

Transcriptomics of the synovium in rheumatoid arthritis

A dissertation presented

by

Kamil Slowikowski

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biomedical Informatics

Harvard University

Cambridge, Massachusetts

April 2019

©2019 – Kamil Slowikowski

All rights reserved.

*Transcriptomics of the synovium in rheumatoid arthritis***Abstract**

Rheumatoid arthritis (RA) is a debilitating chronic inflammatory disease that affects millions of people worldwide. Disease pathogenesis involves the synovium, a thin membrane encapsulating the joint, mostly composed of cells resembling fibroblasts and macrophages. The synovium proliferates and forms a pannus of inflamed tissue that persistently recruits immune cells, degrades collagen, and activates bone-resorbing osteoclasts. The infiltrating immune cells interact with the stromal tissue cells, leading to synovial hypertrophy, and ultimately to joint destruction. The cellular composition and molecular functions of stromal fibroblast cells in the synovium are relevant to understanding disease progression. With small quantities of patient tissues, the latest technologies allow us to assay transcriptomics genome-wide at high resolution. Transcriptomics analysis helps us determine the distinct functions of cellular populations in the synovium and reveal gene regulatory factors activated by cytokines elevated in the disease state. Further understanding of basic biology in these cells may reveal fibroblast-specific targets for the treatment of inflammatory diseases mediated by fibroblasts.

Here, we use computational analyses of bulk tissue and single cell transcriptomic data, and also genetics data, to advance understanding of mechanisms underlying RA. We integrate genome-wide association study (GWAS) results with gene expression data to understand which cell types express the genes that are associated with risk of developing disease. Next, we analyze

transcriptomics data collected from synovial fibroblasts obtained from fresh tissue discarded during joint replacement surgery. Single-cell gene expression profiles reveals the cellular population structure of fibroblasts in the synovium. Within the sublining layer, RA patients have an overabundant fibroblast population negative for CD34 and positive for CD90. The heterogenous synovial cellular populations produce a mixture of signaling molecules that synergistically influence the positive feedback loop between local tissue cells and infiltrating immune cells. To understand the regulation of the synergistic response to tumor necrosis factor (TNF) and interleukin 17 (IL-17A), we use high resolution transcriptomics with time series, dose response, and gene silencing. Our analyses predict and our experiments validate that CUX1 is a key mediator for fibroblast expression of *CXCL1*, *CXCL2*, *CXCL3*. Understanding regulation of transcriptional response to inflammatory factors in synovial fibroblasts has implications for RA as well as other diseases involving chronic inflammation mediated by fibroblasts in connective tissues such as psoriasis. Our approaches for interrogating RA could be applied to these diseases.

Table of Contents

Abstract	iii
Acknowledgments	viii
Citations to previously published works.....	x
Chapter 1 Functional genomics of stromal cells	1
Introduction	3
Fibroblasts mediate inflammation in chronic inflammatory diseases	4
Exploring stromal cell biology with functional genomics	6
Opportunities to advance fibroblast biology with single-cell genomics.....	10
Conclusion.....	13
Chapter 2 Data visualization tools.....	15
Automatically labeling data points	16
Making data accessible to others on the web	20
Gene expression along the T cell innateness gradient	22
Defining inflammatory cell states in rheumatoid arthritis	24
Conclusion.....	28
Chapter 3 Integrating genome wide association study results with gene expression data	29
The SNPSEA algorithm.....	31
Multiple genes implicated by LD	31
Type I error estimates	34
Results for multiple sclerosis, celiac disease, and HDL cholesterol	36
Conclusion.....	50

Chapter 4 Transcriptomics of fibroblast subsets in RA.....	51
Fibroblasts in synovial tissue have distinct surface markers	53
Fibroblast subpopulations have three distinct gene expression signatures	58
Single cell RNA-seq confirmed the presence of three major fibroblast subsets.....	64
Fibroblast subsets localize to specific regions of lymphoid structures in the synovium.....	66
Fibroblast subset proportions are altered in RA.....	69
Transcriptomics predicts distinct fibroblast functions.....	73
Conclusion.....	75
Chapter 5 Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry	86
Generation of parallel mass cytometric and transcriptomic data from synovial tissue	88
Summary of computational data integration strategy to define cell populations.....	93
Flow cytometry features define a set of RA synovia that are leukocyte-rich	96
Single-cell RNA-seq analysis reveals distinct cell subpopulations.....	102
Distinct synovial fibroblasts defined by cytokine activation and MHC II expression	106
Activation states define heterogeneity among synovial monocytes.....	116
Heterogeneity in synovial CD4 and CD8 T cells defined by effector functions.....	120
Autoimmune-associated B cells expanded in RA synovium by single-cell RNA-seq	129
Inflammatory pathways and effector modules revealed by global single cell profiling.....	133
Conclusion.....	138
Chapter 6 Transcription factors mediating response to TNF and IL-17A.....	165
TNF and IL-17A synergistically activate IL-6 and MMP-3 production	168

IL-17A requires TNF to induce a gene expression response in fibroblasts.....	172
Defining the transcriptional response to TNF and IL-17A in fibroblasts with a high resolution time series and dose response study	176
Transcriptional response reveals genes selectively induced by TNF and IL-17A costimulation or TNF alone	179
Transcription factors underlying the synergy between TNF and IL-17A.....	185
Transcriptomics time series after silencing putative mediators	190
I κ B ζ binds NF- κ B p65 and p100/52, and mediates the response to TNF and IL-17A.....	195
CUX1 and I κ B ζ mediate neutrophil and monocyte recruitment.....	196
Genes induced by TNF and IL-17A are highly expressed in RA synovial tissues	198
Conclusion.....	201
Chapter 7 Conclusion	211
Summary of results.....	212
Assaying transcriptomic response to drugs will advance precision medicine	214

Acknowledgments

Funding support for much of the work presented here was provided by the Ruth L. Kirschstein National Research Service Award (F31AR070582) from the National Institute of Arthritis and Musculoskeletal and Skin Diseases.

I want to thank my advisor, Dr. Soumya Raychaudhuri for being a supportive, thoughtful, and patient mentor. He taught me about genetics, statistics, writing, communication, and he helped me to grow as a scientist. As an early member of his laboratory, I have had the pleasure to work and socialize with talented scientists from all over the world. In particular, I would like to thank my colleagues for giving me the opportunity to work with them: Dr. Xinli Hu, Nikola Teslovich, Hyun Kim, Susan Hannes, Jennie Pouget, Dr. Gosia Trynka, Dr. Harm-Jan Westra, Dr. Rachel Knevel, Dr. Buhm Han, Dr. Chikashi Terao, Dr. Emma Davenport, Dr. Dorothee Diogo, Dr. Maria Gutierrez-Arcelus, Dr. Samira Asgari, Dr. Ilya Korsunsky, Dr. Yang Luo, Dr. Fan Zhang, Dr. Yuriy Baglaenko, Dr. Kazuyoshi Ishigaki, Tiffany Amariuta, Chamith Fonseka, and Joseph Mears. Also, I would like to thank Sushil Upadhyayula and Gopal Vashishtha for giving me the opportunity to serve as a mentor.

I also want to thank Dr. Michael B. Brenner, who kindly mentored me in his lab and warmly invited me to spend time with his lab members in his home. It has been wonderful to see the collaboration between Dr. Raychaudhuri and Dr. Brenner flourish. By working together with Dr. Brenner and his talented team, all of us have achieved far more than we could have alone. In particular, I am grateful to Dr. Fumitaka Mizoguchi, Dr. Kevin Wei, Dr. Deepak Rao, Dr. Erika Noss, Dr. Hung N. Nguyen, and Dr. Gerald Watts. I would like to thank these outstanding scientists for inviting me to work with them and for teaching me about immunology.

I would also like to thank Dr. Peter Kharchenko for serving as the chair of my preliminary qualifying examination, and I thank his lab members including Dr. Ruslan Soldatov, Dr. Joe Riggs, Dr. Nikolas Barkas, and Dr. Jean Fan, for many fun gatherings at lunch and happy hour. I want to especially thank Dr. Jean Fan for believing in me and supporting me.

I thank all of the people who have made the Bioinformatics and Integrative Genomics program at Harvard. I want to thank all of them, but in particular Dr. Susanne Churchill, Dr. Peter Park, Dr. Isaac Kohane, Katherine Flannery, and Cathy Haskell for maintaining the bioinformatics training program. I want to thank all of the people whose names I do not know, but who have consistently supported all of the facilities and infrastructure that I have used for several years. I also would like to thank Hyun-Il Paik for answering my questions about the computing infrastructure and inviting me to speak about ggrepel. I want to thank Dr. Alkes Price for teaching the best course I had the privilege to attend at Harvard in medical and population genetics, and for his patience in explaining difficult concepts. I also want to thank Dr. Jennifer Montana and all of the students at The Innovation Institute for giving me the opportunity to serve as an instructor in computer programming.

Finally, I would like to thank my grandparents Dr. Marianna Szwarc and Dr. Stefan Szwarc for taking a leap into the unknown and making all of the sacrifices in their lives that allowed me to go farther than they had imagined. I also would like to thank my parents Mariola and Krzystof for all their love and support, and my brother Karol for being an unwavering ally.

Citations to previously published works

Large portions of Chapter 1 appeared in the following publication:

- **Slowikowski K**, Wei K, Brenner MB & Raychaudhuri S. Functional genomics of stromal cells in chronic inflammatory diseases. *Current Opinion in Rheumatology*. 30, 65–71 (2017). doi: 10.1097/BOR.0000000000000455

Large portions of Chapter 3 appeared in the following publication:

- **Slowikowski K**, Hu X & Raychaudhuri S. SNPSEA: an algorithm to identify cell types, tissues, and pathways affected by risk loci. *Bioinformatics* **30**, 2496–2497 (2014). doi: 10.1093/bioinformatics/btu326

Large portions of Chapter 4 appeared in the following publication:

- Mizoguchi F*, **Slowikowski K***, Wei K, Marshall JL, Rao DA, Chang SK, Nguyen HN, Noss EH, Turner JD, Earp BE, Blazar PE, Wright J, Simmons BP, Donlin LT, Kalliolias GD, Goodman SM, Bykerk VP, Ivashkiv LB, Lederer JA, Hacohen N, Nigrovic PA, Filer A, Buckley CD, Raychaudhuri S, Brenner MB. Functionally distinct disease-associated fibroblast subsets in rheumatoid arthritis. *Nature Communications*. **9**, 789 (2018). doi: 10.1038/s41467-018-02892-y

Large portions of Chapter 5 appeared in the following publication:

- Zhang F*, Wei K*, **Slowikowski K***, Fonseka CY*, Rao DA*, Kelly S, Goodman SM, Tabechian D, Hughes LB, Salomon-Escoto K, Watts GFM, Apruzzese W, Lieb DJ, Boyle DL, Mandelin II AM, Accelerating Medicines Partnership: RA Phase 1, AMP RA/SLE, Boyce BF, DiCarlo E, Gravallese EM, Gregersen PK, Moreland L, Firestein GS, Hacohen N, Nusbaum C, Lederer JA, Perlman H, Pitzalis C, Filer A, Holers M, Bykerk VP, Donlin LT, Anolik JH, Brenner MB, Raychaudhuri S. Defining

inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. *bioRxiv*. (2018). doi: 10.1101/351130

Large portions of Chapter 6 appeared in the following publication:

- **Slowikowski K***, Nguyen HN*, *et al.* CUX1 and IκBζ mediate the synergistic inflammatory response to TNF and IL-17A in stromal fibroblasts. *bioRxiv* 571315 (2019). doi:10.1101/571315

* These authors contributed equally

Chapter 1

Functional genomics of stromal cells

Stroma is a broad term referring to the connective tissue matrix in which other cells reside. It is composed of diverse cell types with functions such as extracellular matrix maintenance, blood and lymph vessel development, and effector cell recruitment. The tissue microenvironment is determined by the molecular characteristics and relative abundances of different stromal cells such as fibroblasts, endothelial cells, pericytes, and mesenchymal precursor cells. Stromal cell phenotypic heterogeneity is due to embryonic developmental lineage, stages of differentiation to other cell types, and activation states. The interaction of the immune system with the stroma is critical to wound healing, cancer, and a wide range of inflammatory diseases. Here we review recent studies of rheumatic diseases that use functional genomics and single-cell technologies to identify and characterize stromal cell types associated with pathogenesis.

High dimensional strategies using mRNA sequencing, mass cytometry, and fluorescence activated cell-sorting with fresh primary tissue samples are already producing detailed views of what is happening in diseased tissue in rheumatoid arthritis, inflammatory bowel disease, and cancer. Fibroblasts positive for CD90 (THY1) are enriched in the synovium of rheumatoid arthritis patients, and also in prostate cancer tumors. Single cell RNA-seq offers potential for even more discoveries about the stroma in the near future.

Stromal cells form the microenvironment of inflamed and diseased tissues. Functional genomics is producing an increasingly detailed view of subsets of stromal cells with pathogenic functions in rheumatic diseases and cancer. Future genomics studies will discover disease mechanisms by perturbing molecular pathways with chemokines and therapies known to affect patient outcomes. Functional genomics studies with large sample sizes of patient tissues will identify patient subsets with different diseases or treatment responses.

Introduction

The stroma is the background of every tissue in the body, and plays a critical role in supporting the normal epithelium, forming the general architecture of the tissue, and modulating the local microenvironment. While many types of cells constitute the stroma in varying proportions across tissues, fibroblasts are the major constituent of stroma in all tissues. One of the first demonstrations of functional fibroblast heterogeneity was done 40 years ago, showing three distinct phenotypes of fibroblast response to prostaglandin¹. Many studies since then have shown how fibroblasts in different systems support tissue homeostasis and shape immune responses².

Fibroblasts are central to tissue biology because they organize the extracellular matrix in which other cells are embedded, and they communicate with these cells². Since cell functions are dependent on the context in which they are embedded, *in vitro* studies may fail to capture physiologically relevant context essential to tissue functions. A major impediment to discovery is the lack of well-defined cell surface markers that distinguish different functional types of fibroblasts. One marker, the GPI-anchored protein Thy-1 (CD90), has been studied in many tissues. Thy-1 is differentially expressed on different types of fibroblasts in the spleen³, lung^{4,5}, female reproductive system^{6,7}, ocular orbit⁸, liver⁹, and prostate cancer tumors¹⁰.

Functional genomics offers a promising path forward to understand the diversity of fibroblast functions in health and disease. By assessing global measures of all genes, it has the potential to find new factors important for production and degradation of extracellular matrix, development of blood and lymph vessels, wound healing, and communication with leukocytes during inflammation.

Currently, most genomic studies have effectively averaged signals in bulk high dimensional assays across millions of cells with diverse phenotypes, which complicates interpretations about the roles of specific cell types. In other words, differences in cell composition cannot be distinguished from differences in gene regulation at the whole tissue level. More recently, studies using low-input genomics on subpopulations sorted by cell surface markers and studies using single-cell genomics demonstrated that the average signals can be explained by independent contributions from different proportions of functionally distinct single cells. Further, single-cell technologies are directly addressing the challenge of discovering cell surface markers that distinguish functional subsets of cells relevant to tissue biology and disease pathology.

Fibroblasts mediate inflammation in chronic inflammatory diseases

The role of stromal cells in orchestrating local inflammatory response is becoming increasingly appreciated. Studies are suggesting that the diversity of stromal cells across anatomical sites may contribute to location-specific disease development^{11–13}. There are consistent patterns of anatomical distribution for diseases like RA, IBD, psoriasis, ankylosing spondylitis, and various types of cancers. For example, ankylosing spondylitis often affects lower extremities or the spine, while RA affects small distal joints of the hands and feet¹¹. These anatomical patterns suggest a few possibilities: site-specific local cell types, local responses to systemic signals, or environmental factors like mechanical stress that affect cells locally¹¹. One recent study shows how expression of developmental HOX genes can influence the TNF induced activation of inflammatory molecular pathways in knee synovial fibroblasts¹².

The synovium is a thin membrane composed mostly of fibroblasts and macrophages that surrounds the joint capsule and contains the synovial fluid. In RA, it is the critical site of chronic inflammation¹⁴. A healthy synovium has a 1 to 2 cell thick lining layer and a sublining layer with blood vessels, lymph vessels, fibroblasts, collagen fibers, nerve fibers, and few leukocytes¹⁵. In RA, the inflamed synovium has a hyperplastic lining layer, dramatic expansion of blood vessels in the sublining, and dense infiltration by inflammatory leukocytes¹⁵. While all of the cells play some role in chronic inflammation, synovial fibroblasts orchestrate joint destruction through production of proteinases that degrade the cartilage and cytokines that activate bone resorbing osteoclasts¹⁶.

Autoimmune diseases, such as rheumatoid arthritis, progress from an initiation phase of generating autoantibodies to an effector phase of continuous feedback between stromal cells and leukocytes^{17,18}. In RA, a vicious cycle of inflammation is fueled by fibroblasts in the synovium producing chemokines that recruit leukocytes to damaged tissue, where these partners then continuously stimulate each other to maintain chronic inflammation^{19,20}. A recent study shows that there is a similar feedback loop between leukocytes and stromal cells in IBD. In IBD fibroblasts are activated by TNF and other chemokines, causing them to produce IL6 and other chemokines to activate leukocytes²¹. In contrast to the cartilage destruction in RA, the fibroblasts of the gastrointestinal tract are responsible for excessive extracellular matrix deposition, causing fibrosis²².

As a result of the feedback loops involving communication between stromal cells and leukocytes, fibroblasts undergo long-lasting epigenetic changes, become resistant to apoptosis, and maintain a proliferative, destructive, invasive, and migratory phenotype²³⁻²⁵. Cultured synovial fibroblasts from inflamed joints of RA patients, but not from healthy joints, express

major histocompatibility complex (MHC) class II molecules in response to IFN-gamma²⁶. This is consistent with a similar demonstration that lung fibroblasts within the human lung express HLA-DR and costimulatory molecules, modulating the CD4 T cell response to pathogens during infection²⁷. Hence, many synovial and lung fibroblasts are actually antigen presenting cells: they are able to extract and present antigens to CD4+ T cells^{26–28}.

Tissue Source	Molecular Assay	Resolution	Study Design
Cell lines Whole tissue • Biopsy • Surgical specimen Phenotypically sorted cells	Genomics • DNA sequencing • Transcriptomics (microarray, RNA-seq, etc.) • Epigenetics (ATAC-seq, ChIP-seq, etc.) Proteomics • Fluorescence-activated cell sorting • Mass cytometry (CyTOF)	Bulk (thousands to millions of cells) Low Input (hundreds of cells) Single-cell	Exploratory (small n) Case control • Disease vs healthy • Disease vs disease • Responder vs nonresponder Case only progression • Early or late disease • Drug response

Figure 1.1 Functional genomics in stromal cells.

All studies involve careful considerations of the costs and benefits for selection of a tissue source, molecular assay, desired resolution, and study design. Critically, effective application of high dimensional molecular assays requires effective strategies to disaggregate whole tissue sources without disturbing biological signal.

Exploring stromal cell biology with functional genomics

Functional genomics, or discovery of biological insights through genome-wide assays, has revolutionized the way in which we understand cell biology and tissue physiology. While there

are several approaches to apply genomics to a target cell type, here we summarize some strategies and glimpse the future of stromal cell genomics (**Figure 1.1**).

Three main approaches are used to apply functional genomics to understand stromal cells in rheumatic disease. First, whole tissue transcriptomics by microarray, RNA-seq, or DNA methylation is used to contrast cases and controls. This type of analysis highlights molecular markers of disease such as cytokines that are elevated in the disease state, or cell surface markers of cell types that are differentially abundant. Second, transcriptomics of purified cell populations by fluorescence activated cell sorting (FACS) after tissue dissociation or laser-dissection reveals the distinct genomic signals underlying the average signals seen in bulk tissues, helping to focus on a smaller set of cell types that are most relevant for disease pathogenesis. Third, the latest single cell RNA-seq techniques can be performed on dissociated tissue samples to provide a completely unbiased view of thousands of cells in patient samples. This overcomes the need for cell surface markers to sort different cell types, and can give a holistic perspective of how abundances and functions of different cell types relate to each other in patient tissues.

To date, many functional genomics studies have been done using cultured cell lines. The high abundance of cellular material from a culture enables a wide range of investigational strategies. However, cell lines are outside the context of the whole tissue, and they are altered by the process of growing *in vitro*. In many instances, studies contrast a small number (<10) of cases and controls, sometimes including perturbations by therapies or gene knockdowns. A major advantage of cultured cell lines derived from patient tissues is that they can be expanded to provide enough RNA, DNA, and other cellular material necessary for most genomics assays. In a study with differential gene expression analysis of microarrays comparing macrophages and fibroblasts from synovial samples from RA and OA, authors found that both cell types are

producing inflammatory cytokines in RA²⁹. POSTN and TWIST1 were identified as key regulators of fibroblast invasiveness in RA, and subsequent siRNA experiments confirmed this finding²⁹. Microarray analysis with OA and RA fibroblasts revealed constitutive upregulation of the TGFB pathway in RA fibroblasts that resulted in greater expression of MMP1³⁰. Genome-wide methylation microarrays and RNA-seq data shows differential methylation and expression of developmental HOX genes between fibroblasts taken from hips and knees from RA and OA individuals¹³. In knee synovial fibroblasts, silencing the *HOTAIR* lncRNA in the HOX cluster resulted in greater expression of constitutive and TNF-induced collagenase MMP1¹². This suggests that the HOX genes not only control stromal cell phenotypes during embryonic development, but also the immunoregulatory phenotypes during disease pathogenesis.

Another common approach is to apply functional genomics to whole tissue samples. Arguably, functional genomics assays with tissue samples are a promising path toward understanding disease mechanisms that may be specific to subsets of cells *in vivo*, and how these mechanisms vary between individuals. However, fresh human derived tissue samples are a more limited resource, and must be used judiciously. For RA, these samples are often derived from joint arthroplasty samples or biopsy samples. In many instances obtaining whole tissue has to be considered in the clinical context of the elective surgical intervention that enabled collection of the tissue. One exciting opportunity for querying tissue samples is the use of research synovial biopsies³¹. While tissue quantities are limited with biopsies, they enable querying tissues under prescribed clinical conditions, and are possible in a trial setting. Transcriptomics by microarrays enabled definition of a rule set for discriminating rheumatoid arthritis from osteoarthritis and healthy controls³². Microarray analysis of RA, OA, and normal donors found that RA patients can be divided into two subsets with high or low expression of PRG4 in the synovium, and low

PRG4 expression is associated with more aggressive stage of disease³³. Pathogenic fibroblast destruction of cartilage, as well as recruitment and retainment of leukocytes is amplified by haematopoietically derived cytokines like TNF and interleukin 6 (IL6). Transcriptomics studies with synovial biopsy tissues revealed that treatment with tocilizumab, methotrexate, adalimumab, and rituximab reduces the mRNA levels of IL6 and many other chemokines and T cell activation genes after 12 weeks of therapy³⁴.

The power of whole tissue transcriptomic profiling was recently revealed in a study of a cohort of patients with Crohn's disease (CD) and ulcerative colitis (UC); in that study West et al found a TNFi-resistant pathway specific to gut-resident stromal cells²¹. Analysis of public transcriptomics data showed that oncostatin M (OSM) is more highly expressed in biopsy tissues from IBD patients. The authors found that OSM is mostly derived from hematopoietic cells, and stromal fibroblasts have highest expression of the OSM receptor (OSMR) in the gut. These PDPN+ fibroblasts are highly abundant in inflamed tissue from Crohn's disease (CD) and ulcerative colitis (UC) patients, and they produce inflammatory chemokines in response to TNF and OSM stimulation. This study highlights the novel system of leukocyte-stromal communication behind the chronic inflammation in IBD.

The OSM result may translate to human therapies. Notably, OSM has phase 1 and 2 trials for RA showing that humanized anti-OSM monoclonal antibodies are well tolerated, and OSMR-deficient mice are healthy and viable. Large scale functional genomics studies with biopsies can help to identify different subsets of patients. In a large scale study with 210 Crohn's disease patients and 35 healthy controls, CD patients were clearly distinguished from healthy controls by biopsy gene expression of 29 genes enriched with genetic risk variants³⁵. Further, the 29 gene signal was also able to identify patients more likely to progress to complicated disease.

Opportunities to advance fibroblast biology with single-cell genomics

Transcriptomic profiling of single cells is a powerful technique to reveal cellular and molecular heterogeneity in tissues with complex and dynamic cellular compositions. These technologies are being applied at scale in efforts such as the human single-cell atlas, with the goal of providing a reference map of all common types of cells in the human body³⁶. These types of large scale studies will be valuable to all researchers. The ability to visualize molecular pathways at a single cell level can transform our understanding of cell identity and function. Single cell RNA-seq overcomes the need for cell surface markers to sort different cell types, and can give a holistic perspective of how abundances and functions of different cell types relate to each other in patient tissues.

Recent advances in immunoprofiling and transcriptomic assay technologies has created an opportunity to assay sorted stromal cells directly from tissues³⁷. Advances in sequencing library construction have lowered the requirements for genomics assays to 10s or 100s of cells per sample³⁸.

Discoveries from single-cell genomics will include new cell types specific to anatomical sites or disease states, dysregulations of molecular pathways exclusive to particular cell types, and genetic effects on those pathways³⁹. One recent study demonstrates that genetic effects on gene expression can be masked when assaying total peripheral blood mononuclear cells (PBMCs). This single-cell eQTL analysis shows that the same variant influences expression of TSPAN13 in CD4 T cells, but not in dendritic cells or other cell types⁴⁰.

Investigators have begun applying single cell technologies to query tissues in pathological conditions. Critical to successful application of single cell technologies include preserving cell viability and effective dissociation of tissue into single cell suspension. Recent advances in droplet-based platforms for single cell RNA-seq have enabled analysis of large number of single cells from a single donor^{41,42}.

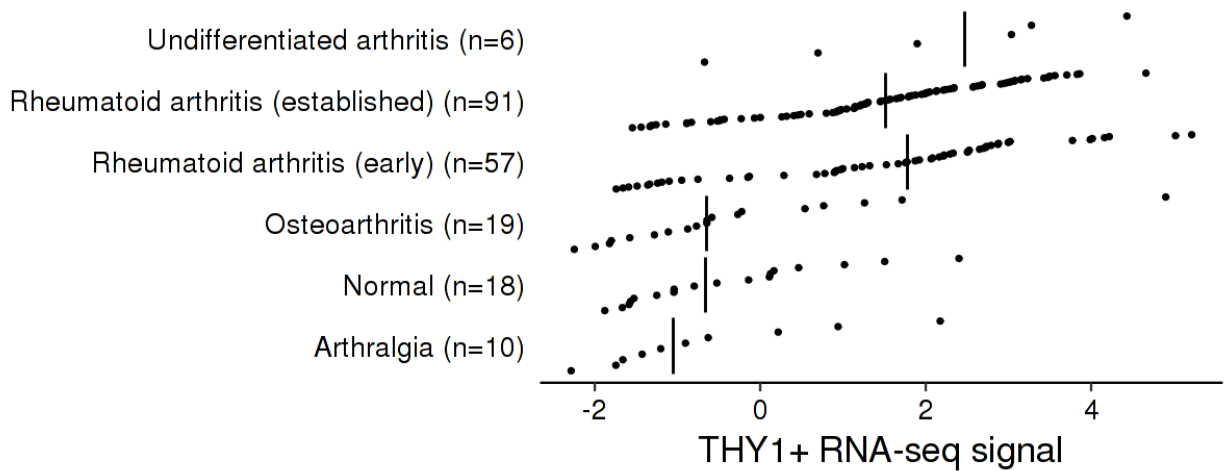


Figure 1.2. THY1+ synovial fibroblast gene expression in RA.

THY1⁺ synovial fibroblast gene expression in RA. First, we selected 4 genes from single-cell RNA-seq data⁴³ that are highly expressed in sublining CD34⁺THY1⁺ fibroblasts (*THY1*, *POSTN*, *TGFBI*, *CCL2*) and 4 genes highly expressed in lining CD34⁺THY1⁻ fibroblasts (*CLIC5*, *TSPAN7*, *SLC2A12*, *SELP*). Next, we took the synovial biopsy RNA-seq expression data from a recent study (GSE89408), computed the average of each gene set, and plotted the ratio of the gene sets. Vertical lines are medians. Early RA is significantly different than OA (Wilcoxon rank sum test) ($P = 2.0e-4$), and established RA is also significantly different than OA ($P = 1.8e-5$). This result is consistent with the idea that synovial biopsy RNA-seq signal is explained by an overabundance of the THY1⁺ fibroblast population in RA patients relative to OA.

Two independent studies have identified Thy-1 positive and negative fibroblasts in the synovium^{41,44}. By applying cell sorting, low-input RNA-seq, and single-cell RNA-seq to synovial biopsies, they identified different functional subsets of fibroblasts⁴⁴. Notably, the Thy-1 positive fibroblasts were found to be significantly overabundant in the synovial tissue of rheumatoid arthritis (RA) relative to osteoarthritis⁴⁴. We checked if we could identify overabundance of Thy-1 positive fibroblasts in a recent study of synovial biopsies⁴⁵, and found that RA has higher relative expression of Thy-1 positive fibroblast genes than Thy-1 negative fibroblast genes (**Figure 1.2**). Expression of Thy-1 influences multiple aspects of fibroblast biology such as propensity to differentiate into adipocytes⁴⁶, and this protein is also involved in neuronal development and oncogenesis⁵.

Despite these exciting initial applications, little is known about the complexity of stromal populations, and single cell genomics offers a powerful way to capture the heterogeneity and compare gene expression of membrane proteins, transcriptional regulators, and signaling pathways that might reveal phenotypic and functional subsets of fibroblasts. Building on the foundation of the human single-cell atlas, new studies will focus on relevant tissues from patients. In the coming years, each new study of single cells from chronic inflammatory diseases will give deeper insights about the commonalities and differences between them. For example, the Accelerating Medicines Partnership (AMP) is a consortium where one focus is generating large scale datasets that include single cell genomics from disease affected tissues in RA and SLE patients. These datasets will show how cell type abundances and functions change over the course of disease and in response to therapy. We envision many future studies will take advantage of the new clarity provided by single-cell technologies.

Conclusion

In the past few years, oncologists have moved from general chemotherapy to personalized treatment based on genetic measurements. Treatment of rheumatic disease can follow a similar path as ongoing and future studies pave the way forward⁴⁷. For example, measuring OSM gene expression in IBD patients may promise to predict successful response to anti-TNF therapy²¹. As synovial biopsy tissue acquisition becomes more common, researchers will have more opportunities to use functional genomics technologies to uncover disease relevant molecular phenotypes. Researchers will assay transcriptomics, DNA methylation, chromatin accessibility, histone modifications, and other molecular signals to find associations with disease activity and response to therapies. Armed with this information, biopsy data may be then used to predict optimal treatments for patients. The advancement of molecular phenotyping will help to deliver on the promise of precision medicine, where therapies are adjusted for maximum benefit to each individual patient.

Functional genomics has the potential to reveal common and differential sets of molecular pathways across multiple anatomical sites and diseases. Stromal biology has thus far been mostly focused on single systems like cancer associated fibroblasts from prostate cancer, gut fibroblasts from IBD patients, or synovial fibroblasts from inflamed joints. Many pathological perturbations of stromal cells are likely to be different across diseases. However, careful analysis of the genome-wide measurements from functional genomics techniques can reveal the commonalities in molecular phenotypes across those diseases. For example, the cytokine gene expression profiles of prostate cancer associated fibroblasts are similar to synovial fibroblasts from rheumatoid arthritis joints as well as fibroblasts from the gut.

We envision that synovial biopsy samples will be treated *in vitro* to study responses to known and novel therapies. We expect that hematopoietic and stromal cells from treatment responders might behave differently than cells from non-responders, and future studies will find the molecular basis for these differences. These findings will give unprecedented resolution about pathogenic functions in the synovium, and will suggest new diagnostic assays and therapeutic targets that might be addressed by existing or future therapies¹⁵.

Chapter 2

Data visualization tools

We describe how a physical simulation of attraction and repulsion between objects can be used to add text annotations to a figure and automatically position them without any overlaps. This work is realized in an R package called `ggrepel`, which has quickly become my most popular software. Each time I start reading a new scientific publication, I now expect to find a figure that makes use of `ggrepel`. Engaging with the R community, accepting pull requests from contributors, and fixing bugs reported by users has been a wonderful experience, and one that I hope to continue for a very long time.

Next, we look at the issue of sharing data with a wide audience and empowering the audience to interact and understand the data. Static images are the default display format for most scientific output. Some journals also make use of videos, either by embedding the videos directly in the HTML web page so viewers can stream them, or by attaching large video files for readers to download to their own computers before viewing. Another medium, the interactive figure, is emerging in journals like Distill (<https://distill.pub>) as well as text books like Immersive Math (<http://immersivemath.com>). I encourage the reader to experience the interactive figures for themselves by visiting the links.

Automatically labeling data points

In a figure, one might plot dots on a two dimensional canvas, where the x-axis represents one quantity of interest, and the y-axis represents another. This type of figure is ubiquitous. Sometimes, one might be interested to call special attention to some of the data points in the figure. While color, shape, and size are appropriate ways to call attention to data points, these

visual features need to be explained in an accompanying legend next to the figure. This allows the reader to understand the meaning of each color, each shape, and each size.

An alternative approach is to directly label some of the data points with text. This approach has the advantage that text is immediately readable and recognizable, so there is not the same need for a legend. This direct labeling approach can be used to convey more information than colors, shapes, and sizes. For example, **Figure 2.1** shows the identity of 11 cars with text labels that indicate the make and model of each car. An alternative figure might use 11 different colors and provide a legend that shows which color indicates which car. This kind of figure would be more difficult for a reader to understand, because they would need to match the color in the legend to the color in the figure. This could quickly get out of hand as we run out of colors, or we run out of combinations of colors and shapes. On the other hand, a significant disadvantage of the direct labeling approach is that text labels are difficult to position in a small figure with limited space. Most often, the operation of positioning text labels is better suited for software with a graphical user interface that allows manually positioning elements with the mouse or keyboard. Many software packages do not offer features for customizing the positions of text labels. So, direct labeling can result in an unsatisfactory figure where text labels are obscuring each other, or where text labels are cut off the display, so the viewer cannot read the labels (**Figure 2.1**).

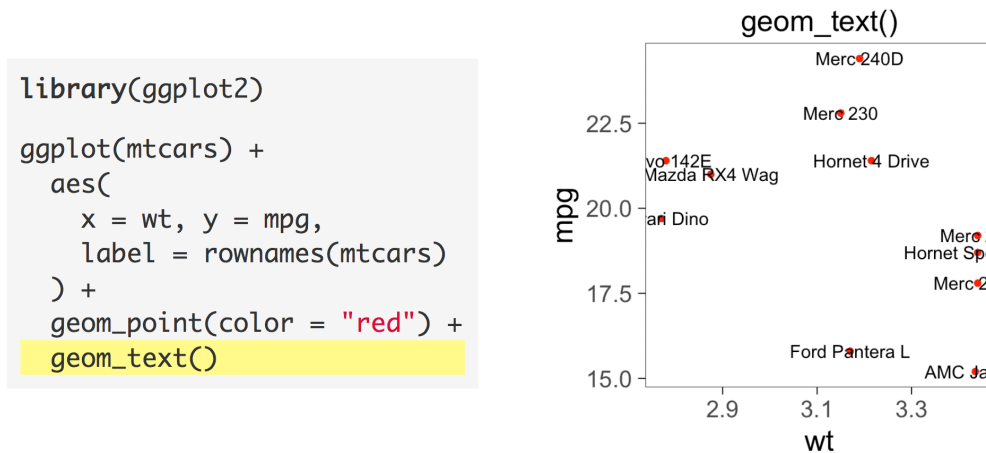


Figure 2.1. Text labels on figures can be difficult to read.

