the-art review. *Nat Rev Rheumatol* **13**, 463-475, doi:10.1038/nrrheum.2017.115 (2017).

199 Gibofsky, A. Epidemiology, pathophysiology, and diagnosis of rheumatoid arthritis: A Synopsis. *Am J Manag Care* **20**, S128-135 (2014).

200 Wolfe, F. *et al.* The mortality of rheumatoid arthritis. *Arthritis Rheum* **37**, 481-494, doi:10.1002/art.1780370408 (1994).

201 Reparon-Schuijt, C. C. *et al.* Secretion of anti-citrulline-containing peptide antibody by B lymphocytes in rheumatoid arthritis. *Arthritis Rheum* **44**, 41-47, doi:10.1002/1529-0131(200101)44:1<41::AID-ANR6>3.0.CO;2-0 (2001).

202 Mulherin, D., Fitzgerald, O. & Bresnihan, B. Synovial tissue macrophage populations and articular damage in rheumatoid arthritis. *Arthritis Rheum* **39**, 115-124, doi:10.1002/art.1780390116 (1996).

203 Kinne, R. W., Brauer, R., Stuhlmuller, B., Palombo-Kinne, E. & Burmester, G. R. Macrophages in rheumatoid arthritis. *Arthritis Res* **2**, 189-202, doi:10.1186/ar86 (2000).

204 Pap, T., Muller-Ladner, U., Gay, R. E. & Gay, S. Fibroblast biology. Role of synovial fibroblasts in the pathogenesis of rheumatoid arthritis. *Arthritis Res* **2**, 361-367, doi:10.1186/ar113 (2000).

205 Noss, E. H. & Brenner, M. B. The role and therapeutic implications of fibroblast-like synoviocytes in inflammation and cartilage erosion in rheumatoid arthritis. *Immunol Rev* **223**, 252-270, doi:10.1111/j.1600-065X.2008.00648.x (2008).

206 Muller-Ladner, U. *et al.* Synovial fibroblasts of patients with rheumatoid arthritis attach to and invade normal human cartilage when engrafted into SCID mice. *Am J Pathol* **149**, 1607-1615 (1996).

207 Dennis, G., Jr. *et al.* Synovial phenotypes in rheumatoid arthritis correlate with response to biologic therapeutics. *Arthritis Res Ther* **16**, R90, doi:10.1186/ar4555 (2014).

208 Orange, D. E. *et al.* Identification of Three Rheumatoid Arthritis Disease Subtypes by Machine Learning Integration of Synovial Histologic Features and RNA Sequencing Data. *Arthritis Rheumatol* **70**, 690-701, doi:10.1002/art.40428 (2018).

209 Lindberg, J. *et al.* Variability in synovial inflammation in rheumatoid arthritis investigated by microarray technology. *Arthritis Res Ther* **8**, R47, doi:10.1186/ar1903 (2006).

210 Schelker, M. *et al.* Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nat Commun* **8**, 2032, doi:10.1038/s41467-017-02289-3 (2017).

211 Papalexi, E. & Satija, R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat Rev Immunol* **18**, 35-45, doi:10.1038/nri.2017.76 (2018).

212 Villani, A. C. *et al.* Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, doi:10.1126/science.aah4573 (2017).

213 Mizoguchi, F. *et al.* Functionally distinct disease-associated fibroblast subsets in rheumatoid arthritis. *Nat Commun* **9**, 789, doi:10.1038/s41467-018-02892-y (2018).

214 Donlin, L. T. *et al.* Methods for high-dimensonal analysis of cells dissociated from cyropreserved synovial tissue. *Arthritis Res Ther* **20**, 139, doi:10.1186/s13075-018-1631-y (2018).

215 De Maesschalck, R., Jouan-Rimbaud, D. & Massart, D. L. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems* **50**, 1-18, doi:10.1016/s0169-7439(99)00047-7 (2000).

216 Krenn, V. *et al.* Grading of chronic synovitis--a histopathological grading system for molecular and diagnostic pathology. *Pathol Res Pract* **198**, 317-325, doi:10.1078/0344-0338-5710261 (2002).

217 Todd, D. J. *et al.* XBP1 governs late events in plasma cell differentiation and is not required for antigen-specific memory B cell development. *J Exp Med* **206**, 2151-2159, doi:10.1084/jem.20090738 (2009).

218 Pillai, S. Now you know your ABCs. *Blood* **118**, 1187-1188, doi:10.1182/blood-2011-06-355131 (2011).

219 Rubtsov, A. V. *et al.* CD11c-Expressing B Cells Are Located at the T Cell/B Cell Border in Spleen and Are Potent APCs. *J Immunol* **195**, 71-79, doi:10.4049/jimmunol.1500055 (2015).

220 Ellebedy, A. H. *et al.* Defining antigen-specific plasmablast and memory B cell subsets in human blood after viral infection or vaccination. *Nat Immunol* **17**, 1226-1234, doi:10.1038/ni.3533 (2016).

221 Wang, S. *et al.* IL-21 drives expansion and plasma cell differentiation of autoreactive CD11c(hi)T-bet(+) B cells in SLE. *Nat Commun* **9**, 1758, doi:10.1038/s41467-018-03750-7 (2018).

222 Pitzalis, C., Kelly, S. & Humby, F. New learnings on the pathophysiology of RA from synovial biopsies. *Curr Opin Rheumatol* **25**, 334-344, doi:10.1097/BOR.0b013e32835fd8eb (2013).

223 Filer, A. The fibroblast as a therapeutic target in rheumatoid arthritis. *Curr Opin Pharmacol* **13**, 413-419, doi:10.1016/j.coph.2013.02.006 (2013).

224 Westra, H. J. *et al.* Fine-mapping and functional studies highlight potential causal variants for rheumatoid arthritis and type 1 diabetes. *Nat Genet* **50**, 1366-1374, doi:10.1038/s41588-018-0216-7 (2018).

225 Al-Mayouf, S. M. *et al.* Loss-of-function variant in DNASE1L3 causes a familial form of systemic lupus erythematosus. *Nat Genet* **43**, 1186-1188, doi:10.1038/ng.975 (2011).

226 Snelling, S. J. *et al.* Dickkopf-3 is upregulated in osteoarthritis and has a chondroprotective role. *Osteoarthritis Cartilage* **24**, 883-891, doi:10.1016/j.joca.2015.11.021 (2016).

227 Lee, E. B. *et al.* Tofacitinib versus methotrexate in rheumatoid arthritis. *N Engl J Med* **370**, 2377-2386, doi:10.1056/NEJMoa1310476 (2014).

228 Zizzo, G., Hilliard, B. A., Monestier, M. & Cohen, P. L. Efficient clearance of early apoptotic cells by human macrophages requires M2c polarization and MerTK induction. *J Immunol* **189**, 3508-3520, doi:10.4049/jimmunol.1200662 (2012).

229 Frara, N. *et al.* Transgenic Expression of Osteoactivin/gpnmb Enhances Bone Formation In Vivo and Osteoprogenitor Differentiation Ex Vivo. *J Cell Physiol* **231**, 72-83, doi:10.1002/jcp.25020 (2016).

230 Jenks, S. A. *et al.* Distinct Effector B Cells Induced by Unregulated Toll-like Receptor 7 Contribute to Pathogenic Responses in Systemic Lupus Erythematosus. *Immunity* **49**, 725-739 e726, doi:10.1016/j.immuni.2018.08.015 (2018).

231 Smolen, J. S. Pharmacotherapy: How well can we compare different biologic agents for RA? *Nat Rev Rheumatol* **6**, 247-248, doi:10.1038/nrrheum.2010.58 (2010).

232 McInnes, I. B. & Liew, F. Y. Cytokine networks--towards new therapies for rheumatoid arthritis. *Nat Clin Pract Rheumatol* **1**, 31-39, doi:10.1038/ncprheum0020 (2005).

233 McInnes, I. B. *et al.* The role of interleukin-15 in T-cell migration and activation in rheumatoid arthritis. *Nat Med* **2**, 175-182, doi:10.1038/nm0296-175 (1996).

234 Witten, D. M., Tibshirani, R. & Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 515-534, doi:10.1093/biostatistics/kxp008 (2009).

235 Parkhomenko, E., Tritchler, D. & Beyene, J. Sparse canonical correlation analysis with application to genomic data integration. *Stat Appl Genet Mol Biol* **8**, Article 1, doi:10.2202/1544-6115.1406 (2009).

236 Hashimshony, T. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol* **17**, 77, doi:10.1186/s13059-016-0938-8 (2016).

237 Bjornson, Z. B., Nolan, G. P. & Fantl, W. J. Single-cell mass cytometry for analysis of immune system functional states. *Curr Opin Immunol* **25**, 484-494, doi:10.1016/j.coi.2013.07.004 (2013).

238 Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* **9**, 171-181, doi:10.1038/nprot.2014.006 (2014).

239 Gonzalez, I., Déjean, S., Martin, P. & Baccini, A. CCA: An R Package to Extend Canonical Correlation Analysis. *Journal of Statistical Software* **23**, doi:10.18637/jss.v023.i12 (2008).

240 Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **21**, 3940-3941, doi:10.1093/bioinformatics/bti623 (2005).

241 Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417-425, doi:10.1016/j.cels.2015.12.004 (2015).

242 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).

243 Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53-65, doi:10.1016/0377-0427(87)90125-7 (1987).

244 Reynolds, A. P., Richards, G., de la Iglesia, B. & Rayward-Smith, V. J. Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms. *Journal of Mathematical Modelling and Algorithms* **5**, 475-504, doi:10.1007/s10852-005-9022-1 (2006).

245 Lebwohl, B., Sanders, D. S. & Green, P. H. R. Coeliac disease. *Lancet* **391**, 70-81, doi:10.1016/S0140-6736(17)31796-8 (2018).

246 Green, P. H., Lebwohl, B. & Greywoode, R. Celiac disease. *J Allergy Clin Immunol* **135**, 1099-1106; quiz 1107, doi:10.1016/j.jaci.2015.01.044 (2015).

247 Quarsten, H. *et al.* Staining of celiac disease-relevant T cells by peptide-DQ2 multimers. *J Immunol* **167**, 4861-4868, doi:10.4049/jimmunol.167.9.4861 (2001).

248 Bodd, M. *et al.* HLA-DQ2-restricted gluten-reactive T cells produce IL-21 but not IL-17 or IL-22. *Mucosal Immunol* **3**, 594-601, doi:10.1038/mi.2010.36 (2010).

249 Bodd, M. *et al.* Direct cloning and tetramer staining to measure the frequency of intestinal gluten-reactive T cells in celiac disease. *Eur J Immunol* **43**, 2605-2612, doi:10.1002/eji.201343382 (2013).

250 Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486-490, doi:10.1038/nature14590 (2015).

251 Blanco, P., Viallard, J. F., Pellegrin, J. L. & Moreau, J. F. Cytotoxic T lymphocytes and autoimmunity. *Curr Opin Rheumatol* **17**, 731-734, doi:10.1097/01.bor.0000179942.27777.f8 (2005).

252 Andersen, M. H., Schrama, D., Thor Straten, P. & Becker, J. C. Cytotoxic T cells. *J Invest Dermatol* **126**, 32-41, doi:10.1038/sj.jid.5700001 (2006).

253 de la Roche, M., Asano, Y. & Griffiths, G. M. Origins of the cytolytic synapse. *Nat Rev Immunol* **16**, 421-432, doi:10.1038/nri.2016.54 (2016).

254 Halle, S., Halle, O. & Forster, R. Mechanisms and Dynamics of T Cell-Mediated Cytotoxicity In Vivo. *Trends Immunol* **38**, 432-443, doi:10.1016/j.it.2017.04.002 (2017).

255 Barry, M. & Bleackley, R. C. Cytotoxic T lymphocytes: all roads lead to death. *Nat Rev Immunol* **2**, 401-409, doi:10.1038/nri819 (2002).

256 Wagner, H., Starzinski-Powitz, A., Jung, H. & Rollinghoff, M. Induction of I region-restricted hapten-specific cytotoxic T lymphocytes. *J Immunol* **119**, 1365-1368 (1977).

257 Wagner, H., Gotze, D., Ptschelinzew, L. & Rollinghoff, M. Induction of cytotoxic T lymphocytes against I-region-coded determinants: in vitro evidence for a third histocompatibility locus in the mouse. *J Exp Med* **142**, 1477-1487, doi:10.1084/jem.142.6.1477 (1975).

258 Maimone, M. M., Morrison, L. A., Braciale, V. L. & Braciale, T. J. Features of target cell lysis by class I and class II MHC-restricted cytolytic T lymphocytes. *J Immunol* **137**, 3639-3643 (1986).

259 Billings, P., Burakoff, S., Dorf, M. E. & Benacerraf, B. Cytotoxic T lymphocytes specific for I region determinants do not require interactions with H-2K or D gene products. *J Exp Med* **145**, 1387-1392, doi:10.1084/jem.145.5.1387 (1977).

260 Hintzen, R. Q. *et al.* Regulation of CD27 expression on subsets of mature T-lymphocytes. *J Immunol* **151**, 2426-2435 (1993).

261 van Leeuwen, E. M. *et al.* Emergence of a CD4+CD28- granzyme B+, cytomegalovirus-specific T cell subset after recovery of primary cytomegalovirus infection. *J Immunol* **173**, 1834-1841, doi:10.4049/jimmunol.173.3.1834 (2004).

262 Stuller, K. A. & Flano, E. CD4 T cells mediate killing during persistent gammaherpesvirus 68 infection. *J Virol* **83**, 4700-4703, doi:10.1128/JVI.02240-08 (2009).

263 Aslan, N. *et al.* Cytotoxic CD4 T cells in viral hepatitis. *J Viral Hepat* **13**, 505-514, doi:10.1111/j.1365-2893.2006.00723.x (2006).

264 van de Berg, P. J., van Leeuwen, E. M., ten Berge, I. J. & van Lier, R. Cytotoxic human CD4(+) T cells. *Curr Opin Immunol* **20**, 339-343, doi:10.1016/j.coi.2008.03.007 (2008).

265 Takeuchi, A. & Saito, T. CD4 CTL, a Cytotoxic Subset of CD4(+) T Cells, Their Differentiation and Function. *Front Immunol* **8**, 194, doi:10.3389/fimmu.2017.00194 (2017).

266 Cheroutre, H. & Husain, M. M. CD4 CTL: living up to the challenge. *Semin Immunol* **25**, 273-281, doi:10.1016/j.smim.2013.10.022 (2013).

267 Christophersen, A. *et al.* Distinct phenotype of CD4(+) T cells driving celiac disease identified in multiple autoimmune conditions. *Nat Med* **25**, 734-737, doi:10.1038/s41591-019-0403-9 (2019).

268 Schaier, M. *et al.* The extent of HLA-DR expression on HLA-DR(+) Tregs allows the identification of patients with clinically relevant borderline rejection. *Transpl Int* **26**, 290-299, doi:10.1111/tri.12032 (2013).

269 Kisielewicz, A. *et al.* A distinct subset of HLA-DR+-regulatory T cells is involved in the induction of preterm labor during pregnancy and in the induction of organ rejection after transplantation. *Clin Immunol* **137**, 209-220, doi:10.1016/j.clim.2010.07.008 (2010).

270 Baecher-Allan, C., Wolf, E. & Hafler, D. A. MHC class II expression identifies functionally distinct human regulatory T cells. *J Immunol* **176**, 4622-4631, doi:10.4049/jimmunol.176.8.4622 (2006).

271 Ashley, C. W. & Baecher-Allan, C. Cutting Edge: Responder T cells regulate human DR+ effector regulatory T cell activity via granzyme B. *J Immunol* **183**, 4843-4847, doi:10.4049/jimmunol.0900845 (2009).

272 Ahmed, A. *et al.* Circulating HLA-DR+CD4+ effector memory T cells resistant to CCR5 and PD-L1 mediated suppression compromise regulatory T cell function in tuberculosis. *PLoS Pathog* **14**, e1007289, doi:10.1371/journal.ppat.1007289 (2018).

273 Gutierrez-Arcelus, M. *et al.* Lymphocyte innateness defined by transcriptional states reflects a balance between proliferation and effector functions. *Nat Commun* **10**, 687, doi:10.1038/s41467-019-08604-4 (2019).

274 Satpathy, A. T. *et al.* Transcript-indexed ATAC-seq for precision immune profiling. *Nat Med* **24**, 580-590, doi:10.1038/s41591-018-0008-8 (2018).

275 Han, A., Glanville, J., Hansmann, L. & Davis, M. M. Linking T-cell receptor sequence to functional phenotype at the single-cell level. *Nat Biotechnol* **32**, 684-692, doi:10.1038/nbt.2938 (2014).

276 Genshaft, A. S. *et al.* Multiplexed, targeted profiling of single-cell proteomes and transcriptomes in a single reaction. *Genome Biol* **17**, 188, doi:10.1186/s13059-016-1045-6 (2016).

277 Frei, A. P. *et al.* Highly multiplexed simultaneous detection of RNAs and proteins in single cells. *Nat Methods* **13**, 269-275, doi:10.1038/nmeth.3742 (2016).

278 Stubbington, M. J. T. *et al.* T cell fate and clonality inference from single-cell transcriptomes. *Nat Methods* **13**, 329-332, doi:10.1038/nmeth.3800 (2016).

279 Singh, M. *et al.* High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. *Nat Commun* **10**, 3120, doi:10.1038/s41467-019-11049-4 (2019).

280 Kinslow, J. D. *et al.* Elevated IgA Plasmablast Levels in Subjects at Risk of Developing Rheumatoid Arthritis. *Arthritis Rheumatol* **68**, 2372-2383, doi:10.1002/art.39771 (2016).

281 Horowitz, A. *et al.* Genetic and environmental determinants of human NK cell diversity revealed by mass cytometry. *Sci Transl Med* **5**, 208ra145, doi:10.1126/scitranslmed.3006702 (2013).