R code and the resulting figure. By default, the `geom_text()` function will position text directly at the coordinates specified by the data. In this case, the x-axis represents the weight of each car, and the y-axis represents the miles per gallon (mpg). Notice that some of the text labels overlap each other (“Volvo 142E” and “Mazda RX4 Wag”). Also notice that many labels are cut off the plotting area, so they are not fully visible.

To address this limitation, I created a software package for R called `ggrepel` (<https://cran.r-project.org/package=ggrepel>). The “gg” prefix stands for Grammar of Graphics, a systematic way to specify the components of a graphic or figure, and I chose this prefix to indicate that `ggrepel` is an extension of the popular R package `ggplot2` by Hadley Wickham *et al.* The word “repel” is supposed to concisely summarize the approach to solving the problem of overlapping text labels: a physical simulation with forces that repel text labels away from each other. By running a physical simulation with objects that can collide and repel one another (**Figure 2.2**), we can very quickly converge, after just a few hundred iterations, to a satisfactory solution where text labels are legible (**Figure 2.3**). By extending `ggplot2`, we gain a lot of flexibility to accommodate the options implemented in `ggplot2` (**Figure 2.4**).

$O(n^2)$ N-body physical simulation

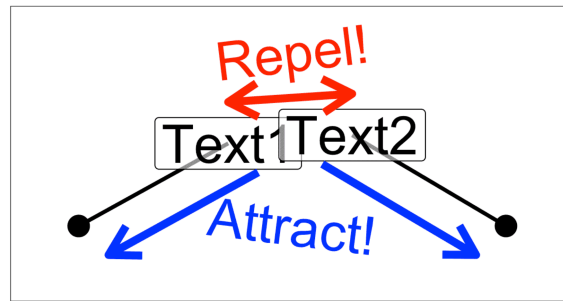


Figure 2.2. The ggrepel algorithm.

The algorithm underlying ggrepel is a naïve $O(n^2)$ physical simulation. We loop through all pairs of text labels and iteratively calculate the repulsive and attractive forces between all of the objects. After running the algorithm for hundreds or thousands of iterations, up to some user-specified limit, the algorithm halts and the resulting layout often meets the requirement of no overlaps.

```
library(ggrepel)

ggplot(mtcars) +
  aes(
    x = wt, y = mpg,
    label = rownames(mtcars)
  ) +
  geom_point(color = "red") +
  geom_text_repel()
```

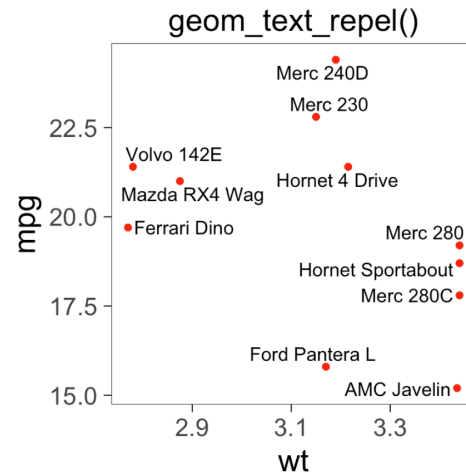


Figure 2.3. ggrepel automatically positions text labels.

The ggrepel package implements a function called `geom_text_repel()` that is easy to use. The resulting figure has clearer text labels than the figure produced with `geom_text()`. All of the text labels are fit within the boundaries of the plotting area. The text labels repel away from each other, and away from their respective data points. Finally, the labels are attracted toward their respective data points.

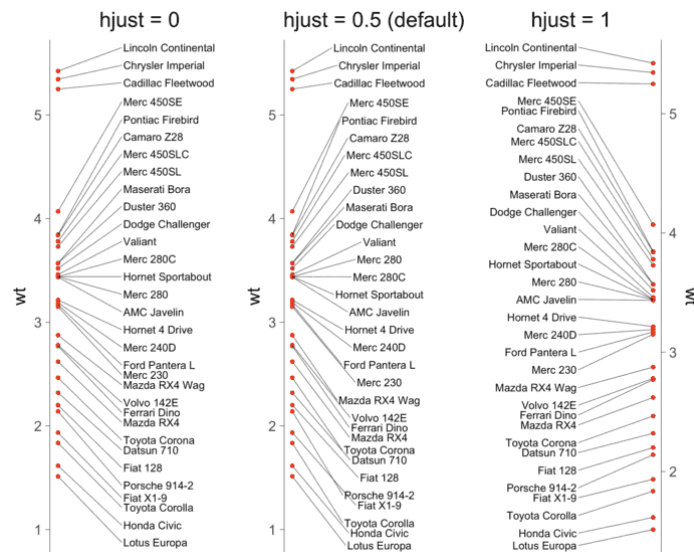


Figure 2.4. The **ggrepel** package provides flexibility to accommodate many different figures.

Since **ggrepel** extends the functionality of **ggplot2**, it inherits the flexibility built into **ggplot2**. Here, we show that the user can control the direction of repulsive forces. In this case, text labels only move up or down, but they do not move left and right. Further, the text labels can be aligned on the left edge, centered, or right-aligned to suit the style of the figure.

Making data accessible to others on the web

Functional genomics data such as RNA-seq is high dimensional, including measurements on thousands of genes across hundreds of cells or samples. To take full advantage of this type of rich data, it is necessary to write code. This is because code can be used to express arbitrarily complex algorithms and ideas to highlight any aspect of the data, only limited by the author's skill and imagination. However, writing code is laborious, and it takes a long time. If we could create an interface that provides a glimpse of such data without requiring the viewer to write any

code at all, then this would lower the barrier to access the data. Further, if such access were available quickly through a web based interface, then users could have data at their fingertips. Since mobile devices with web browsers have become a standard item in many individuals' pockets, there is a great opportunity to deliver functional genomics data visualizations to these individuals.

Today, we expect instant and free access to search engines whenever we have any desire to find something. The comfort in knowing that we have such access is, in some sense, transformative. When a conversation turns to some topic, someone can open a search engine and find information relevant to the conversation in an instant. There is an opportunity to develop the same kind of tools to enhance scientific discussion. For example, when I presented a poster at a scientific meeting, an attendee asked about the expression of a gene that was not displayed on my poster. With my mobile device in hand, I was able to query that gene and share the expression data within a few seconds. In the future, I hope to continue striving to increase accessibility to functional genomics data that researchers can instantly and freely access whenever they are reading papers, discussing with colleagues, or thinking about biological systems.

To start building web sites for data sharing, I rented a server from Amazon EC2, bought a domain name from Gandi.net, and installed Ubuntu, Apache, R, and R shiny server. I used the shiny R package to make several web pages that display RNA-seq data for my collaborators. Shiny provides an easy path forward for programmers familiar with R to transition from making figures locally on their own computers, to making figures appear in a web browser.

Gene expression along the T cell innateness gradient

Maria Gutierrez-Arcelus, *et al* describe an innateness gradient defined by transcriptional states across 7 sorted subsets of T cells ⁴⁸. To increase accessibility to the RNA-seq data underlying this finding, I built a web page hosted at (<https://immunogenomics.io/itc>) that provides the viewer with figures showing gene expression (**Figure 2.5**).

The page includes multiple components intended to help the viewer explore the data. There is a table with summary statistics for all genes, summarizing the relationship between expression level of the gene across seven subsets of T cells versus the position of each subset along the gradient of innateness. Upon clicking on one of the rows in the table, a gene is selected. The selected gene is plotted as a box plot, giving a glimpse of the distribution of gene expression across low-input RNA-seq samples from the seven T cell subsets. At the same time, the instance of R running on the server automatically correlates this selected gene against all other measured genes. The top 20 genes with greatest absolute value for Pearson correlation are selected and displayed below as a heatmap with row-normalized expression, giving the viewer some clues about other genes that might be related to the selected gene. Below the table, we show the expression data collected from 2,036 T cells assayed with single-cell RNA-seq technology. The cells are positioned on a two-dimensional map with the UMAP algorithm⁴⁹, and two adjacent plots show expression level of the gene in each cell as well as cluster identity of each cell. We also query PubMed automatically to retrieve a list of gene aliases, the description of the gene, and a list of recent scientific publications related to the selected gene. Finally, the bottom of the page includes links to download the original copy of the text, code, and data, as well as links to the lab websites for investigators involved in this work.

<https://immunogenomics.io/itc>



Figure 2.5. R Shiny app with RNA-seq data for T cells.

Screenshot of the web page available at (<https://immunogenomics.io/itc>). This Shiny app provides visitors with simple tools to quickly query a single gene of interest and view the expression of that gene in sorted low-input RNA-seq, single-cell RNA-seq, a description of the gene, papers related to the gene, and other genes correlated with the selected gene.

Defining inflammatory cell states in rheumatoid arthritis

Fan Zhang, Kevin Wei, Kamil Slowikowski, Chamith Fonseca *et al* describe a comprehensive study of synovial tissue cell populations that applies single-cell RNA-seq, mass cytometry, bulk RNA-seq, and flow cytometry to 51 synovial tissue samples from patients with rheumatoid arthritis (RA) or osteoarthritis (OA) ⁵⁰. To increase accessibility to the single-cell RNA-seq, bulk RNA-seq, and mass cytometry data, I built a web page (<https://immunogenomics.io/ampra>) (Figure 2.6).

Multiple components show different aspects of the data. The first page shows links to lab websites, code, and data downloads, as well as information about the scientific publication that describes this project. We also include a figure that gives an overview of all of the data collected from each synovial tissue sample, highlighting the fact that we focused on 4 cell types in the synovium: fibroblasts, monocytes, T cells, and B cells. The next page shows single-cell RNA-seq data and bulk RNA-seq data. The top panels show cells arranged in two dimensions with the tSNE algorithm ⁵¹. The left panel shows expression of a selected gene in each of the 5,265 cells, and the right panel shows the cluster membership of each cell. The next section shows the single-cell RNA-seq data again with a different representation. The bars indicate the percentage of cells in each single-cell cluster that have some non-zero level of expression. This is useful to determine how many cells of each type actually transcribe a given gene. The adjacent panel shows the distribution of expression in each single-cell cluster. To select a gene, the visitor can click on a row in the table of summary statistics. The main statistic we highlight is the area under receiver-operator curve (AUC), which indicates how well a single gene's expression values can predict membership of cells to a single cluster versus all other cells. For example, an AUC of 1.0

would indicate that all of the cells in a given cluster express the given gene, and all cells outside that cluster do not express the gene at all. An AUC of 0.5 indicates that the gene is not informative at all. The table is sortable, filterable, and searchable, to help visitors find any gene quickly. Bulk RNA-seq data is displayed below the table with box plots and individual points that represent samples from each tissue. Finally, the last page shows mass cytometry data, with expression of a selected protein on the left and the cluster membership of each cell on the right.

This page can help to provide evidence for which cell types are the sources of key effector molecules such as the inflammatory cytokine IL-6. By querying for “IL6” in the search box, we can see that *IL6*, the gene encoding the cytokine IL-6, is expressed in 36.7% of HLA+ sublining fibroblasts (SC-F2) (**Figure 2.7A**). We also see that 27.6% of IGHD+ CD270 naïve B cells (SC-B1) express *IL6*. We also measured the transcriptional profiles of bulk tissue samples after sorting them into 4 cell types (T cell, B cell, fibroblast, monocyte) with flow assisted cell sorting (FACS). We can see that fibroblasts from inflamed RA tissues with high abundance of T cell, B cell, and monocyte infiltration (leukocyte-rich RA) have the greatest expression of IL6 (**Figure 2.7C**).

<https://immunogenomics.io/ampra>

Links to lab websites, code, data.

scRNA-seq data.

Mass cytometry data.

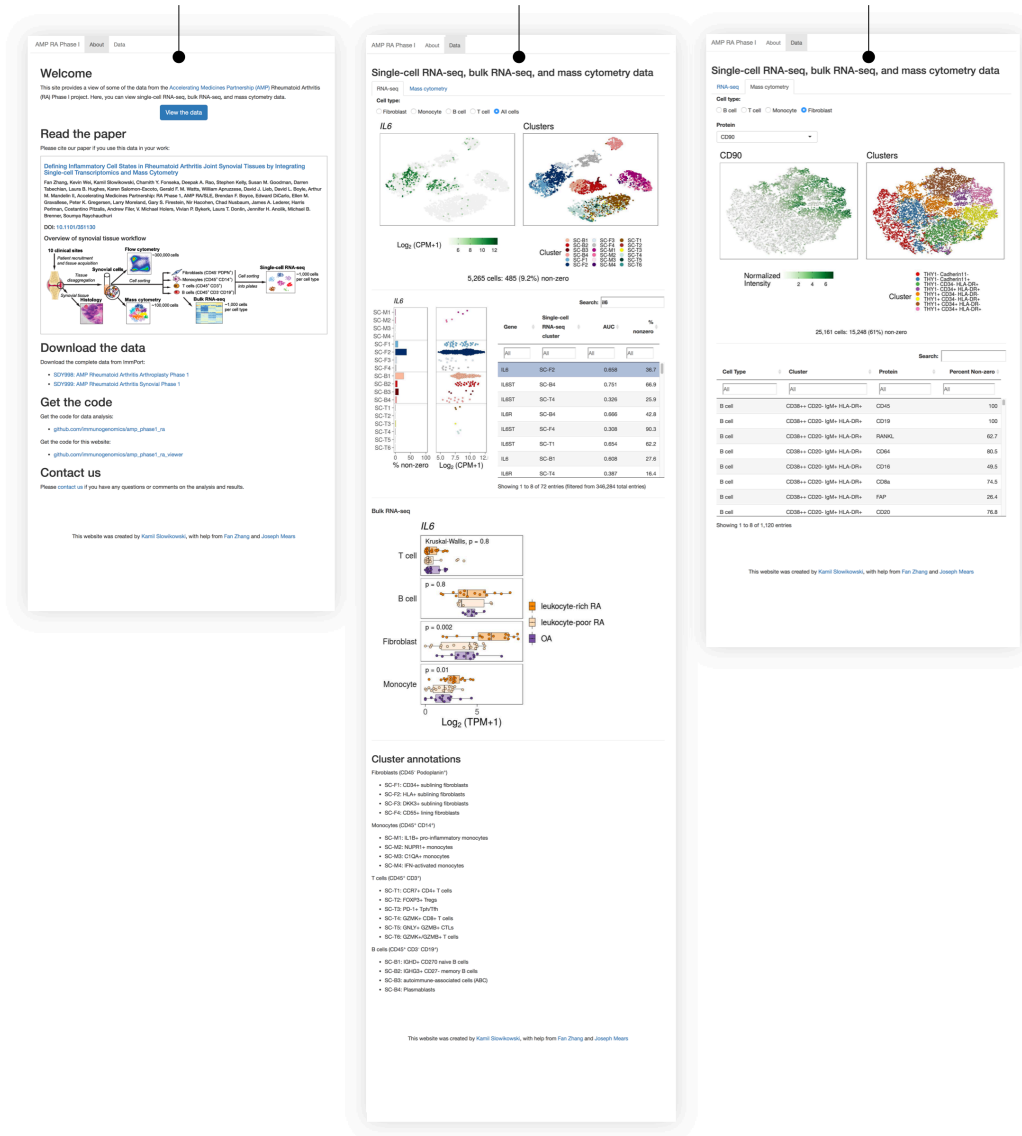


Figure 2.6. R shiny app for scRNA-seq and mass cytometry data from synovial tissues. Screenshot of the web pages available at (<https://immunogenomics.io/ampra>). This website provides visitors with access to single-cell RNA-seq data and mass cytometry data.

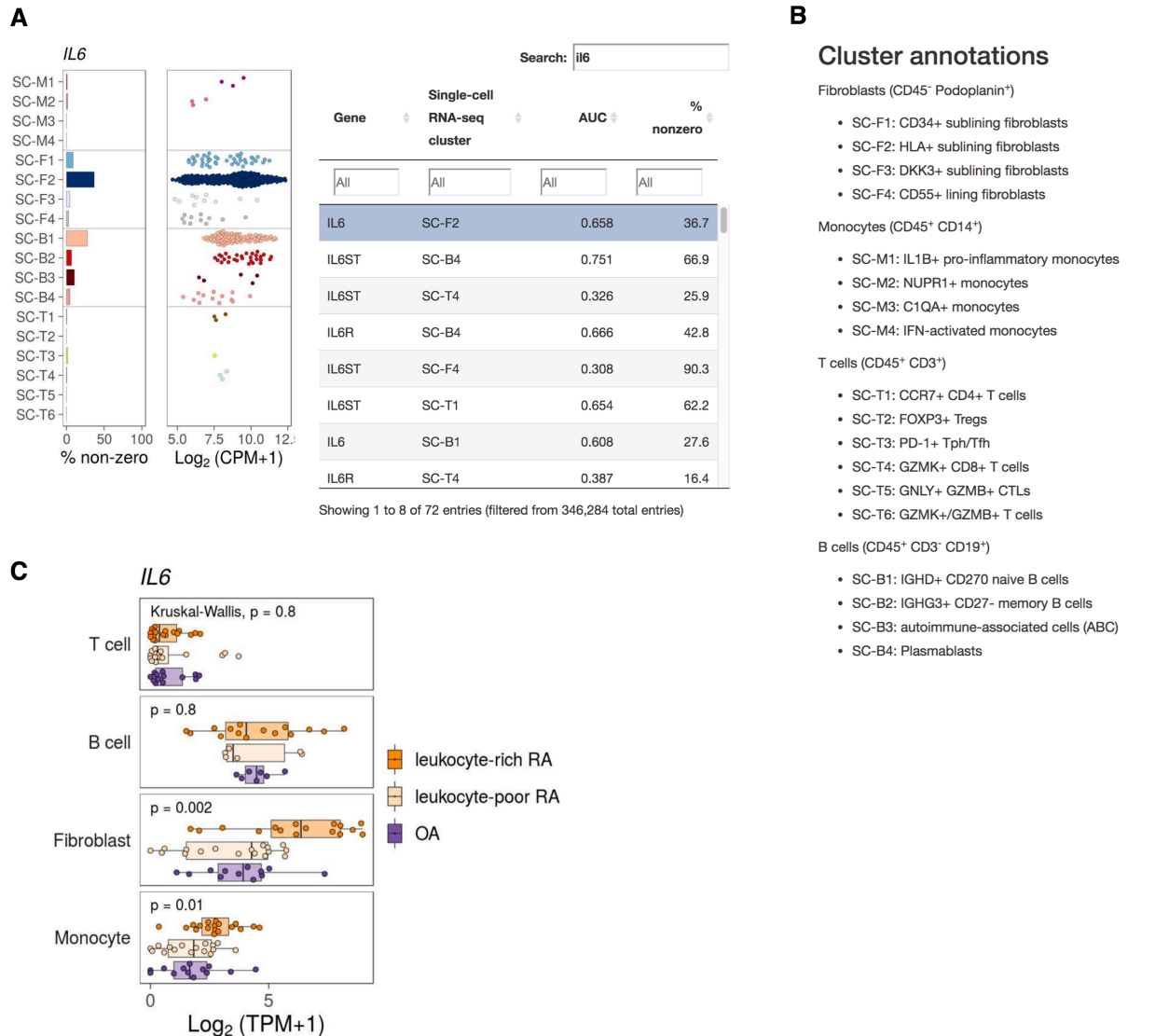


Figure 2.7. HLA⁺ sublining fibroblasts and IGHD⁺ CD270 naive B cells transcribe *IL6*.

Single-cell RNA-seq expression data measured in cells isolated from samples of inflamed joint tissue from patients with RA and OA. **(A)** Single cells with expression level in Log₂ counts per million (CPM), with evidence for *IL6* transcription in 485 cells, mostly fibroblasts and B cells. The percent of cells in each cluster with expression of *IL6* and the distribution of expression values. The adjacent table shows the area under the curve (AUC) for the gene as a predictor of cluster membership to the listed cluster versus all other cells. **(B)** A description of each cluster name. SC-F2 are HLA⁺ sublining fibroblasts and SC-B1 are IGHD⁺ CD270 naive B cells. **(C)** Bulk RNA-seq expression levels of *IL6* in sorted cells from synovial tissues. Fibroblasts from leukocyte-rich RA samples have the highest expression.

Conclusion

I showed how we can use physical simulations to improve the layout of objects in a figure. By implementing a simple algorithm and distributing it as an R package called `ggrepel`, I have enabled other data analysts to take advantage of my approach in their own figures. I also showed that functional genomics data such as RNA-seq can be presented on web pages. By increasing accessibility to the data, researchers can use these web pages to answer some of their own questions. For example, I showed that we can quickly form a hypothesis for which cell type is the dominant source of IL-6 in the inflamed synovial tissue by querying for “IL6” and inspecting the figures.

Chapter 3

Integrating genome wide association study results with gene expression data

As genome-wide association studies (GWAS) continue to find disease alleles, investigators seek to identify the set of pathways and tissue-types affected by these alleles, and the physiological conditions under which they act^{52–56}. For example, we have previously presented statistical methods to identify immune cell types for further functional investigation by finding cell type-specific expression of genes in linkage disequilibrium (LD) with autoimmune disease-associated SNPs⁵⁷. Presumably, alleles influence disease risk through pathways specific to these cell types.

We wanted to create a general implementation of these methods to leverage data from high-throughput functional assays that assess genome-wide transcription, protein binding, epigenetic modifications and other functional parameters across diverse cellular conditions and tissue types. Each of these diverse data types can be represented as a continuous matrix of genes and *conditions* (e.g. cell types, tissues, pathways, experimental conditions, etc.). Databases such as Gene Ontology offer expert-defined pathways and complementary gene annotations which can be represented as binary values.

We created a fast, robust, and general C++ implementation of a single-nucleotide polymorphism (SNP) set enrichment algorithm to identify cell types, tissues, and pathways affected by risk loci. It tests trait-associated genomic loci for enrichment specificity to conditions (cell types, tissues, pathways) in a matrix of genes and conditions. We use a nonparametric statistical approach to compute empirical p-values by comparison to null SNP sets. As a proof of concept, we present novel applications of our method to four sets of genome-wide significant SNPs associated with red blood cell count, multiple sclerosis, celiac disease, and HDL cholesterol. Investigators have already described strategies to assess enrichment of GWA results for pathways or gene sets, but not for condition-specificity^{58,59}. In contrast to these methods, we do not require genotypes, p-values, *a priori* gene sets or pathways, or *a priori* definitions of

gene-SNP associations. We require only a list of SNP identifiers, use LD structure to identify plausibly influential genes, and employ a simple sampling approach to identify the conditions they influence. SNPSEA is a general algorithm to identify the conditions relevant to a trait by assessing the genes within associated loci for enrichment with condition-specificity.

The SNPSEA algorithm

For a given set of SNPs, SNPSEA tests genes implicated by LD, in aggregate, for enrichment of specificity to a condition in a given matrix of genes and conditions. The matrix must be normalized so that conditions are comparable.

First, we identify genes implicated by each SNP using LD from reference genomes. Second, we calculate a specificity score for each condition with those genes. Finally, we compare those scores to scores obtained with null sets of matched SNP sets to calculate an empirical p-value for each condition.

We calculate p-values empirically because we previously found that analytical distributions can result in inaccurate p-values⁵⁷. SNP linkage intervals, gene densities, gene sizes and gene functions are correlated across the genome and are challenging to model analytically. The online reference manual details compilation and installation procedures; we also provide executable files for immediate use on select platforms.

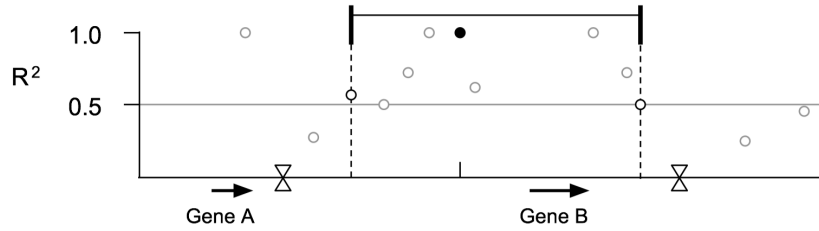
Multiple genes implicated by LD

Accurate analyses must address the critical issue that SNPs from GWA studies frequently implicate more than one gene. When we looked at 11,561 SNPs from the GWAS Catalog, we found that 50% of them are in linkage disequilibrium (LD) with 2 or more genes (**Figure 3.2**). We defined LD intervals with SNPs from the 1000 Genomes Project (EUR) and a previously described strategy (**Figure 3.1**)⁶⁰. A SNP implicates genes overlapping its LD interval, defined by the furthest SNPs in a 1 Mb window with $r^2 > 0.5$. To ensure that all associated genes are included, we extend each interval to the nearest recombination hotspots with recombination rate >3 cM/Mb (HapMap3)⁶¹. We merge SNPs with shared genes into a single locus.

By default, SNPSEA will run with the assumption that each associated locus harbors a single influential gene rather than multiple genes. However, we also provide an alternative scoring method to account for multiple genes – in practice, this alternative method produces similar results to the single-gene method in the four traits we tested (**Figure 3.8**).

Since interval lengths depend on the choice of r^2 threshold, we looked for an effect of this choice (**Figure 3.7**). The significant result for the Gene Atlas and blood cell count SNPs is robust to different thresholds. Similarly, the choice of r^2 threshold has little effect on the significant GO enrichment result for those SNPs.

1. For each SNP, find neighbors with $R^2 \geq 0.5$ within a 1 Mb window.



2. Extend to nearest recombination hotspots with rate > 3 cM / Mb.

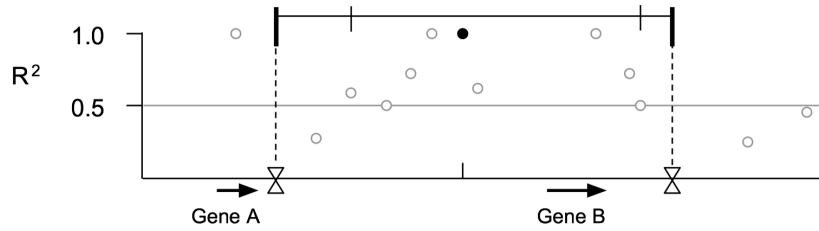


Figure 3.1. Determining SNP linkage intervals.

We calculated r^2 values for all pairs of SNPs within a 1 Mb sliding window along each chromosome. Next, we assigned each of the SNPs from the 1000 Genomes Project Phase I⁶² to a linkage interval by identifying each SNP's furthest upstream and downstream neighbors with $r^2 \geq 0.5$. Finally, we extended each interval to recombination hotspots reported by HapMap⁶³ with recombination rate > 3 cM/Mb.

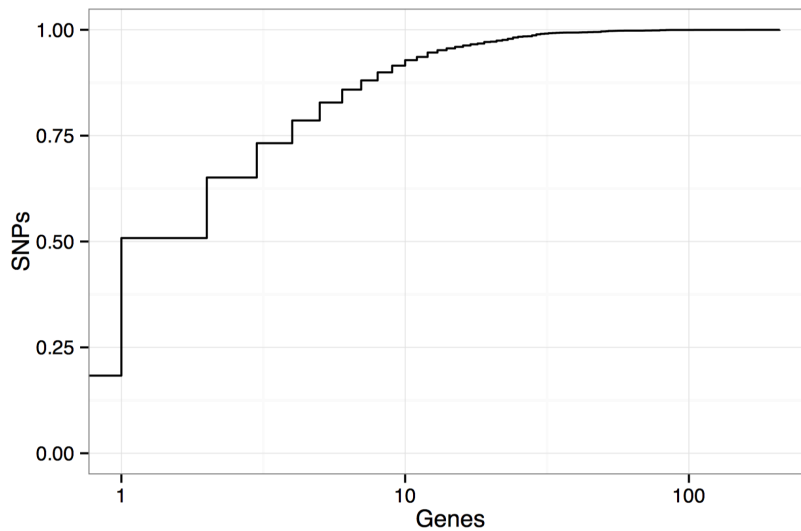


Figure 3.2. Counting genes in GWAS SNP linkage intervals.

A cumulative density plot of the number of genes overlapped by the linkage intervals of GWAS SNPs. We downloaded the GWAS Catalog SNPs on January 17, 2014 and selected the 11,561 SNPs present in the 1000 Genomes Project⁶². Of these SNPs, 2,119 (18%) of them have linkage disequilibrium (LD) intervals that overlap 0 genes, and 3,756 (32%) overlap a single gene. The remaining 50% of SNPs overlap 2 or more genes. This illustrates the critical issue that many SNPs implicate more than one gene.

Type I error estimates

We tested 10,000 sets of 100 randomly selected LD pruned SNPs. For each condition (tissue or GO term), we observed appropriate proportions of p-values less than 0.5, 0.1, 0.05, 0.01, and 0.005 (**Figure 3.3**).

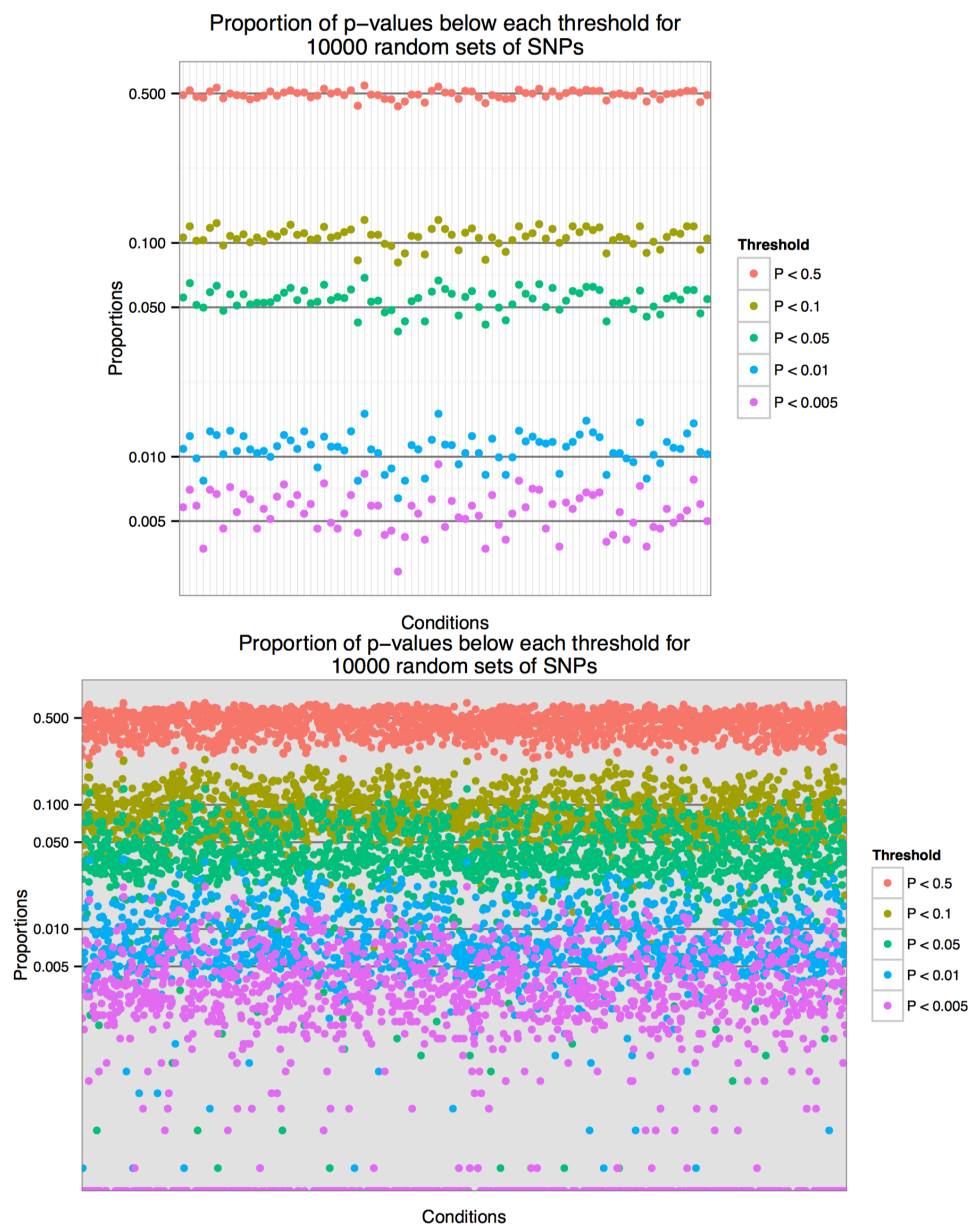


Figure 3.3. Type 1 error estimates.

We sampled 10,000 sets of 100 SNPs uniformly from a list of LD-pruned SNPs⁶⁴. We tested each of the 10,000 sets for enrichment of tissue-specific expression in the Gene Atlas⁶⁵ gene expression matrix (top) and for enrichment of annotation with Gene Ontology terms (bottom). For each condition, we show the proportion of the 10,000 enrichment p-values that are below the given thresholds. We observe that the p-values are near the expected values, so the type 1 (false positive) error rate is well-calibrated.

Results for multiple sclerosis, celiac disease, and HDL cholesterol

We used SNPSEA to identify tissues relevant to blood cell count by testing 45 genome-wide significant SNPs⁶⁶ with expression data (Gene Atlas) for 17,581 genes across 79 human tissues⁶⁷.

Bone marrow CD71+ early erythroid cells are significantly enriched for cell type-specific expression of the genes within the trait-associated loci ($P = 2 \times 10^{-7}$) (**Figure 3.4**).

The genes in these loci are enriched for the term *hemopoiesis* (GO:0030097) ($P = 2 \times 10^{-5}$) (**Figure 3.5**), suggesting that blood cell count may be influenced by the genes expressed specifically in early erythroid cells and involved in forming blood cellular components.

We provide additional examples for SNPs associated with multiple sclerosis, celiac disease, and HDL cholesterol. Each includes Gene Atlas and GO enrichments, r^2 comparisons, and comparisons of results assuming a single causal gene or multiple causal genes (**Figure 3.6**, **Figure 3.7**, **Figure 3.8**).

45 Red blood cell count-associated SNPs (Harst 2012) Gene Atlas (Su 2004)

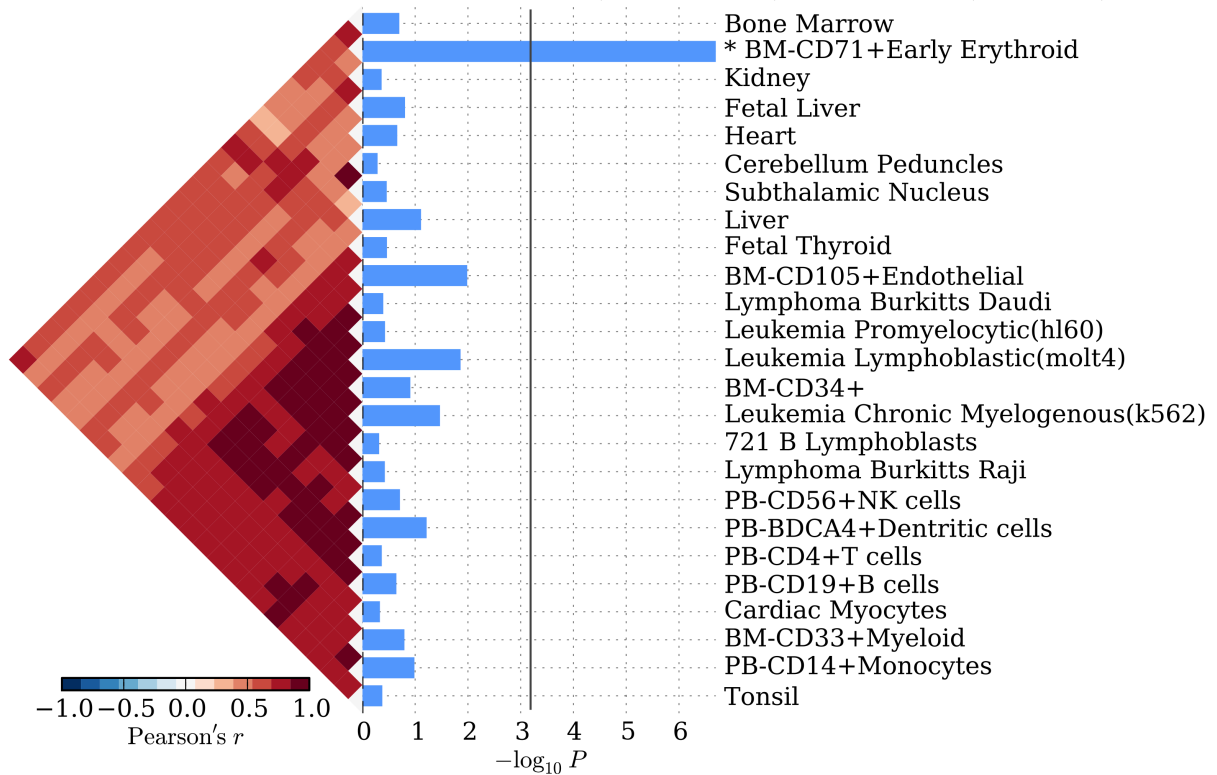


Figure 3.4. SNPSEA p-values for 45 red blood cell count-associated SNPs.

Empirical p-values for specificity to each condition. Heatmap shows Pearson correlation coefficients for pairs of gene expression profiles clustered with the unweighted pair group method with arithmetic mean (UPGMA).

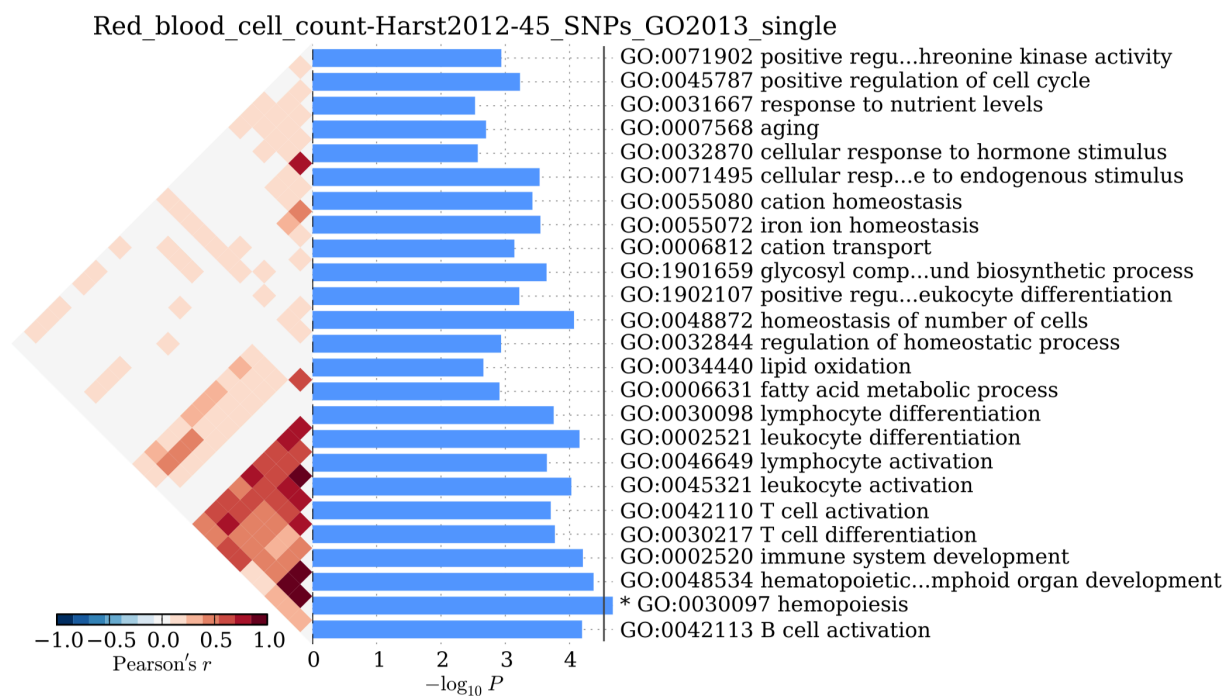


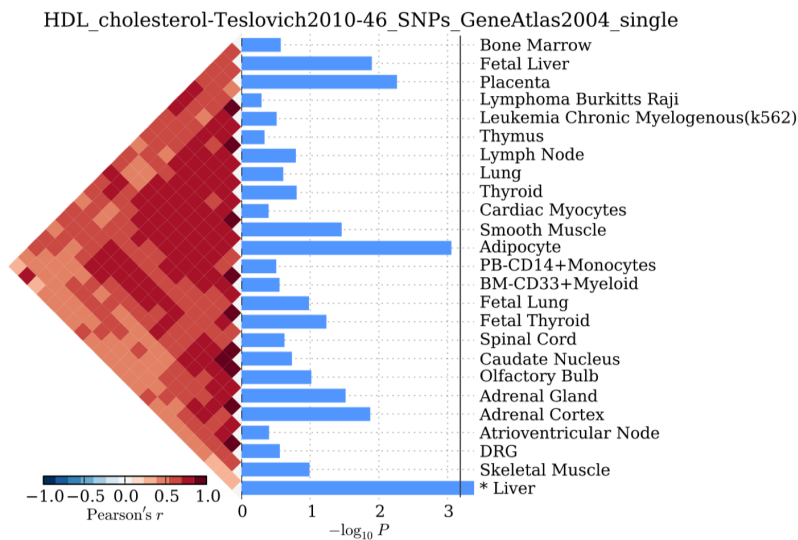
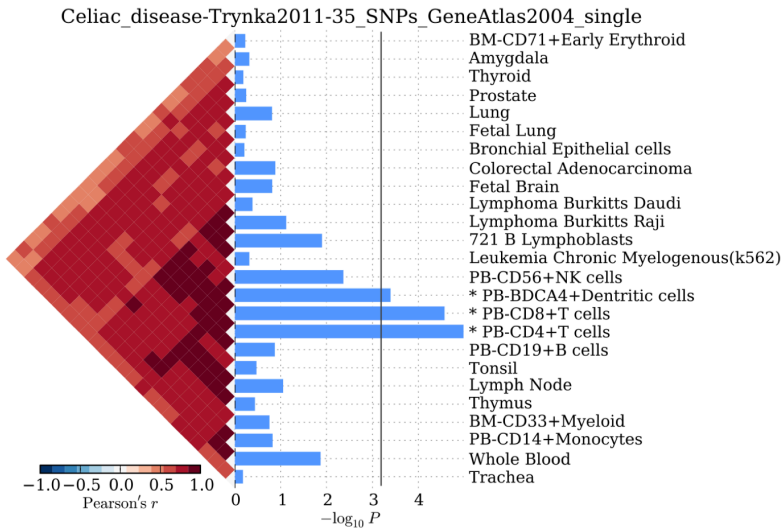
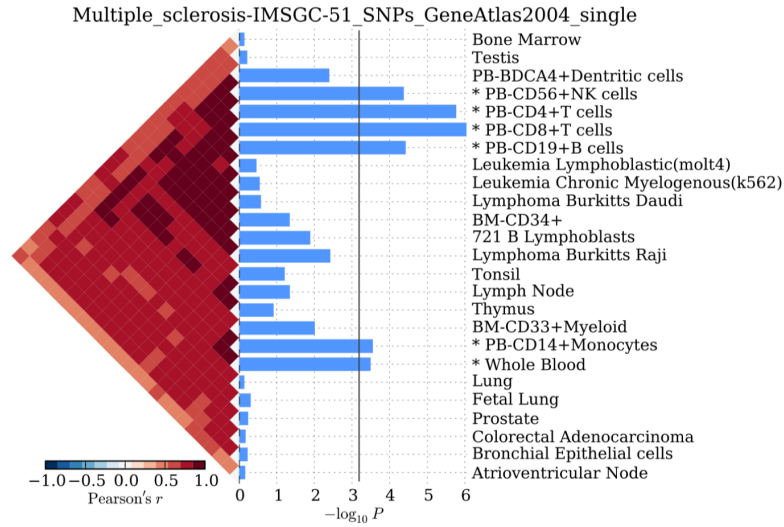
Figure 3.5. Red blood cell count GO enrichment.

SNPSEA results for the top 25 terms from Gene Ontology (GO). We observed significant enrichment for hemopoiesis ($2e-5$) using 45 SNPs from a GWAS associating SNPs to red blood cell count.

Figure 3.6. SNPSEA results for Gene Atlas and Multiple Sclerosis, Celiac disease, and HDL cholesterol.

SNPSEA results for 3 GWAS studies. For example, we see that SNPs associated with risk for developing multiple sclerosis are in linkage disequilibrium with the genes preferentially expressed in NK cells, T cells, B cells, and monocytes. Similarly, celiac disease is associated with CD4⁺ T cells, CD8⁺ T cells, and dendritic cells. SNPs associated with HDL cholesterol are linked to genes expressed in liver cells.

Figure 3.6 (Continued)



Materials and Methods

SNPSEA tests if genes implicated by risk loci (e.g., those discovered through genome-wide association (GWA) studies) are specifically expressed in some conditions over others, and if this specificity is statistically significant. The program requires two inputs:

1. A list of SNP identifiers: rs123, chr6:123456, ...
2. A matrix of genes and conditions, such as:
 - a. Expression profiles of different cell types.
 - b. Ontology terms and binary presence/absence values for each gene in each term.

For example, SNPSEA can be used to find tissues or cell types whose function is likely to be influenced by genes in risk loci. If the genes in risk loci are expressed in relatively few cell types, we hypothesize that they are likely to affect those cell types' unique functions. This assumes that expression specificity is a good indicator of a gene's importance to the unique function of the cell type.

For a given set of SNPs associated to some phenotype, SNPSEA tests whether all implicated genes, in aggregate, are enriched for specificity to a condition in a user-provided matrix of genes and conditions/annotations. The algorithm consists of three steps:

1. Assigning genes to each SNP.

We use linkage disequilibrium (LD) to identify the genes implicated by each SNP.

2. Calculating specificity scores.

We look up implicated genes in a user-provided matrix and calculate a specificity score for each annotation/condition based on the values of these genes.

3. Testing significance.

We compare the specificity scores to a null distribution of scores obtained with random sets of matched SNP sets and compute an empirical p-value.

Assigning genes to each SNP. Accurate analyses must address the critical issue that SNPs frequently implicate a region with multiple different genes (**Figure 3.2**). The challenge is to find evidence to show which of those genes are associated with a given trait.

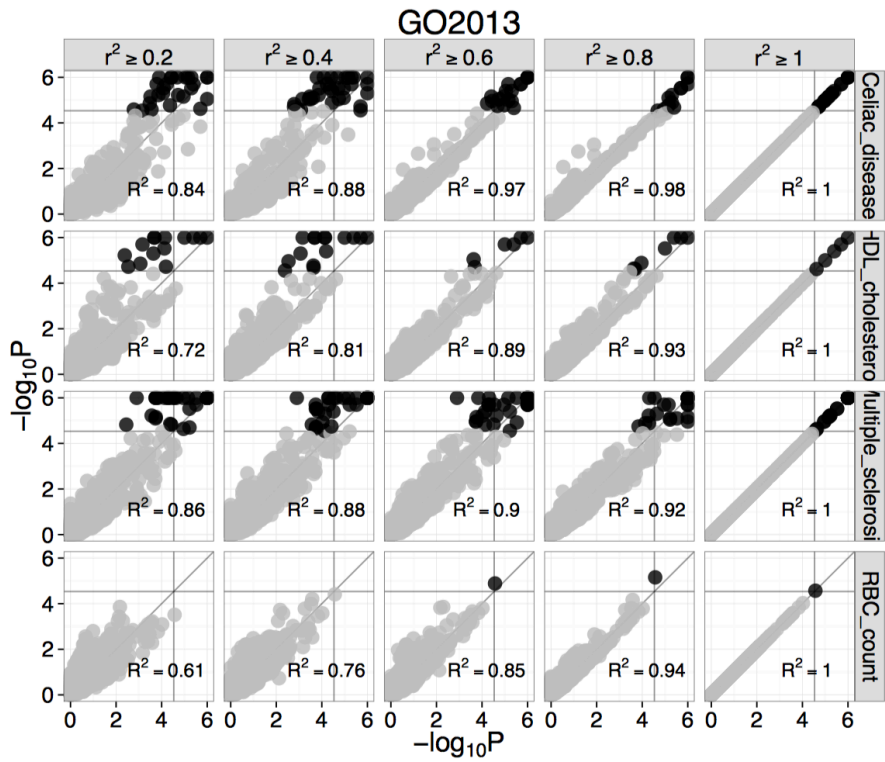
We determine the genes plausibly implicated by each trait-associated SNP using a previously described strategy⁶⁸ (**Figure 3.1**). First, we define the linkage interval for a given SNP as the span between the furthest correlated SNPs $r^2 > 0.5$ within a 1 Mb window, using genotypes from individuals with European ancestry (EUR) from the 1000 Genomes Project⁶⁹. Next, we extend the interval to the nearest recombination hotspots with recombination rate >3 cM/Mb⁶³. To address the case when no genes overlap an interval, we provide an option for SNPSEA to extend the interval up- and downstream (by default 10 Kb).

Most frequently, we find multiple genes ($m_k > 1$) in a single SNP locus k . We expect many loci with multiple genes because of regions with high LD across long stretches of a chromosome. Less frequently, a locus has a single gene ($m_k = 1$), and loci with no genes ($m_k = 0$) are discarded. After each SNP has been assigned an interval and a set of genes overlapping the interval, we merge SNPs with shared genes into a single locus to avoid multiple-counting of genes.

Figure 3.7. Choosing the r^2 threshold for linkage intervals.

Gene Atlas and Gene Ontology (top and bottom). Each subplot has $-\log_{10} P$ for $r^2 = 1$ on the x-axis and $-\log_{10} P$ on the y-axis for the r^2 threshold marked above. Grey lines are significance thresholds after correction testing multiple conditions (cell types, GO annotations). Black points are significant and grey are not. We used the 'score single' option. Red blood cell count SNPs are enriched for hemopoiesis (GO:0030097) ($P = 2 \times 10^{-5}$) for linkage intervals with $r^2 = (0.6, 0.8, 1.0)$. This result falls below the multiple testing threshold at $r^2 \geq 0.4$, but remains significant at $r^2 \geq 0.5$ (see main text). To investigate if the choice of r^2 threshold influences SNPSEA results, we repeated the analysis of four traits using 5 different thresholds ($r^2 \geq 0.2, 0.4, 0.6, 0.8, 1.0$). SNPSEA results seem to be robust to the choice of threshold, mostly due to the fact that we extend linkage intervals for each SNP to the nearest recombination hotspots as described in Supplementary Figure 1. We chose to use $r^2 \geq 0.5$ for our analysis and for our provided data files due to this result, and also due to previous experience⁶⁸.

GeneAtlas2004



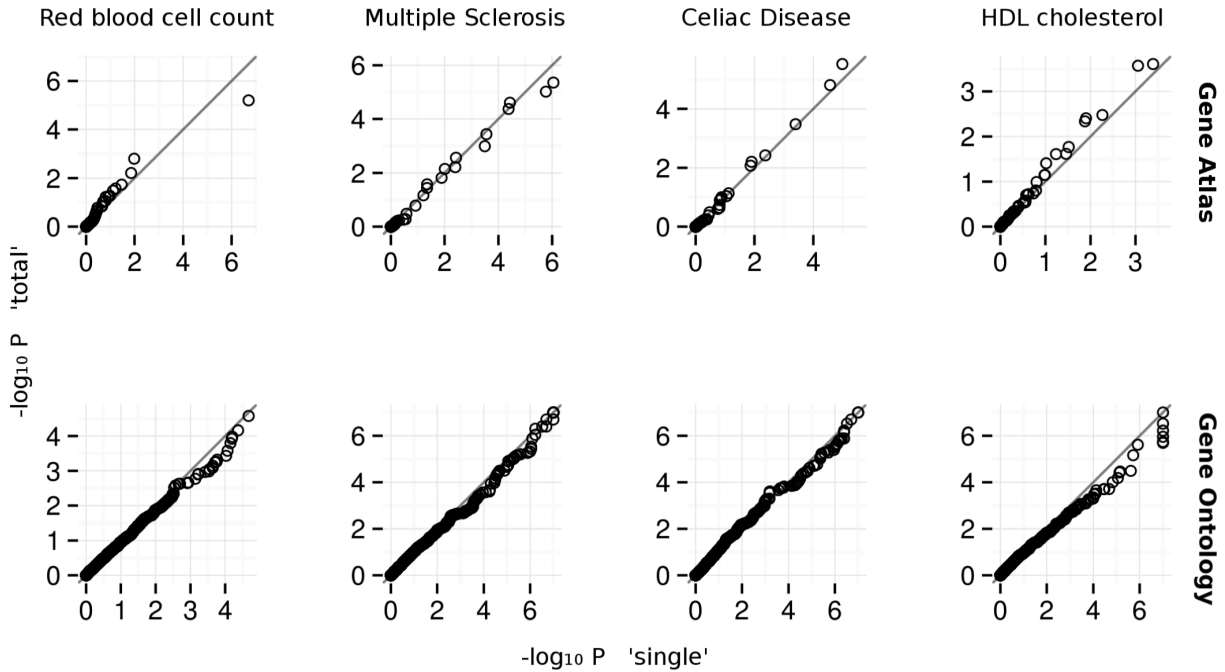


Figure 3.8. Comparison of SNPSEA results assuming one gene or many genes per locus.

Quantile-quantile plots for Gene Atlas (top) and Gene Ontology (bottom). The x and y axes represents the $-\log_{10}$ p-values obtained by running SNPSEA with either ‘—score single’ (x-axis) or ‘—score total’ (y-axis). The results look similar for four traits we tested: Red blood cell count, Multiple Sclerosis, Celiac Disease, and HDL cholesterol.

Calculating specificity scores. SNPSEA uses different algorithms for matrices with continuous or binary values. By default, SNPSEA assumes one gene in each associated locus is associated with the given trait. We also include the option to assume all genes within a locus are associated. We compare results of the two options with four phenotypes (**Figure 3.8**). The ‘—score single’ option (default) assumes that a single gene in each locus is associated with the given phenotype. For each condition, we choose the gene in each locus with the greatest specificity to that

condition. The ‘--score total’ option assumes that all genes in a SNP’s linkage interval are associated. We account for all linked genes when calculating scores.

Specificity for a matrix of continuous numeric values. Before running SNPSEA, a matrix with continuous values must be normalized so that columns are directly comparable. It is not appropriate to use this method on a “raw” matrix of expression values. We extend an approach we have previously described in detail ⁷⁰. Let A denote a continuous gene expression matrix with m genes and n conditions. First, we normalize the expression of each gene by dividing each value by the L2 norm of the gene’s values across all conditions.

$$A'_{i,j} = \frac{A_{i,j}}{\sqrt{A_{i,1}^2 + A_{i,2}^2 + \dots + A_{i,n}^2}}$$

The resulting matrix A' has values between 0 and 1 indicating specificity of gene i to condition j . A value of 1 indicates that gene i is exclusively expressed in condition j with no expression in other conditions. A value of 0 indicates that gene i is not expressed in condition j at all.

Next, we transform A' into a matrix of condition-specificity percentiles as follows. For each condition j , we rank the values of A' in ascending order and divide them by the number of genes m , resulting in percentiles between 0 and 1 where a lower value indicates greater specificity to the given condition.

Locus scores for a matrix of continuous numeric values. We create a new matrix P , where each value is a score for a SNP locus k and a condition j . We define the locus scores for a single condition to be approximately uniformly distributed for a set of randomly selected loci under the null hypothesis of no association to the condition. We assume that, for a set of genes in a given SNP locus, the A'' values for genes in that locus are random, independent, and approximately

uniformly distributed. If there is an actual association to a condition j , we will observe an unexpectedly small value for the condition score P_j .

The default approach in SNPSEA tests for the association of one gene in each SNP locus. For each locus-condition pair (k, j) , we choose the single gene i in locus k with greatest specificity to condition j among the m_k genes in that locus, as previously described in Hu et al.⁷⁰. Let g_k denote this most specific gene, so that $A''_{g_k,j} = \text{Min}_{i \in I_k}(A''_{i,j})$ where I_k denotes the set of genes in locus k . If we assume values of $A''_{i,j}$ are uniformly distributed for a given condition j and genes $i \in I_k$, then the probability to randomly draw a value equal to or less than $A''_{g_k,j}$ is:

$$P_{k,j} = 1 - (1 - \text{Min}_{i \in I_k}(A''_{i,j}))^{m_k}$$

We also provide an alternative approach to test for the association of all genes in each SNP locus to condition j – we consider this model to be unlikely in most situations. The product (log sum) of uniform values between (0,1) follows a gamma distribution⁷¹. If the genes I_k have no specificity to a condition j , then the values $A''_{i \in I_k}$ are approximately uniformly distributed. So, we compute the probability to randomly draw values with a smaller product as the upper tail of the gamma distribution:

$$P_{k,j} = \int_x^\infty \Gamma(m_k, 1) \text{ for } x = \sum_{i \in I_k} -\ln A''_{i,j}$$

Locus scores for a matrix of binary values. Let B denote a binary matrix (1=present, 0=absent) with m genes and n conditions. Let m_j denote the number of genes present in condition j . Let m_k denote the number of genes in locus k and $m_{k,j} \leq m_k$ denote the number of genes in locus k that are present in condition j .

We provide two options to calculate locus scores. By default, we account for presence or absence of any of the m_k genes in condition j , as shown below ('--score single'). Alternatively, we account for the number of genes in a given locus ('--score total').

$$\begin{array}{ccc}
 & \text{'--score single'} & \text{'--score total'} \\
 p(x) = \frac{\binom{m_j}{x} \binom{m-m_j}{m_k-x}}{\binom{m}{m_k}} & P_{k,j} = \begin{cases} 1 - p(0) & m_{k,j} > 0 \\ 1 & m_{k,j} = 0 \end{cases} & P_{k,j} = \begin{cases} 1 - \sum_{x=0}^{m_{k,j}-1} p(x) & m_{k,j} > 0 \\ 1 & m_{k,j} = 0 \end{cases}
 \end{array}$$

Condition specificity scores. For both continuous and binary matrices, we define a specificity score S_j for each condition j as the aggregate of $P_{k,j}$ values across SNP loci:

$$S_j = \sum_k -\log P_{k,j}$$

Testing significance with permutation p-values. We previously found that aggregating the $P_{k,j}$ scores and determining a P-value analytically from a distribution results in inaccurate p-values [4]. A 00 values may be relatively uniform i,j genome-wide, but proximate genes often have shared functions. The genome has a complex correlation structure of linkage disequilibrium, gene density, gene size and function that is challenging to model analytically. We use the sampling strategy described below instead.

For each condition, we use a sampling approach to calculate an empirical p-value. This is the tail probability of observing a condition-specificity score greater or equal to $\underline{S}_i$. We obtain the distribution empirically with null SNP sets.

We compute specificity scores S for random SNP sets. Each SNP in a null set is matched to a SNP in the user's set on the number of linked genes. To adequately sample genes from the

entire genome, we sample SNP sets from a list of LD-pruned SNPs (subset of SNPs in 1000 Genomes Project).

For each condition j , we calculate an exact permutation p-value⁷². Let a_j denote the number of sampled SNP sets (e.g. 10,000) and let b_j denote how many null specificity scores are greater than or equal to the user's score S_j :

$$p_j = \frac{b_j + 1}{a_j + 1}$$

We implemented adaptive sampling to calculate p-values efficiently. As each condition is tested for significance, we increase the number of iterations to resolve significant p-values and save computation by using fewer iterations for less significant p-values. Two options allow the user to control the adaptive sampling:

1. '--max-iterations N' The maximum number of iterations for each condition. We stop testing a condition after sampling N SNP sets.
2. '--min-observations N' The minimum number of observed null specificity scores greater than or equal to S_j required to stop sampling SNP sets for a condition j .

Data. Please find all of the data required to reproduce this analysis at FigShare

(<http://dx.doi.org/10.6084/m9.figshare.871430>). Details about each dataset are listed below.

Gene Atlas gene expression matrix. We downloaded the data from BioGPS:

<http://plugins.biogps.org/download/gnflh-gcrma.zip> We averaged the expression values for tissue replicates. For each gene, we selected the single probe with the largest minimum value. Finally, we converted the file to GCT format.

Gene Ontology binary presence/absence matrix. We downloaded the OBO file from Gene Ontology (data-version: 2013-06-29, CVS revision: 9700, <http://www.geneontology.org>). For each gene, we climbed the hierarchy of ontology terms and applied parental terms. If a gene is annotated with some term T, we also add all of the terms that are parents of T. We copy terms between homologous genes using Homologene data (<http://www.ncbi.nlm.nih.gov/homologene>). If a mouse gene is annotated with some term and the human homolog is not, then we copy the term to the human gene. We discard all GO terms assigned to fewer than 100 or to more than 1000 genes. This leaves us with a matrix of 19,111 genes and 1,751 terms.

1000 Genomes Project. We downloaded a filtered (biallelic and 5 or more copies of the minor allele) set of markers from the BEAGLE website and calculated pairwise LD (EUR) for all SNPs in a 1 Mb sliding window (<http://bochet.gcc.biostat.washington.edu/beagle>).

Conclusion

We have shown how SNPSEA can integrate GWAS results with a gene expression matrix to discover which cell types express genes linked to the GWAS results. We used a permutation approach to simulate null results of a GWAS, and showed that this can be used to estimate the statistical significance of enrichment. New methods such as LD score regression⁷³ have more statistical power because they utilize genome-wide summary statistics instead of a short list of top SNPs. Still, SNPSEA serves as an interesting proof of concept that we can make similar inferences even from a list of SNP identifiers.

Chapter 4

Transcriptomics of fibroblast subsets in RA

Fibroblasts are key mediators of end organ pathology and inflammation in many diseases. Although they mediate normal matrix deposition and inflammation in wound healing, chronically activated fibroblasts can differentiate into myofibroblasts that lay down collagen and are responsible for fibrosis in lung, liver, gut, skin, and other tissues ⁷⁴. Conversely, chronically activated fibroblasts are responsible for excessive matrix degradation that destroys cartilage and causes permanent joint damage in rheumatoid arthritis ^{16,75,76}. Moreover, recent studies have emphasized the role of fibroblasts as stromal cells that regulate immune responses in lymph nodes and tumor stroma ^{77,78}. Unlike hematopoietic cell types that are known to be comprised of a variety of functionally different cellular types and subsets, fibroblasts have generally been considered to have little heterogeneity: functionally distinct subpopulations have yet to be clearly defined.

Recent advances in high-throughput technologies enable investigators to query complex diseases in humans in new ways. For example, global transcriptomic analysis revealed distinct activation states and cellular subsets of immune cells ⁷⁹. These approaches offer an opportunity to determine how stromal cells mediate various types of local tissue pathology. Transcriptomics of small numbers of cells, and even single cells, from human pathological samples can advance the understanding of tissue dynamics in disease. For example, single cell RNA-seq has revealed heterogeneity of tumor cells and identified a mechanism for drug resistance in cancer ^{80,81}.

Here, we focus on mesenchymal cell heterogeneity in RA, a complex autoimmune disease affecting up to 1% of the world's population ⁸². In RA, the synovium changes dramatically: the thin membrane encapsulating the joint becomes an inflamed, hyperplastic, and invasive tissue mass that causes joint destruction ⁷⁶. Synovial fibroblasts secrete inflammatory

cytokines and chemokines, invade and degrade cartilage, and stimulate osteoclasts that cause bone erosion ^{16,76}.

We hypothesized that these different functions might be carried out by distinct cellular subsets of fibroblasts, analogous to functionally distinct subsets of leukocytes. If fibroblast subsets exist, we thought that altered proportions of fibroblast subsets might underlie pathological observations in tissues ⁸³. Since there are no approved drugs that directly target fibroblasts, identifying pathogenic fibroblast subsets may reveal therapeutic targets broadly applicable across a range of diseases. In RA, targeting fibroblast subsets might complement anti-inflammatory therapies that target leukocytes and their cytokines ⁷⁶.

Fibroblasts in synovial tissue have distinct surface markers

To examine the heterogeneity of fibroblasts in joint tissue, we isolated cells by enzymatic digestion of synovial specimens from joint replacement surgery. We compared RA samples to osteoarthritis (OA) samples. OA samples have been widely used as controls or comparators in the studies of RA, since OA is a degenerative joint disease while RA is an autoimmune disease associated with synovial hypertrophy and inflammation. We first examined protein expression of a variety of surface markers on fibroblasts including podoplanin (PDPN), cadherin 11 (CDH11), THY1 (CD90), and CD34 by flow cytometry (see Materials and Methods). To isolate fibroblasts, we first gated PTPRC(CD45)⁻GYPA(CD235a)⁻PECAM1(CD31)⁻ cells to remove hematopoietic lineage cells, red blood cells (RBCs), and endothelial cells, and further removed MCAM(CD146)⁺ pericytes. The remaining cells had high PDPN expression, consistent with fibroblasts in the synovium (**Figure 4.1A**). We identified two major populations based on the

expression of CD34. In 42 donors (26 OA and 16 RA), we observed medians of 34.7% CD34⁺ and 54.7% CD34⁻ cells (**Figure 4.1A, Table 4.1**). CD34⁻ and CD34⁺ fibroblasts could further be divided into 4 and 3 populations, respectively, based on THY1 and CDH11 expression. However, we remained uncertain about which of these divisions were functionally meaningful. This result prompted us to further investigate if the gated fibroblast populations might have different gene expression profiles and effector functions.

Table 4.1. Clinical characteristics of patients evaluated with flow cytometry.

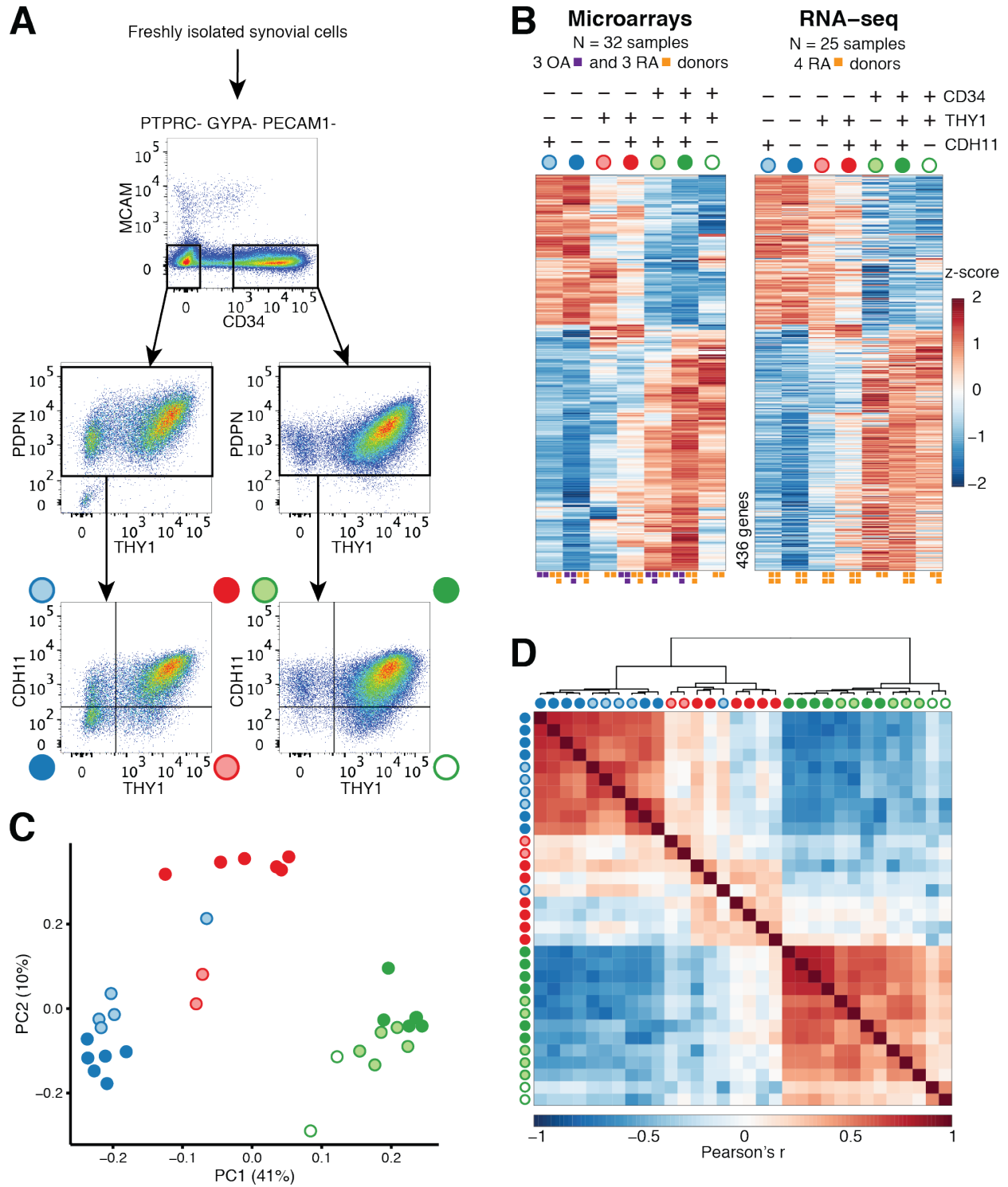
Disease	OA	RA
Number	26	16
Age (years $\pm$ SD)	70 $\pm$ 9	61 $\pm$ 14
Female (%)	54	75
Joint (%)		
Knee	100	50.0
Wrist		25.0
MCP		6.25
MTP		6.25
Elbow		12.5
RF or ACPA (%)		
Positive		56
Negative		6
Unknown		38
CRP (mg/l)		13 $\pm$ 22.7
Methotrexate (%)		56.3
Other synthetic (%)		18.8
Biologic (%)		68.8
Glucocorticoid (%)		31.3

MCP; metacarpophalangeal joints, MTP; metatarsophalangeal joint, RF; rheumatoid factor, ACPA; anti-citrullinated protein antibody, CRP; C-Reactive Protein. Other synthetic includes leflunomide and sulfasalazine. Biologic includes TNF inhibitors, abatacept, rituximab, tocilizumab and tofacitinib.

Figure 4.1. Fibroblast subsets have distinct protein and mRNA expression.

(A) Gating strategy for synovial fibroblasts with heterogeneous expression of surface proteins. (B) Analysis of variance (ANOVA) reveals 436 genes with significant (1% FDR) variation across 7 gated populations that are measured and statistically significant in both microarray and RNA-seq datasets. Squares beneath each column indicate number of samples for that phenotypic subset. (C) Principal components analysis (PCA) with 2,986 genes (1% FDR, ANOVA) in microarray data separates the 32 microarray samples into 3 subsets: CD34⁺THY1⁺, CD34⁺THY1⁻, and CD34⁻. (D) Pairwise Pearson correlation of microarrays also suggests 3 major subsets of fibroblasts.

Figure 4.1 (Continued)



Fibroblast subpopulations have three distinct gene expression signatures

We used two complementary strategies to investigate synovial fibroblast populations obtained after tissue disaggregation: (1) fluorescence sorting using a set of candidate protein markers followed by bulk transcriptomics of gated populations and (2) unbiased single cell transcriptomics without gating.