282 Huang, J. *et al.* Detection, phenotyping, and quantification of antigen-specific T cells using a peptide-MHC dodecamer. *Proc Natl Acad Sci U S A* **113**, E1890-1897, doi:10.1073/pnas.1602488113 (2016).

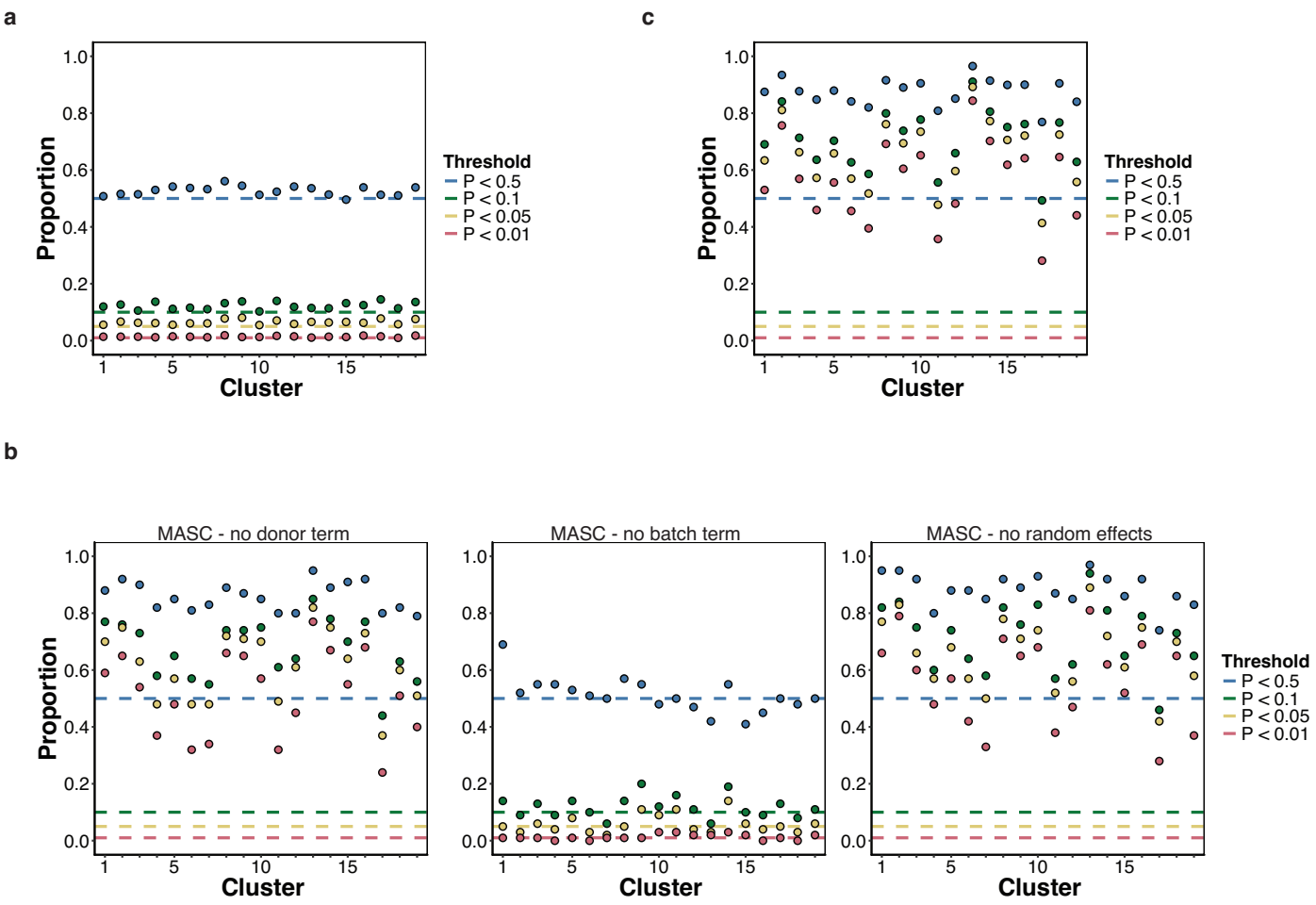
283 Esfandiary, L. *et al.* Single-cell antibody nanowells: a novel technology in detecting anti-SSA/Ro60- and anti-SSB/La autoantibody-producing cells in peripheral blood of

rheumatic disease patients. *Arthritis Res Ther* **18**, 107, doi:10.1186/s13075-016-1010-5 (2016).

284 Bentzen, A. K. *et al*. Large-scale detection of antigen-specific T cells using peptide-MHC-I multimers labeled with DNA barcodes. *Nat Biotechnol* **34**, 1037-1045, doi:10.1038/nbt.3662 (2016).

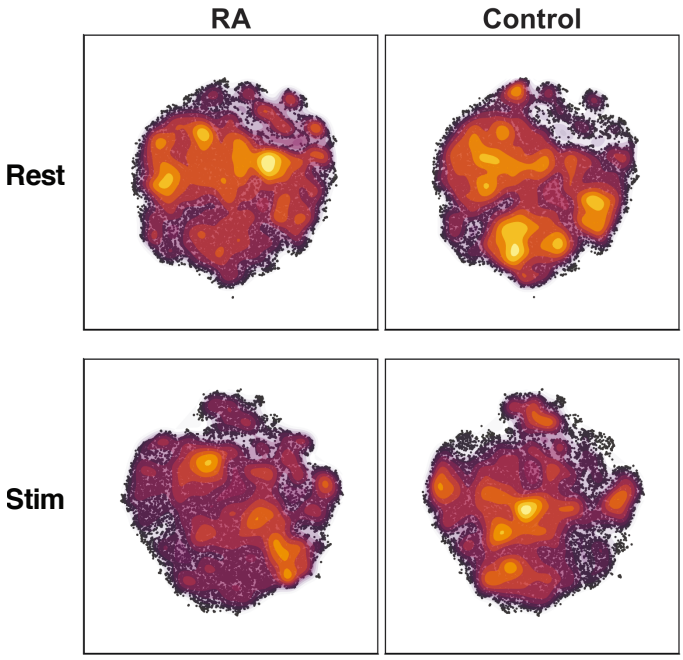
Appendix I: Supplemental Material for Chapter 3

List of Supplementary Materials



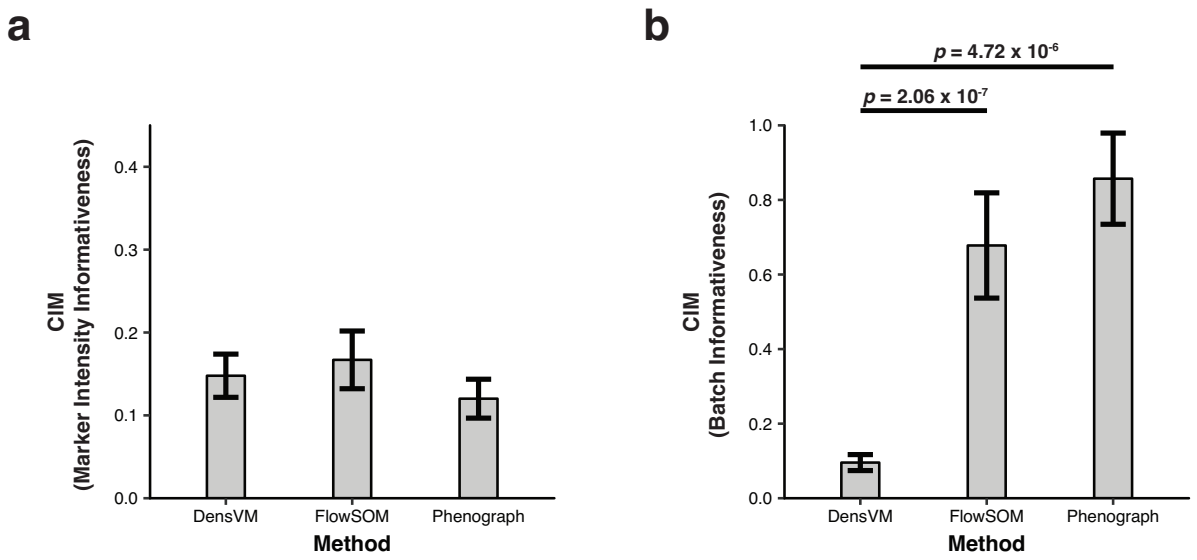
Supplementary Figure 1 – MASC Type 1 Error

(a) MASC demonstrates well-controlled type 1 error rates. MASC was run on the resting dataset after randomizing case-control labels 10000 times to eliminate any case-control associations. The proportion of p-values at different thresholds are plotted for each cluster. (b) MASC p-values obtained in the same manner as previous, but without donor or batch specific random effect terms. (c) P-values obtained in the same manner for binomial association tests on clusters found in the resting dataset.



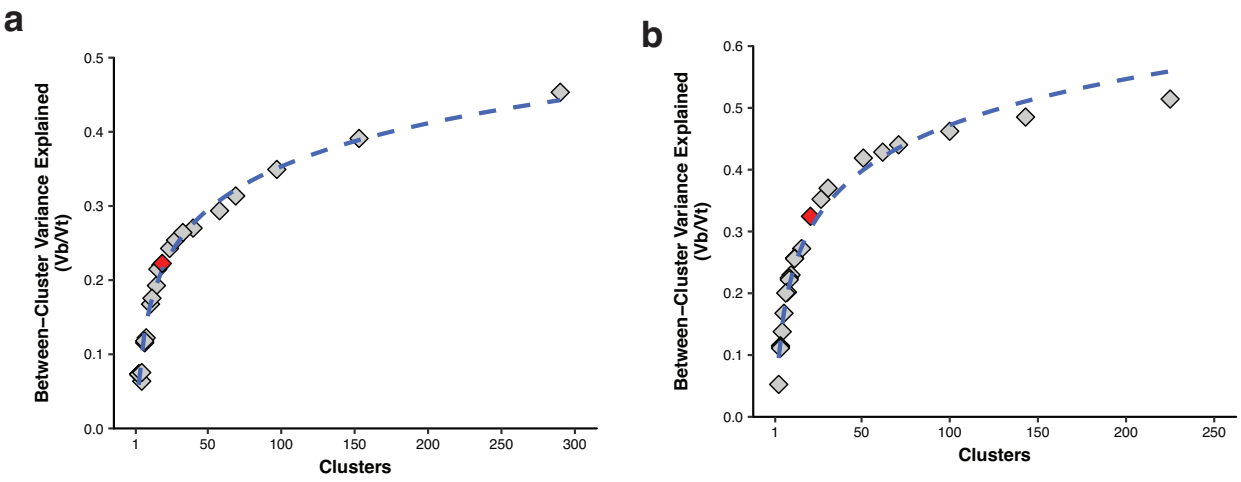
Supplementary Figure 2 – SNE Projection Density

(a) SNE projections of datasets before (top) and after (bottom) stimulation, split by case-control status. Coloring the SNE projections by density identifies regions that are differentially abundant between RA and control samples.



Supplementary Figure 3 – Cluster Informativeness Metric Analysis of Clustering Approaches

We clustered the same dataset using three different clustering algorithms, DensVM, Phenograph, and FlowSOM. These algorithms identified 19 (DensVM and FlowSOM) or 21 (Phenograph) clusters. (a) Clusters found by DensVM, Phenograph, and FlowSOM had similar average CIM scores when considering marker expression, indicating that the clusters found by these algorithms were similarly informative. That is, marker intensities were different from the average marker expression profile across clusters to the same extent. (b) Clusters found by Phenograph and FlowSOM had a significantly higher CIM score when considering batch than those found by DensVM, indicating that the Phenograph and FlowSOM clusters were more affected by batch effects. We assessed significance using a Wilcoxon rank sum test and p-values were Bonferroni adjusted to control for multiple testing.



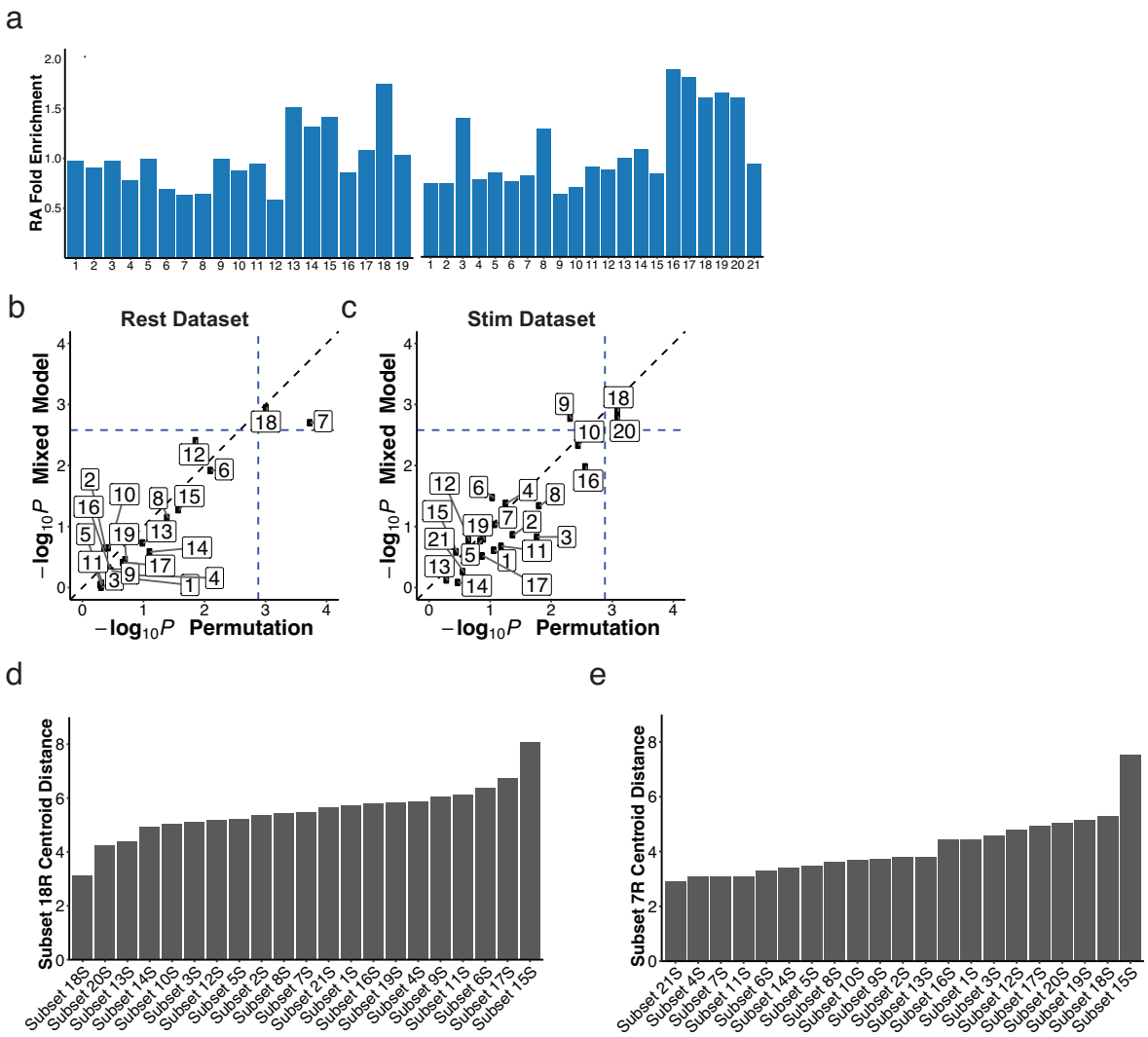
Supplementary Figure 4 – DensVM Clustering Elbow Plots

In order to define the optimal number of clusters, we use the the elbow strategy. We clustered data with DensVM across a range of bandwidth values, yielding different numbers of clusters at each of the 25 bandwidth values chosen. We then took the ratio of between-cluster variance to total variance to measure the amount of variance explained by each set of clusters. The set of clusters used in our analyses is marked in red and an exponential fit to the points shown is plotted as a dashed blue line. (a) DensVM clustering of the resting dataset produced 3-290 clusters across different bandwidths. The bandwidth producing 19 clusters (red) is at an inflection point for the amount of between-cluster variance explained. (b) DensVM clustering of the stimulated dataset produced 3-225 clusters across different bandwidths. The bandwidth producing 21 clusters (red) is at an inflection point for the amount of between-cluster variance explained.