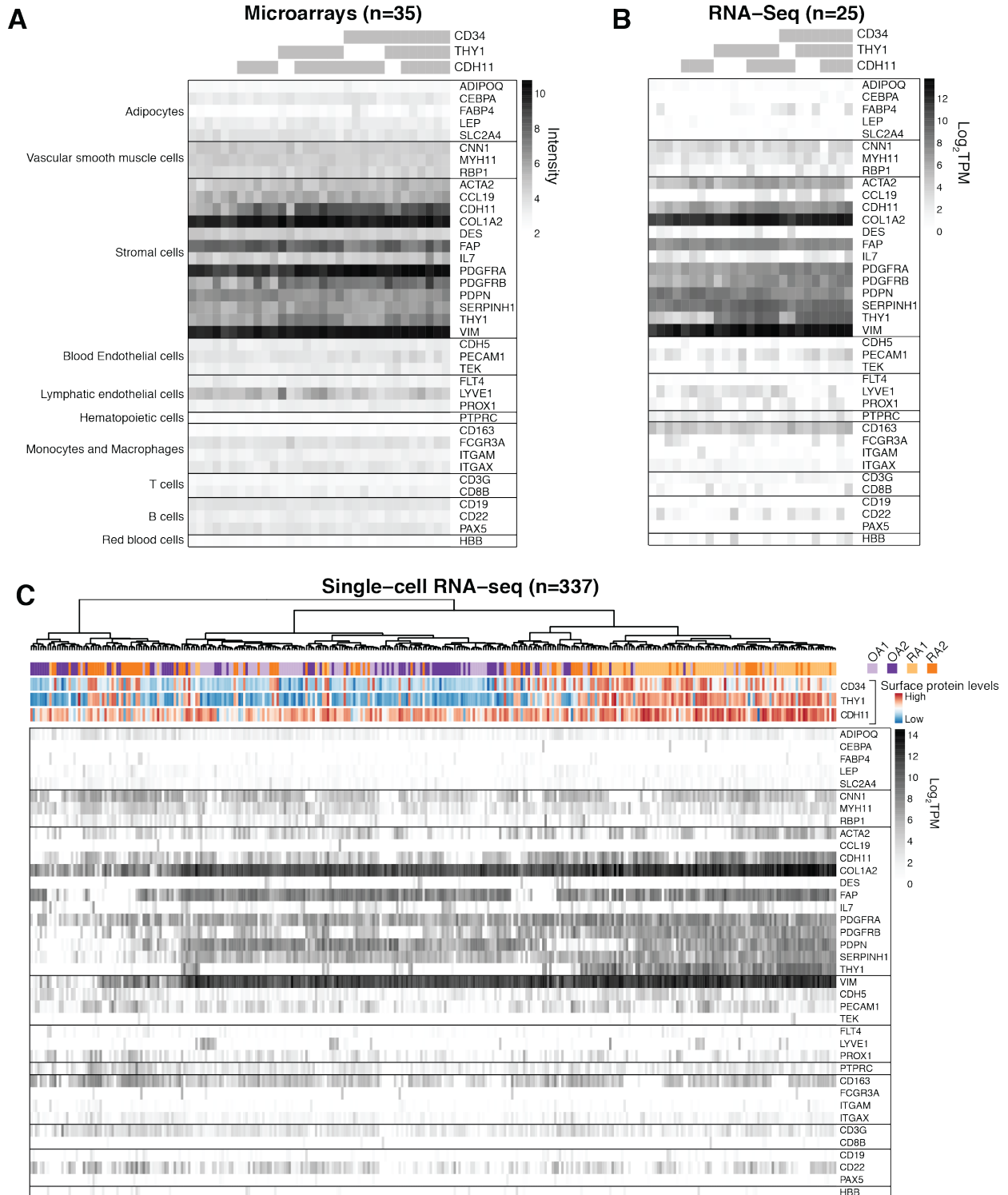
First, we assayed each of the 7 gated populations from 3 OA and 3 RA donors with the Affymetrix HuGene 2.0 ST microarray, using Robust Multichip Average to normalize 53,617 probesets and 20,452 genes. As expected, we observed that all of the samples expressed genes typically expressed in fibroblasts and lacked expression of other lineage specific genes (**Figure 4.2A**). However, after controlling for variation between donors, we observed 2,986 genes with significant variation across the 7 distinct populations gated based on surface protein marker expression (FDR<1%, analysis of variance (ANOVA)), suggesting notable transcriptional differences across these putative subsets. To validate these microarray findings, we applied RNA-seq to the same 7 subsets from 4 independent RA donors. We prepared libraries with Smart-Seq2, sequenced to an average depth of 5.6M fragments per sample, and quantified expression for 19,532 genes. These samples were also enriched with fibroblast lineage genes (**Figure 4.2B**). Those genes that were significantly differentially expressed in the microarray experiment and that overlapped with the RNA-seq experiment (n=2,659 genes) had similar expression profiles across the 7 subsets (**Figure 4.3**). In total, 436 genes were measured on both platforms and had significant variation across the 7 putative subpopulations in both experiments (1% FDR, ANOVA) (**Figure 4.1B**). The concordance of these results suggests that the gene expression differences reflect biological variation rather than technical or stochastic artifacts. We

reasoned that these expression profiles could serve as proxies for molecular functions to define putative cellular subsets with distinct biological roles.

Figure 4.2. Heatmap of expression data for genes characteristic of cell types in synovial tissues.

We used fluorescence activated cell sorting (FACS) to enrich samples for mesenchymal stromal cells (MSCs) and deplete other cell types. **(A)** Microarray data (n=32) and **(B)** RNA-seq data (n=25) are annotated with 3 tracks indicating FACS sorting gates for CD34, THY1, and CDH11. **(C)** Single-cell RNA-seq data (n=346) is annotated with FACS data for CD34, THY1, and CHD11, but these cells were not sorted on these markers (see Methods for details). Most samples show high expression of lineage markers for stromal cells, and low expression of lineage markers for adipocytes, vascular smooth muscle cells, blood endothelial cells, lymphatic endothelial cells, hematopoietic cells, monocytes and macrophages, T cells, B cells, and red blood cells.

Figure 4.2 (Continued)



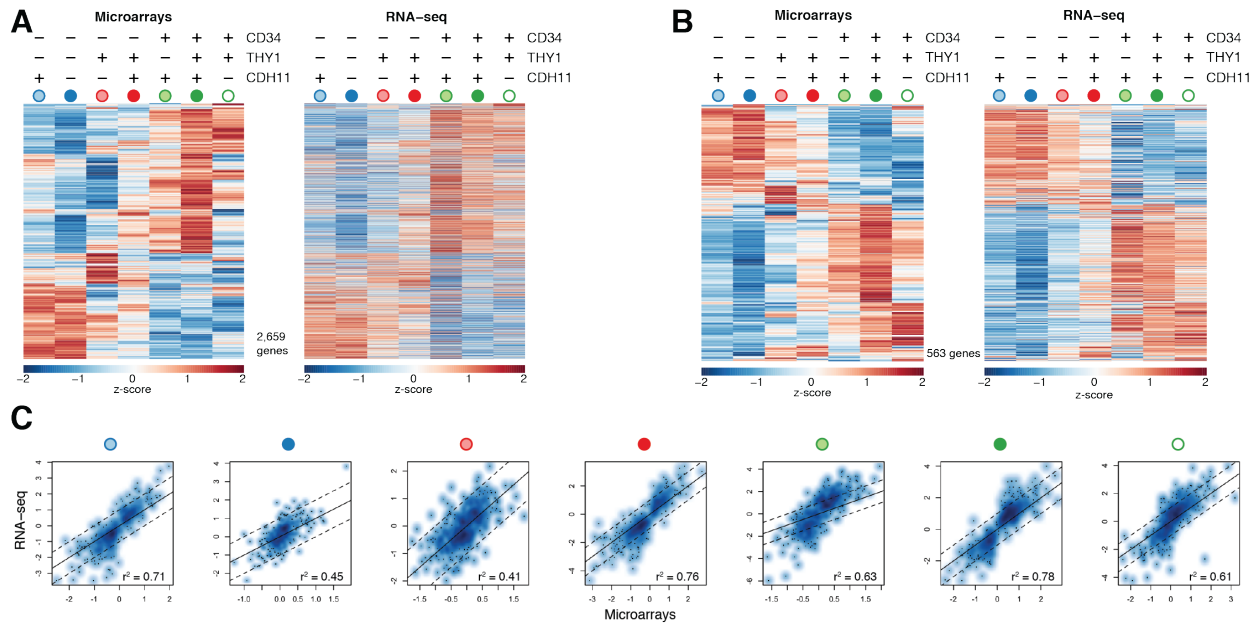


Figure 4.3. Concordance of differential expression between microarrays and low-input RNA-seq.

(A) ANOVA reveals 2,659 genes that have significant F statistic (1% FDR) across 7 gated subpopulations of cells in the microarray data and are measured in the RNA-seq data. These genes show a consistent pattern across the 7 gated subpopulations. (B) 563 genes significant in the RNA-seq data and measured in the microarray data. (C) 436 genes significantly differentially expressed in both datasets. The x-axis shows normalized expression in microarray data and y-axis shows normalized expression in RNA-seq data.

Principal component analysis revealed that 7 phenotypic populations fall into 3 distinct major subsets: CD34⁻THY1⁻, CD34⁻THY1⁺, and CD34⁺. Principal component 1 (PC1) clearly separates CD34⁻ and CD34⁺ samples and PC2 separates the CD34⁻ samples that are THY1⁺ and CDH11⁺ (**Figure 4.1C**). The 3 subsets were also clearly apparent by hierarchical clustering on pairwise Pearson correlations (**Figure 4.1D**). CD34⁻ and CD34⁺ samples were positively correlated within subsets and negatively correlated between subsets. CD34⁻THY1⁺ sample correlations were less consistent between the 3 subsets and within this subset, indicating that this subset may be more heterogeneous than the other two (**Figure 4.1D**). We decided to group

CD34⁺THY1⁺CDH11⁺ and CD34⁺THY1⁺CDH11⁻ samples because they had similar gene expression profiles overall.

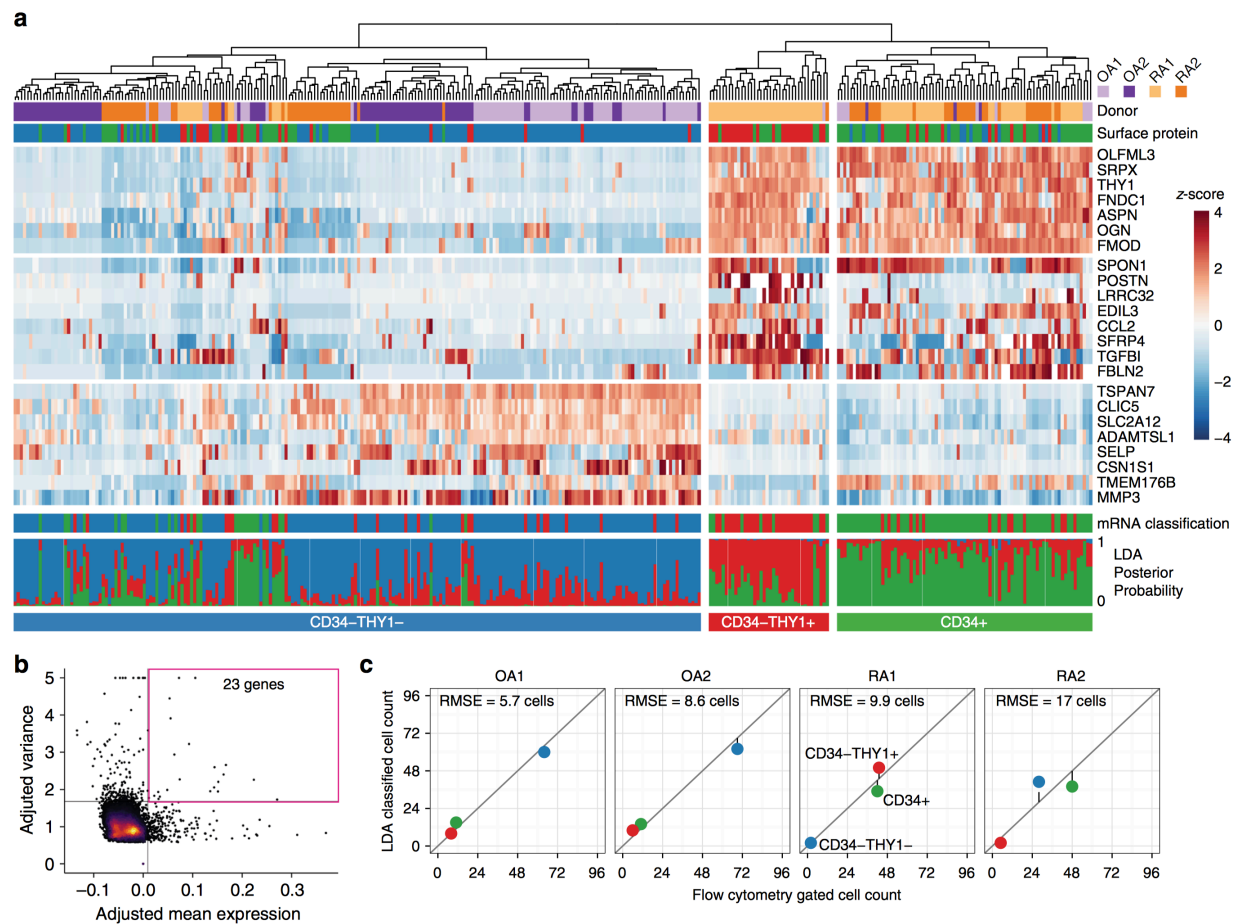


Figure 4.4. Unbiased clustering of fibroblasts by single-cell RNA-seq.

(a) Hierarchical clustering is concordant with our surface protein gating. The first track indicates which donor the cell came from. The second track indicates which cell sorting gate the cell falls into, by surface protein expression. The heatmap indicates mRNA expression, scaled across cells to reveal variation between cells. Below the heatmap, the first track shows linear discriminant analysis (LDA) classification by mRNA expression of 968 genes. The last track shows posterior probabilities of LDA classification. (b) For the heatmap, we selected 23 genes with greatest 1% mean and greatest 1% variance in expression levels. (c) The number of cells in each subset as determined by surface protein levels (x-axis) and by LDA classification (y-axis) is consistent. RMSE, root mean squared error.

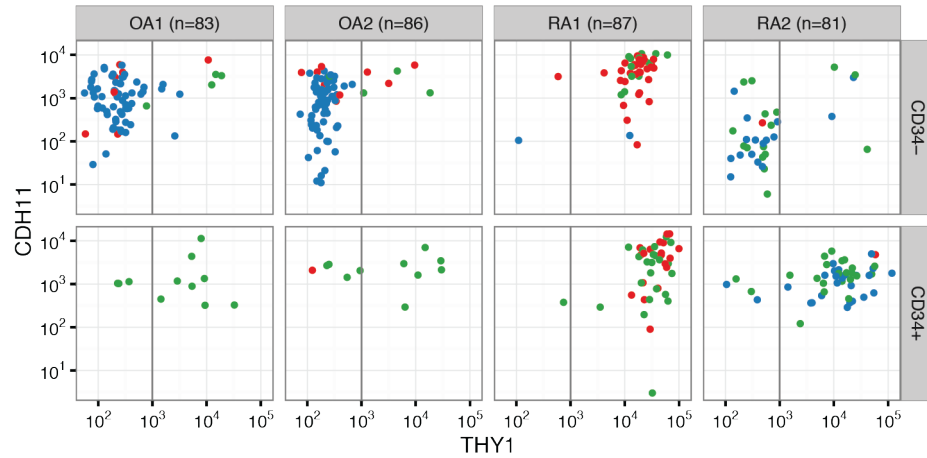


Figure 4.5. Agreement between protein gating and LDA classification based on single cell RNA-seq.

Single cells were colored by mRNA classification and displayed as they would have been gated by protein levels. For example, CD34⁻THY1⁺ cells are expected to be red if the two methods are in perfect agreement. Agreement between protein gating and LDA classification for OA1, OA2, RA1, RA2 is 82%, 87%, 62%, 44%, respectively. Protein level boundaries were set at 600 for CD34 and 1,000 for THY1.

Single cell RNA-seq confirmed the presence of three major fibroblast subsets

Next, since our observation from gated fibroblast populations may be potentially biased by our a priori selection of surface markers, we performed single cell mRNA sequencing to obtain an unbiased characterization of transcriptional heterogeneity in fibroblasts. We depleted hematopoietic cells, RBCs, and endothelial cells, and selected cells positive for PDPN (PTPRC⁻ GYPA⁻ PECAM1⁻ PDPN⁺) (**Figure 4.2C**). We used Illumina Smart-Seq2 to prepare single cell libraries from fibroblasts from 4 additional donors, 2 RA and 2 OA; we sequenced to an average depth of 5.4M fragments per cell, and detected an average of 8,842 genes per cell with at least 1

transcript per million (TPM) (**Figure 4.6**). We selected 337 cells with at least 5,000 detected genes, and then used Single Cell Differential Expression (SCDE) to estimate error models for each cell, normalize expression values, remove aspects of variation due to library complexity, and correct for batch effects across donors ⁸⁴. We measured protein levels on the single cells by flow cytometry at the time of single cell sorting for comparison with the single cell transcriptomic analyses. Remarkably, unbiased subsets (defined by hierarchical clustering of single cells with 23 genes that show high mean and variance across cells) were concordant with the three subsets defined by protein levels of CD34 and THY1 surface markers (permutation $P < 10^{-4}$) (**Figure 4.4A,B**).

Next, we assessed whether the bulk and single cell RNA-seq profiles were marking similar cellular subpopulations. We trained a linear discriminant analysis (LDA) classifier on 968 genes in the bulk RNA-seq data and predicted the classes of single cells (see **Supplementary methods**). The LDA classifier produced a probability of belonging to each class (CD34⁻THY1⁻, CD34⁻THY1⁺, and CD34⁺). The classifications were confident (median probability 0.8) and consistent with the three subsets (**Figure 4.4A**). We compared mRNA classification with flow cytometric protein marker data recorded at the time of unbiased single cell sorting and found concordant cell identities and proportions (**Figure 4.4C**, **Figure 4.5**). These single-cell RNA-seq results provide an unbiased and independent validation of the 3 major subsets defined with bulk transcriptomics of samples gated by protein surface markers.

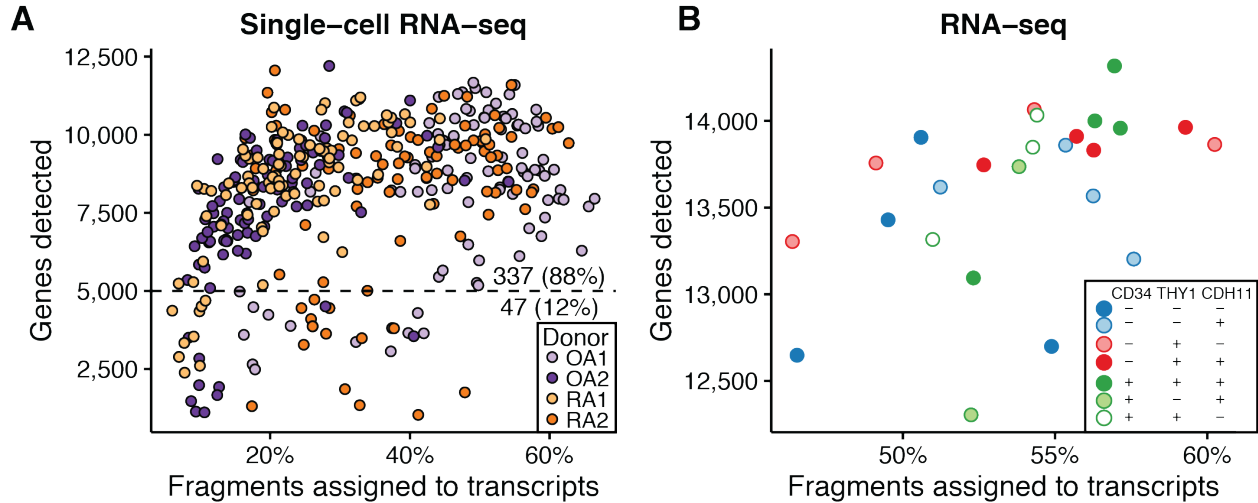


Figure 4.6. Assessing quality of RNA-seq data.

Distributions of number of cDNA fragments assigned to mRNA transcripts and the number of genes detected with at least 1 transcript per million (TPM). (A) 337 of 384 (88%) cells have 5,000 or more genes detected. We excluded 47 cells with fewer than 5,000 genes detected from analysis. These cells also have low abundance of mitochondrial ribosomal protein genes and stromal cell lineage genes. (B) As a point of reference, 25 bulk RNA-seq samples are all high quality in terms of fragments assigned to transcripts and number of genes detected.

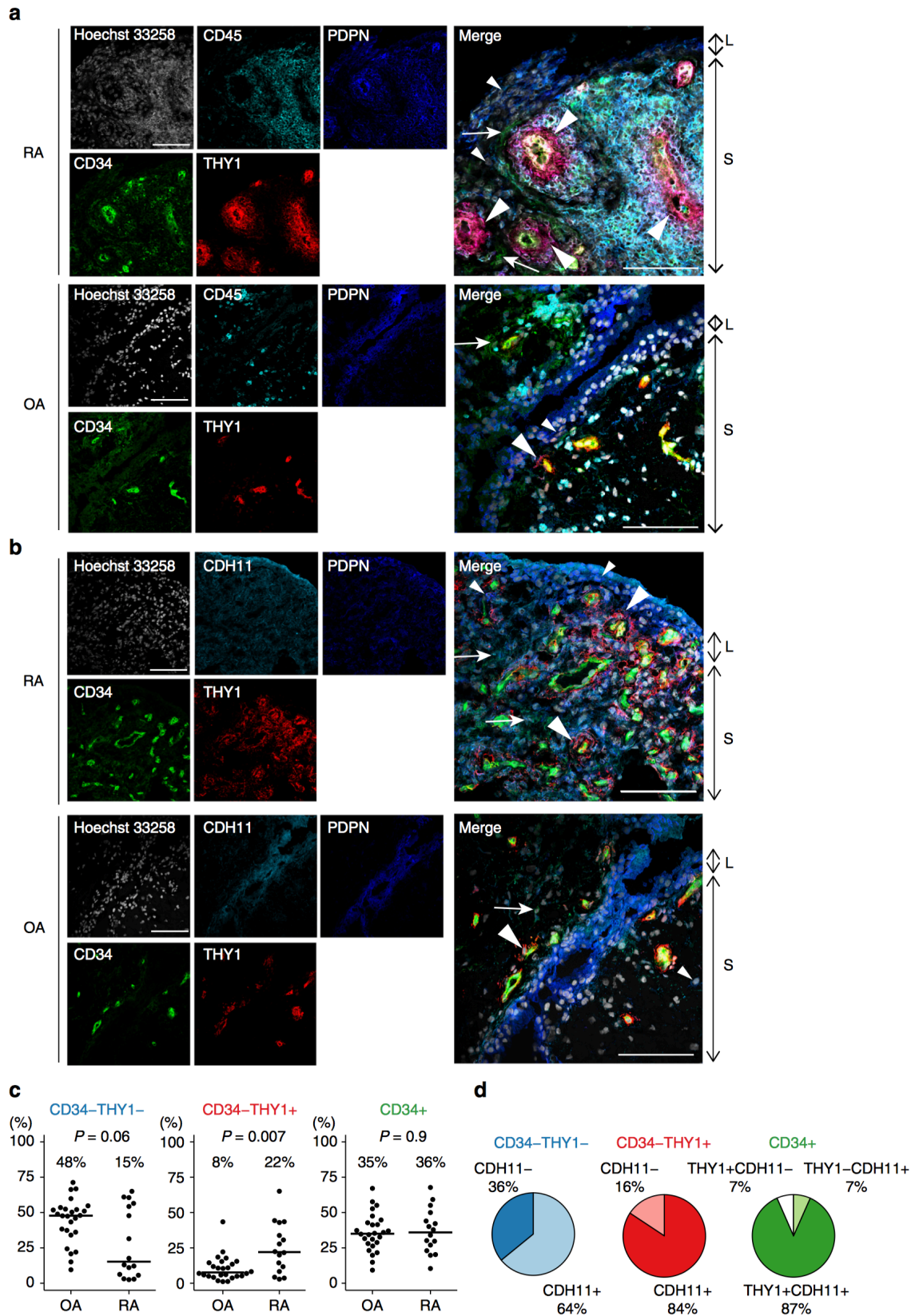
Fibroblast subsets localize to specific regions of lymphoid structures in the synovium

Next, we characterized the anatomical localization of fibroblast subsets in synovial tissue and assessed differences between OA and RA (**Figure 4.7A,B**). CD34⁻THY1⁺ fibroblasts in RA samples appeared as expanded masses of cells in the deeper sublining layer of the synovium surrounding blood vessels, especially near accumulations of lymphocytes. In OA samples, the CD34⁻THY1⁺ fibroblasts comprise a thin layer with fewer cells surrounding blood vessels than in RA samples. CD34⁺ fibroblasts were observed in both superficial lining and deeper sublining areas of the synovium. CD34⁻THY1⁻ fibroblasts were mostly observed in lining area.

Figure 4.7. Anatomical localization and proportions of fibroblast subsets.

(a) Anatomical localization of fibroblast subsets and leukocytes in RA and OA synovial tissue. Hoechst 33258: white, CD45: cyan, PDPN: blue, CD34: green, THY1: red. Small arrowheads: CD34 – THY1 – fibroblasts. Big arrowheads: CD34 – THY1 + fibroblasts. Arrows: CD34 + fibroblasts. Scale = 100 μ m. (b) Expression of fibroblast subset markers in RA and OA synovial tissue. Hoechst 33258: white, CDH11: cyan, PDPN: blue, CD34: green, THY1: red. (c) Proportions of fibroblast subsets in synovial tissue in OA (n = 26) and RA (n = 16) evaluated by flow cytometry. P values from the Wilcoxon's rank-sum test. d Proportions of CDH11 + cells in CD34 – THY1 – fibroblasts, CD34 – THY1 + fibroblasts, and CD34 + fibroblasts. L lining area, S sublining area.

Figure 4.7 (Continued)



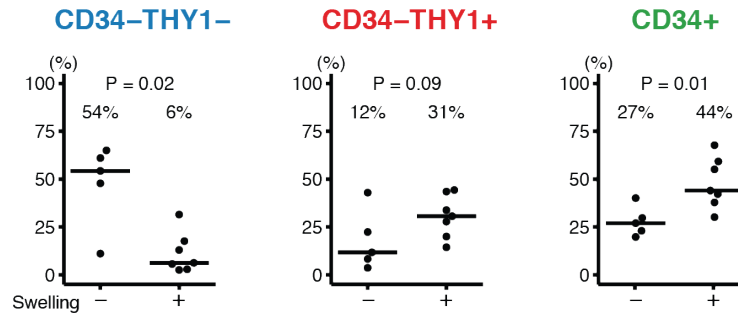


Figure 4.8. Proportion of fibroblast subsets in swollen and non-swollen joints in RA patients.

Proportion of cells were evaluated by flow cytometry. Swollen joint: n=7. Non-swollen joint: n=5.

Fibroblast subset proportions are altered in RA

We hypothesized that if pathological fibroblast subsets exist, then some subsets might be more or less abundant in RA relative to OA synovial tissues. Indeed, proportions of fibroblast subsets defined by flow cytometry with protein surface markers were different between RA synovial tissue (n=16) and OA (n=26) (**Figure 4.7C, Table 4.1**). The proportion of CD34⁻THY1⁺ fibroblasts comprised a median of 22% of total fibroblasts in RA compared to 8% in OA (OR=3 (95% CI 1.33-6.48), P=0.007). By contrast, CD34⁻THY1⁻ cells were less abundant in RA at 15% compared to 48% in OA (OR=0.48 (95% CI 0.23-1.03), P=0.06). Within 12 RA samples, the 7 samples that were obtained from swollen joints (according to a rheumatologist's assessment) had different subpopulation proportions than 5 samples from non-swollen joints (**Figure 4.8**). The swollen joints had fewer CD34⁻THY1⁻ (P=0.02), more CD34⁻THY1⁺ (P=0.09), and more CD34⁺ fibroblasts (P=0.01). Notably, most cells in the expanded CD34⁻THY1⁺ population in RA also were positive for surface protein levels of CDH11 (median 84%), and CDH11 was also expressed on the other fibroblast subsets including CD34⁻THY1⁻ (median 64%) and CD34⁺

(median 94%) cells. These results suggest that synovial fibroblast subpopulation proportions are closely related to disease type and activity.

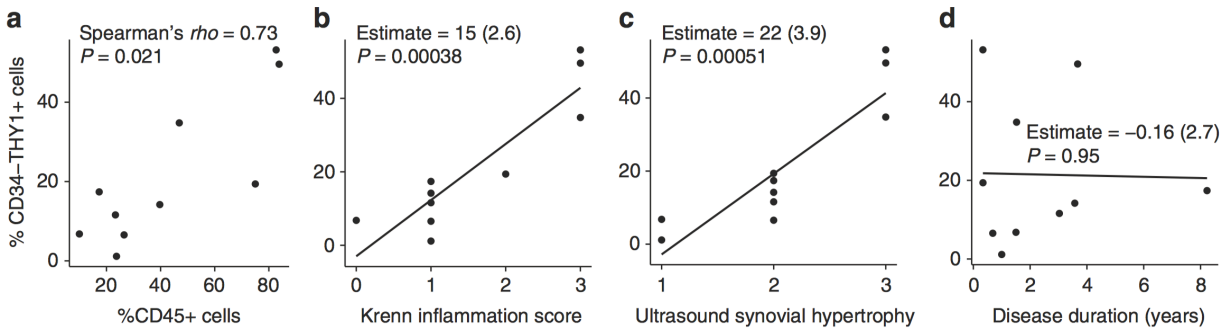


Figure 4.9. Correlation between proportion of fibroblast subsets and synovial inflammation.

(a) The Spearman's correlation between the proportion of CD34 – THY1 + fibroblasts in total fibroblasts and the proportion of CD45 + cells in total live cells in synovial tissue biopsies from RA knee joints. (b–d) The linear correlation coefficient between the proportion of CD34 – THY1 + fibroblasts and Krenn inflammation score (b), ultrasound synovial hypertrophy score (c), and disease duration (years) (d). Proportion of cells were evaluated by flow cytometry. Plots b–d show linear model coefficient estimates and t-test P values.

We note that all OA samples were taken from the knee, while RA samples included those from the knee (n=8) as well as other smaller joints (n=8) (**Table 4.1**). To confirm that the altered proportion of fibroblast subsets in RA reflects the level of tissue inflammation, rather than joint location of origin, we collected independent RA synovial tissue biopsies from only knee joints and examined the proportion of fibroblast subsets and infiltrated leukocytes by flow cytometry (**Table 4.2**). We selected only samples with synovial hypertrophy on ultrasound images. Similar to our RA versus OA comparison, we observed that the proportion of CD34⁺THY1⁺ fibroblasts is positively correlated with the proportion of infiltrated leukocytes by flow cytometry (**Figure 4.9A**). In addition, the proportion of CD34⁺THY1⁺ fibroblasts correlated with both histological

synovitis and synovial hypertrophy assessed by ultrasound (**Figure 4.9B, C**). These results indicate that the altered proportion of fibroblast subsets in RA reflects tissue inflammation at both the molecular and clinical level. In contrast, the correlation between the proportion of CD34-THY1+ fibroblasts and disease duration was not observed, suggesting that the altered proportion of fibroblast subsets is not a secondary effect of chronic tissue damage (**Figure 4.9D**).

Table 4.2. Clinical characteristics of patients who donated synovial biopsy samples from knee joint.

	Average $\pm$ SD
Number	10
Age (years)	55.4 $\pm$ 11.3
Female (%)	40
CDAI	41.7 $\pm$ 12.3
RF or ACPA	
Positive (%)	70
Negative (%)	30
CRP (mg/l)	17.0 $\pm$ 22.4
ESR (mm/hr)	13.3 $\pm$ 9.6
Disease duration (years)	2.4 $\pm$ 2.4
Methotrexate (%)	60
Other syntethic (%)	20
Biologic (%)	20
Glucocorticoid (%)	0

CDAI; Clinical Disease Activity Index, CRP; C-Reactive Protein, ESR; Erythrocyte Sedimentation Rate. Other syntethic includes leflunomide and sulfasalazine. Biologic is etanercept.

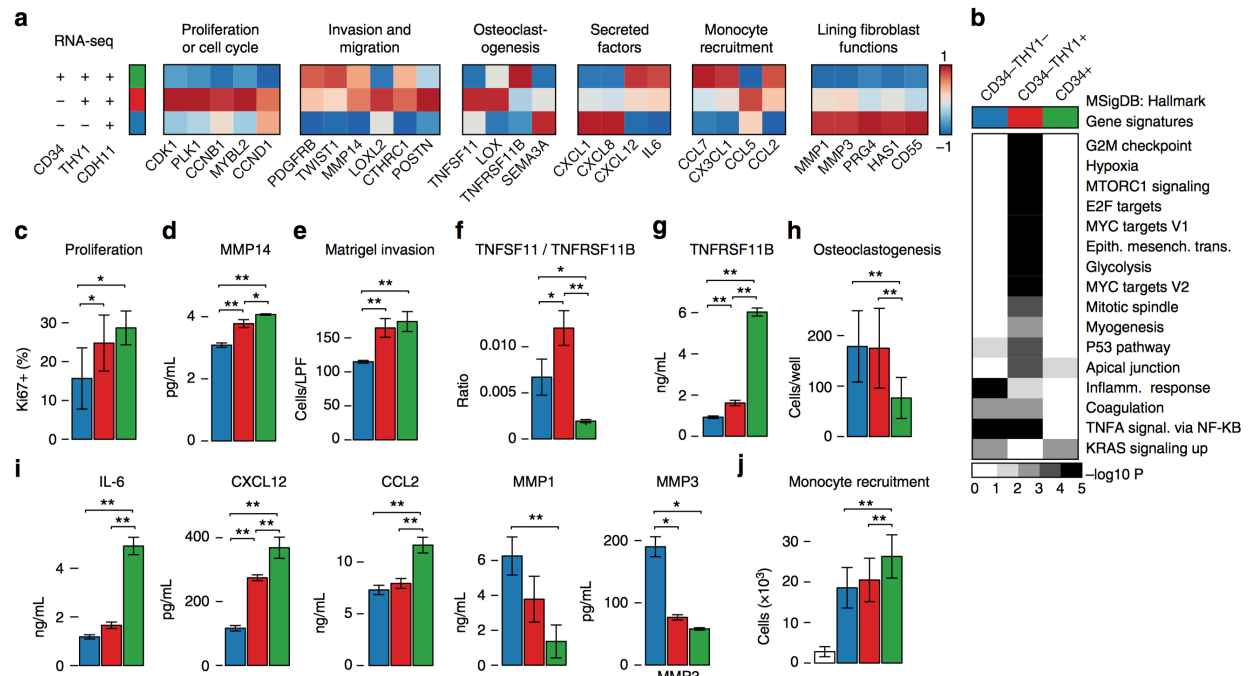


Figure 4.10. Functional differences between fibroblast subsets.

(a) Gene expression pattern of key effector molecules to predict the function of fibroblast subsets. (b) Gene set enrichment analysis with RNA-seq samples and MSigDB Hallmark Gene Signatures. Terms with <5% FDR are displayed. (c) The proportion of Ki67-positive cells in freshly isolated synovial fibroblast subsets. (d) Protein expression of MMP14 in the cell lysates of expanded fibroblast subsets in the presence of 1 ng/mL TNF- α . (e) The number of migrated and invaded cells in transwell matrix invasion assay. Fifty nanogram per mL of PDGFB was used to promote migration and invasion. (f) The ratio of TNFSF11 (RANKL) and TNFRSF11B (OPG) in mRNA level in expanded fibroblast subsets. (g) Protein expression of TNFRSF11B in the supernatant of expanded fibroblast subsets. (h) Number of osteoclasts developed in the co-culture of osteoclast progenitors with fibroblast subsets in the presence of 20 ng/mL M-CSF and 5 ng/mL RANKL. (i) Protein expression of cytokines and MMPs in the supernatant of expanded fibroblast subsets in the presence of 1 ng/mL TNF- α . (j) The number of migrated monocytes in monocyte recruitment assay using supernatant of cultured fibroblast subsets. In bar graphs, blue, red and green color coding corresponds to the fibroblast subsets shown in a. White bar in j indicates culture media without fibroblasts. Bars show means and error bars show standard deviations. One-way ANOVA, Tukey's post hoc: *P < 0.05; ** P < 0.005.

Transcriptomics predicts distinct fibroblast functions

We used gene expression profiles as proxies for molecular functions to investigate functional differences between fibroblast subsets. Gene set enrichment analysis prompted us to investigate how the expression patterns of cytokines are different between fibroblast subsets (**Figure 4.10A**)^{85,86}. Each fibroblast subset expresses distinct types of cytokines and matrix metalloproteinases (MMPs) (**Figure 4.10B**). CD34⁻THY1⁺ fibroblasts express *TNFSF11* (*RANKL*), and lowly express *TNFRSF11B* (*OPG*), a decoy receptor for *TNFSF11* (**Figure 4.10B**). The increase of CD34⁻THY1⁺ fibroblasts in RA may explain the increased bone destruction and accumulation of lymphocytes in RA synovial tissue, since RANKL is essential for osteoclastogenesis and is involved in T cell trafficking in autoimmune inflammation^{87,88}.

CD34⁻THY1⁻ fibroblasts express high levels of *CXCL1* and *CXCL8*, essential for neutrophil recruitment and activation (**Figure 4.10B**)⁸⁹. They also express *BMP6*, known to promote osteoblastic bone formation (**Figure 4.10B**)⁹⁰. Depletion of CD34⁻THY1⁻ fibroblasts may explain the decreased bone formation activity in RA. CD34⁺ fibroblasts express a variety of cytokines including *IL6*, *CXCL12* and *CCL2*, crucial cytokines that cause tissue inflammation (**Figure 4.10B**)^{16,76}.

To validate our transcriptional analysis, we expanded sorted fibroblast subsets *in vitro* and analyzed their protein expression by ELISA and mRNA expression by qPCR. Consistent with microarray and RNA-seq data in freshly isolated cells, CD34⁻THY1⁻ fibroblasts express high levels of MMP1 and MMP3 in response to TNF stimulation; whereas, CD34⁺ fibroblasts express high levels of IL6 and CXCL12 in response to TNF stimulation (**Figure 4.10C,D** and **Figure 4.11**). CD34⁻THY1⁺ fibroblasts express high levels of *TNFSF11* (*RANKL*) and low levels

of *TNFRSF11B* (*OPG*) relative to CD34⁻THY1⁻ or CD34⁺ fibroblasts (**Figure 4.10E** and **Figure 4.11**).

We hypothesized that the differential expression of *TNFSF11* and *TNFRSF11B* between freshly isolated fibroblast subsets would affect their ability to promote osteoclastogenesis, an important effector function for fibroblasts. Indeed, when we co-cultured fibroblast subsets with monocytes in osteoclastogenic conditions *in vitro*, we observed that the number of osteoclastic cells developed from monocytes was greater in the co-culture with CD34⁻THY1⁻ fibroblasts or CD34⁻THY1⁺ fibroblasts compared to CD34⁺ fibroblasts (**Figure 4.10F**). This suggests that fibroblast subsets differ in their ability to promote osteoclastogenesis and that this function can be detected during short term *in vitro* culture.

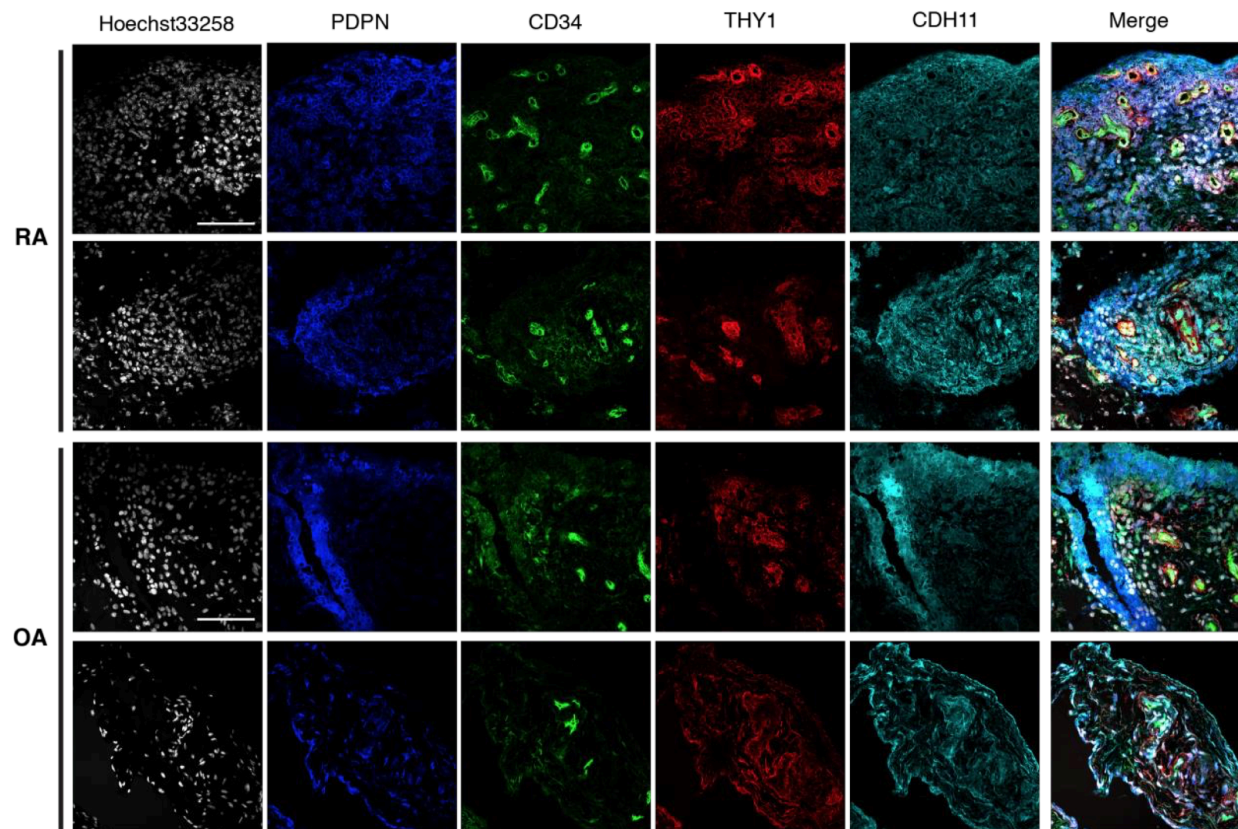


Figure 4.11. Expression of fibroblast subset markers in RA and OA synovial tissue.

Hoechst 33258: White, CDH11: Cyan, PDPN: Blue, CD34: Green, THY1: Red. Scale = 100 μ m.

Conclusion

This study identifies subsets of fibroblasts in fresh human synovium, including a distinct subset of PDPN⁺CD34⁻THY1⁺ fibroblasts that is expanded in RA and may be pathogenic. These cells are enriched around blood vessels in RA synovium, and their expression profile reveals potential pathogenic roles in lymphocyte recruitment and osteoclastogenesis. We note that almost all of

these cells are positive for CDH11, which we have previously shown to be associated with pathological behavior of fibroblasts in *in vitro* studies and RA mouse models ⁷⁵.

Expansion of fibroblasts is a dynamic component of RA synovitis. An increase in synovial lining fibroblasts was noted previously to correlate with disease activity score in 28 joints (DAS28), disease duration, and the level of macrophage infiltration ⁹¹. Expansion of synovial sublining fibroblasts is also observed in RA, but no previous literature reports significant correlation between sublining fibroblasts and other clinical or pathological findings, except for a negative correlation with DAS28 in one study ⁹¹. Here, by separating fibroblasts into subsets based on the expression patterns of multiple markers, we found that the increase of CD34⁻THY1⁺ fibroblasts around blood vessels in the sublining area is a dominant change in fibroblasts in RA synovium. Moreover, this expansion distinguishes RA from OA, reflects RA disease activity, and correlates with lymphocyte infiltration in the synovium. Previous studies have shown that mesenchymal CDH11 determines adhesion between fibroblasts, increases their migration and invasion and synergizes in the activation of fibroblasts to produce MMPs, cytokines, and chemokines ^{75,92,93}. Since a large majority of the expanded fibroblast population expresses this cadherin, it also may contribute to their pathologic behavior in RA (**Figure 4.7E**).

There is a real need for strategies to define fibroblast heterogeneity and pathogenic fibroblast populations in order to understand the complex nature of tissue pathology and the role of tissue resident cells in end organ damage. Our approach involves an integrative analysis of cell surface markers, bulk transcriptomes, single-cell transcriptomics, and histological imaging of human tissues that identified a disease-related fibroblast subpopulation that may ultimately serve as a specific target for therapy. We anticipate that such an approach may serve as a template for future studies to identify pathogenic subsets of tissue cells in other human diseases.

Materials and Methods

Patient recruitment and isolation of synovial cells

We obtained synovial tissue from joint replacement surgeries, synovectomy surgeries or synovial biopsies of OA or RA patients with appropriate informed consent as required. The study protocols are Institutional Review Board approved at Partners HealthCare, Hospital for Special Surgery, and the University of Birmingham Local Ethical Review Committee. We modified previously described protocols to isolate synovial cells^{94,95}. Briefly, we obtained tissue immediately after the surgeries. We removed bone and adipose tissues with scissors. We cut synovial tissues into small pieces, and then subjected these pieces to enzymatic digestion. For microarray analysis, low input bulk RNA-seq and *in vitro* assays, we digested tissues with 4 mg/ml collagenase type 4 (Worthington, NJ), 0.8 mg/ml DispaseII, 0.1 mg/ml DNaseI (Roche) in DMEM at 37°C. After 15 minutes, we collected the supernatant and replaced with fresh enzyme mix. We repeated these procedures every 15 minutes for total 1 hour. For the analysis of synovial biopsy samples, we digested the tissues with 0.05 mg/ml Liberase TM (Roche) and 0.04 mg/ml DNaseI at 37°C for 30 minutes. For mass cytometry analysis, we digested tissues with 0.2 mg/ml Liberase TL (Roche), 0.1 mg/ml DNaseI in RPMI at 37°C for 20 minutes to minimize the cleavage of surface markers of lymphocytes during the enzymatic digestion. The same protocol was used to isolate synovial cells for single cell RNA-seq. After lysing red blood cells with ACK-lysing buffer, we stained cells with antibodies, and sorted by FACS Aria Fusion (BD) with 100 µm nozzle at 20 psi. For the analysis with microarray and low input RNA-seq, the cells were sorted into 2% FBS HBS+ buffer, spun down, and lysed with TRIzol (Invitrogen). We extracted

RNA and cleaned up by RNeasy micro kit (QIAGEN) with DNaseI treatment. For single cell RNA-seq, the cells were stained with antibodies and directly sorted into 5 µl of TCL buffer (QIAGEN) with 1% β-mercaptoethanol (SIGMA) in 96 well plates. For the analysis with mass cytometry, we froze isolated synovial cells in CryoStor CS10 (STEMCELL Technologies, Vancouver, Canada) for batched analysis.

Fibroblast markers.

Selection of markers. We selected fibroblast markers for separation of putative fibroblast subsets on the basis of published literatures. PDPN is known to be a marker for fibroblastic reticular cells (FRCs) in lymph nodes and is highly expressed in fibroblasts in synovial tissues, especially in lining fibroblasts^{37,96}. CDH11 is also highly expressed in FRCs and lining fibroblasts in RA synovial tissue^{37,97}. THY1 has been widely used as a fibroblast marker in cultured cells, and is highly expressed in sublining fibroblasts in synovial tissue^{96,98}. Expression of CD34 has been documented in fibroblasts and endothelial cells^{99,100}. CD146 has been used as a marker for pericytes¹⁰⁰.

Antibodies and reagents. The following antibodies and reagents were used for the analysis of synovial cells with flow cytometry and cell sorting: anti-CD45-APC-H7 (2D1, BD Pharmingen), anti-CD235a-APC-Alexa Fluor750 (11E4B-7-6(KC16), Beckman Coulter), anti-CD31-PE-Cyanine7 (WM-59, eBioscience), anti-CD146-BV450 (P1H12, BD Horizon), anti-CD34-PE (4H11, eBioscience), anti-Podoplanin-PerCP-eFluor710 (NZ-1.3, eBioscience), anti-THY1-FITC (5E10, BD Pharmingen), anti-Cadherin11-Biotin (23C6), human TruStain FcX (Biolegend),

Streptavidin-APC (Jackson ImmunoResearch), Live/Dead fixable aqua dead cell stain kits (Molecular probes).

For immunofluorescence staining of synovial tissue, following antibodies and reagents were used: anti-CD45 (135-4C5, AbD serotec), anti-CD34 (EP373Y, Abcam), anti-Podoplanin (NZ-1.3, eBioscience), anti-THY1 (F15-42-1, Merck Millipore), anti-mouse IgG1-FITC (Southern Biotech), anti-mouse IgG2b-Alexa Fluor 647 (Life Technologies), anti-rat IgG-Alexa Fluor 594 (Life Technologies), anti-rabbit IgG-Alexa Fluor 546 (Life Technologies), and Hoechst 33258 (Life Technologies).

Transcriptional analysis.

Gene expression microarrays. We evaluated the integrity of RNA with Bioanalyzer or by TapeStation (Agilent). We used only RNA with more than RIN score of 7. We prepared cDNA from 38.1 µg total RNA using Ovation Pico WTA (NuGEN), followed by labeling 5 µg cDNA using Biotin Module (Nugen). We assayed gene expression using the GeneChip Human 2.0 ST microarray (Affymetrix). We normalized expression by Robust Multiarray Averaging (RMA). We identified and removed two outlier arrays by principal components analysis (PCA). We assigned probes to Entrez Gene IDs using Ensembl BioMart on March 17, 2015. In the instance where there were multiple probesets assigned to a single gene, we averaged them to obtain a single gene value.

RNA-seq library preparation and sequencing. We used 1,250 cells per sample for library preparation. We prepared sequencing libraries using the Smart-Seq2 protocol. We pooled and

sequenced libraries were pooled and sequenced with the Illumina HiSeq 2500 to a depth of 8-14M reads per library.

Single-cell RNA-seq sequencing. We assayed 384 fibroblasts from 4 donors, 2 with OA and 2 with RA. For each donor, we collected fresh synovial tissue, isolated synovial cells by enzymatic digestion, and stained with antibodies against CD45, CD235a, CD31, CD146, PDPN, CD34, THY1 and CDH11. We sorted 96 single CD45-CD235a-CD31-PDPN⁺ cells by FACS Aria Fusion (BD), and assayed mRNA expression with the Smart-Seq2 protocol ¹⁰¹. Single-cell libraries were also prepared with the same protocol, and we aimed to sequence to a depth of 200K-12M reads per library ¹⁰². On average, we sequenced 5.4M fragments and detected 8,153 genes per cell with at least 1 TPM. We discarded 47 cells (12%) with fewer than 5,000 genes detected from further analysis.

RNA-seq gene expression quantification. We quantified cDNAs on canonical chromosomes in Ensembl release 83 with Kallisto v0.42.4 ¹⁰³ in transcripts per million (TPM) and summed to get gene-level expression values. We quantified gene expression the same way for bulk and single cell RNA-seq samples. For differential expression analysis, hierarchical clustering, and principal components analysis (PCA), we log (base 2) transformed TPM values.

Lineage marker analysis of microarrays, bulk RNA-seq, and single-cell RNA-seq. We selected lineage markers for fibroblast, endothelial, and hematopoietic cells ⁷⁸ and checked their expression levels to confirm that our samples are from the fibroblast lineage (Fig. S1).

Differential expression analysis with microarrays. We used the R package limma to assess differential expression analysis on RMA normalized expression values ¹⁰⁴ Before performing differential expression analysis, donor-specific variation was regressed out by obtaining the residuals from linear models. Each gene was modeled as a linear combination of donor-specific effects, and the residuals from these models were used for differential expression analysis, hierarchical clustering, and PCA. We expected differences in gene expression between RA and OA. However, we lacked power to see these differences within fibroblast subsets.

Differential expression analysis with RNA-seq. RNA-seq data was analyzed the same way as the microarray data, starting with log2 transformed TPM expression values. Donor-specific variation was similarly regressed out before differential expression analysis, hierarchical clustering, and PCA.

Gene set enrichment analysis. Terms from MSigDB Hallmark Gene Signatures were used for enrichment analysis with Gene Set Enrichment Analysis ⁸⁵. We tested gene sets for enrichment with differential expression (DE) between the 3 major subsets in order to assess how they might differ from each other in terms of molecular pathways. MSigDB hallmark pathways enriched with DE signal for CD34-THY1+ population, expanded in RA joints, include "epithelial to mesenchymal transition", "hypoxia", and "glycolysis" ⁸⁶.

Principal components analysis (PCA). After gene selection by ANOVA or differential expression analysis, we scaled each sample and then scaled each gene across the samples to

obtain a specificity of the gene to each sample. Next, we used the `prcomp` function in R to perform PCA with centered and scaled log2 expression values.

Linear discriminant analysis (LDA). We checked if transcriptional profiles of single cells are similar to profiles of the 3 major subsets we defined by bulk transcriptomics. First, we selected 1,171 genes with significant (5% FDR) differential expression between any pair of the 3 subsets in bulk RNA-seq data. We selected a subset of 968 genes with high expression in single cells ($\text{mean log}_2(\text{TPM}) > \text{log}_2(10)$). We used the bulk RNA-seq data to train a linear discriminant analysis (LDA) model with these genes, and then classified each single cell's expression profile to predict each single cell's identity. The confidence of each classification is represented by a posterior probability.

Histological analysis. RA synovial tissues were obtained by biopsies from RA patients in the BEACON Birmingham early arthritis cohort, which is an early arthritis cohort recruiting patients with new onset arthritis prior to treatment with disease-modifying antirheumatic drugs. Synovial tissues for staining were frozen in OCT compound. Sections were made in 6 μm thickness, fixed with acetone, and frozen prior to use. Slides were rehydrated in PBS, blocked with 10% normal goat serum in PBS for 10 minutes, and then incubated with primary antibodies, followed by secondary antibodies. Slides were mounted using ProLong Diamond (Life Technologies), and imaged using a Zeiss LSM 780 confocal microscope. Images were processed using Zen Black and Zen lite (Zeiss). Representative images were shown. Synovial tissues for histological analysis and hematoxylin and eosin staining were fixed in formaldehyde then mounted in paraffin, sectioned and stained by the Hospital Pathology service.

Histological evaluation of inflammatory infiltrate. Hematoxylin and eosin-stained sections of knee synovial biopsy samples were examined histologically for the severity of inflammatory infiltrate using the inflammatory component of the Krenn synovitis score (37). Inflammatory infiltrates were graded from 0-3 (0=no inflammatory infiltrate, 1=few mostly perivascular situated lymphocytes or plasma cells, 2=numerous lymphocytes or plasma cells sometimes forming follicle-like aggregates, and 3= dense band-like inflammatory infiltrate or numerous large follicle-like aggregates). The tissues were graded in a blinded manner by two trained individuals then consensus agreed.

Clinical evaluation of synovitis by ultrasound. The joint to be biopsied was assessed using ultrasound immediately prior to the procedure using a Siemens Acuson Antares scanner (Siemens PLC, Bracknell, UK) and multifrequency (5-13MHz) linear array transducers. Synovitis and power Doppler (PD) positivity were defined using consensus OMERACT definitions (38). Greyscale synovial hypertrophy and Power Doppler ultrasound variables were graded on 0-3 semi-quantitative scales as previously reported (39).

Functional analysis

Cell culture. We sorted CD34-THY1⁻ fibroblasts, CD34-THY1⁺CDH11⁺ fibroblasts, and CD34⁺THY1⁺CDH11⁺ fibroblasts, and cultured them in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum (FBS; Gemini), 2 mM L-glutamine, antibiotics (penicillin and streptomycin), and essential and nonessential amino acids (Life

Technologies). The cells were expanded for 3-20 days for assays *in vitro*. The cells with one or two passages were used. The cells were cultured in the presence or absence of 1 ng/ml of TNF- α (R and D) for 24 hours for ELISA of IL-6, CXCL12, MMP-1, MMP-3 and MMP-14, or for 72 hours for qPCR of *TNFSF11* and *TNFRSF11B*.

ELISA. The levels of IL-6, CXCL12, MMP-1, MMP-3 and TNFRSF11B in the supernatant or the levels of MMP-14 in the cell lysate were evaluated by ELISA kit as described in manufacturer's instructions (Duo Set, R and D).

Quantitative Real-Time PCR (qPCR). cDNA was synthesized with QuantiTect Reverse Transcription kit (QIAGEN). qPCR was performed with Brilliant III Ultra-Fast SYBR Green qPCR master mix (Agilent Technologies) on a Mx3000 (Stratagene). Following primers were used; *TNFSF11*, forward; 5'-GGA GAG GAA ATC AGC ATC GAG, reverse; 5'-CCA AAC ATC CAG GAA ATA CAT AAC AC, *TNFRSF11B*, forward; 5'-CAA CAC AGC TCA CAA GAA CAG, reverse, 5'-GAA GGT GAG GTT AGC ATG TCC, *GAPDH* forward; 5'-AAT CCC ATC ACC ATC TTC CAG, reverse; 5'-AAA TGA GCC CCA GCC TTC.

Osteoclastogenesis assay. Osteoclast progenitors were prepared by culturing PBMC in the presence of 20 ng/ml of M-CSF (PeproTech) in DMEM supplemented with 10% FBS (GE Healthcare), 2 mM L-glutamine and antibiotics for 5-6 days. Expanded fibroblasts were seeded at 5,000 cells/well in 96 well plate. On the next day, osteoclast progenitors were added at 5,000 cells/well, and were co-cultures with fibroblasts in the presence of 20 ng/ml M-CSF and 5-20 ng/ml RANKL (PeproTech). The media was replaced every 2 days. After 6 days of the co-

culture, the cells were fixed by 4% PFA. After TRAP staining, TRAP positive multinucleated cells were counted as osteoclasts.

Mass cytometry. After thawing cryopreserved synovial cells, the cells were first stained with cisplatin for the assessment of viability, and then stained with antibody cocktails as described previously (40). The cells were fixed and permeated with intracellular fixation and permeabilization buffer (eBioscience), stained for intracellular markers, and re-fixed with formalin (Sigma). The stained cells were analyzed with Helios (Fluidigm, CA, USA). Data were normalized using EQ™ Four Element Calibration Beads (Fluidigm), and were analyzed with FlowJo 10.1 (FlowJo, OR, USA).

Chapter 5

**Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by
integrating single-cell transcriptomics and mass cytometry**

Rheumatoid arthritis (RA) is an autoimmune disease with chronic inflammation in the synovium of the joint tissue^{31,82,105}. This inflammation leads to joint destruction, disability and shortened life span¹⁰⁶. Defining key cellular subsets and their activation states in the inflamed tissue is a critical step to define new therapeutic targets for RA. CD4⁺ T cell^{107,108}, B cells¹⁰⁹, monocytes^{110,111}, and fibroblasts^{112,113} 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^{31,114–116}. 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^{117–119}. These technologies have already been used to discover roles for T peripheral helper (Tph) cells¹²⁰ and HLA-DR⁺CD27⁻ cytotoxic T cells¹²¹ in RA pathogenesis. Studies using scRNA-seq have defined myeloid cell heterogeneity in human blood¹²² and identified overabundance of PDPN⁺CD34⁻THY1⁺ (THY1, also known as CD90) fibroblasts in RA synovial tissue^{43,117}.

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 pipeline¹²³ 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.

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 5.1**) and obtained synovial tissues from ultrasound-guided biopsies or joint replacements (see Materials and Methods, **Figure 5.1a**). 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 sorting¹²³ (**Figure 5.1a**) to isolate B cells (CD45⁺CD3⁻CD19⁺), T cells (CD45⁺CD3⁺), monocytes (CD45⁺CD14⁺), and stromal fibroblasts (CD45⁻CD31⁺PDPN⁺) (**Figure 5.2a**). We applied bulk RNA-seq to all four sorted subsets for all 51 samples. For samples with sufficient cell yield (see Materials and Methods), we also measured single-cell protein expression using a 34-marker mass cytometry panel (n=26, **Table 5.2**), and single-cell RNA expression in sorted cell populations (n=21, **Figure 5.1b**).

Table 5.1 Clinical characteristics of 51 recruited patients.

		OA	leukocyte-poor RA	leukocyte-rich RA
		(n=15)	(n=17)	(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		15.7	5.5
	(range)		(<1-51)	(<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).

RF = Rheumatoid Factor.

CCP = Cyclic Citrullinated Peptide.

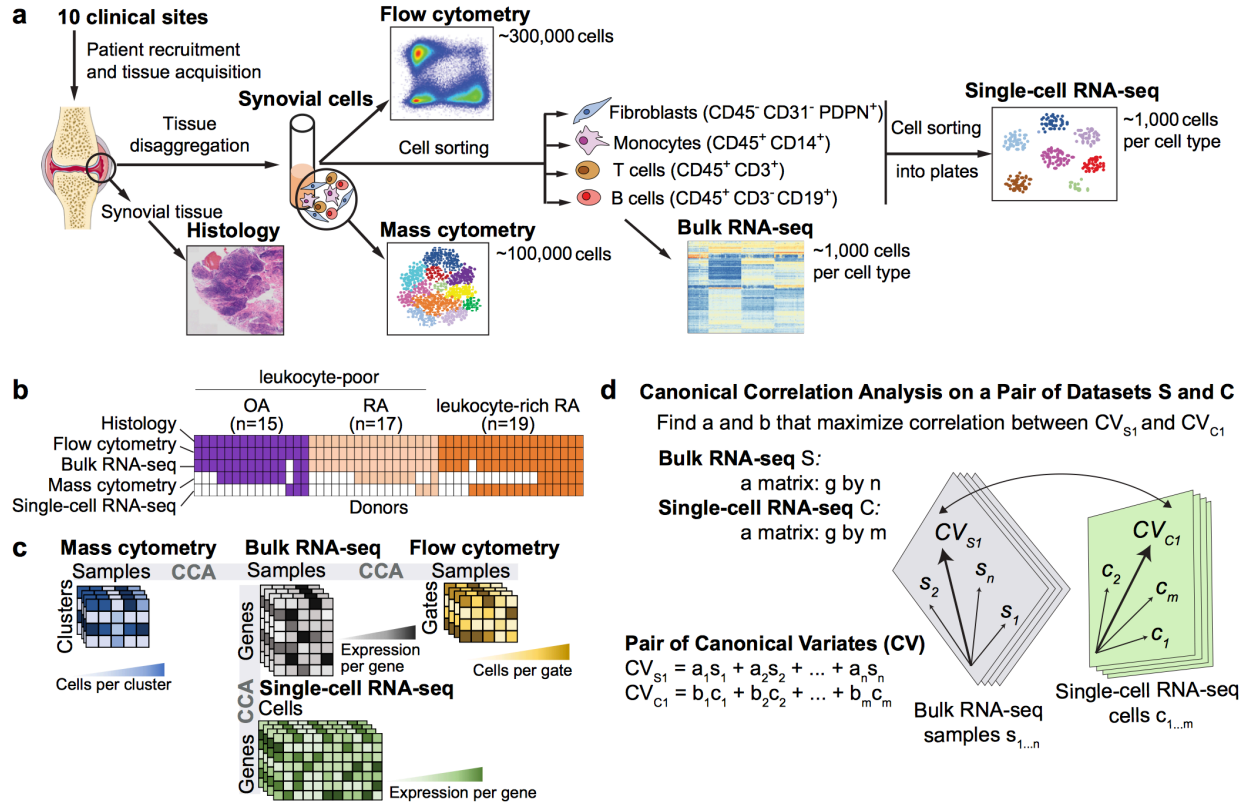


Figure 5.1 Overview of synovial tissue workflow and pairwise analysis of high-dimensional data.

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 schematic of the joint. **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.

Table 5.2 Panel of antibodies and metals for mass cytometry.

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

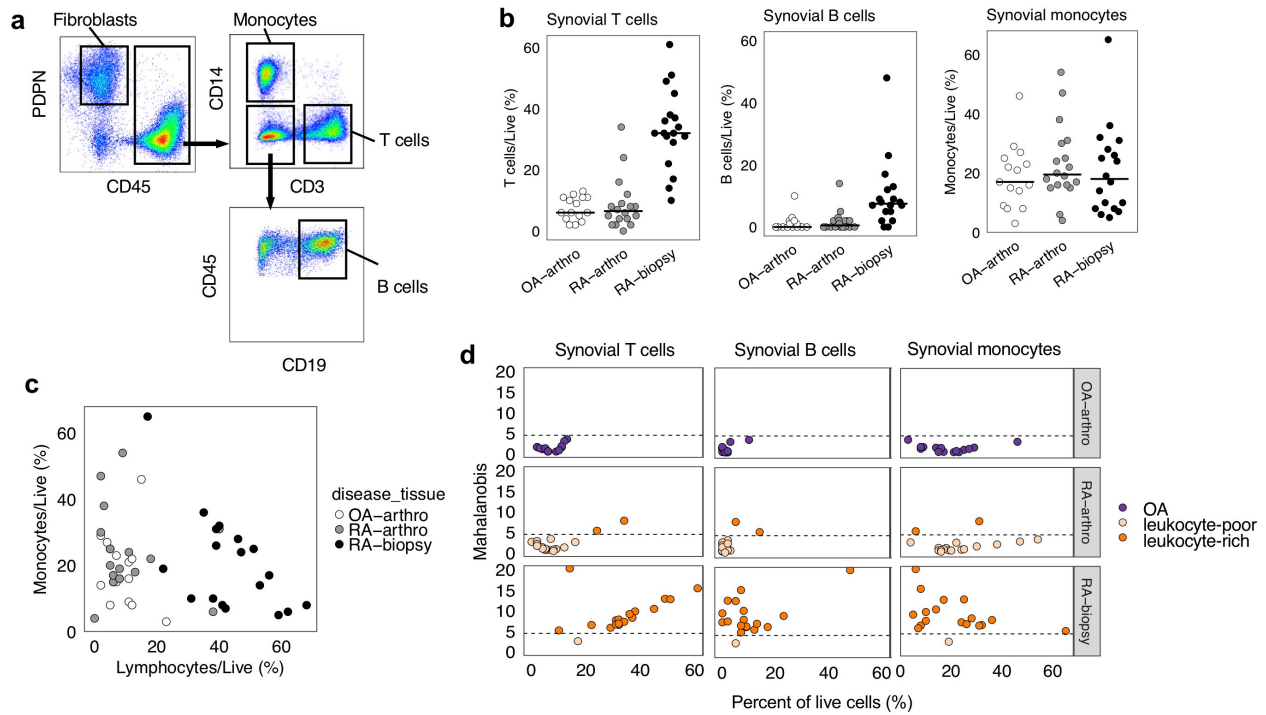


Figure 5.2 Flow cytometry gating scheme and a data-driven approach to separate samples based on flow cytometry data.

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.

Summary of computational data integration strategy to define cell populations

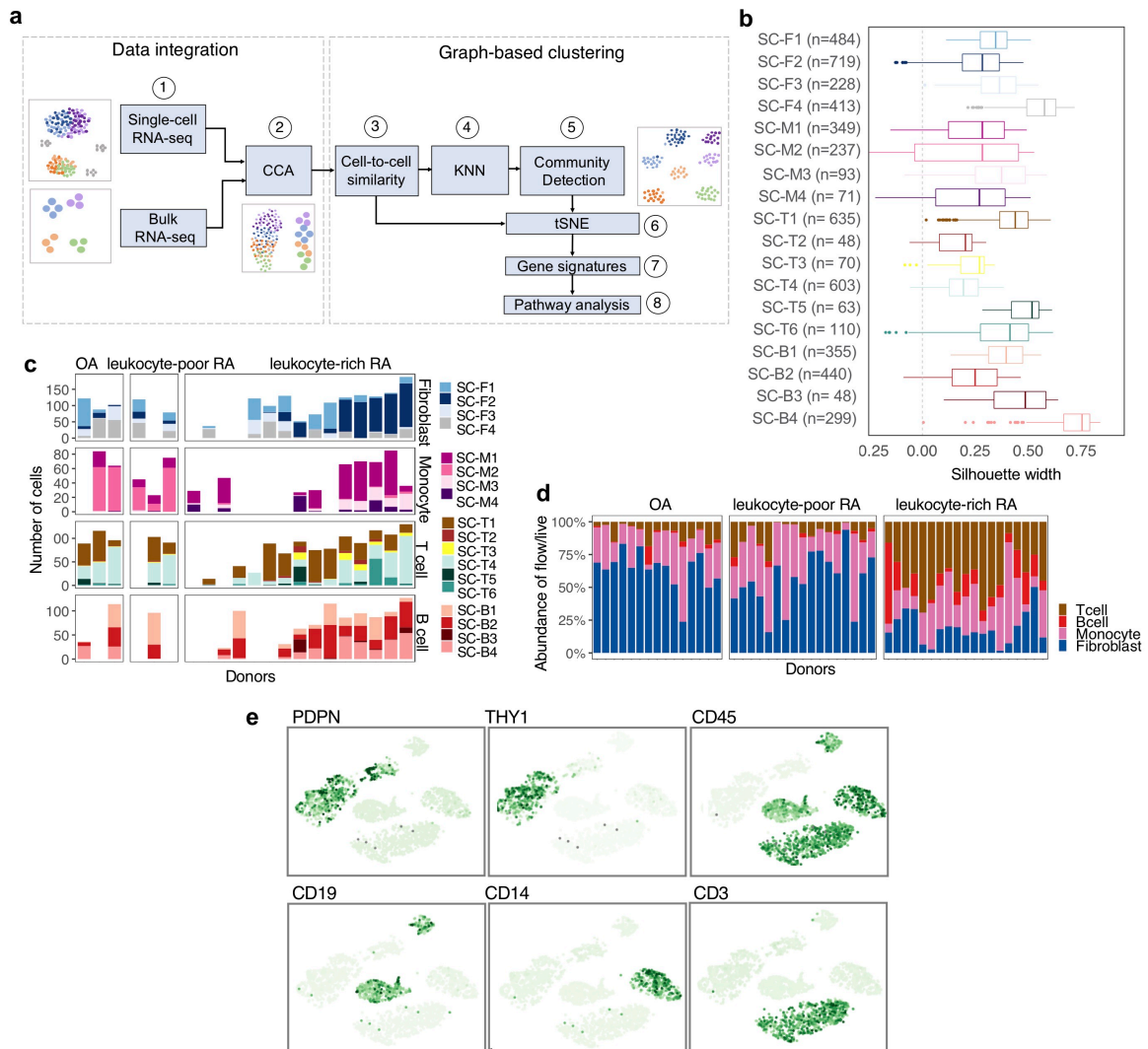
To confidently define RA-associated cell populations, we integrated multiple data modalities (**Figure 5.1b, c**). We use bulk RNA-seq data as the reference point because it was available for all of the donors and most of the cell types, it had the highest dimensionality and least sensitive to technical artifacts (**Figure 5.1b**).

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 5.1c, 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 (see Materials and Methods, **Figure 5.3b**).

Figure 5.3 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 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.

Figure 5.3 (Continued)



We identified cell clusters in mass cytometry data with density-based clustering¹²⁴. 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 (see Materials and 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 5.4a,b, Figure 5.3c,d**). 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 5.2b,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 OA¹²⁵. 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 (see Materials and Methods, **Figure 5.2d**). 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 5.4c**). Synovial monocyte abundances were similar between RA and OA (**Figure 5.4c**).

To test if our classification indicates inflammation, we assessed tissue histology and assigned each sample a Krenn inflammation score¹²⁶. Samples we classified as leukocyte-rich RA had a significantly higher Krenn inflammation score than leukocyte-poor RA or OA (**Figure 5.4d**). In contrast, synovial lining membrane hyperplasia was not significantly different between leukocyte-rich RA, leukocyte-poor RA, and OA samples (**Figure 5.4d**). We observed significant correlation between synovial leukocyte infiltration measured by flow cytometry and the histological Krenn inflammation score (**Figure 5.4e**). 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 5.4f, Figure 5.5a**).