[data shown in separate file: Appendix_I_Supplementary_Figure_5.pdf]

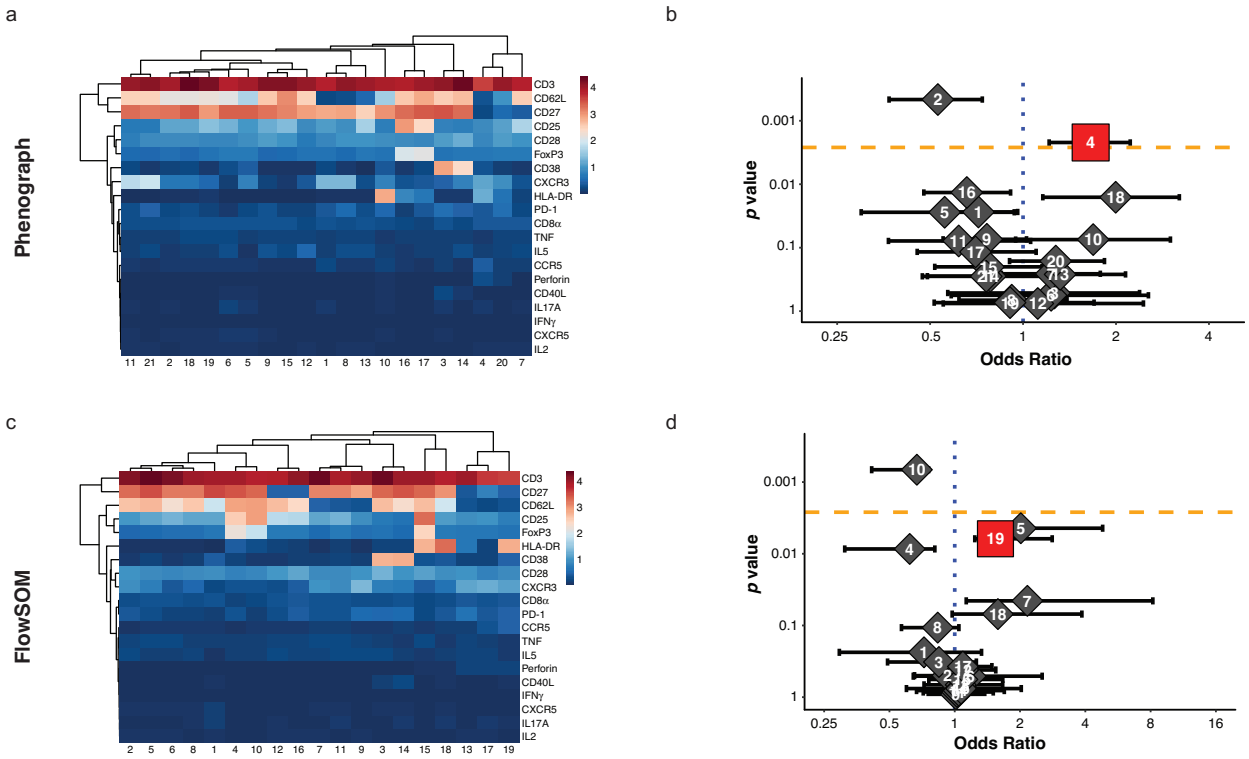
Supplementary Figure 5 – Marker Expression Distribution Plots for DensVM Clusters

For each cluster in the resting (n = 19) and stimulated (n = 21) datasets, we plotted the distribution of marker expression for cells in the cluster against the expression distribution for that marker for cells across the entire dataset. Expression specific to the cluster is colored in red, while dataset expression is colored in dark grey. Mass cytometry expression values are shown after applying a standard arcsinh transformation.



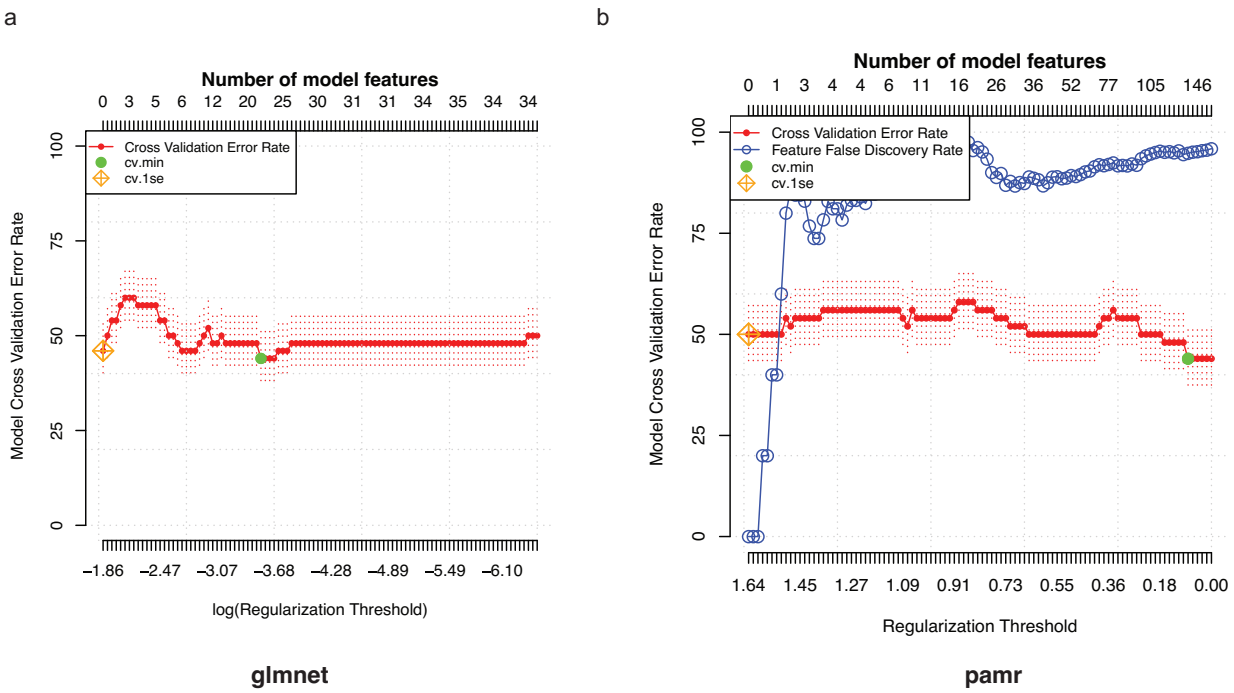
Supplementary Figure 6 – Association Permutation Testing and Cluster Alignment

(a) The enrichment or depletion of RA cells relative to the overall proportion is shown for all clusters identified in the resting (left) and stimulated (right) datasets. (b, c) Association p-values as calculated by MASC (y-axis) and by explicit permutation (x-axis) correlate in both resting and stimulated datasets. Spearman's correlation coefficients for (b) and (c) were $r_s = 0.82$ and $r_s = 0.86$, respectively. (d) Clusters in the stimulated dataset ranked by their overall distance from cluster 18. After normalizing marker expression in each cluster, cluster centroids were created and Euclidean distances were calculated between all clusters in the stimulated dataset and cluster 18 in the resting dataset. (e) Same as (d), but for distance from cluster 7 in the resting dataset.



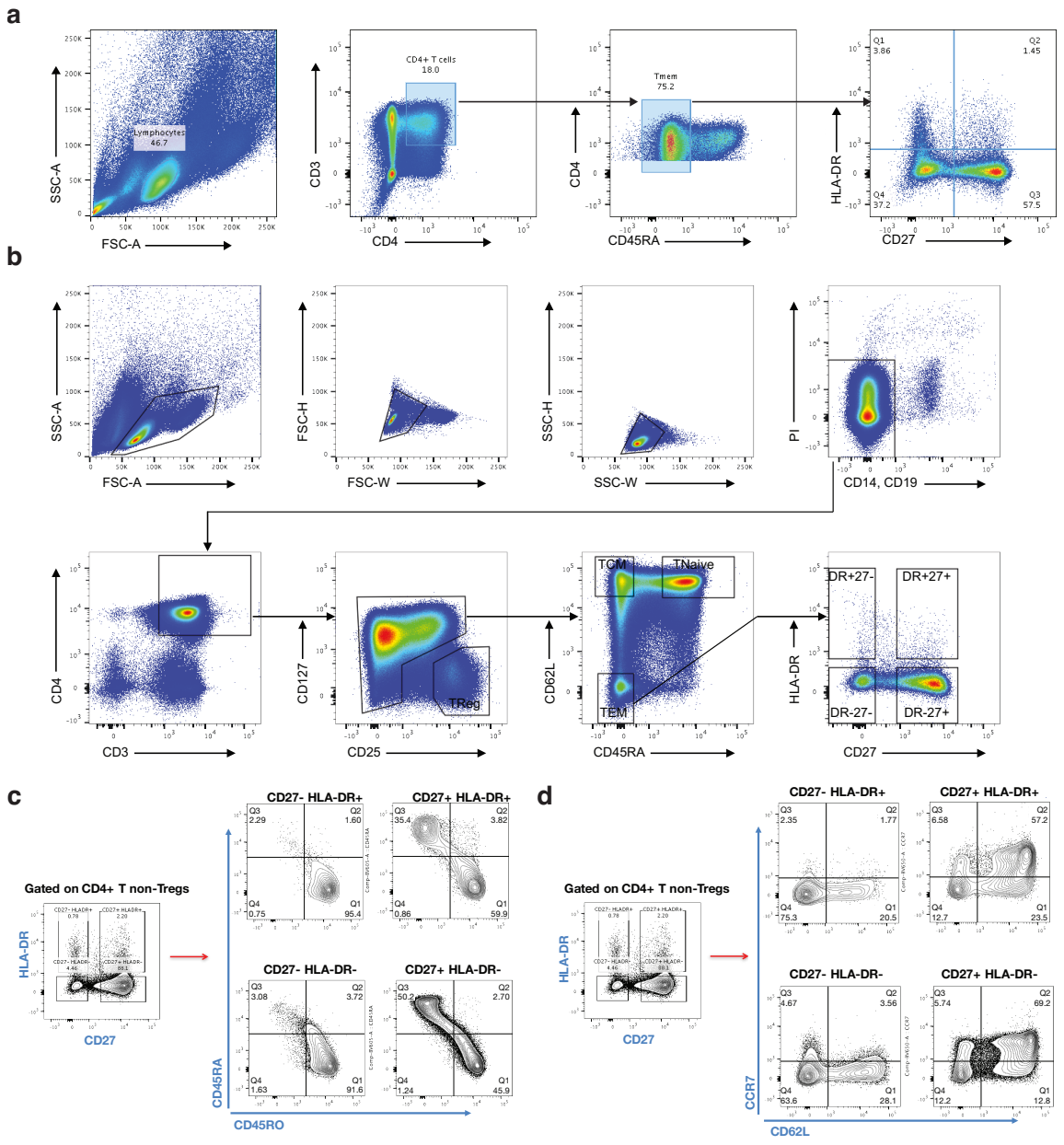
Supplementary Figure 7 – Phenograph and FlowSOM Clustering.

(a) Phenograph identified 21 clusters in the resting dataset, including an CD27- HLA-DR+ T_{EM} population (cluster 4) that is significantly expanded in RA. (b) Odds ratios and association p-values were calculated by MASC for each cluster identified by Phenograph. The yellow line indicates the significance threshold after applying the Bonferroni correction for multiple testing. (a) FlowSOM identified 19 clusters in the resting dataset, including an CD27- HLA-DR+ T_{EM} population (cluster 19) that is nominally expanded in RA. (b) Odds ratios and association p-values were calculated by MASC for each cluster identified by FlowSOM. The yellow line indicates the significance threshold after applying the Bonferroni correction for multiple testing.



Supplementary Figure 8 – Association Testing with Citrus

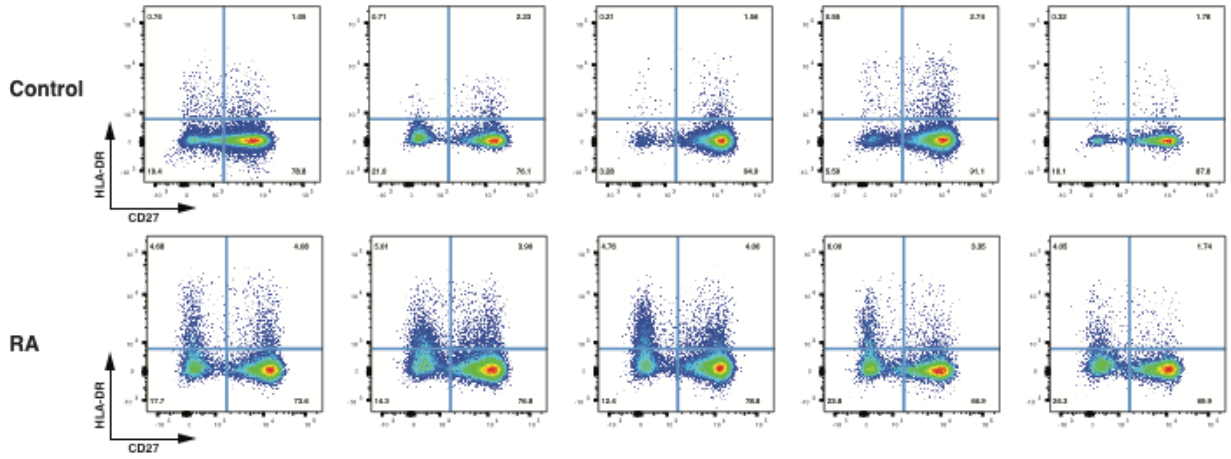
Citrus was run on the resting dataset but failed to produce models with acceptable error rates using either L1-penalized regression (a) or nearest shrunken centroid (b) methods. Model features found to be associated with case-control status by either method are unlikely to be meaningful given the extremely high cross-validation error.



Supplementary Figure 9 – Flow Cytometry and RNA-seq Gating Strategies

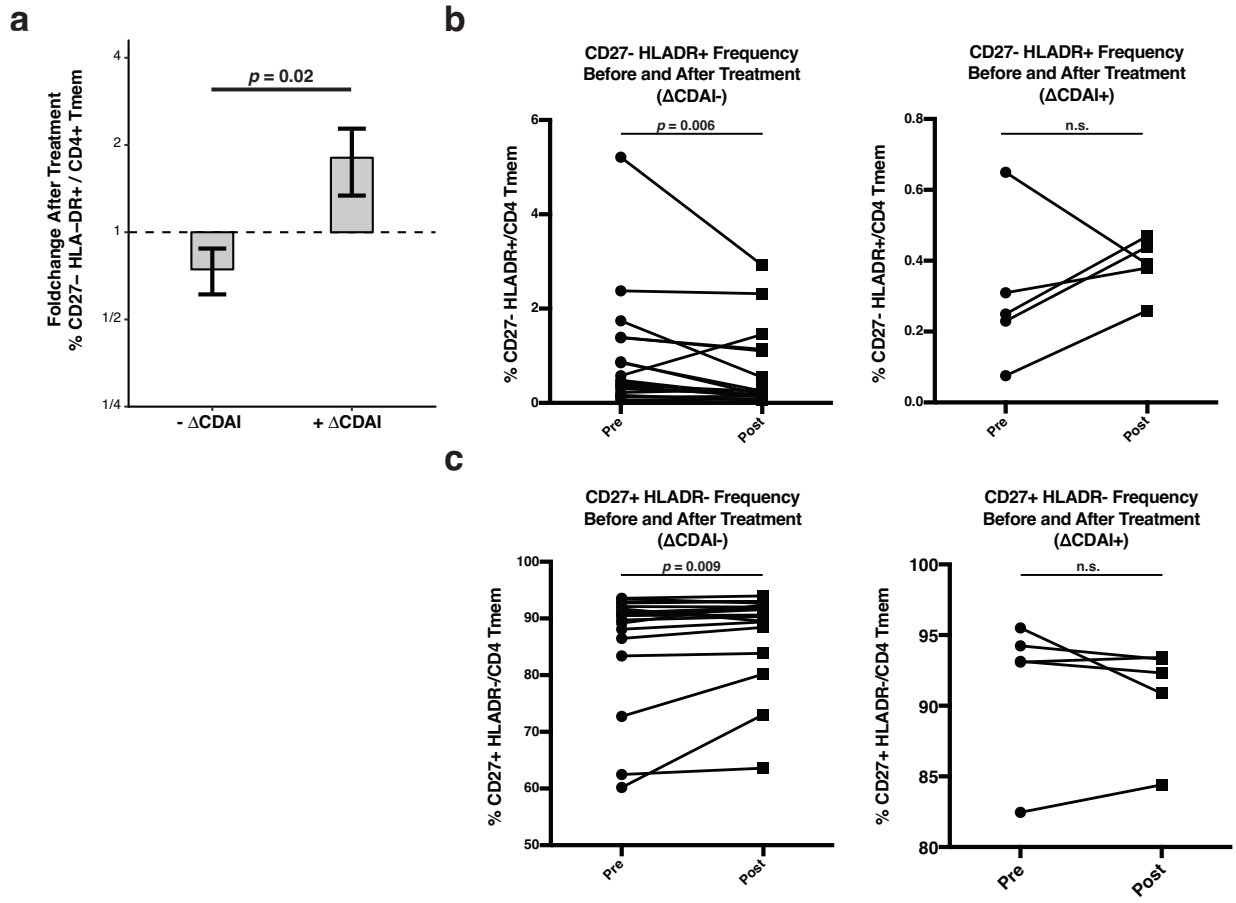
(a) Gating strategy used to isolate CD27- HLA-DR+ cells for flow cytometry quantification. Cells were first gated to lymphocytes using forward and side scatter parameters, then to CD4+ memory T cells before being split into four populations based upon the expression of CD27 and HLA-DR. (b) Gating strategy used to isolate populations for RNA sequencing. Cells were gated to lymphocytes using forward and side scatter parameters, then gated as CD14- CD19- to remove any non T cell lymphocytes. Cells were then gated to CD4+ T cells before isolating the following populations: regulatory T cells (CD25+ CD127-), central memory T cells (CD62L+ CD45RA-), naïve T cells (CD62L+ CD45RA+) and effector memory T cells (CD62L- CD45RA-).

Supplementary Figure 9 (continued) Effector memory T cells were then split into four populations based upon the expression of CD27 and HLA-DR. (c) Expression of CD45RO and CD45RA is shown for all four effector memory populations analyzed by RNA-seq. CD27- HLA-DR+ cells are uniformly CD45RA- CD45RO+. (d) Same as (c), but the expression of CD62L and CCR7 are shown.