Figure 5.4 Distinct cellular composition in synovial tissue from OA, leukocyte-poor RA, and leukocyte-rich RA patients.

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 $P = 4 \times 10^{-9}$, 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 $P = 1 \times 10^{-3}$, t-value = 3.50, df = 20.56) compared to leukocyte-poor RA and OA. Center value is mean. Statistical significance levels: **** $P < 1 \times 10^{-4}$ and *** $P < 1 \times 10^{-3}$. **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 $P = 1 \times 10^{-3}$, 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 $P = 3 \times 10^{-9}$, 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.

Figure 5.4 (Continued)

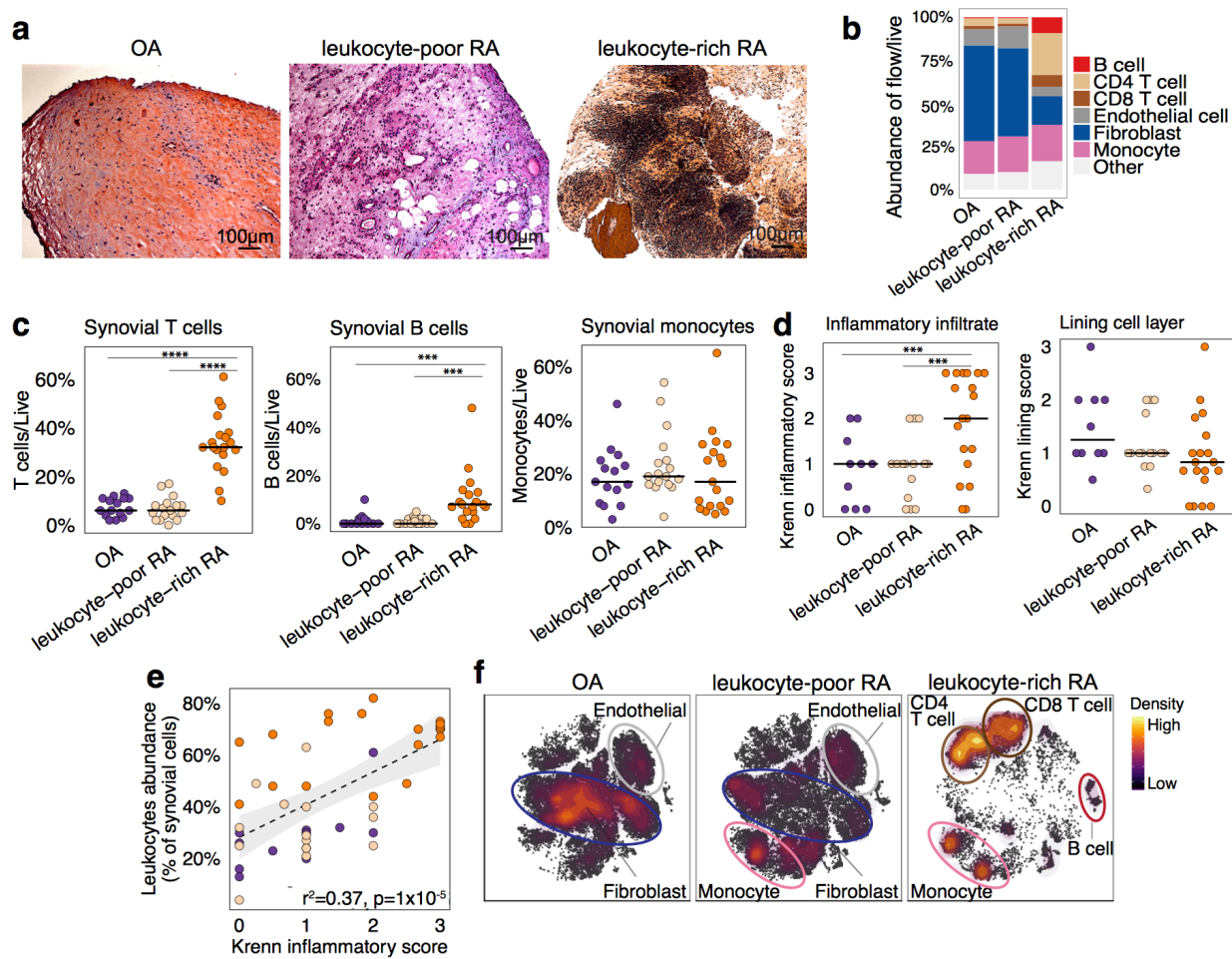
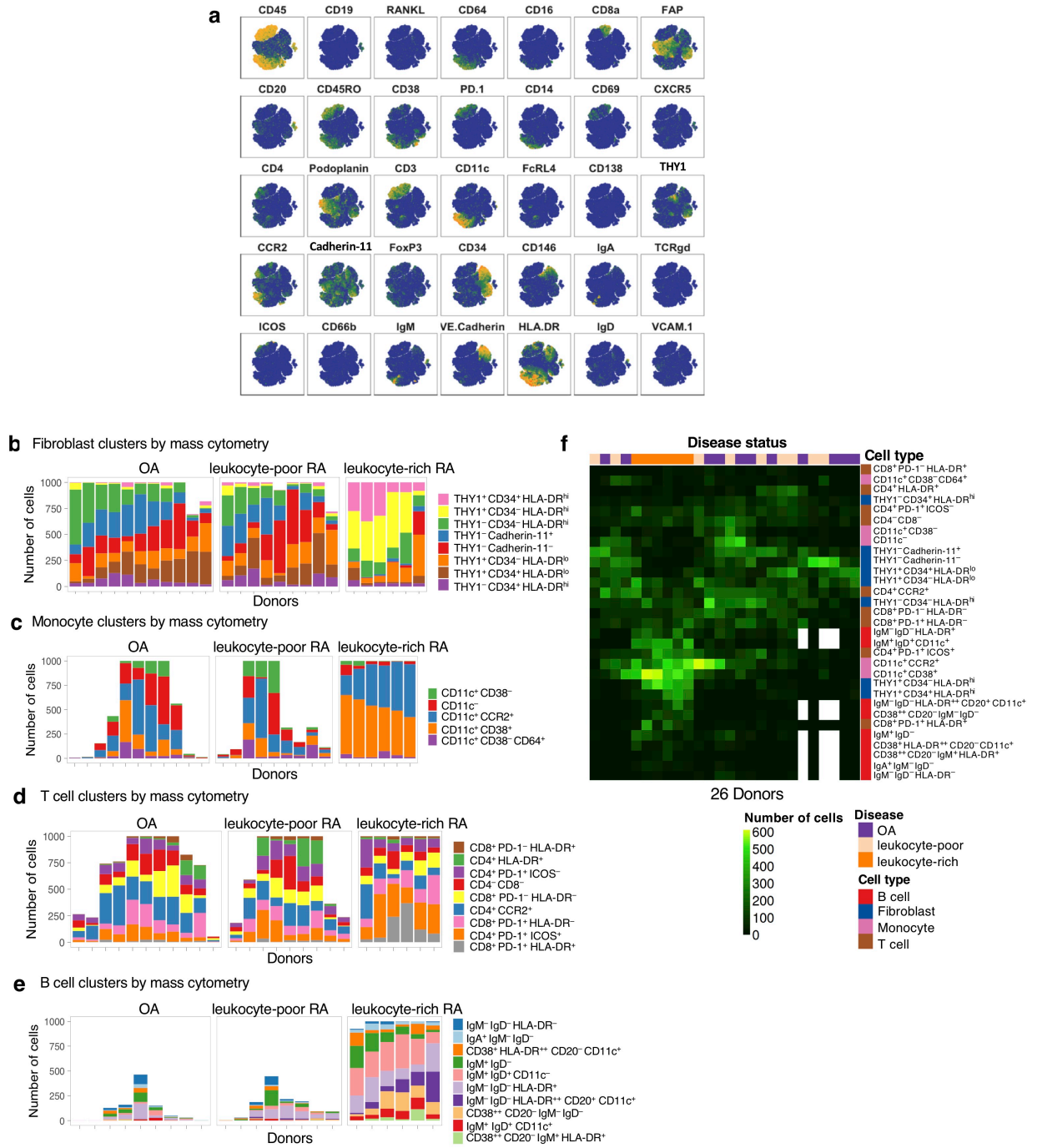


Figure 5.5 Mass cytometry data analysis.

a. Protein markers of synovial cells (3,000 subsampled) 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⁺).

Figure 5.5 (Continued)



Single-cell RNA-seq analysis reveals distinct cell subpopulations

Next, we analyzed 5,265 scRNA-seq profiles passing quality control (see Materials and 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 5.6a-b, Figure 5.3c, Figure 5.7a**). In contrast, conventional PCA-based clustering led to clusters that were confounded by batch effects (**Figure 5.7b**). 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 (see Materials and Methods, **Supplementary Table 4**). For selected genes, we show expression values in each cell positioned in a t-distributed Stochastic Neighbor Embedding (tSNE¹²⁷) (**Figure 5.6c-f**). Among fibroblasts, we identified four putative subpopulations (**Figure 5.6c**): *CD34*⁺ sublining fibroblasts (SC-F1), *HLA-DRA*^{hi} sublining fibroblasts (SC-F2), *DDIT3*⁺ sublining fibroblasts (SC-F3), and *CD55*⁺ lining fibroblasts (SC-F4). In monocytes (**Figure 5.6d**), we identified *IL1B*⁺ pro-inflammatory monocytes (SC-M1), *NUPR1*⁺ monocytes (SC-M2), *CIQA*⁺ monocytes (SC-M3), and interferon (IFN) activated monocytes (SC-M4). In T cells (**Figure 5.6e**), we identified three CD4⁺ clusters: *CCR7*⁺ T cells (SC-T1), *FOXP3*⁺ regulatory T cells (T_{reg} cells) (SC-T2), and *PDCDI*⁺ 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 5.6f**), 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 *XBPI*, a transcription factor for plasma cell differentiation¹²⁸.

We assessed protein fluorescence measurements of typical cell type markers, which were consistent with our identified scRNA-seq clusters (**Figure 5.3e**). Cell density quantified from 10 histology samples was correlated with the lymphocyte flow cytometric cell yields, suggesting that samples with the greatest number of single cell measurements are those with the best yields and the most inflammation (**Figure 5.8**).

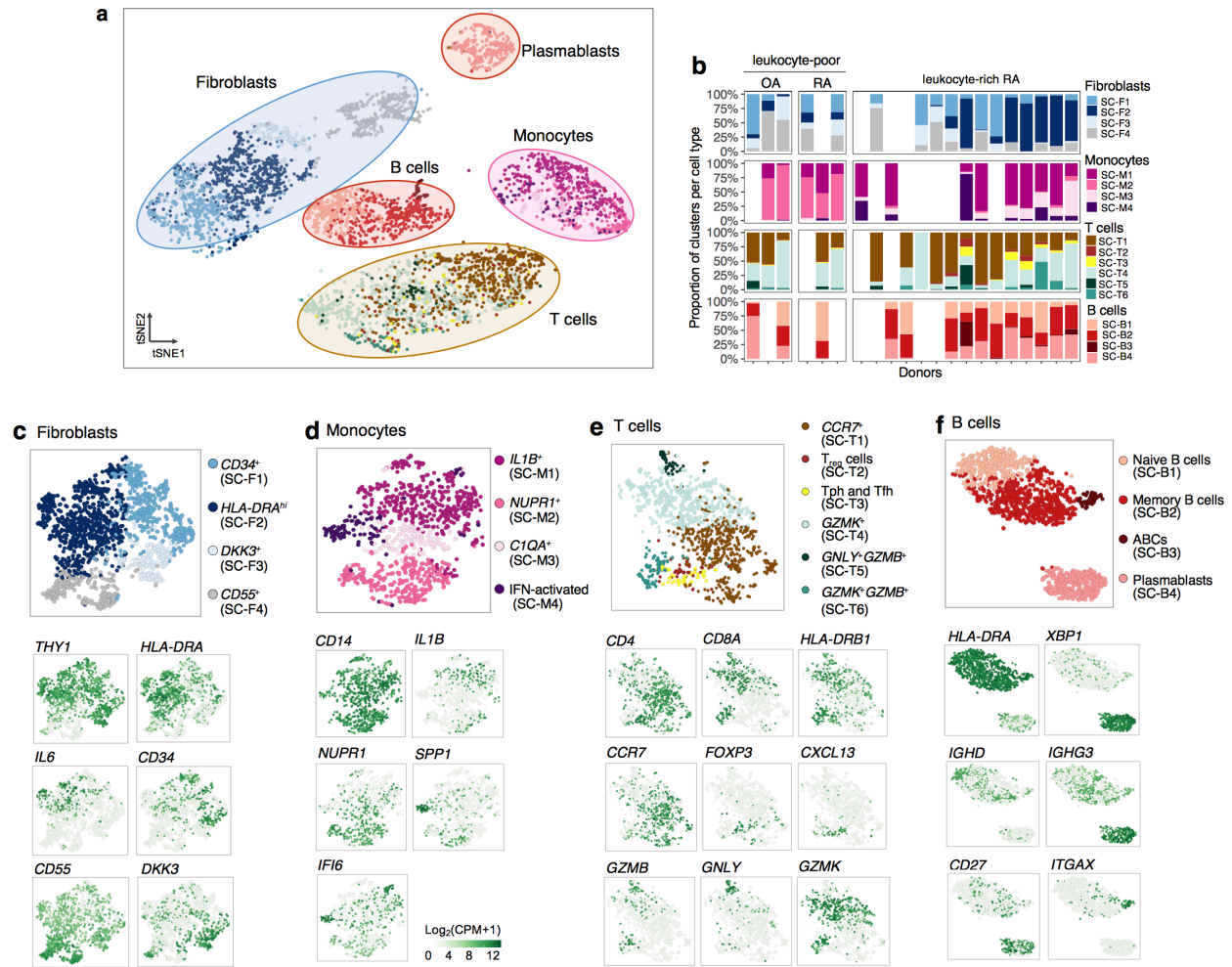


Figure 5.6 High-dimensional transcriptomic scRNA-seq clustering reveals distinct cell type subpopulations.

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**).

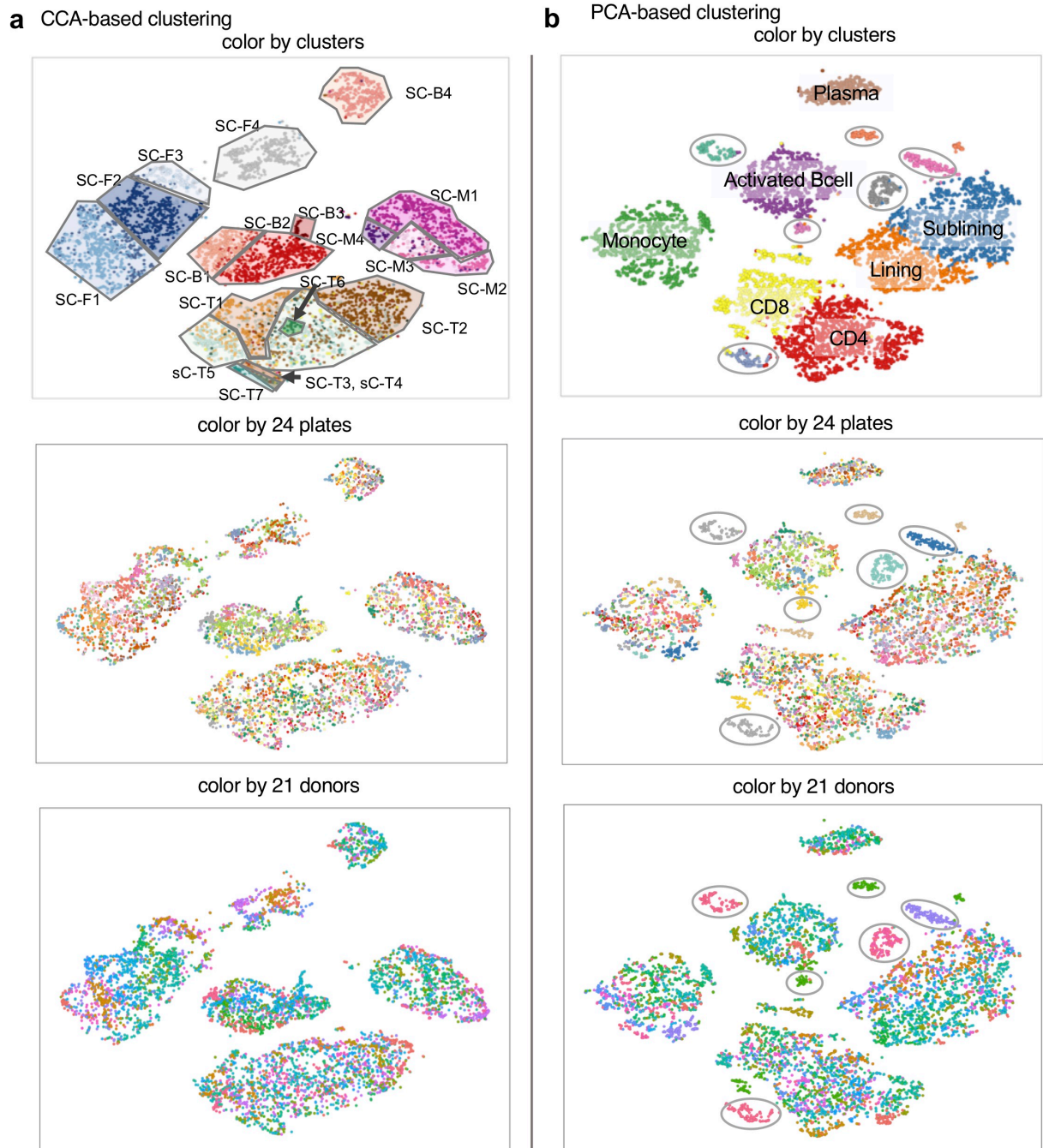


Figure 5.7 Comparison of CCA-based clustering and PCA-based clustering.

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.

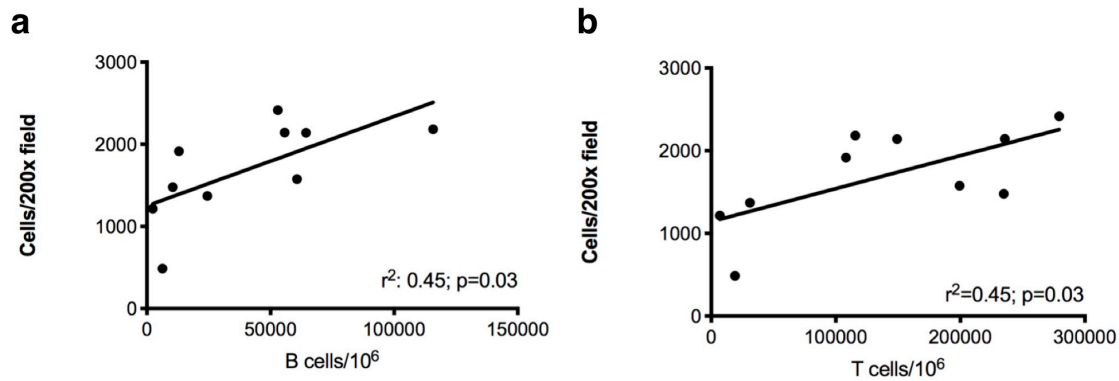


Figure 5.8 Cell density quantification on 10 histological synovial samples.

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.

Distinct synovial fibroblasts defined by cytokine activation and MHC II expression

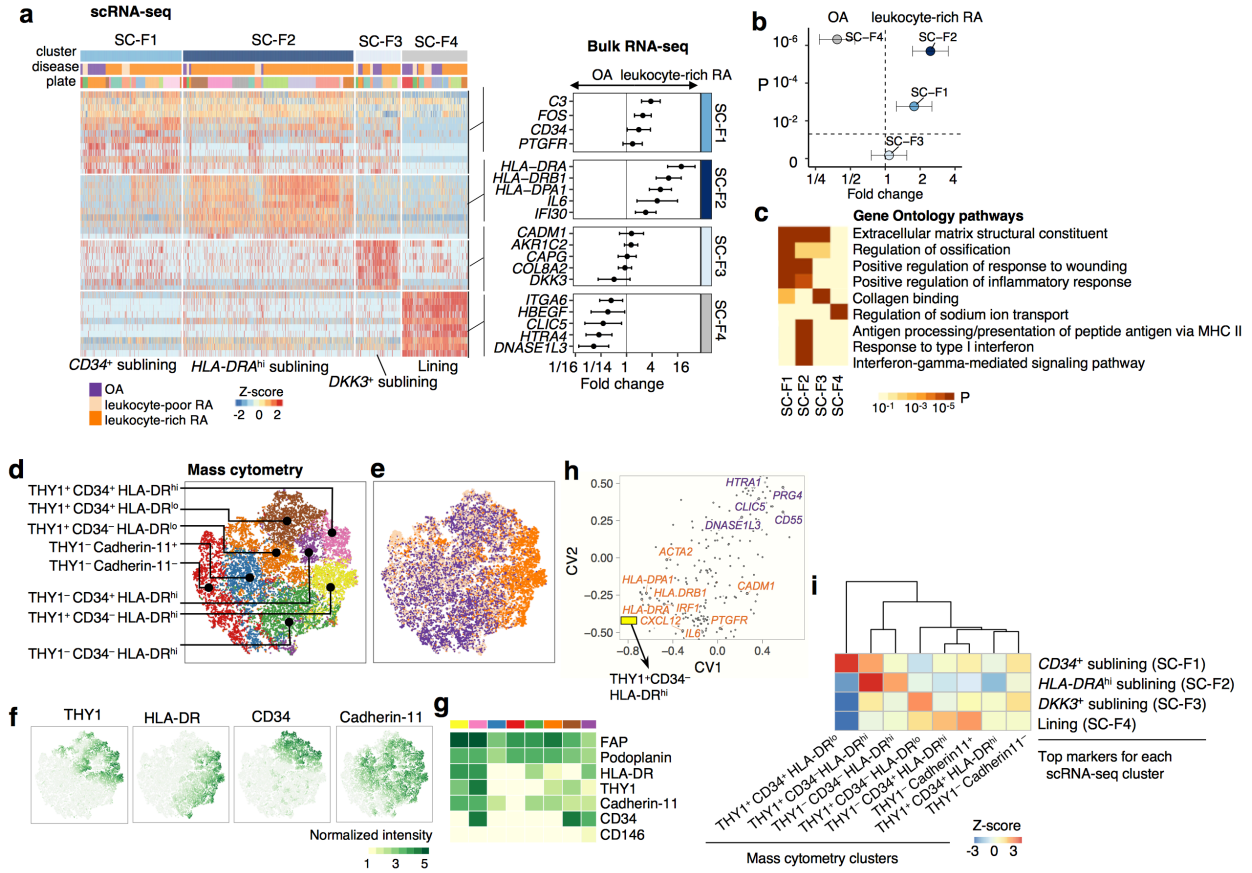
To identify the fibroblast subpopulations overabundant in leukocyte-rich RA synovia, we selected marker genes for each cluster and assessed their expression levels in bulk RNA-seq from sorted fibroblasts ($CD45^-PDPN^+$) from RA and OA patients. For example, genes associated with $HLA-DRA^{hi}$ (SC-F2) fibroblasts were more highly expressed in bulk RNA-seq samples from leukocyte-rich RA than OA (t -test $p < 1 \times 10^{-3}$ for *HLA-DRA*, *IFI30*, and *IL6*) (**Figure 5.9a**). 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 $p < 1 \times 10^{-3}$ for *HBEGF*, *CLIC5*, *HTRA4*, and *DNASE1L3*) (**Figure 5.9a**). *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 5.9a**). 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 $p = 2 \times 10^{-6}$ and $p = 2 \times 10^{-3}$, respectively), while the gene averages for *CD55*⁺ lining fibroblasts (SC-F4) were higher in OA than leukocyte-rich RA (t -test $p = 5 \times 10^{-7}$) (**Figure 5.9b**).

Figure 5.9 Distinct synovial fibroblast subsets defined by cytokine activation and MHC II expression.

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 $P=2 \times 10^{-6}$, t-value=6.2, df = 23.91 and $P=2 \times 10^{-3}$, 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 $P=5 \times 10^{-7}$, 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 HLA^{hi} sublining fibroblasts by scRNA-seq are strongly correlated with $THY1^+CD34^+HLA-DR^{hi}$ fibroblasts by mass cytometry.

Figure 5.9 (Continued)



Consistent with the role of synovial fibroblasts in matrix remodeling, the sublining fibroblast subsets (SC-F1-3) expressed genes encoding extracellular matrix constituents (**Figure 5.9c**). *HLA-DRA*^{hi} sublining fibroblasts (SC-F2) expressed genes related to MHC class II presentation and the interferon gamma-mediated signaling pathway (*IFI30*) (**Figure 5.9a,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 5.9a**).

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 5.9d-g**, **Figure 5.5b**). 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 5.9h**). 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 5.9i**, see Materials and 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 5.3**). 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 $p=1.9 \times 10^{-5}$) (**Table 5.3**).

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 5.10a**). 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-DR*^{hi} (SC-F2) and the THY1⁻HLA-DR⁻ population was similar to *THY1*⁻ (SC-F4) (**Figure 5.11a-d**). Genes upregulated in the sorted THY1⁺HLA-DR⁺ fibroblasts included the interleukin *IL6* and the chemokine *CXCL12*, consistent with the scRNA-seq data.

Table 5.3 Connection between cell populations determined by mass cytometry and scRNA-seq clusters and disease associations.

scRNA-seq cluster	Mass cytometry cluster	leukocyte-poor RA and OA	leukocyte-rich RA	One-sided MASC p value	leukocyte-rich OR (CI)
Lining (SC-F4)	THY1 ⁺ Cadherin-11 ⁻	21%	4%	1.00	0.04 (0-0.2)
	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)
NUPRI ⁺ (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)

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 subsampled 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.

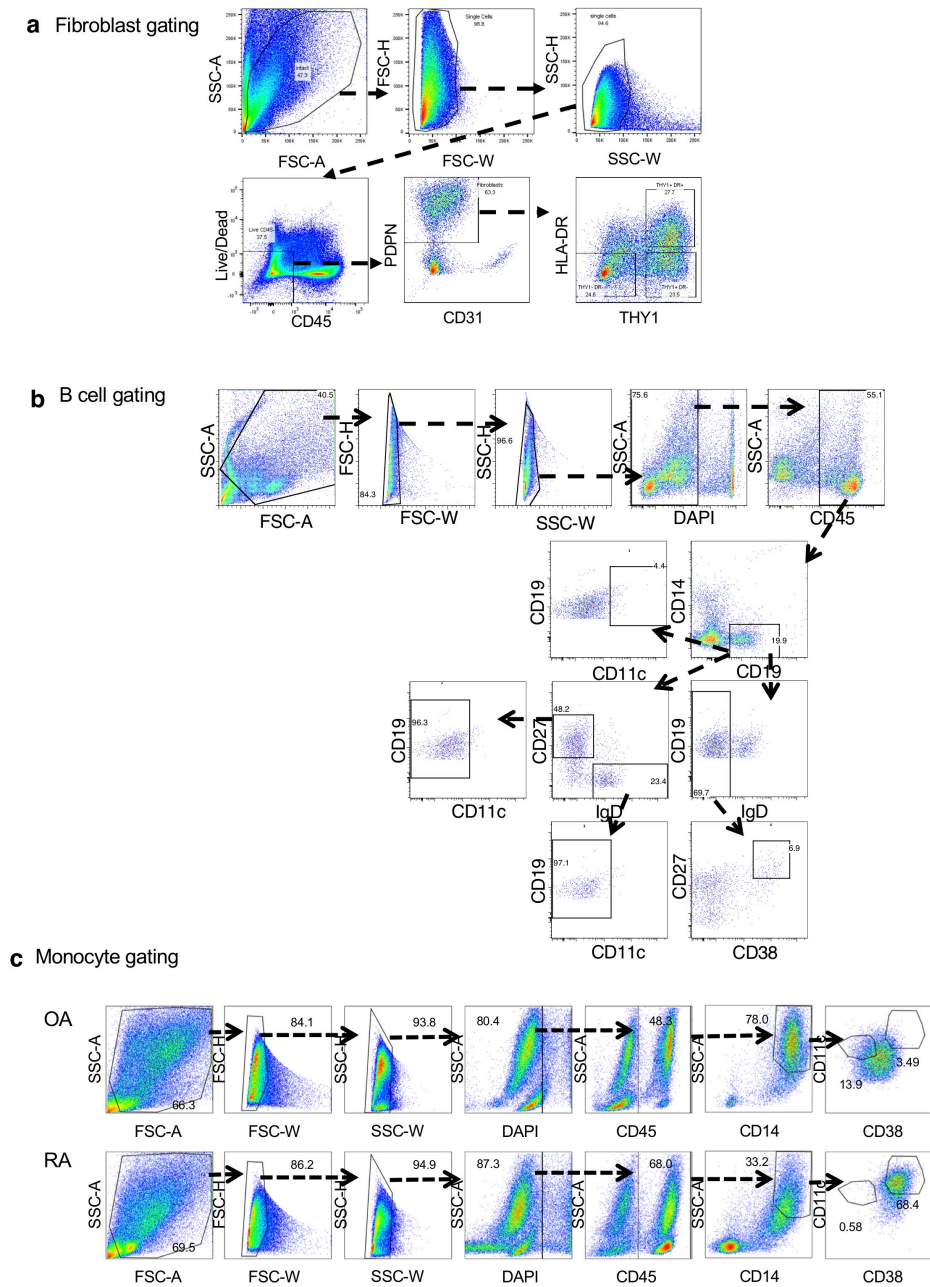


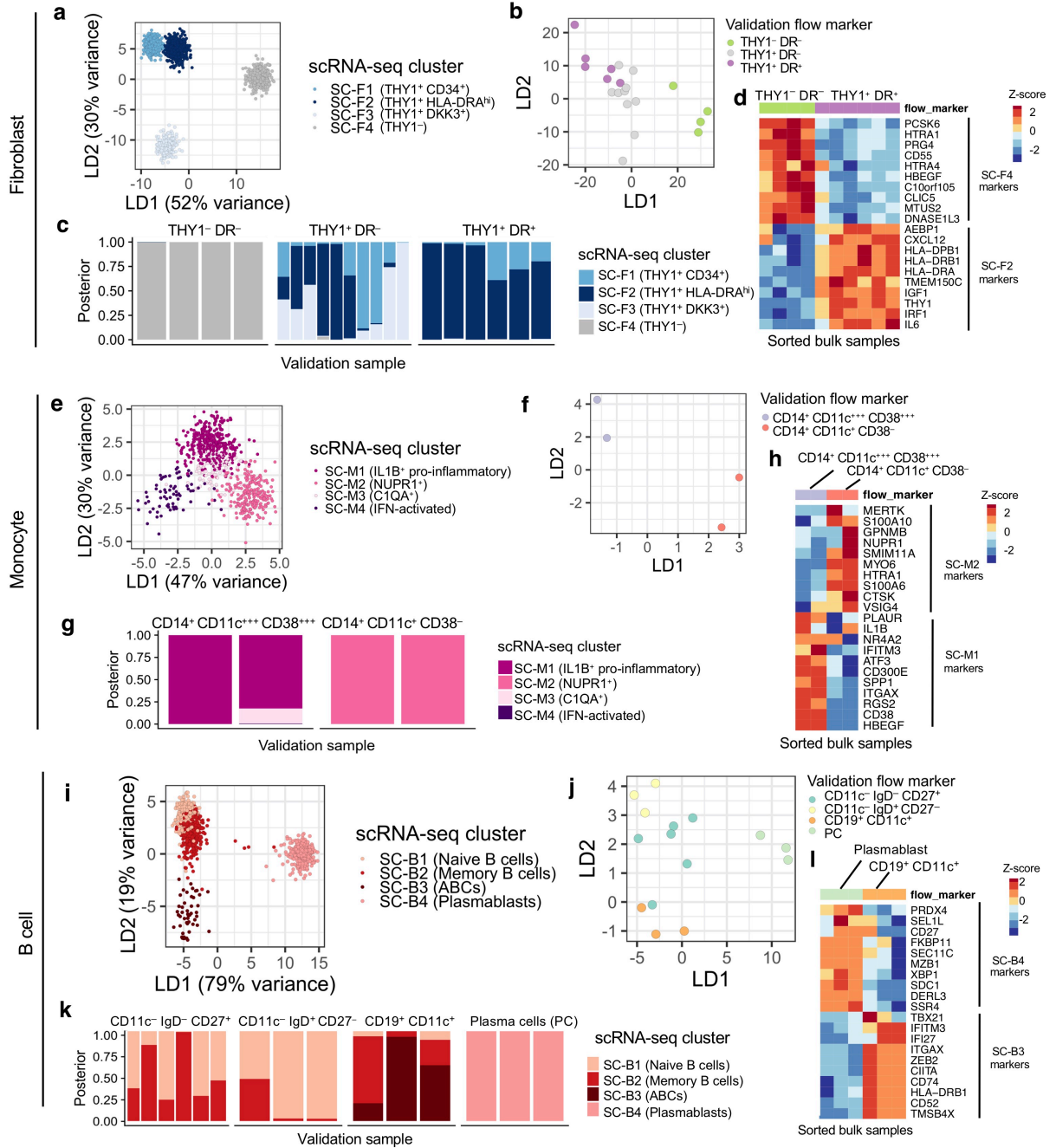
Figure 5.10 Flow cytometry gating schema for experimental validations.

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 Materials and Methods for more details.

Figure 5.11 Validation of clusters identified in scRNA-seq by bulk RNA-seq analysis of flow-sorted subpopulations of synovial fibroblasts, monocytes, and B cells.

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 (**Figure 5.10**). 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, 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.

Figure 5.11 (Continued)



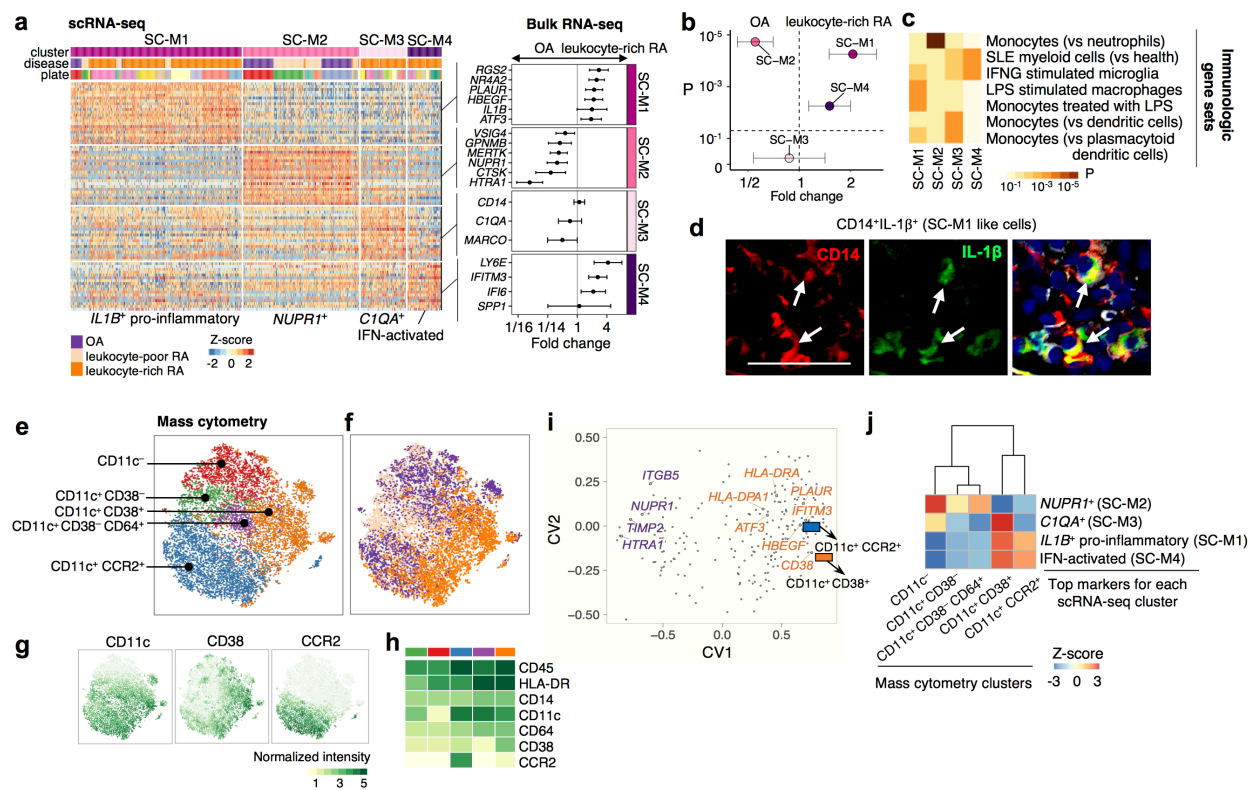
Activation states define heterogeneity among synovial monocytes

We identified four transcriptionally distinct monocyte subsets in the scRNA-seq data: *IL1B*⁺ pro-inflammatory monocytes (SC-M1), *NUPRI*⁺ monocytes (SC-M2), *CIQA*⁺ monocytes (SC-M3) and IFN-activated *SPPI*⁺ monocytes (SC-M4) (**Figure 5.12a**). 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 $p < 1 \times 10^{-4}$). In contrast, marker genes associated with *NUPRI*⁺ monocytes (SC-M2) were downregulated in leukocyte-rich RA relative to OA (**Figure 5.12a**). 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 $p = 6 \times 10^{-5}$) and IFN-activated monocytes (*t*-test $p = 6 \times 10^{-3}$), but lower abundance of *NUPRI*⁺ monocytes (*t*-test $p = 2 \times 10^{-5}$) (**Figure 5.12b**). These data suggest that cytokine activation drives expansion of unique monocyte populations in active RA synovia.

Figure 5.12 Unique activation states define synovial monocytes heterogeneity.

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 $P=6 \times 10^{-5}$, t-value=4.56, df=26.33) and IFN-activated monocytes (two-sided Student's t-test $P=6 \times 10^{-3}$, 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 $P=2 \times 10^{-5}$, 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-1 β in inflamed synovium by multicolor immunofluorescent staining with antibodies CD14 (red), IL-1 β (green), and counterstained with DAPI (blue) identified CD14⁺IL-1 β ⁺ 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 μ m. **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 CD11c⁺CD38⁺ by mass cytometry are significantly associated with *IL1B*⁺ pro-inflammatory (SC-M1) monocytes.

Figure 5.12 (Continued)



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 5.12b**). 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 5.13a**), including *IFITM3* and *IFI6* (**Figure 5.12a**). 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-1 β positive cells in 6 tissue samples, consistent with an enrichment of the *IL1B*⁺ pro-inflammatory monocytes (SC-M1) phenotype in RA synovium (**Figure 5.12d, Figure 5.15a,b**).

In the mass cytometry data, we identified five CD14⁺ monocyte clusters (**Figure 5.12e-h, Figure 5.5c**). 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 5.12i**). 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 5.12j, Table 5.3**). 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 5.3**). Conversely, *NUPRI*⁺ monocytes (SC-M2) correspond to CD11c⁻ monocytes in mass cytometry and are inversely correlated with inflammatory monocyte populations (z-score=2.7) (**Figure 5.12j, Table 5.3**).

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 5.10c**). 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 *NUPRI*⁺ (SC-M2) cells, respectively (**Figure 5.11e-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 *NUPRI*⁺ (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 5.14a**). *CCR7*⁺ T cells (SC-T1) expressed genes in the MSigDB immunologic gene set for central memory T cells (**Figure 5.14a, c**). The two other CD4⁺ populations, *FOXP3*⁺ T_{reg} cells and *PDCDI*⁺ Tph and Tfh cells, were marked by high expression of *FOXP3* (SC-T2) and *CXCL13* (SC-T3) by examining differentially expressed genes between these two clusters¹²⁰ (**Figure 5.13c**). *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 $p=1.2 \times 10^{-4}$) (**Figure 5.14a**). 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 5.14b**), 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 5.14a**). 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 5.14a,c**).

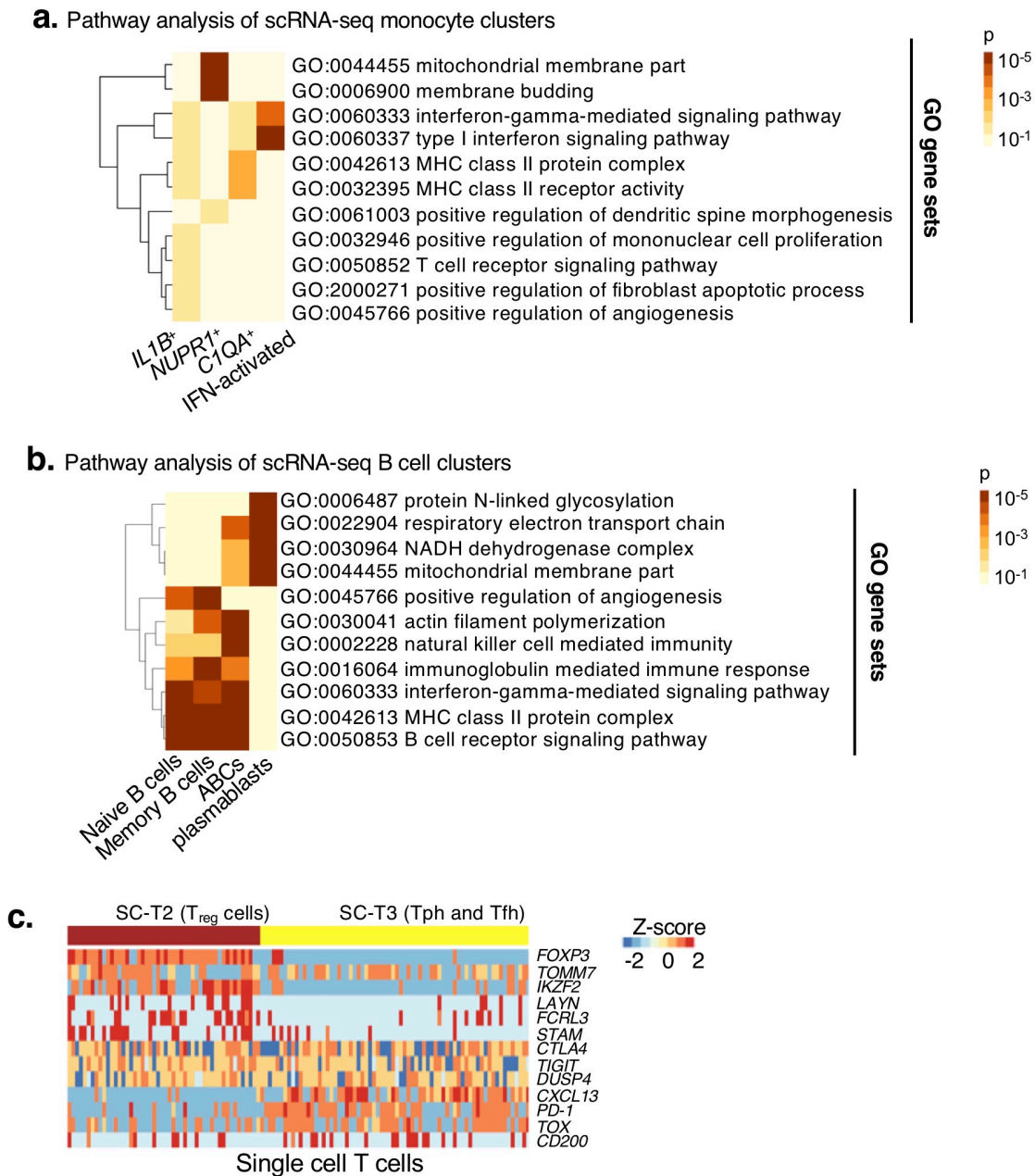


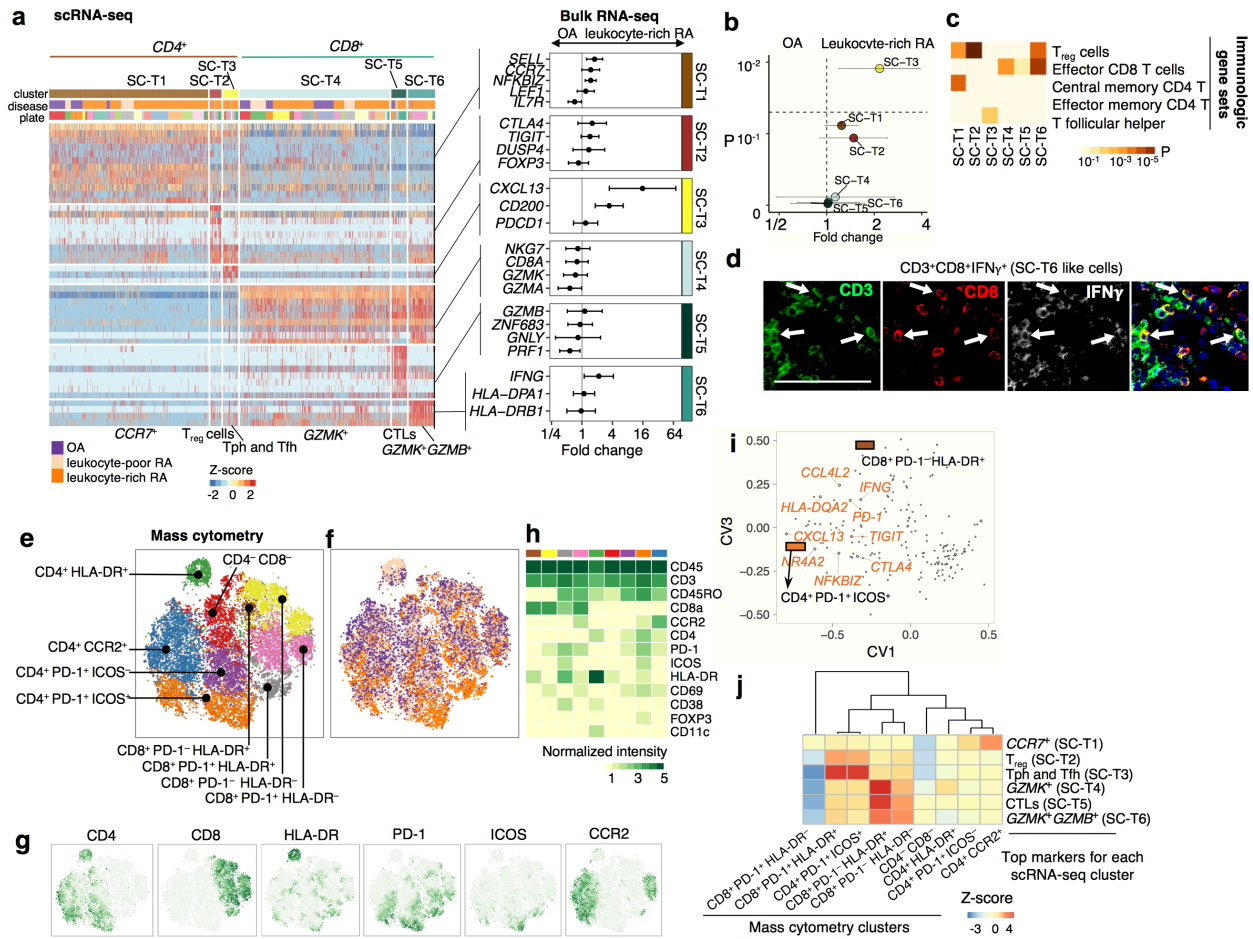
Figure 5.13 Pathway enrichment analysis for scRNA-seq clusters.

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.

Figure 5.14 Synovial T cells display heterogeneous CD4 and CD8 T cell subpopulations in RA synovium.

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 10⁵ 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 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-1⁺ICOS⁺ and CD8⁺PD-1⁺HLA-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-1⁺ICOS⁺ and CD8⁺PD-1⁺HLA-DR⁺ by mass cytometry are highly associated with Tph and Tfh (SC-T3) by scRNA-seq; CD8⁺PD-1⁺HLA-DR⁺ T cells by mass cytometry are highly associated with CD8⁺ T cells (SC-T4, SC-T5, and SC-T6).

Figure 5.14 (Continued)



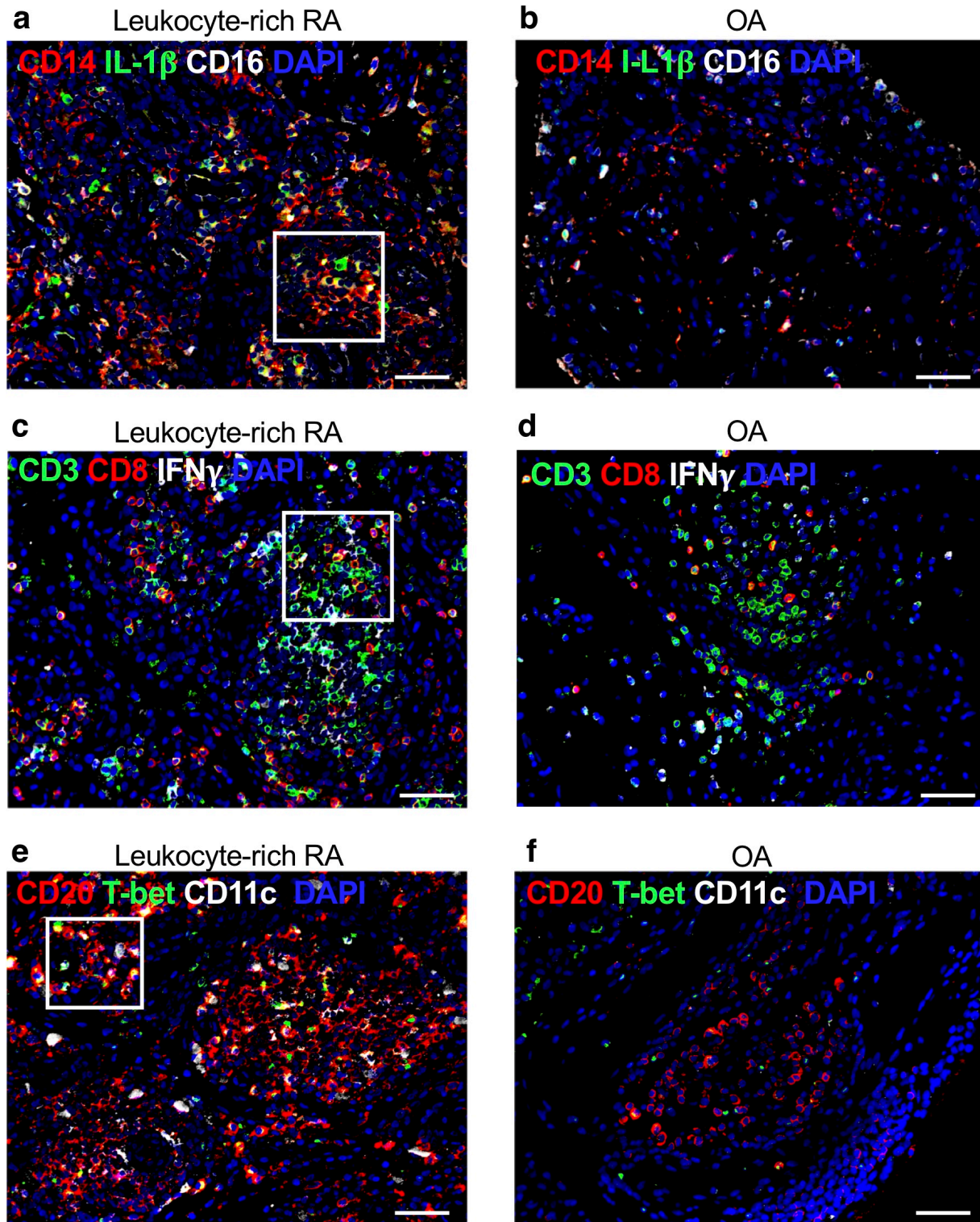
To confirm these findings, we applied intracellular staining to tissues from RA samples and RNA-seq to sorted CD8 T cells. Intracellular staining of GZMK and GZMB proteins in disaggregated tissue samples from patients with RA revealed that the majority of CD8 T cells in synovial tissue express GZMK (**Figure 5.16a**). Furthermore, we found that most HLA-DR⁺ CD8 T cells express both GZMB and GZMK by intracellular protein staining (**Figure 5.16b**). In a comparison of 7 synovial tissue samples, CD8 T cells had higher proportion of IFN γ ⁺ cells than CD4 T cells from the same sample (**Figure 5.16c,d**). We also applied immunofluorescence to 6 synovial tissue samples and found that IFN γ ⁺CD3⁺CD8⁺ T cells were more frequent in RA than OA (**Figure 5.14d, Figure 5.15c,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 5.14e-h, Figure 5.5d**). 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-1⁺ICOS⁺ mass cytometry cluster. Greater abundance of CD8⁺ PD-1⁺HLA-DR⁺ cells was associated with greater expression of *IFNG* (**Figure 5.14i**). We found correspondence between Tph and Tfh cells (SC-T3) and CD4⁺PD-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-1⁺HLA-DR⁺ T cells by mass cytometry (**Figure 5.14j, Table 5.3**). In addition, CD4⁺PD-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 5.3**).

Figure 5.15 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-1 β ⁺ co-localization in inflamed RA synovium tissue (denoted by white box). **b.** Very few CD14⁺IL-1 β ⁺ cells in the OA synovium. **c.** Numerous CD3⁺CD8⁺IFN γ ⁺ T cells in RA biopsy (denoted by white box). **d.** Low numbers of CD3⁺CD8⁺IFN γ ⁺ T cells in the synovium of OA. **e.** An increased number of CD20⁺T-bet⁺CD11c⁺ B cells in the inflamed RA synovium (denoted by white box). **f.** very scared B cells and specially T-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 μ m.

Figure 5.15 (Continued)



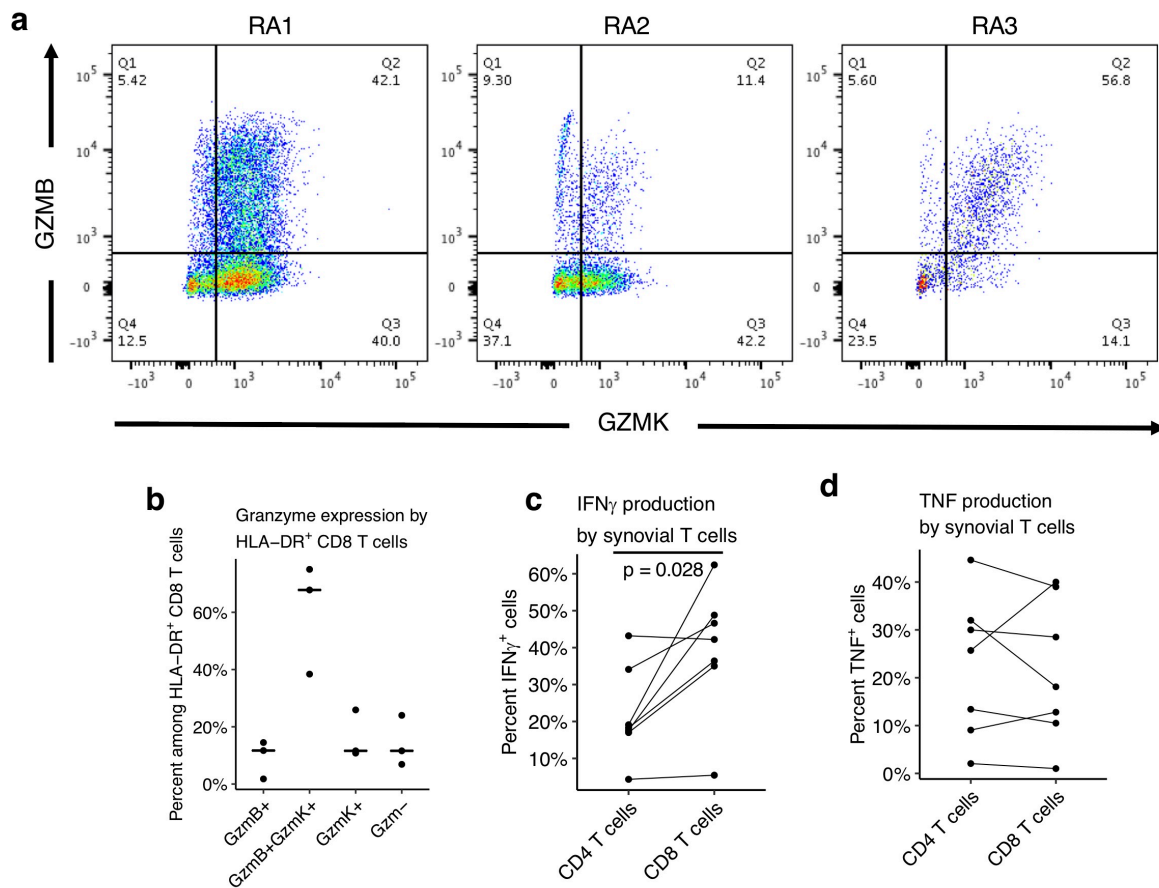


Figure 5.16 Granzyme expression and cytokine production by synovial tissue CD8 T cells.

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.** IFN γ 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.

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 5.17a**). GSEA with Gene Ontology pathways suggested that SC-B1, SC-B2, and SC-B3 clusters represent activated B cells (**Figure 5.13b**). 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 5.17b**). SC-B3 cells express high levels of *ITGAX* and *TBX21* (*T-bet*), which are markers of autoimmunity-associated B cells (**Figure 5.6f** and **Figure 5.17a**)^{129,130}, as well as markers of recently activated B cells including *ACTB*¹³¹. High expression of *AICDA* is consistent with the recently reported transcriptomic analysis of CD11c⁺ B cells from SLE peripheral blood¹³². Interferon stimulated genes (*GBP1* and *ISG15*) are also expressed in ABCs (SC-B3) and upregulated in leukocyte-rich RA (**Figure 5.17a**). 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 5.6b**). 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 5.17c**, **Figure 5.15e, f**).

We identified 10 putative B cell clusters in the mass cytometry data (CD45⁺CD3⁻CD14⁻CD19⁺) (**Figure 5.17d-g**, **Figure 5.5e**). CCA analysis showed that samples with higher gene expression of *CD38*, *MZB1*, and plasma cell differentiation factor *XBPI* had greater abundance of CD38⁺⁺CD20⁻IgM⁻IgD⁻ plasmablasts (**Figure 5.17h**). Plasmablasts (SC-B4) corresponded

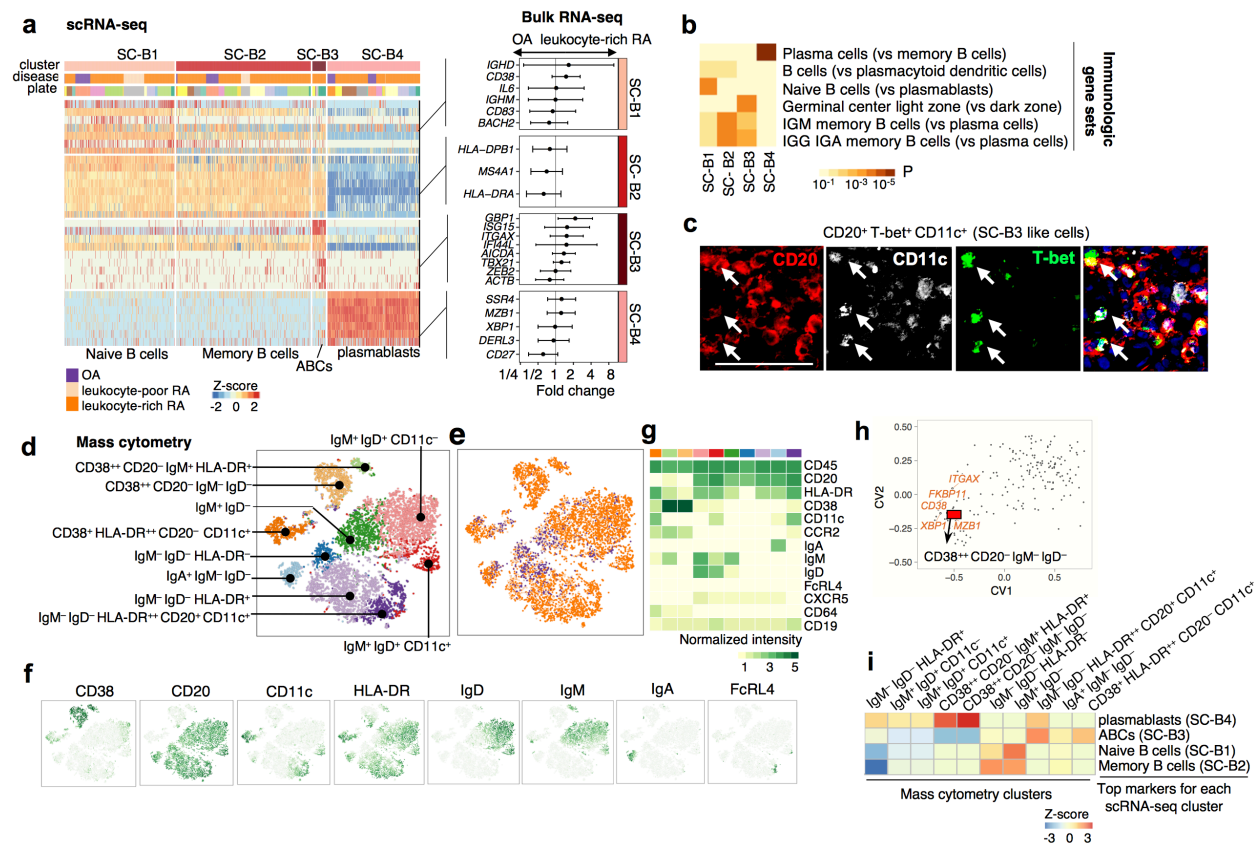
with CD38⁺⁺CD20⁻IgM⁻IgD⁻ B cells (z-score=2.7) (**Figure 5.17i, Table 5.3**). ABCs (SC-B3) corresponded with the IgM⁻IgD⁻HLA-DR⁺⁺CD20⁺CD11c⁺ mass cytometry cluster (z-score=1.6), which is significantly overabundant in leukocyte-rich RA (OR = 5.7 (95% CI: 1.8-22.3), one-sided MASC $p=2.7 \times 10^{-3}$) (**Figure 5.17i, Table 5.3**). 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 5.10b**). We show that these RNA-seq profiles are most consistent with ABC cells (**Figure 5.11i-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 5.11l**).

Figure 5.17 Synovial B cells display heterogeneous subpopulations in RA synovium.

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 $CD20^+T\text{-bet}^+CD11c^+$ (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 $CD38^{++}CD20^-IgM^+HLA-DR^+$ and $CD38^{++}CD20^-IgM^-IgD^-$ are significantly associated with plasmablast (SC-B4); $IgM^-IgD^-HLA-DR^{++}CD20^+CD11c^+$ B cells are associated with ABCs (SC-B3).

Figure 5.17 (Continued)



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 5.18a,b**) showed that cell type accounted for most of the data variance. Each cell type expressed specific marker genes, *PDGFRA* for fibroblasts, *CIQA* for monocytes, *CD3D* for T cells, and *CD19* for B cells (**Figure 5.18c**). Within each cell type, PCA showed that leukocyte-rich RA samples separated from OA and leukocyte-poor RA samples (**Figure 5.18d-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 5.18h-i**), Fc receptor signaling (monocytes), NF-kappa B signaling (fibroblasts), and interferon gamma (T cells) (**Figure 5.19a**). 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 5.19b**), 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 5.19b**)

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 5.19c**); CD8 T cells, rather than CD4 T cells, were the dominant source of *IFNG* transcription in leukocyte-rich synovia.

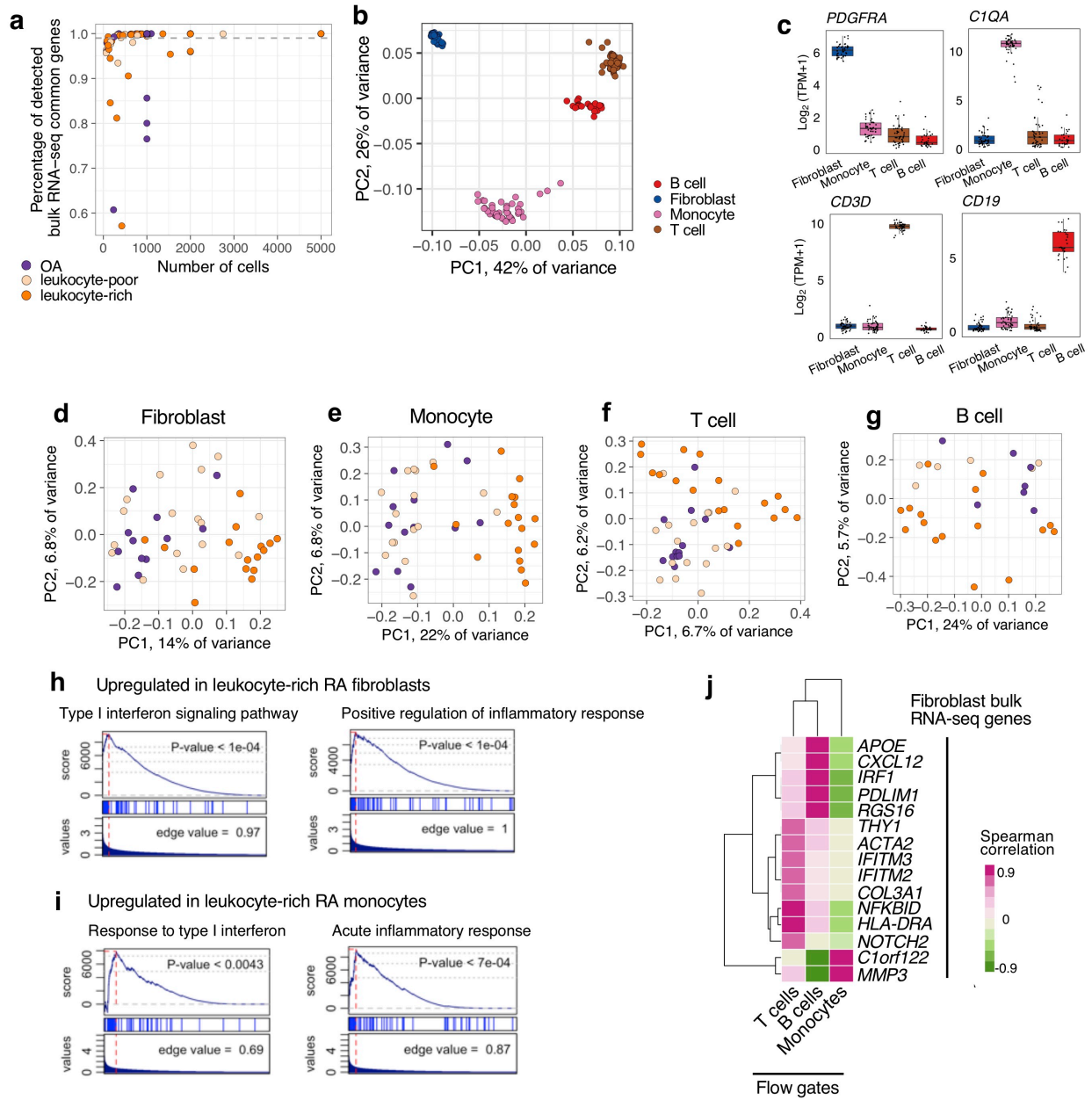
We also observed cell subset-specific responses to inflammatory pathways. Toll-like receptor signaling pathway was enriched in B cells and monocytes in leukocyte-rich RA tissues (**Figure 5.19a**). At the single cell level, we observed that *TLR10* was only expressed by activated B cells, indicating 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 5.5f**).

Figure 5.18 Analysis of sorted fibroblast, monocyte, T cell, and B cell bulk RNA-seq data.

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 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.

Figure 5.18 (Continued)



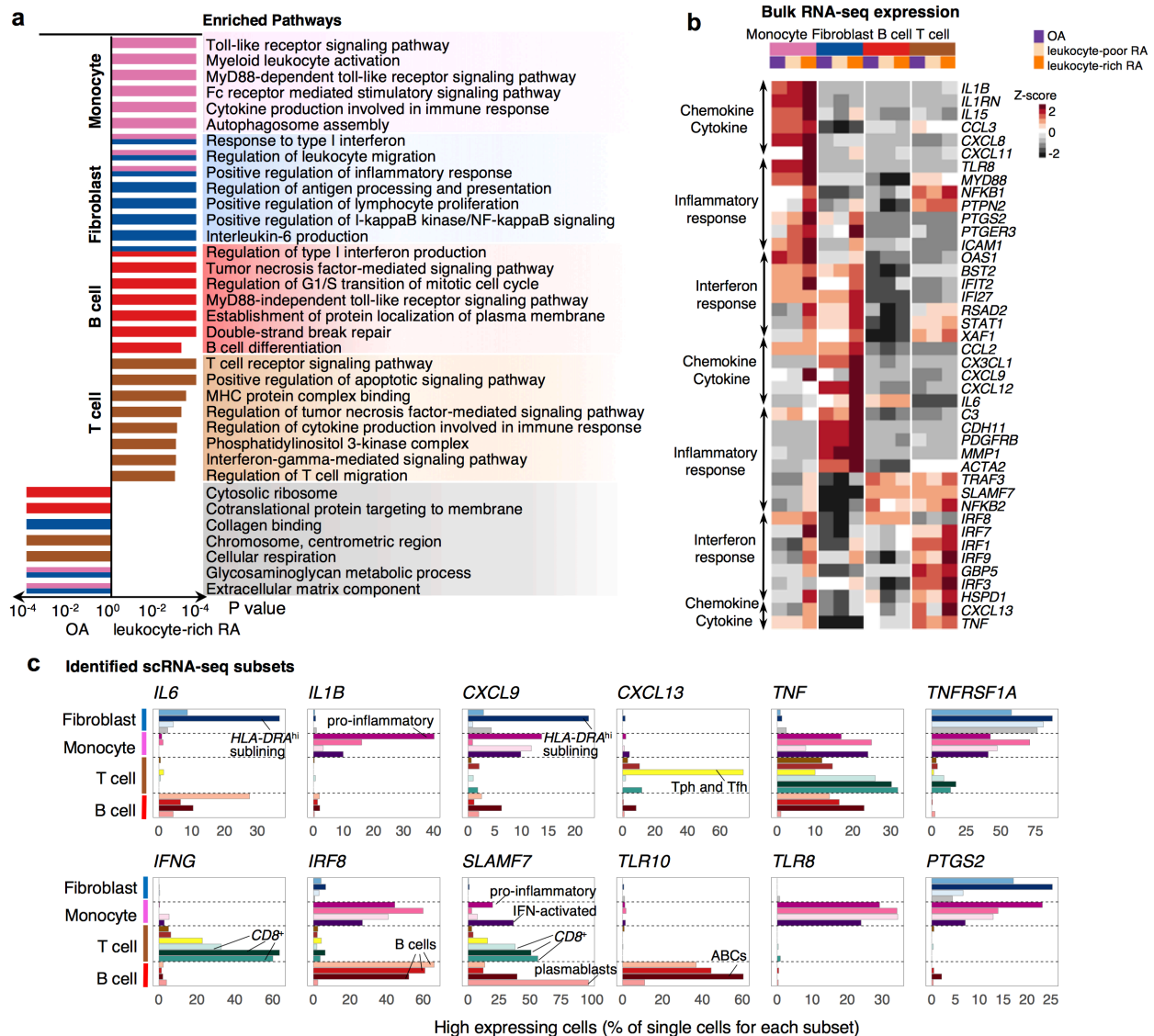


Figure 5.19 Transcriptomic profiling of synovial cells reveals upregulation of inflammatory pathways in RA synovium.