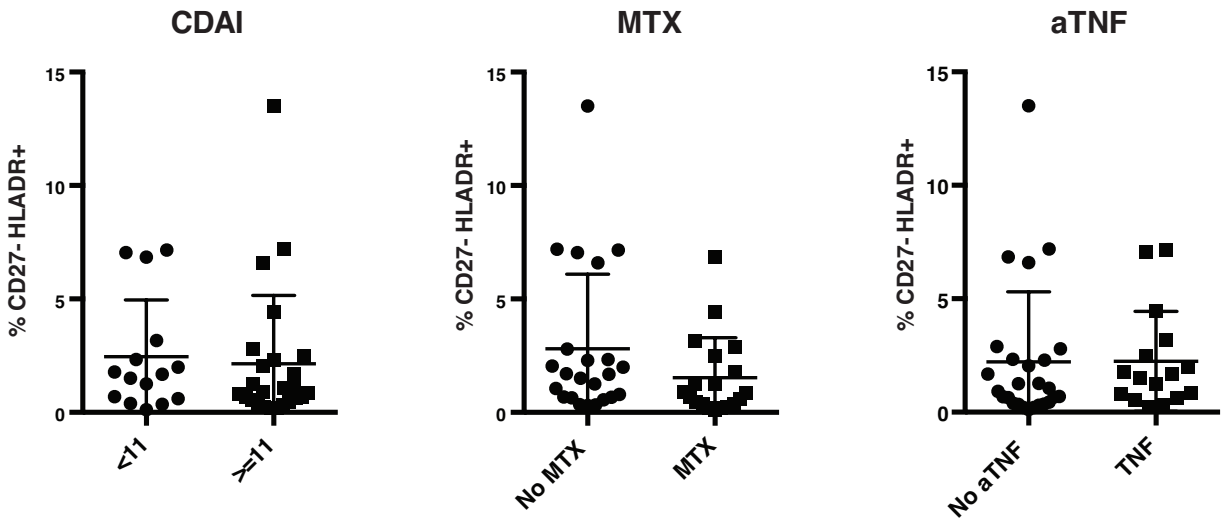
Supplementary Figure 10 – CD27 and HLA-DR Expression in Flow Cytometry Cohort

The expansion of the CD27- HLA-DR+ T cell population in RA patients was validated in an independent cohort of 39 seropositive RA patients and 27 controls using flow cytometry. The frequency of CD27 and HLA-DR cells among CD4+ memory T cells is shown for 10 representative donors, 5 cases and 5 controls.



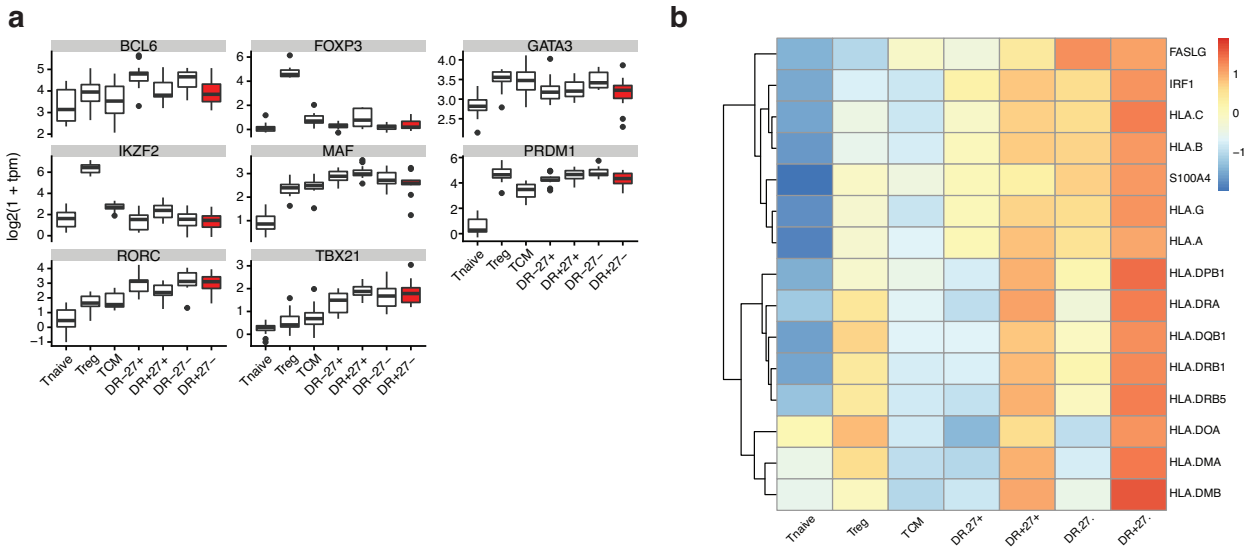
Supplementary Figure 11 – CD4+ Effector Memory T Cell Populations in a Clinical Response Cohort

We quantified the frequency of CD27- HLA-DR+ T cells in 23 RA patients before and 3 months after initiation of a new medication for RA. Patients were separated into those who experienced a clinical response (n = 18) versus those that did not (n = 5), defined as a reduction (-**ΔCDAI**) or an increase in CDAI scores (+**ΔCDAI)**. (a) The fold-change in CD27- HLA-DR+ frequency was significantly different between the two groups of patients (p=0.02, Wilcoxon rank sum test). (b) We quantified CD27- HLA-DR+ cell frequencies in patients who experienced a reduction in disease activity after initiation of a new medication for RA and those who did not. The frequency of the CD27- HLA-DR+ subset significantly decreased in -**ΔCDAI** individuals (p=0.006, Wilcoxon rank sum test), but did not significantly change among +**ΔCDAI** individuals. (c) Same as (b), except that the frequency of CD27+ HLA-DR- was quantified in **ΔCDAI** and +**ΔCDAI** individuals. We calculated frequencies of CD27- HLA-DR+ and CD27+ HLA-DR- cells from all CD4+ memory T cells, and assessed significance with Wilcoxon signed-rank tests.



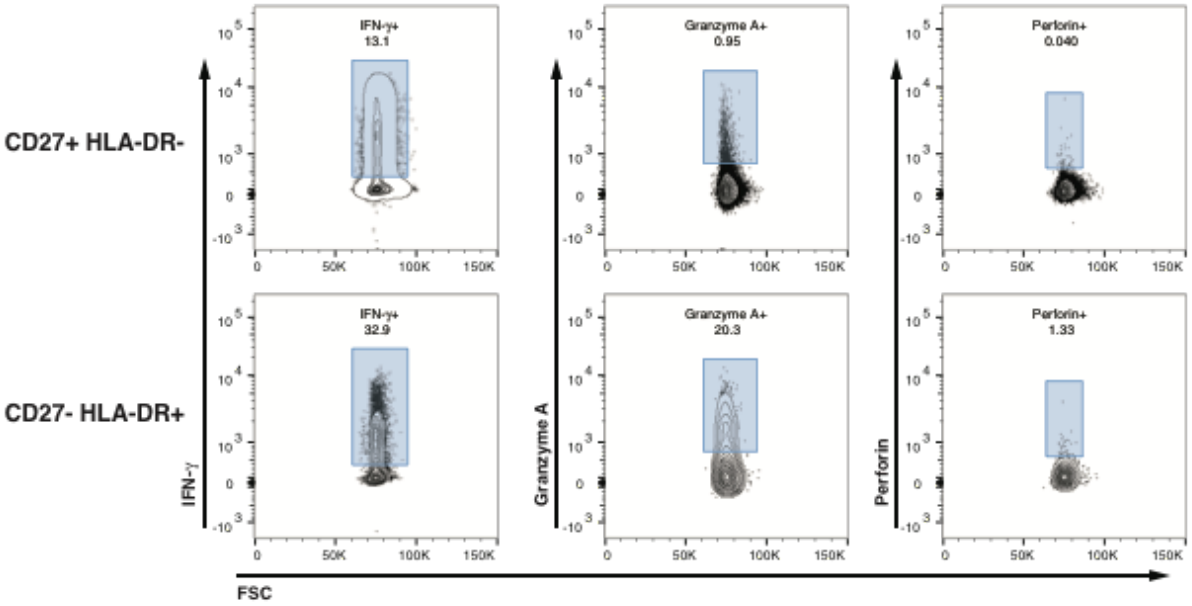
Supplementary Figure 12 – CD27- HLA-DR+ Frequency and Clinical Characteristics

The frequency of CD27- HLA-DR+ cells was quantified as the percentage of memory CD4+ T cells in an independent cohort of 39 seropositive RA patients and 27 controls using conventional flow cytometry. RA patients were then dichotomized by clinical disease activity index (CDAI) scores, methotrexate use (MTX) or anti-TNF therapy use (aTNF). The frequency of CD27- HLA-DR+ cells was not significantly different between groups in any comparison.



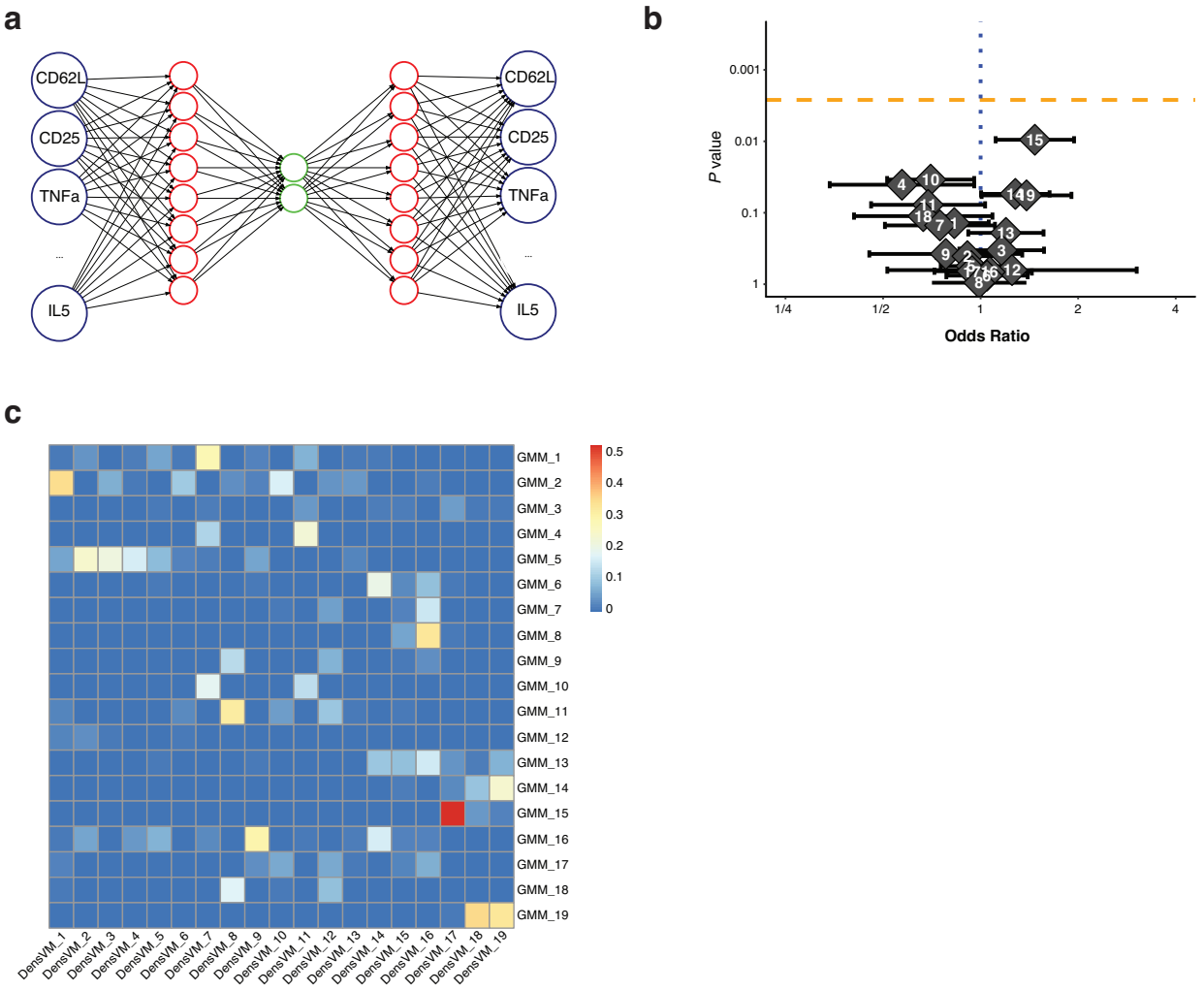
Supplementary Figure 13 – RNA-seq Analysis of CD4+ T Cell Subsets

(a) Expression of lineage-defining transcription factors for CD4+ T helper subsets shown for each population analyzed by RNA-seq. Populations are ordered by principal component 1 loadings, from naïve to effectors. (c) The expression of selected targets of transcription factor CIITA is shown for each sequenced T cell population.



Supplementary Figure 14 – Flow Cytometry Expression Quantification

The expression of markers granzyme A, perforin, and IFN-γ across all samples are displayed, with the gating used to to define percent positivity for those markers. The same gates were used to analyze CD27+ HLA-DR- and CD27- HLA-DR+ populations. The expression plots of granzyme A and perforin show the concatenation of six samples (3 RA, 3 OA), while the expression plot of IFN-γ shows concatenation of 12 samples (6 RA, 6 OA).



Supplementary Figure 15 – Using a Neural-Net Auto-encoder to Cluster Mass Cytometry Data

(a) Schematic of the deep auto-encoder that was trained upon the rest data using 3 hidden layers. Clustering was then performed on the middle layer (2 node) projection (highlighted in green). (b) Clusters identified by the auto-encoder were tested for case-control associations using MASC. None of the clusters reached significance after correcting for multiple hypothesis testing. (c) The Jaccard index was calculated between each cluster identified by DensVM (x-axis) and the autoencoder (y-axis). The most significant disease-associated auto-encoder clusters share significant overlap with the original DensVM clusters.

Isotope	Marker	Clone
Nd143Di	IL-5	TRFK5
Nd144Di	CCR5	NP-6G4
Nd145Di	CD4*	RPA-T4
Nd146Di	CD8α	RPA-T8
Sm147Di	CD45RO*	UCHL1
Nd148Di	CD28	CD28.2
Sm149Di	CD25	2A3
Eu151Di	PD-1	EH12.2H7
Sm152Di	TNF	MAb11
Eu153Di	CD62L	DREG-56
Sm154Di	CD3	UCHT1
Gd155Di	CD27	L128
Gd156Di	CXCR3	G025H7
Gd158Di	IL-2	MQ1-17H12
Dy162Di	FoxP3	PCH101
Ho165Di	IFN-γ	B27
Er167Di	CD38	HIT2
Er168Di	CD40L	24-31
Tm169Di	IL-17A	BL168
Yb171Di	CXCR5	51505
Yb174Di	HLA-DR	L243
Lu175Di	Perforin	B-D48

Supplementary Table 1: Panel design for mass cytometry experiments. Markers that are starred were only used for gating purposes to confirm the purity of CD4 memory T cell isolation and were not including in clustering or downstream analyses.

Cluster	Cell Number	RA Proportion	Permutation *p* value	MASC *p* value	Odds Ratio	Odds Ratio, 2.5% CI	Odds Ratio, 97.5% CI
1	6000	0.493	2.43E-01	7.12E-01	1.0	0.8	1.3
2	4506	0.475	4.18E-01	2.36E-01	0.8	0.6	1.1
3	3994	0.492	4.71E-01	9.60E-01	1.0	0.8	1.3
4	2776	0.437	1.24E-01	6.40E-01	0.9	0.7	1.2
5	2205	0.498	4.69E-01	8.78E-01	1.0	0.7	1.3
6	1578	0.406	7.68E-03	9.31E-03	0.7	0.5	0.9
7	2100	0.385	1.80E-04	8.78E-04	0.6	0.5	0.8
8	2138	0.389	3.94E-02	6.20E-02	0.6	0.3	1.0
9	3856	0.497	3.27E-01	5.39E-01	1.1	0.8	1.6
10	2317	0.467	3.64E-01	1.81E-01	0.8	0.5	1.1
11	790	0.486	4.89E-01	8.96E-01	1.0	0.6	1.5
12	1421	0.367	1.34E-02	2.03E-03	0.5	0.4	0.8
13	965	0.602	9.90E-02	1.54E-01	1.4	0.9	2.1
14	2781	0.568	7.59E-02	2.56E-01	1.3	0.8	1.9
15	1048	0.586	2.57E-02	5.29E-02	1.5	1.0	2.2
16	7507	0.461	3.54E-01	4.89E-01	0.9	0.7	1.2
17	1020	0.520	2.09E-01	3.30E-01	1.2	0.9	1.5
18	1184	0.636	9.40E-04	5.59E-04	1.9	1.3	2.7
19	1814	0.507	1.91E-01	2.17E-01	1.2	0.9	1.6

Supplementary Table 2: MASC analysis of the 19 clusters identified in the resting dataset.

Cluster	Cell Number	RA Proportion	Permutation *p* value	MASC *p* value	Odds Ratio	Odds Ratio, 2.5% CI	Odds Ratio, 97.5% CI
1	2518	0.430	8.35E-02	2.45E-01	0.7	0.4	1.2
2	1757	0.429	4.10E-02	1.37E-01	0.7	0.5	1.1
3	2981	0.586	1.64E-02	1.48E-01	1.3	0.9	1.8
4	2086	0.440	5.32E-02	4.15E-02	0.7	0.4	1.0
5	3202	0.463	1.38E-01	1.71E-01	0.8	0.5	1.1
6	2031	0.435	8.83E-02	3.35E-02	0.6	0.4	1.0
7	3687	0.452	8.10E-02	9.26E-02	0.7	0.5	1.1
8	3834	0.565	1.52E-02	4.56E-02	1.2	1.0	1.5
9	1160	0.389	4.66E-03	1.68E-03	0.5	0.4	0.8
10	2908	0.416	3.47E-03	4.67E-03	0.7	0.5	0.9
11	4742	0.478	6.32E-02	2.11E-01	0.9	0.7	1.1
12	3934	0.469	2.17E-01	1.67E-01	0.8	0.5	1.1
13	2436	0.501	4.97E-01	7.52E-01	1.0	0.7	1.3
14	4462	0.522	3.26E-01	8.27E-01	1.0	0.8	1.3
15	755	0.458	3.45E-01	2.61E-01	0.7	0.4	1.3
16	1459	0.657	2.66E-03	1.05E-02	1.6	1.1	2.3
17	1701	0.646	1.29E-01	3.04E-01	1.4	0.8	2.4
18	1182	0.619	7.90E-04	1.28E-03	1.7	1.2	2.2
19	1483	0.626	1.23E-01	1.60E-01	1.4	0.9	2.4
20	911	0.618	7.90E-04	1.61E-03	1.7	1.2	2.3
21	2771	0.487	2.68E-01	5.48E-01	0.9	0.7	1.2

Supplementary Table 3: MASC analysis of the 21 clusters identified in the stimulated dataset.

Gene Set	*Pathway*	*p value*	*q value*	*Size*	*Enrichment*
GSE22886	NAÏVE CD4 T CELL VS NK CELL	8.45E-13	1.01E-11	159	NK CELL
GSE3982	CENT MEMORY CD4 T CELL VS NK CELL	3.53E-08	1.74E-07	159	NK CELL
GSE27786	CD4 T CELL VS NK CELL	1.59E-04	6.38E-04	170	NK CELL
GSE3039	CD4 T CELL VS NKT CELL	5.07E-03	1.22E-02	166	NKT CELL
GSE3982	EFF MEMORY CD4 T CELL VS NK CELL	1.67E-01	3.35E-01	149	NK CELL

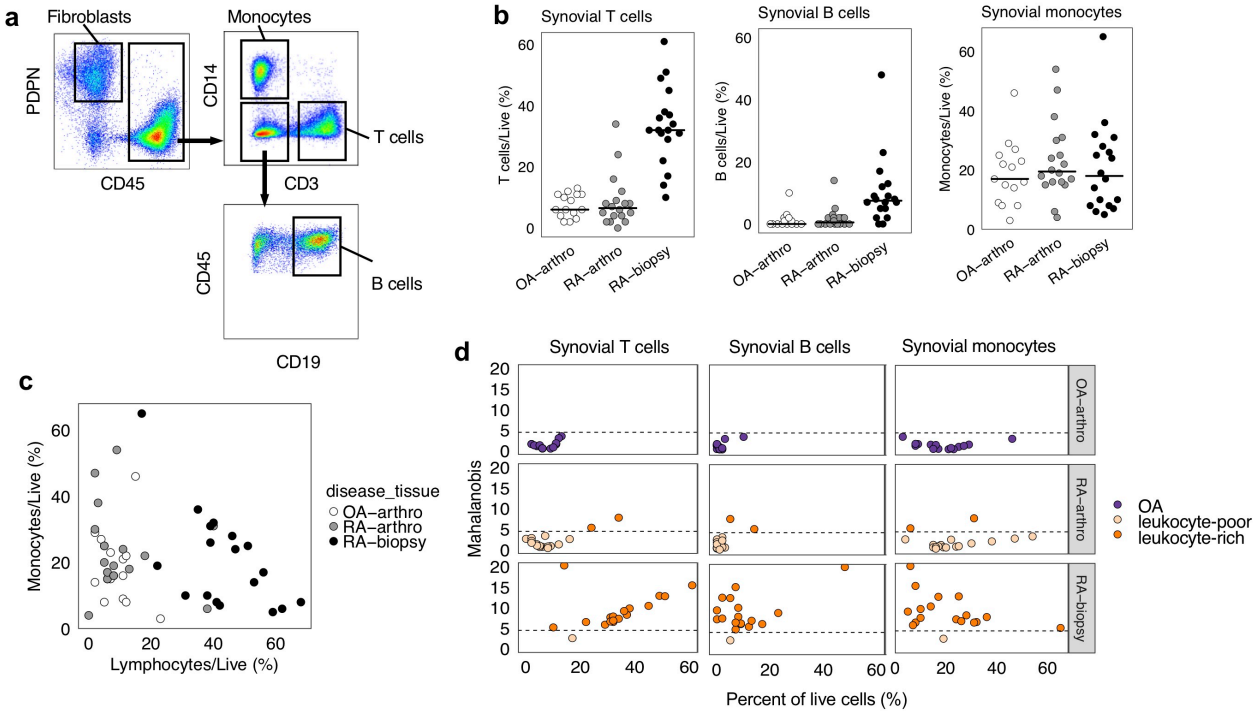
Supplementary Table 4: Gene set enrichment analysis of genes differentially expressed in CD27- HLA-DR+ cells. Q values represent FDR corrected p-values, using an FDR of 5%. The number of genes in each set is listed as size. Enrichment indicates which cell type gene signature was enriched in genes specific for CD27- HLA-DR+ cells.

Appendix II: Supplemental Material for Chapter 4

List of Supplementary Materials

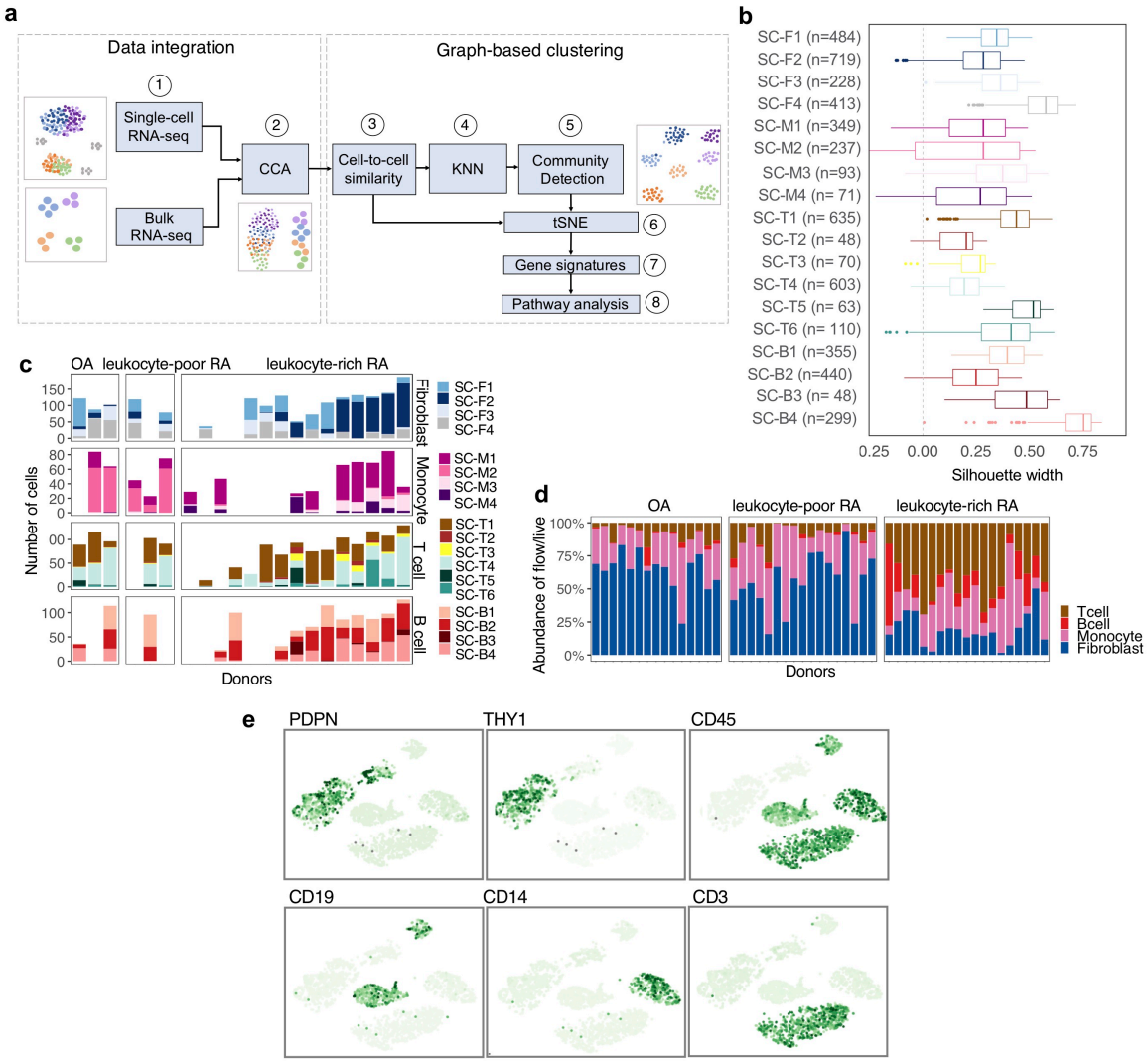
Table S3. Identified mass cytometry populations with proportion of cells from each disease cohort and on tailed FDR q value.

Table S4. Top 20 marker genes for each single-cell RNA-seq cluster.



Supplementary Figure 1 – Flow cytometry gating scheme and a data-driven approach to separate samples based on flow cytometry data.

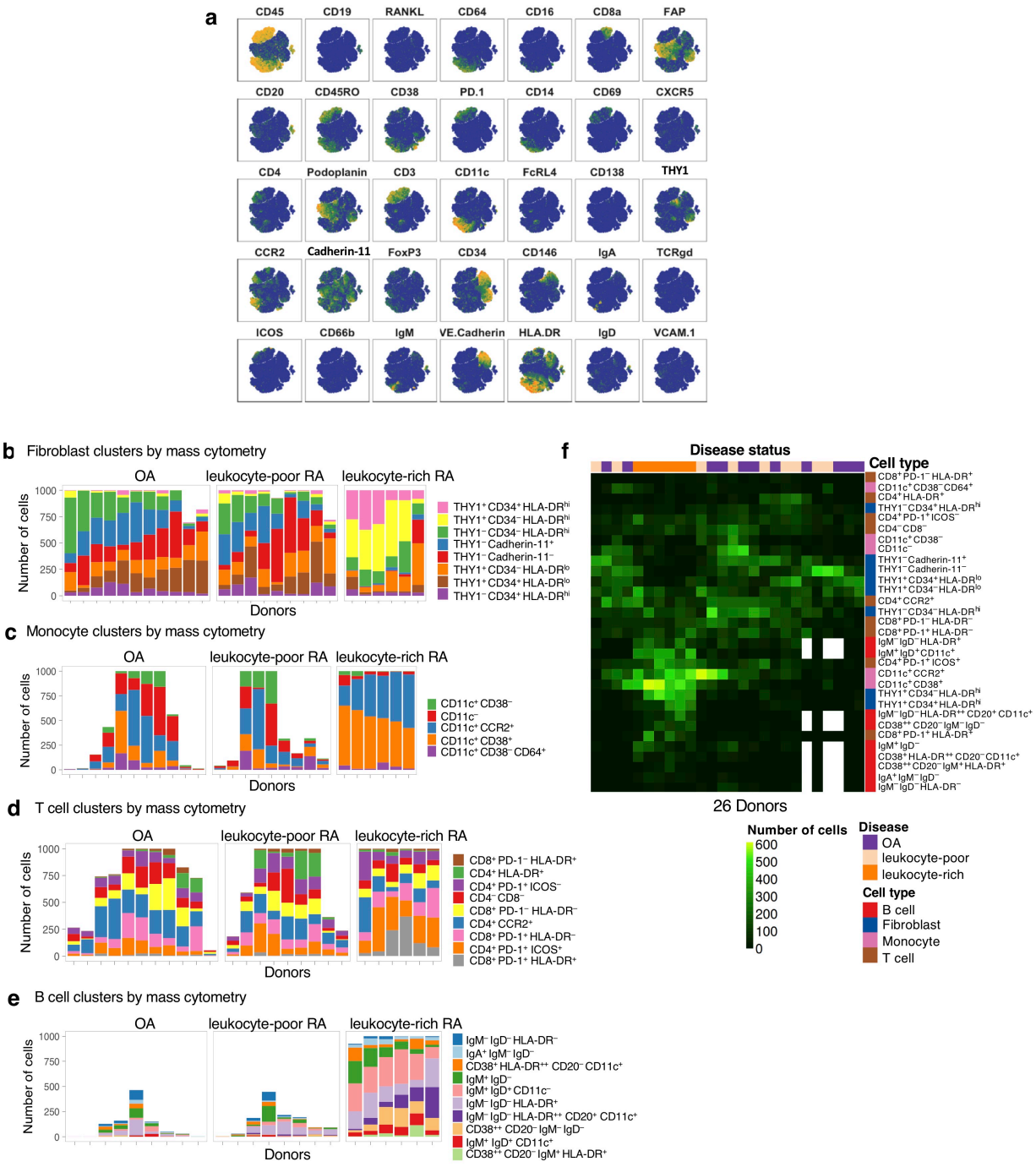
a. Flow cytometry gating: stromal fibroblasts ($CD45^{-}PDPN^{+}$), monocytes ($CD45^{+}CD14^{+}$), T cells ($CD45^{+}CD3^{+}$), and B cells ($CD45^{+}CD3^{-}CD19^{+}$). **b.** As a percentage of live cells: synovial T cells, B cells, and monocytes for OA-arthro (OA arthroplasty), RA-arthro (RA arthroplasty), and RA-biopsy (RA biopsy) by flow cytometry. **c**. Comparison of lymphocytes (T and B cells) and monocytes, as a percentage of live cells by flow cytometry. **d**. Mahalanobis distance from OA samples. Each dot represents a donor. Each panel highlights the contribution of T cells, B cells, and monocytes to the distance (y-axis). We defined leukocyte-rich RA samples as those with Mahalanobis distance from OA greater than 4.5 (dashed line). We identified 19 leukocyte-rich RA, 17 leukocyte-poor RA, and 15 OA samples in our cohort.



Supplementary Figure 2 – scRNA-seq analysis pipeline and distribution of identified subsets by scRNA-seq, flow cytometry, and protein fluorescence on each cell.

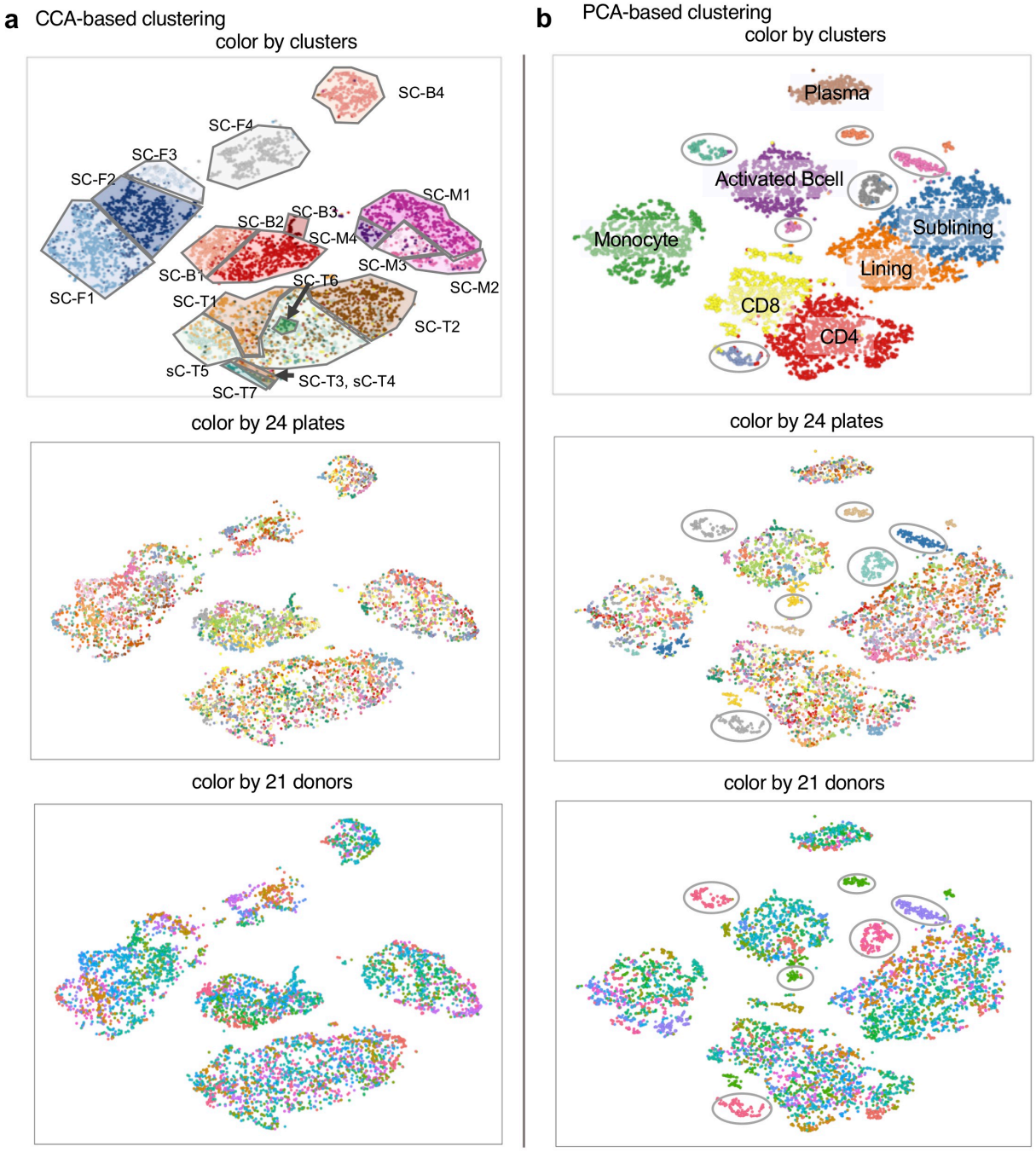
a. CCA-based integrative pipeline of scRNA-seq analysis. 1) We first select the highly variable genes from both scRNA-seq and bulk RNA-seq; 2) We integrate single cells with bulk samples based on the selected genes from both sides and learn a linear projection that the correlation between both sides are maximized using CCA; 3) we then calculate a cell-to-cell similarity matrix based on the top 10 canonical variates from CCA; 4) We build a K-nearest neighbors (KNN) network on the cell-to-cell similarity matrix and then convert it into an adjacency matrix; 5) we cluster the cells using the Infomap community detection algorithm to identify major groups on the cell-to-cell adjacency matrix; 6) we visualize the cells with tSNE; 7) We perform differential gene expression analysis on the identified cell type clusters and report three statistics: AUC, percent of non-zero expressing cells, and fold change; 8) finally, we perform gene set enrichment analysis to find pathways associated with each identified cell cluster. **b**. Silhouette analysis of 18 scRNA-seq clusters. The measure range is [-1, 1], where a value near 1 indicates a cell is far from neighboring clusters, a value near 0 indicates a cell is near a decision boundary, and a negative value indicates a cell is closer to a neighboring cluster. The features of the boxes are as follows. The box represents the 25th and 75th quantiles. The

Supplementary Figure 3 (continued) center line represents the 50th quantile. The low whisker is the lowest value greater than -1.5 times the inter-quartile range plus the 25th quantile. The high whisker is the greatest value less than 1.5 times the inter-quartile range plus the 75th quantile. The points are values outside the range of the whiskers. **d.** Cellular composition of major synovial cell types for each donor by flow cytometry. **e.** Flow cytometry protein fluorescence of cell type markers on each single cell: PDPN, THY1 (CD90), CD45, CD19, CD14, and CD3.



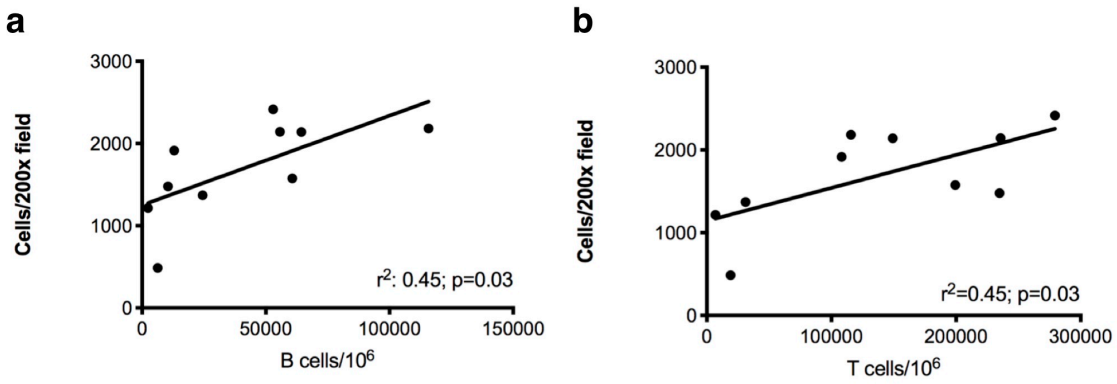
Supplementary Figure 4 – Mass cytometry data analysis.

a. Protein markers of synovial cells (3,000 downsampled) from all donors by mass cytometry. Color represents intensity of expression level. **b-e.** Distribution of identified subpopulations for each cell type. **f**. Cell counts of all clusters by comparing all the 26 donors reveal that leukocyte-rich donors show high cell abundance of HLA-DR$^+$ fibroblasts (THY1$^+$ CD34$^-$ HLA-DR$^+$ and THY1$^+$ CD34$^+$ HLA-DR$^+$), Tph cells (CD4$^+$ PD-1$^+$ ICOS$^+$), two CD14$^+$ monocytes subpopulations (CD11c$^+$ CCR2$^+$ and CD11c$^+$ CD38$^+$), and a B cell subpopulation (IgM$^+$ IgD$^+$ CD11c$^+$).



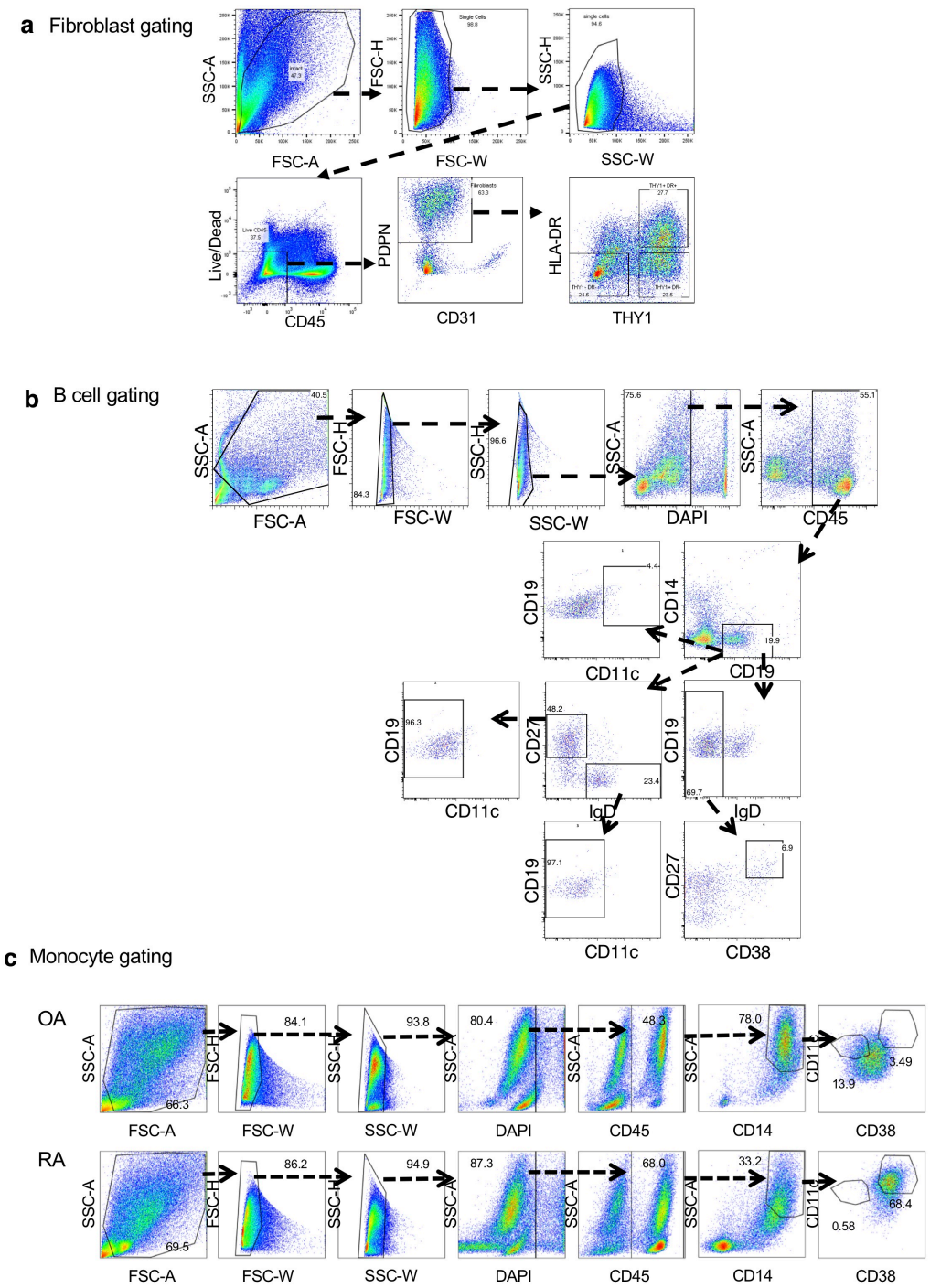
Supplementary Figure 5 – Comparison of CCA-based clustering and PCA-based clustering on batch effect correction performance.

a. Cells colored by 18 scRNA-seq clusters (top), 24 384-well plates (middle), and 21 donors (bottom) using the CCA-based integrative pipeline. **b**. Cells colored by scRNA-seq clusters, plates, and donors using PCA-based clustering by Seurat R package. Small clusters of cells from single donors are circled.



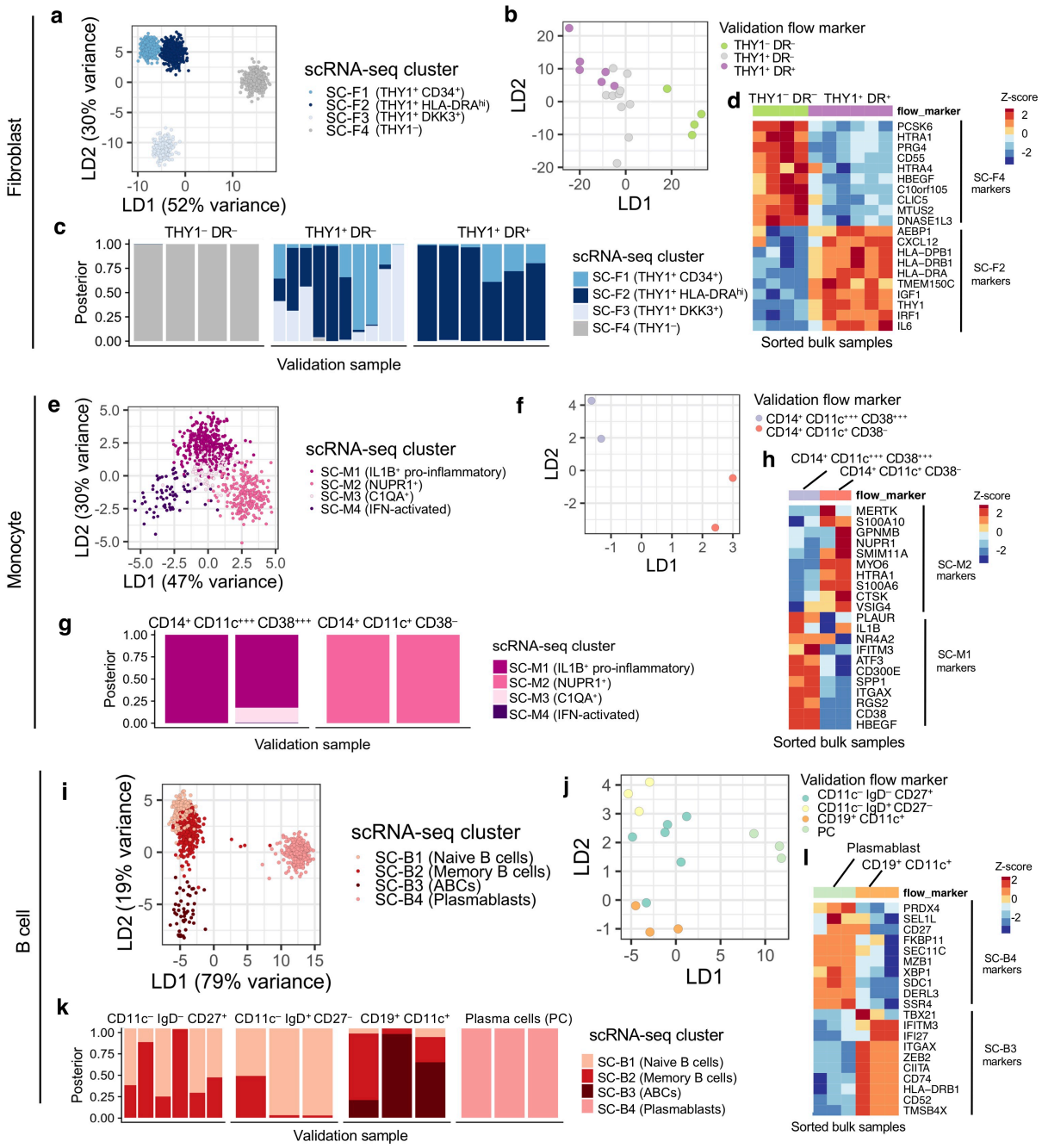
Supplementary Figure 5 – Cell density quantification on 10 histological synovial samples.

a. Correlation between cell density (cell counts per 200x field) and flow cytometric cell yields on B cells. **b**. Correlation between cell density (cell counts per 200x field) and flow cytometric cell yields on T cells. In general, we observed that the samples where we get the most single cell measurements are exactly the samples with the best yield and also the ones with the most inflammation.



Supplementary Figure 6 – Flow cytometry gating schema for experimental validations.

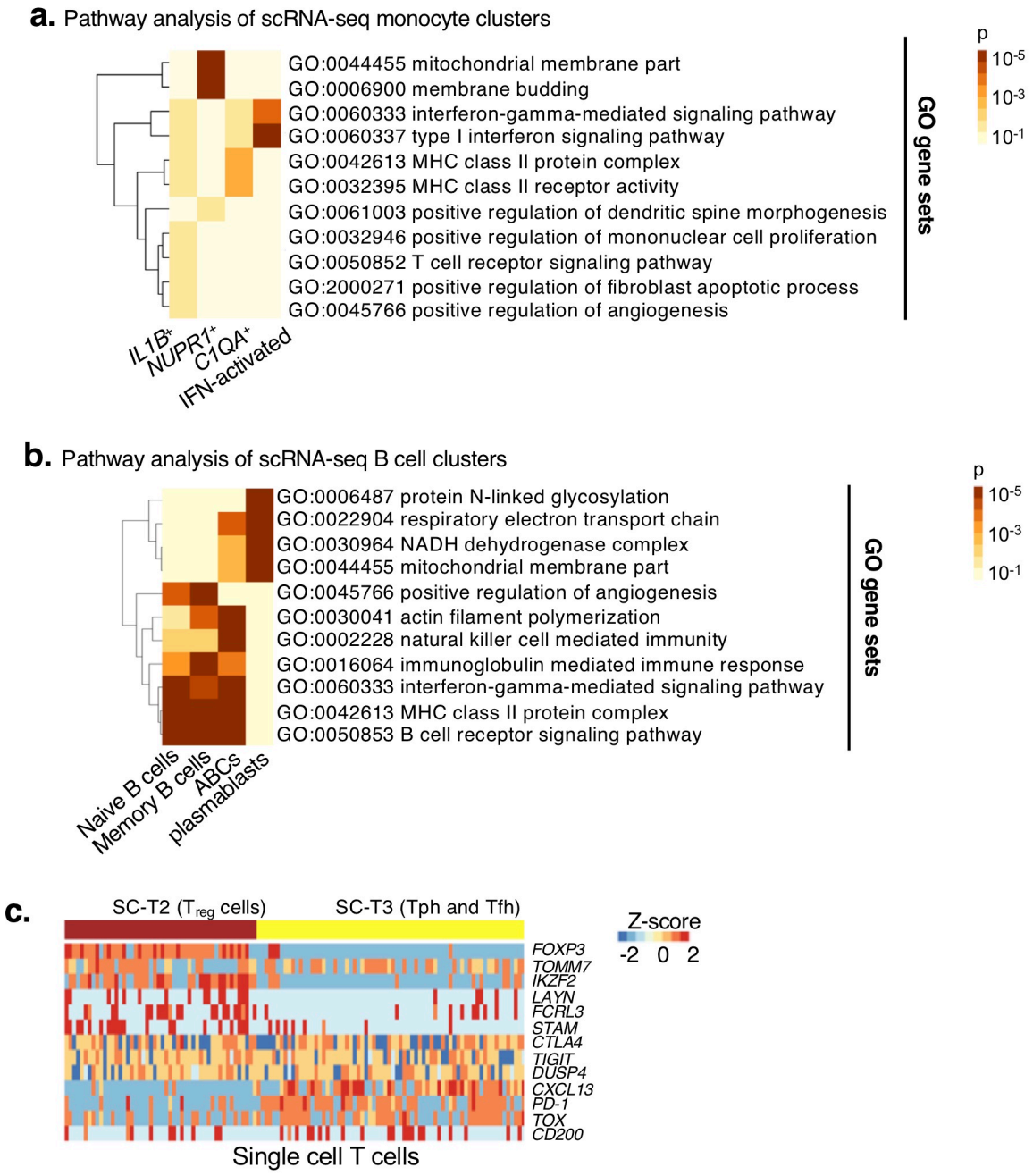
To identify and subset the immune and stroma populations that emerged from the scRNA-seq analyses, we sorted synovial cell subsets and disaggregated synovial tissues based on markers revealed by the scRNA-seq analysis. **a**. Flow gating strategy for synovial fibroblasts. **b**. Flow gating strategy for synovial B cells. **c**. Flow gating strategy for synovial monocytes. See **Methods** for more details.



Supplementary Figure 7 – Bulk RNA-seq analysis for flow sorted subpopulations of synovial fibroblasts, monocytes, and B cells to validate identified scRNA-seq clusters.

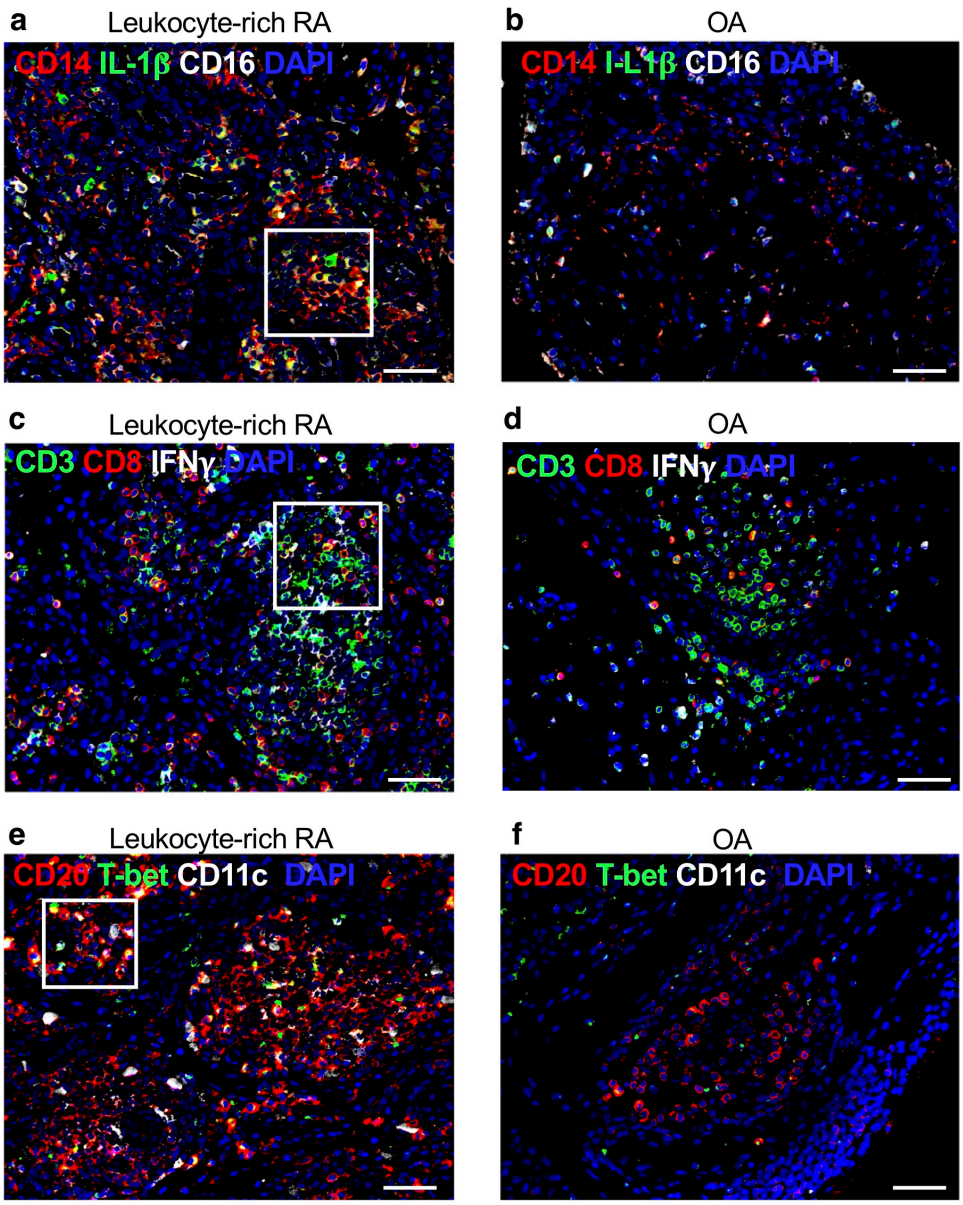
For each cell type, we trained a linear discriminant analysis (LDA) model on the scRNA-seq clusters. Next, we applied this LDA model to classify each bulk RNA-seq sample **(Supplementary Figure 6)**. After discovering scRNA-seq cluster markers (top 500 genes sorted by AUC for each cluster), we wanted to test if we could sort new cells from independent samples and see the same gene expression profiles in the new bulk samples as the original scRNA-seq samples. **a**. LDA projection of training data on single-cell fibroblasts (SC-F1-4). **b**. LDA projection of bulk RNA-seq samples that include sorted THY1$^-$DR$^-$ populations from 4 OA,

Supplemental Figure 7 (continued) $THY1^+DR^-$ population from 4 OA and 6 RA, and $THY1^+DR^+$ population from 6 RA. **c**. Posterior probabilities showing confidence of assigning each sorted fibroblast bulk sample to fibroblast scRNA-seq clusters. **d**. Genes (top 10 by Z-score) that are differentially expressed between two scRNA-seq clusters (SC-F2 and SC-F4) are also differentially expressed in the sorted bulk RNA-seq. **e**. LDA projection of training data on single-cell monocytes (SC-M1-4). **f**. LDA projection of bulk RNA-seq samples that include sorted CD14+CD11c+++CD38+++ population from 2 RA and $CD14^+CD11c^+CD38^-$ population from 2 OA. **g**. Posterior probabilities showing confidence of assigning each sorted monocyte bulk sample to monocyte scRNA-seq clusters. **h**. Genes (top 10 by Z-score) that are differentially expressed between two scRNA-seq clusters (SC-M1 and SC-M2) are also differentially expressed in the sorted bulk RNA-seq. **i**. LDA projection of training data on single-cell B cells (SC-B1-4). **j**. LDA projection of bulk RNA-seq samples that include sorted $CD11c^-$ IgD^- $CD27^+$ population from 6 RA, $CD11c^-$ IgD^+ $CD27^-$ population from 3 RA, $CD19^+CD11c^+$ population from 3 RA, and plasma cells from 3 RA. **k**. Posterior probabilities showing confidence of assigning each sorted B cell bulk sample to B cell scRNA-seq clusters. **l**. Genes (top 10 by Z-score) that are differentially expressed between two scRNA-seq clusters (SC-B3 and SC-B4) are also differentially expressed in the sorted bulk RNA-seq.



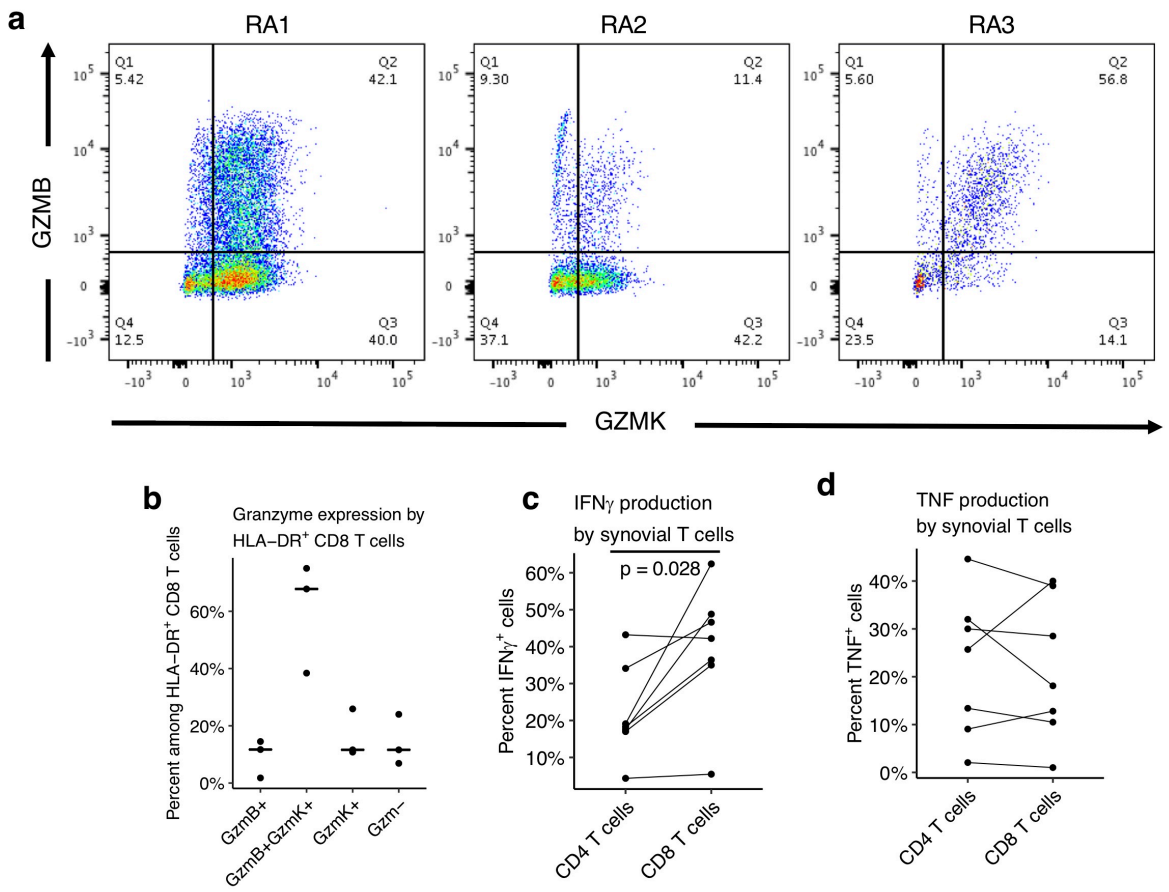
Supplementary Figure 8 – Pathway enrichment analysis for identified scRNA-seq clusters.

a. Enriched pathways on each monocyte cluster by scRNA-seq. **b**. Enriched pathways on each B cell cluster by scRNA-seq. **c**. Identified T_{reg} (SC-T2) and Tph and Tfh (SC-T3) scRNA-seq clusters. We used hierarchical clustering with R functions hclust() and cutree(k=2) to pinpoint the previously characterized rare cell populations.



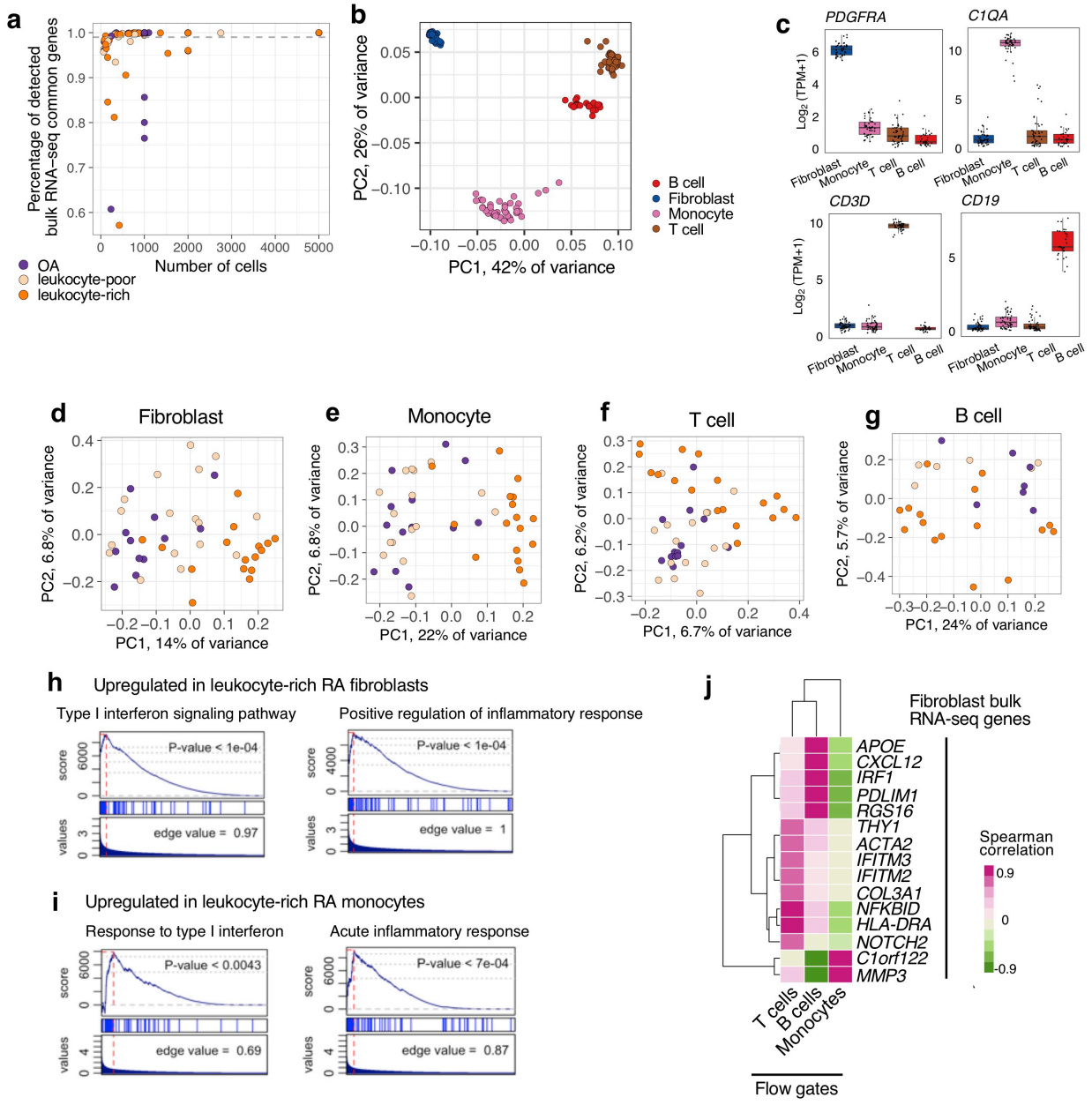
Supplementary Figure 9 – Multi-color immunofluorescent staining of paraffin synovial tissue from target RA and OA patient samples.

We performed multi-color immunofluorescence images to validate the unique cellular and molecular signatures revealed by the scRNA-seq analysis and show the contrasting cellular and molecular features of the microenvironment in the inflamed RA and OA synovia. **a**. Numerous $CD14^{+}IL\text{-}1\beta^{+}$ co-localization in inflamed RA synovium tissue (denoted by white box). **b**. Very few $CD14^{+}IL\text{-}1\beta^{+}$ cells in the OA synovium. **c**. Numerous $CD3^{+}CD8^{+}IFN\gamma^{+}$ T cells in RA biopsy (denoted by white box). **d.** Low numbers of $CD3^{+}CD8^{+}IFN\gamma^{+}$ T cells in the synovium of OA. **e**. An increased number of $CD20^{+}T\text{-}bet^{+}CD11c^{+}$ B cells in the inflamed RA synovium (denoted by white box). **f**. very scared B cells and specially $T\text{-}bet^{+}CD20^{+}$ B cells in synovial tissues of OA. Globally, we found enrichment for the populations of interest in the biopsies of RA inflamed synovium, compared to specific populations in OA synovia. Images were acquired at 200x magnification. Scale bar is 100 μm.



Supplementary Figure 10 – Granzyme expression and cytokine production by synovial tissue CD8 T cells.

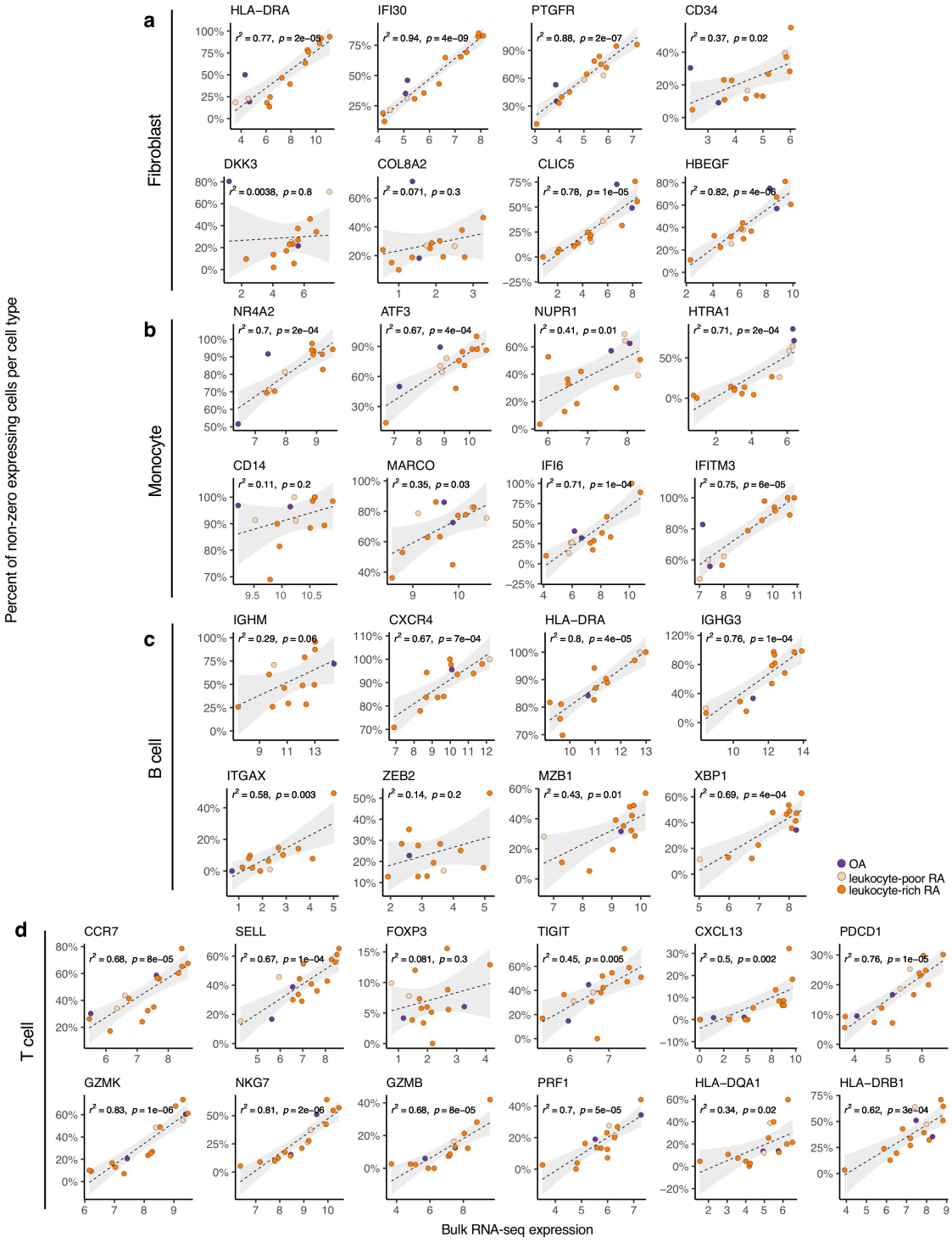
a. RA synovial tissue samples were disaggregated, stained for surface markers and intracellular granzyme B (GZMB) and granzyme K (GZMK), and analyzed by flow cytometry. Shown are plots of GZMB versus GZMK expression by CD8 T cells from three representative tissue specimens out of eight total tissues analyzed. **b**. GZMK and GZMB expression patterns by HLA-DR+ CD8 T cells. **c**. IFNγ production by CD4 and CD8 T cells from RA synovial tissue, measured by intracellular flow cytometry after stimulation with PMA/ionomycin. Cells from the same synovial tissue sample are connected by a line. (one–tailed Student's *t*-test p = 0.028, t-value = 2.1, df = 10.94). **d**. TNF production by CD4 and CD8 T cells from RA synovial tissue



Supplementary Figure 11 – Bulk RNA-seq data analysis.

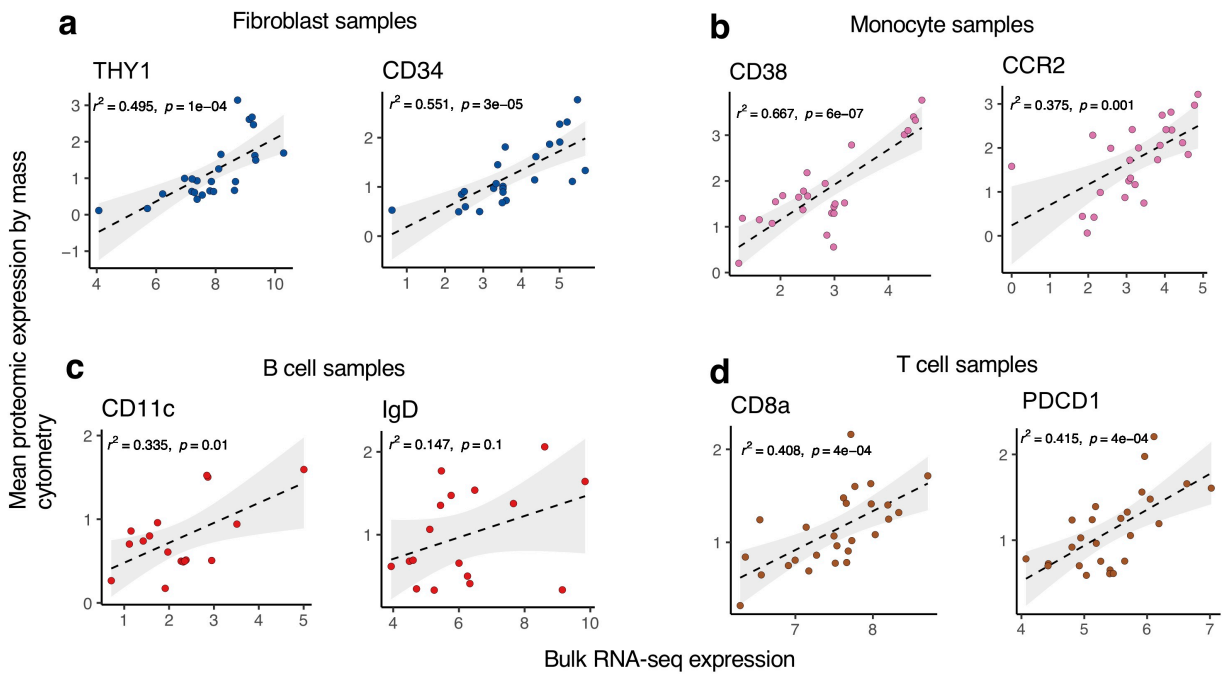
a. Quality control of bulk RNA-seq samples. Common genes are defined as the set of genes detected with at least 1 mapped fragment in 95% of the samples (13,041 genes). X-axis is the number of cells for each bulk RNA-seq sample. Y-axis is the percentage of detected common genes for each sample. We discarded 25 low quality samples that have less than 99% (dashed line) of common genes detected, resulting 167 post-QC samples in all. **b**. PCA analysis on all the post-QC samples shows that most of the variance in the bulk RNA-seq data is due to cell type. **c**. Cell type marker genes show that there is no obvious contamination in the bulk RNA-seq data. **d-g**. PCA analysis on samples from each cell type. The samples from leukocyte-rich

Supplementary Figure 11 (continued) RA appear distinct from leukocyte-poor RA and OA samples. This difference in transcriptional signatures in inflamed tissue is largely determined by altered cellular composition. **h-i.** Distribution of significantly enriched GO terms in leukocyte-rich RA by GSEA. Leukocyte-rich fibroblasts and monocytes share the common pathways of Type I interferon and inflammatory response. **j.** Correlation between bulk RNA-seq genes and immune cell type abundances in RA synovial fibroblasts. Integrating bulk RNA-seq samples from fibroblasts with multiple cell type flow gates reveals that T cells, B cells, and monocytes that are abundant in RA synovial tissue directly influence the expression of fibroblasts in the RA synovium.



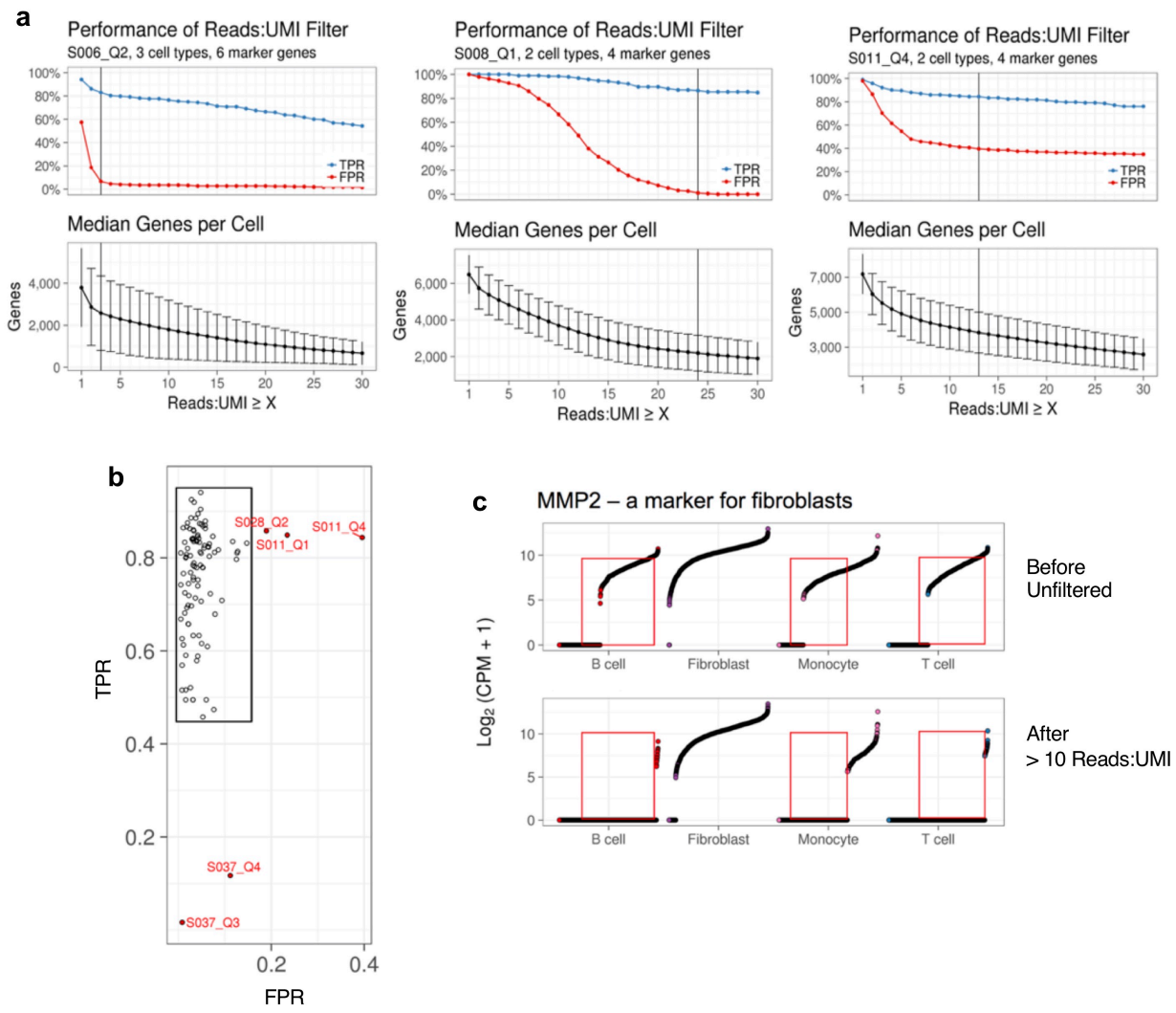
Supplementary Figure 12 – Correlation between bulk RNA-seq expression and proportion of non-zero expressing cells on scRNA-seq cluster markers.

Supplementary Figure 12 (continued) The expression level of a gene in a bulk RNA-seq sample can indicate the abundance of cells expressing that gene in the bulk sample. In other words, it is possible to infer the abundance of some cellular subpopulations from bulk RNA-seq data. We depict two marker genes per scRNA-seq cluster and show the bulk RNA-seq expression (x-axis) is correlated with the percent of non-zero expressing cells over the total number of cells (y-axis) for the overlapped **a**. fibroblast samples, **b**. monocyte samples, **c**. B cell samples, and **d**. T cell samples. Pearson's R-squared and p value are given using the limma R package for each correlation scenario. The grey zone represents 95% confidence level interval for predictions from a linear model.



Supplementary Figure 13 – Correlation between mean proteomic expression by mass cytometry and transcriptomic expression by bulk RNA-seq on the overlapped samples.

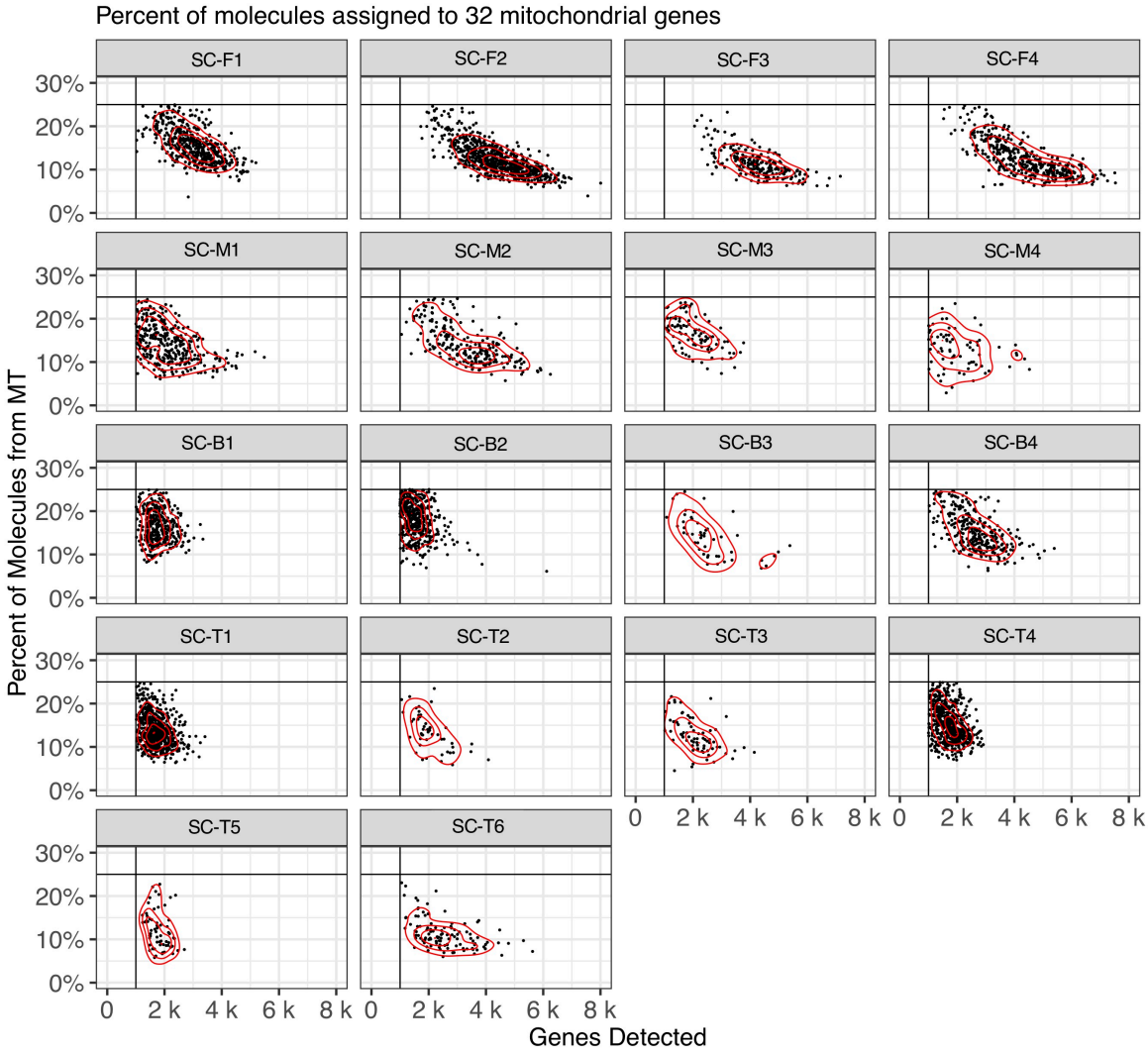
Two typical protein/gene markers per cell type were shown for **a**. fibroblast samples on THY1 and CD34, **b**. monocyte samples on CD38 and CCR2, **c**. B cell samples on CD11c and IgD, and **d**. T cell samples on CD8a and PDCD1. Pearson's R-squared and p value are given using the limma R package for each correlation scenario. The grey zone represents 95% confidence level interval for predictions from a linear model.



Supplementary Figure 14 - Dynamic filtering strategy for scRNA-seq quality control.

We selected 2 marker genes expected to be exclusively expressed in each of the 4 cell types: *PDGFRA* and *ISLR* for fibroblasts, *CD2* and *CD3D* for T cells, *CD79A* and *RALGPS2* for B cells, and *CD14* and *C1QA* for monocytes. We counted nonzero expression of these genes in the correct cell type as a true positive and nonzero expression in the incorrect cell type as a false positive. **a**. Estimated optimal thresholds for reads per unique molecular identifier (UMI) shown for three example quadrants (Q2, Q1, Q4) of three scRNA-seq plates (S006, S008, S011). The reader per UMI threshold determines which UMIs we discard. By discarding UMIs with few supporting reads, we can reduce the false positive rate (FPR), but this comes at the cost of also reducing the true positive rate (TPR). **b**. After selecting the optimal threshold that maximizes the ratio of TPR to FPR for each quadrant, we visualize the FPR and TPR for each quadrant. We noticed that some quadrants had poor performance even after optimizing the reads per UMI, so we discarded the cells from the red outlier quadrants. **c**. An example of gene expression for *MMP2* across all cells before and after filtering the reads per UMI threshold. This matrix metalloproteinase is known to be expressed by fibroblasts, but not by other cell types, so we did not expect to see *MMP2* expression in cell types other than fibroblasts. After applying the

Supplementary Figure 14 (continued) reads per UMI threshold to discard UMIs that are probably contaminants, *MMP2* gene expression is closer to our *a priori* expectation: very few B cells, monocytes, and T cells express this gene. Note that we did not use *MMP2* to determine the optimal reads per UMI threshold.



Supplementary Figure 15 – Assessment of quality of scRNA-seq data for each identified cluster.

We excluded cells with less than 1000 genes detected (at least 1 read). We also excluded cells with more than 25% of UMIs coming from 32 mitochondrial genes. This figure shows the post-QC single cells used in our scRNA-seq analyses.

Supplemental Table 1. Clinical characteristics of 51 recruited patients.

		OA (n=15)	**leukocyte-poor RA** (n=17)	**leukocyte-rich RA** (n=19)
Demographic variables	Age, mean	71	64.2	57.3
	(Range)	(64-81)	(42-79)	(36-71)
	Females, n (%)	10 (66.7)	15 (82.4)	14 (73.7)
RA-related variables	Mean years of disease duration (range)		15.7 (<1-51)	5.5* (<1-29)
	RF positive, n (%)		8 (47.1)	16* (84.2)
	CCP positive, n (%)		10 (58.8)	14 (73.7)
DMARDs	Prednisone, n (%)		10 (55.6)	4* (22.2)
	Methotrexate, n (%)		7 (41.2)	3 (15.8)
	TNFi, n (%)		4 (23.5)	2 (10.5)
	Rituximab, n (%)		0 (0)	1 (5.3)
	Abatacept, n (%)		1 (5.9)	1 (5.3)
	Tofacitinib, n (%)		2 (11.8)	1 (5.3)

DMARDs = Disease-Modifying Antirheumatic Drugs.
TNFi = TNF inhibitors (infliximab, etanercept, adalimumab, golimumab).
RhF = Rheumatoid Factor. CCP = Cyclic Citrullinated Peptide.
*Significant p-value between leukocyte-poor RA and leukocyte-rich RA.

Supplemental Table 2. Antibody staining and fixation of mass cytometry panel.

marker	**clone**	Metal	Dilution
CD45	HI30	141Pr	1:100
CD19	HIB19	142Nd	1:100
RANKL	MIH24	143Nd	1:50
CD64	10.1	144Nd	1:100
CD16	3G8	145Nd	1:100
CD8a	RPA T8	146Nd	1:100
FAP	Poly	147Sm	1:50
CD20	2H7	148Nd	1:100
CD45RO	UCHL1	149Sm	1:100
CD38	HIT2	150Nd	1:100
CD279/PD-1	EH12.2H7	151Eu	1:100
CD14	M5E2	152Sm	1:100
CD69	FN50	153Eu	1:100
CD185/CXCR5	J252D4	154Sm	1:100
CD4	RPA T4	155Gd	1:100
Podoplanin	NC-08	156Gd	1:100
CD3	UCHT1	158Gd	1:100
CD11c	Bu15	159Tb	1:100
CD307d/FcRL4	413D12	160Gd	1:100
CD138	MI15	161Dy	1:100
CD90	5.00E+10	162Dy	1:50
CCR2	K036C2	163Dy	1:100
Cadherin 11	3C10	164Dy	2:25
FoxP3	PCH101	165Ho	1:50
CD34	581	166Er	1:100
CD146/MCAM	SHM-57	167Er	1:50
IgA	9H9H11	168Er	1:100
ICOS	C398.4A	170Er	1:100
CD66b	G10F5	171Yb	1:100
IgM	MHM-88	172Yb	1:200
CD144/VE-Cadherin	BV9	173Yb	1:100
HLA-DR	L243	174Yb	1:100
IgD	IA6-2	175Lu	1:100
CD106/VCAM-1	STA	176Yb	1:100

Supplemental Table 3: Identified mass cytometry populations with proportion of cells from each disease cohort and on tailed FDR q value.

[data shown in separate file: Appendix_II_Supplementary_Table_3.pdf]

Supplemental Table 4: Top 20 marker genes for each single-cell RNA-seq cluster.

[data shown in separate file: Appendix_II_Supplementary_Table_4.pdf]