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.

Conclusion

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 scores¹³³, 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 5.2**, **Figure 5.18d-g**).

This study and a previous study¹³⁴ 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 5.9**), 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 5.18j**). We found *DNASE1L3*, a gene whose loss of function is associated with RA¹³⁵ and systemic lupus erythematosus¹³⁶ to be highly expressed in *CD55⁺* lining fibroblasts (SC-F4) (**Figure 5.9a**). We identified a novel fibroblast subset (SC-F3) with high expression of *DKK3* (**Figure 5.9**), encoding Dickkopf3, a protein upregulated in OA that prevents cartilage degradation in vitro¹³⁷.

Transcriptional heterogeneity in the synovial monocytes indicated that distinct RA-enriched subsets are driven by inflammatory cytokines and interferons (**Figure 5.12**). 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 inhibitors¹³⁸. The *NUPRI*⁺ (SC-M2) monocytes were inversely correlated with tissue inflammation, and expressed high levels of monocyte tissue remodeling factors such as *MERTK* (**Figure 5.12**)¹³⁹.

Alternatively, *NUPRI*⁺ markers such as osteoactivin (*GPNMB*) and cathepsin K (*CTSK*) may indicate a subset of osteoclast progenitors that control bone remodeling (**Figure 5.12**)^{138,140}.

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 5.14a**). 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^{132,141}. 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 5.4b**), which limited our ability to identify RA-associated B cell subsets through case-control comparisons (**Figure 5.17g**).

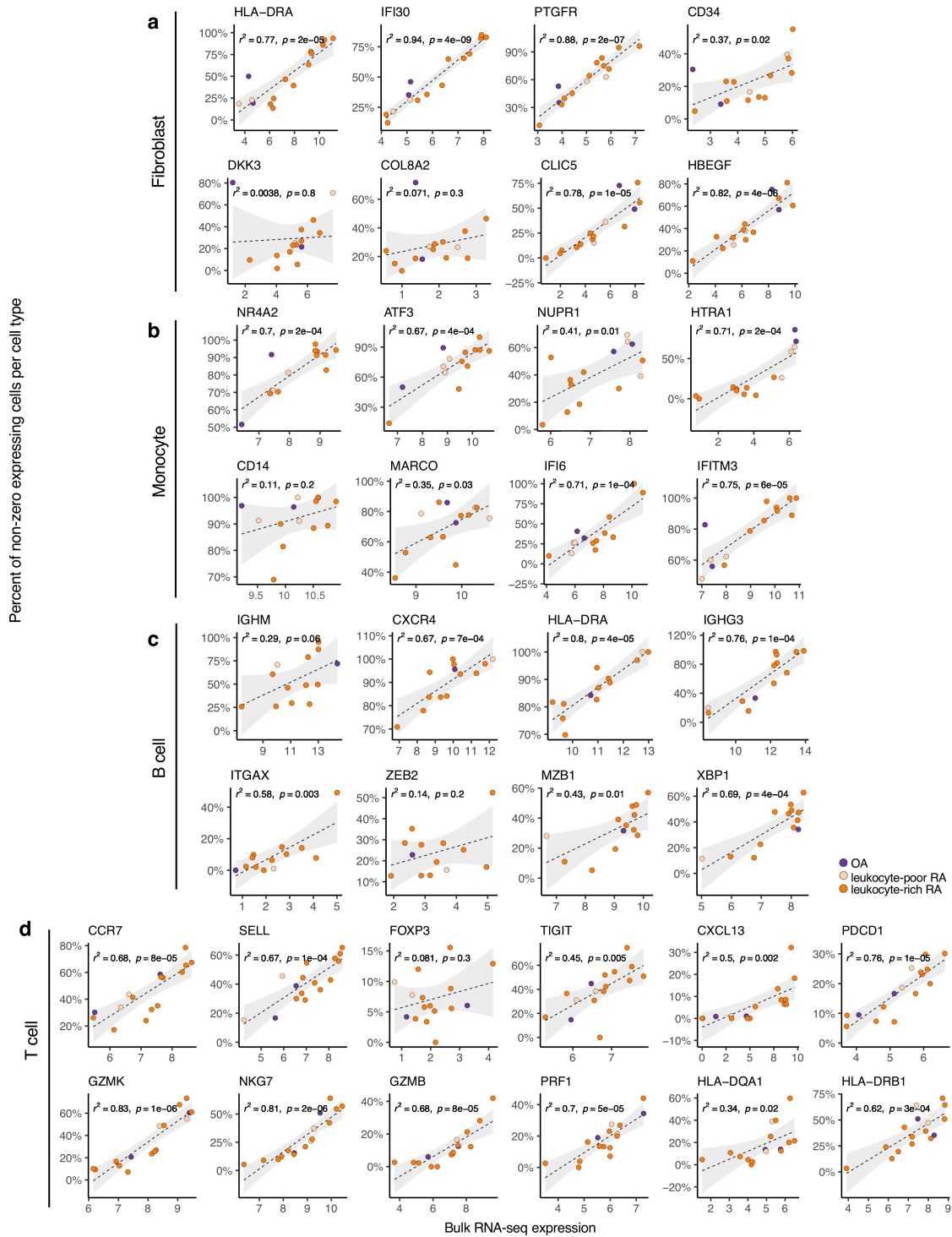
A critical unmet need in RA is identifying therapeutic targets for patients failing to respond to disease-modifying antirheumatic drugs (DMARDs)¹⁴². We observed upregulation of chemokines (*CXCL8*, *CXCL9*, and *CXCL13*), cytokines (*IFNG* and *IL15*^{143,144}), and surface receptors (*PDGFRB* and *SLAMF7*) 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 tissue¹²³.

We developed advanced strategies to integrate multiple molecular datasets by modulating technical artifact from single cell technologies¹⁴⁵, while emphasizing biological signals. CCA has been successfully employed in other contexts to integrate high-dimensional biological data^{146,147}. 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 5.20** and **Figure 5.21**).

Figure 5.20 Correlation between bulk RNA-seq expression and proportion of non-zero expressing cells on scRNA-seq cluster markers.

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.

Figure 5.20 (Continued)



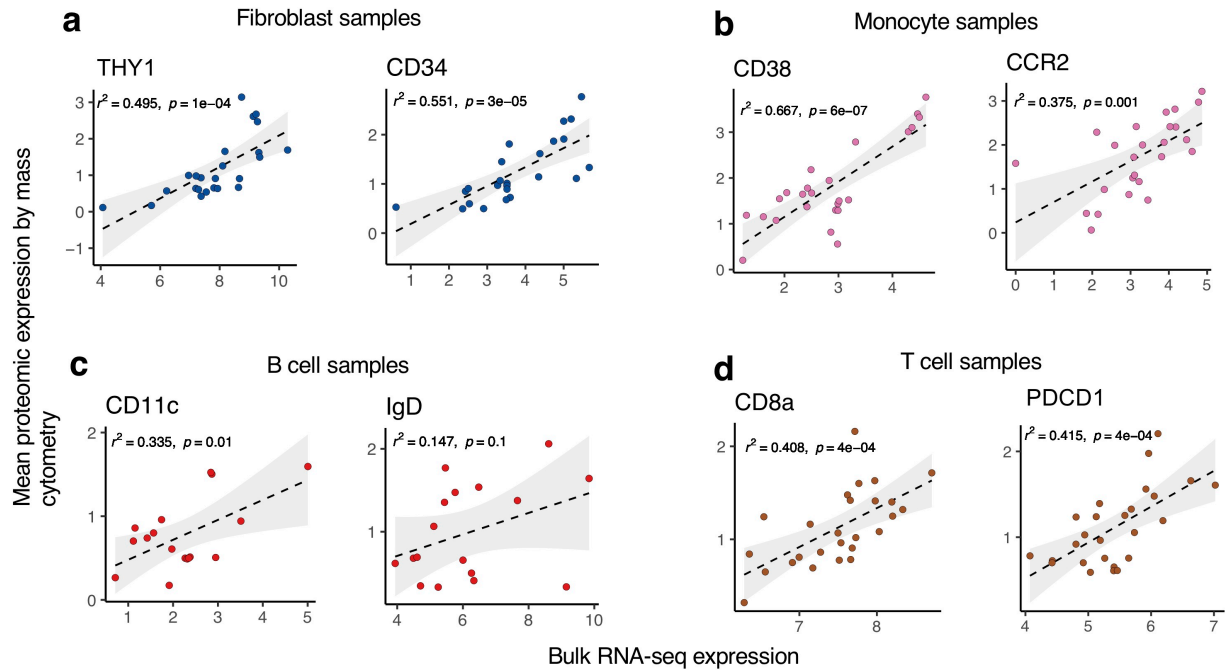


Figure 5.21 Correlation between mean proteomic expression by mass cytometry and transcriptomic expression by bulk RNA-seq on the overlapped samples.

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.

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 data¹⁴⁸. Mass cytometry captures hundreds of thousands of individual cells, but measures a limited number (~ 40)¹⁴⁹ of pre-selected markers. However, since markers are backed with decades of experimental experience they can be effective at defining cellular heterogeneity¹⁵⁰. 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 simultaneously¹⁵¹ 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 $\geq$ 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 pathologists¹²⁶.

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 described¹²³. 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, 3 mM KCl, 1 mM 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 **Figure 5.10**. 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) model 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 IFN γ ⁺ T cells include goat anti-CD3 epsilon (clone M-20, Santa Cruz Biotechnology), mouse anti-human CD8 (clone 144B, GeneTex), and rabbit anti-human IFN γ (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 (pixel²): 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:

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

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 (**Figure 5.2**).

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 protocol¹⁰¹. 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.

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 (**Figure 5.18a**). 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*, *CIQA*, *CD3D*, and *CD19*, are expressed selectively by fibroblasts, monocytes, T cells, and B cells, respectively (**Figure 5.18c**).

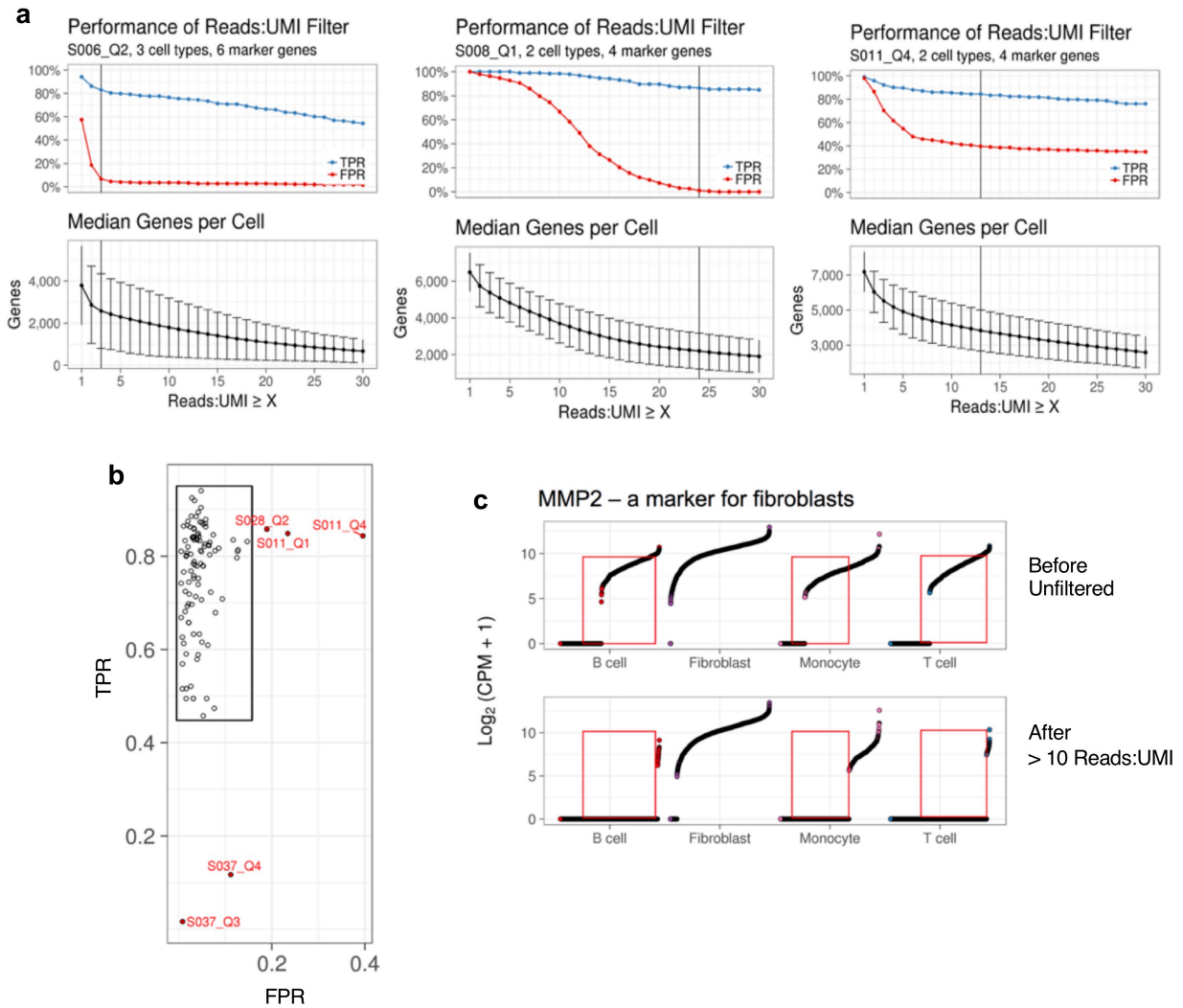
Single-cell RNA-seq gene expression quantification. Single-cell RNA-seq was performed using the CEL-Seq2 method¹⁴⁸ with the following modifications. Single cells were sorted into 384-well plates containing 0.6 μ L 1% NP-40 buffer in each well. Then, 0.6 μ L dNTPs (10mM each; NEB) and 5 nl of barcoded reverse transcription primer (1 μ g/ μ L) 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 $\text{Log}_2(\text{CPM}+1)$ (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 *CIQA* 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 (**Figure 5.22**). 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 (**Figure 5.23**).

Figure 5.22. Dynamic UMI 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 *CIQA* 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 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.

Figure 5.22 (Continued)



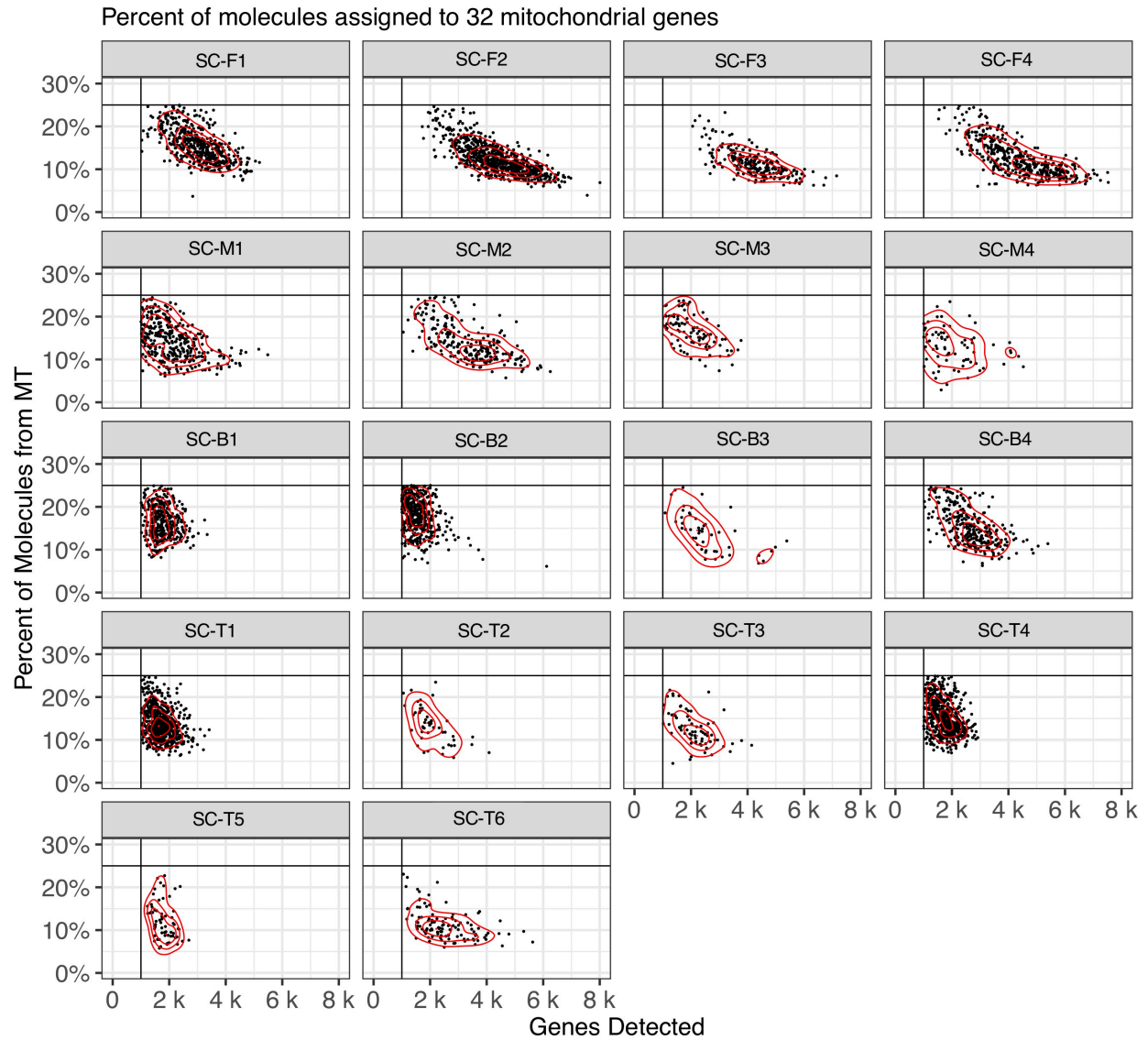


Figure 5.23 Assessment of quality of scRNA-seq data for each identified cluster.

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.

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 previously¹²³. Briefly, we analyzed samples on a Helios instrument (Fluidigm) after antibody staining and fixation (**Table 5.2**). Mass cytometry data were normalized using EQ™ Four Element Calibration Beads (Fluidigm), as previously described¹⁵². 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 **Figure 5.3**. 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 80th quantiles of their respective density distributions. Let the single-cell RNA-seq matrix be C (g genes by m cells, $c_1, \dots, c_m$) and the bulk RNA-seq matrix be S (g genes by n samples, $s_1, \dots, s_n$), respectively. We then retain only the genes detected in both scRNA-seq and bulk RNA-seq datasets.
2. Based on the shared highly variable genes, we use CCA to integrate 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 maximized¹⁵³. CCA finds two vectors a and b that maximize the linear correlations $cor(CV_{S1}, CV_{C1})$, where:

$$CV_{S1} = a_1s_1 + a_1s_1 + \cdots a_ns_n$$

$$CV_{C1} = b_1c_1 + b_1c_1 + \cdots b_mc_m$$

So, 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 $s_1, \dots, s_n$ arranges bulk genes along the canonical variate CV_{S1} and the linear combination of all cells $c_1, \dots, c_m$ arranges single-cell genes along CV_{C1} . CCA defines the coefficients $a_1, \dots, a_n$ and $b_1, \dots, b_m$ 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.

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.
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 score¹⁵⁴, and fold-change.

8. For pathway analysis, we downloaded gene sets from Gene Ontology (GO) terms on April 2017^{85,155}. This included 9,797 GO terms and 15,693 genes. We also used the immunological signatures from 4872 hallmark gene sets from MSigDB¹⁵⁶ to test enrichment of all the tested genes sorted by decreased AUC scores for each cluster by 10^5 permutation tests⁸⁵. 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^{157,158} on the cell-to-cell Euclidean distance matrix to evaluate our clustering results (**Figure 5.3b**). For each cell, the silhouette width $s(i)$ is defined as follows:

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

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 i 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 matrices, we ran our pipeline on the shared donors for which both datasets were available (**Figure 5.1b**). 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 subsampled 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) 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 modified version of DensVM¹²⁴. 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 **Table 5.4**.

Table 5.4 Cell populations identified in mass cytometry data.

Cell type	Mass cytometry cluster	# Cells	Cells from OA (%)	Cells from leukocyte-poor RA (%)	Cells from leukocyte-rich RA (%)	One-sided FDR q value
Fibroblast	THY1 ⁻ Cadherin-11 ⁻	4,333	48.3	45.7	6.0	1.00
	THY1 ⁻ Cadherin-11 ⁺	3,752	55.8	40.9	3.3	1.00
	THY1 ⁻ CD34 ⁺ HLA-DR ^{hi}	1,474	42.5	44.2	13.3	1.00
	THY1 ⁻ CD34 ⁻ HLA-DR ^{hi}	4,171	50.2	28.3	21.5	1.00
	THY1 ⁺ CD34 ⁻ HLA-DR ^{hi}	2,577	7.1	9.4	83.4	5.8E-04
	THY1 ⁺ CD34 ⁻ HLA-DR ^{lo}	3,794	38.1	38.6	23.3	1.00
	THY1 ⁺ CD34 ⁺ HLA-DR ^{lo}	3,514	52.0	40.9	7.1	1.00
	THY1 ⁺ CD34 ⁺ HLA-DR ^{hi}	1,546	9.8	10.2	80.0	1.2E-03
Monocyte	CD11c ⁻	2,801	58.0	33.4	8.6	1.00
	CD11c ⁺ CCR2 ⁺	5,602	32.6	24.2	43.2	0.72
	CD11c ⁺ CD38 ⁻	1,313	37.0	54.8	8.2	1.00
	CD11c ⁺ CD38 ⁻ CD64 ⁺	1,031	37.9	43.7	18.3	1.00
	CD11c ⁺ CD38 ⁺	4,551	20.0	13.1	66.9	9.5E-04
T cell	CD4 ⁻ CD8 ⁻	2,475	44.6	34.5	20.9	1.00
	CD4 ⁺ CCR2 ⁺	4,093	51.7	28.7	19.6	1.00
	CD4 ⁺ HLA-DR ⁺	1,178	29.3	62.3	8.4	1.00
	CD4 ⁺ PD-1 ⁺ ICOS ⁻	2,368	33.7	37.1	29.1	1.00
	CD4 ⁺ PD-1 ⁺ ICOS ⁺	3,138	22.9	28.6	48.5	1.7E-03
	CD8 ⁺ PD-1 ⁻ HLA-DR ⁻	2,535	46.1	31.6	22.3	1.00
	CD8 ⁺ PD-1 ⁻ HLA-DR ⁺	400	42.3	35.5	22.3	1.00
	CD8 ⁺ PD-1 ⁺ HLA-DR ⁻	2,710	40.0	29.3	30.8	1.00
	CD8 ⁺ PD-1 ⁺ HLA-DR ⁺	1,088	9.8	9.4	80.8	9.5E-04
B cell	CD38 ⁺⁺ CD20 ⁻ IgM ⁻ IgD ⁻	770	3.6	6.5	89.9	0.04
	CD38 ⁺⁺ CD20 ⁻ IgM ⁺ HLA-DR ⁺	220	1.8	5.5	92.7	0.04
	IgM ⁻ IgD ⁻ HLA-DR ⁻	379	42.0	39.1	19.0	1.00
	CD38 ⁺ HLA-DR ⁺⁺ CD20 ⁻ CD11c ⁺	550	16.7	18.2	65.1	1.00
	IgM ⁻ IgD ⁻ HLA-DR ⁺⁺ CD20 ⁺ CD11c ⁺	771	1.0	6.6	92.3	0.01
	IgM ⁻ IgD ⁻ HLA-DR ⁺	1,774	15.7	25.3	59.0	1.00
	IgA ⁺ IgM ⁻ IgD ⁻	336	20.2	11.0	68.8	1.00
	IgM ⁺ IgD ⁻	1,126	18.7	22.6	58.8	1.00
	IgM ⁺ IgD ⁺ CD11c ⁻	1,762	5.6	7.3	87.2	0.05
	IgM ⁺ IgD ⁺ CD11c ⁺	491	8.8	7.5	83.7	0.69

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 MASC¹²¹, 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 5.9h**).
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 ρ between the scRNA-seq average CV and the CV for mass cytometry clusters.
5. Next, we generated a null distribution for the Spearman correlation by shuffling the scRNA-seq gene names and then repeating Steps 2-4 10,000 times.
6. Then, we counted how many times the correlation of each pair was greater than the observed value from Step 4 and computed a permutation p-value:

$$p = \frac{1 + \sum_{i=1}^{10,000} \rho_i \geq \rho}{1 + 10,000}$$

7. Finally, we converted the permutation p-value 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 $\text{Log}_2(\text{CPM}+1)$. We prioritized cluster marker genes using three criteria: 1) percent of non-zero expressing cells $> 60\%$; 2) are under the receiver-operator curve (AUC)¹⁵⁴ > 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). Cells 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% CO₂ 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 **Figure 5.16**.

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. Genes with Benjamini-Hochberg FDR < 0.01 and fold-change > 2 were considered to be statistically significant where appropriate.

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 (accession codes SDY998 and SDY999). The raw single-cell RNA-seq and mass cytometry data are deposited in dbGAP (accession code phs001457.v1.p1).

Source code for the computational and statistical analysis is available at:

- https://github.com/immunogenomics/amp_phase1_ra.

Data can also be viewed online at these websites:

- <https://immunogenomics.io/ampra>
- <https://immunogenomics.io/cellbrowser>
- https://portals.broadinstitute.org/single_cell/study/amp-phase-1

Chapter 6

Transcription factors mediating response to TNF and IL-17A

In rheumatoid arthritis (RA), stromal fibroblasts are key components of the hyperplastic and invasive synovial tissue that mediates joint destruction and bone resorption ¹⁶. Interleukin 17 (IL-17A), a cytokine produced by CD4⁺ Th17 cells ¹⁵⁹ and other cells ¹⁶⁰, is crucial to perpetuate the vicious inflammatory cycle localized to the joint tissue that contributes to induction, progression, and chronicity of RA ¹⁶¹. Briefly, leukocytes activate stromal cells to produce inflammatory cytokines and chemokines such as IL-6, IL-8, CCL2 ¹⁶², and GM-CSF ¹⁶³. In turn, these cytokines attract more leukocytes to infiltrate the synovial tissue and activate them to produce more IL-17A. The infiltrating leukocytes produce tumor necrosis factor alpha (TNF), and IL-17A synergizes with TNF to further amplify the pathways underlying chronic inflammatory disease pathogenesis ¹⁶⁰.

The combination of TNF and IL-17A synergistically activates synovial fibroblasts to produce IL-6 ¹⁶², and these cells are a dominant source of IL-6 in the joint tissue ^{50,164}. While other cytokines are involved in RA, clinical efficacy of targeting TNF and IL-6 has clearly established their pivotal roles in pathogenesis ¹⁶⁵. Clinical relevance of the synergistic activity of TNF and IL-17A has also been shown: measurements of TNF and IL-17A mRNA in RA synovial tissue predict early joint damage progression by magnetic resonance imaging (MRI) ¹⁶⁶. Joint damage progresses as TNF and IL-17A induce synovial fibroblasts to produce matrix metalloproteinases (MMPs) for digesting extracellular matrix and invading the cartilage ¹⁶⁷. In addition, IL-17A plays a role in differentiation of osteoclast progenitors into bone-resorbing osteoclasts via increased expression of RANKL in osteoblasts ¹⁶⁸.

However, most studies of IL-17A transcriptional response have used qPCR to measure only the most critical genes with known roles in the context of inflammation, osteoclastogenesis, angiogenesis, or neutrophil recruitment ^{162,167,169–174}. One study used microarrays to measure

genome-wide gene expression response in a single sample 12 hours after stimulation with IL-17A and IL-17F alone or in combination with TNF ¹⁶². An expanded view with robust measurements of transcriptional dynamics over time would help to reveal mediators in regulatory pathways downstream of TNF and IL-17A. While these regulatory pathways are well studied in leukocytes ¹⁷⁵, there remains much to be learned about the transcriptional regulation of inflammatory functions in mesenchymal cells such as synovial fibroblasts. For example, we recently described an autocrine loop driving IL-6 production that is only used by fibroblasts, and not by leukocytes, involving LIF, LIFR, and STAT4 ¹⁶⁴. Many different cell types have been shown to use the NF- κ B signaling pathway for inflammatory response, but additional components of this pathway are being discovered that modulate cell type and condition-specific responses to inflammatory signals ¹⁷⁶. We aimed to discover additional factors that regulate the inflammatory response (**Figure 6.1**).

Surprisingly, IL-17A in the absence of TNF has a minimal transcriptional effect in fibroblasts, and defining its role with TNF requires a highly sensitive study. To identify key mediators of the response to TNF and IL-17A in human synovial fibroblasts, we use transcriptomics data from a high-resolution time series with multiple IL-17A doses. By measuring the *in vitro* transcriptomic effects caused by stimulation with TNF or costimulation with TNF and IL-17A at three doses of IL-17A (0, 1, 10 ng/mL), we identify genes with transcriptional changes proportional to IL-17A dose. We also find that transcription of specific sets of chemokines is controlled by different mediators. A molecular complex of NF- κ B p65 and CUX1 mediates transcription of CXC chemokines (*CXCL1*, *CXCL2*, *CXCL3*, *CXCL6*). In contrast, LIFR mediates transcription of CC chemokines (*CCL7*, *CCL8*, *CCL20*). We distinguish these transcriptional responses by silencing each mediator with short interfering RNAs (siRNAs)

and then assaying the downstream effects. Furthermore, we find that transcription of atypical I κ B factor I κ B ζ (NFKBIZ) is responsive to IL-17A dose, and I κ B ζ mediates IL-6, IL-8, and MMP-3 production only after costimulation with TNF and IL-17A, but not after TNF alone. Finally, we demonstrate that fibroblasts require these mediators to recruit neutrophils and monocytes in an *in vitro* assay.

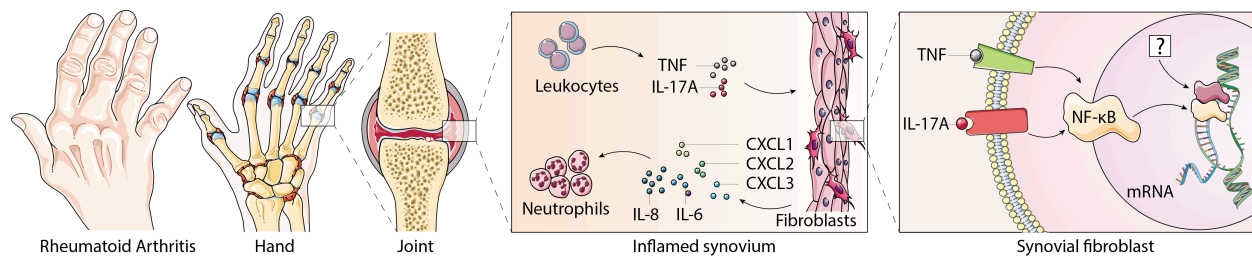


Figure 6.1. Overview of the fibroblast response to TNF and IL-17A in the synovial tissue.

Our study is focused on the role of synovial fibroblasts in the inflamed joint from a patient with rheumatoid arthritis. We know leukocytes infiltrate the tissue and produce inflammatory cytokines such as TNF and IL-17A. We also know that stromal cells such as synovial fibroblasts respond to TNF and IL-17A by producing inflammatory cytokines such as IL-6 and chemoattractants such as IL-8 (CXCL8), CXCL1, CXCL2, CXCL3. The NF- κ B pathway mediates the response to TNF, and we want to discover other mediators.

TNF and IL-17A synergistically activate IL-6 and MMP-3 production

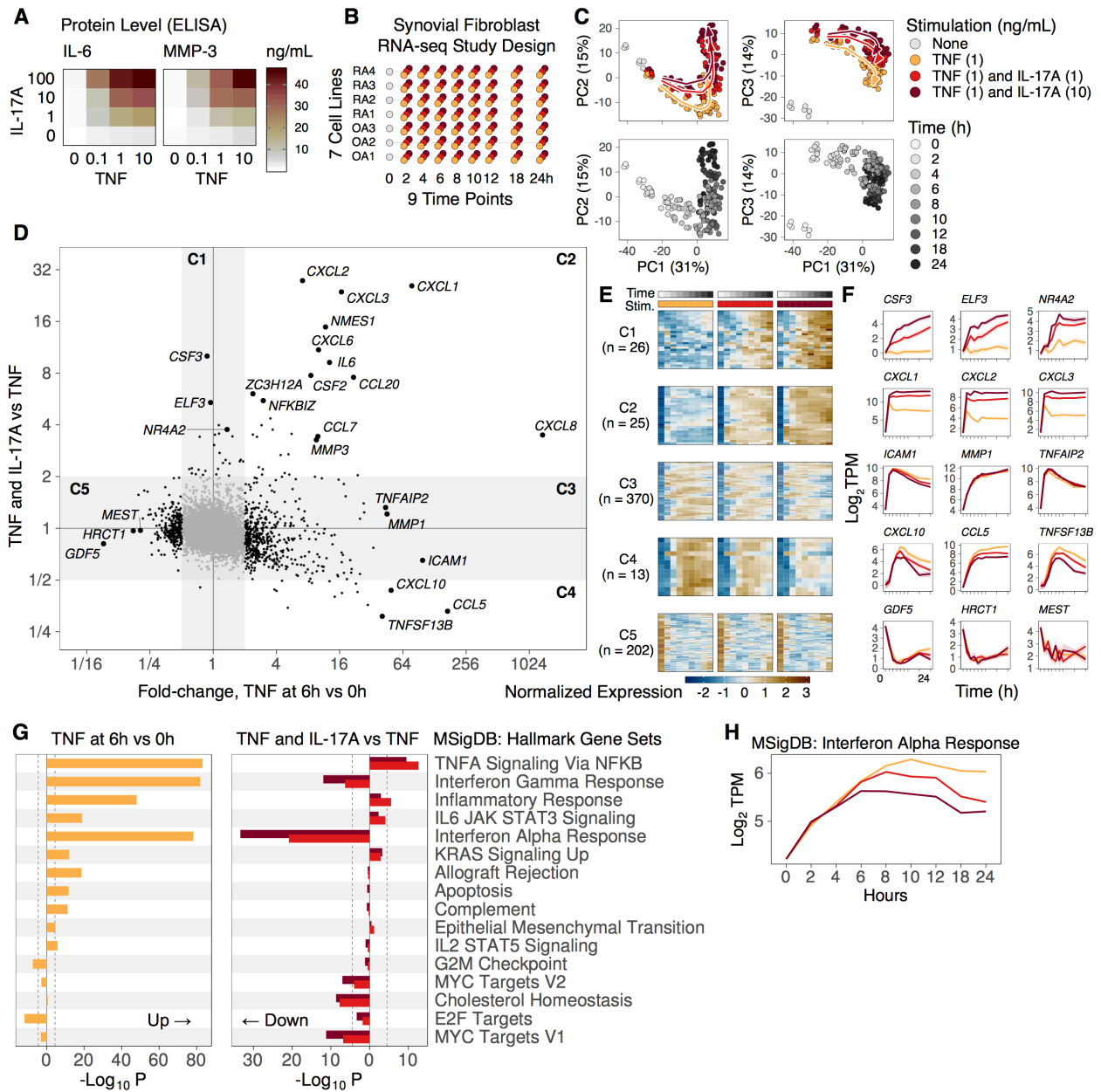
To demonstrate the synergy between TNF and IL-17A, we stimulated synovial fibroblasts with 16 combinations of dosages of TNF (0, 0.1, 1, 10 ng/mL) and IL-17A (0, 1, 10, 100 ng/mL). We cultured primary fibroblasts from two RA donors and measured IL-6 protein levels in the supernatant after stimulation by enzyme-linked immunosorbent assay (ELISA). Stimulation with

one cytokine (TNF or IL-17A) is significantly less effective at inducing IL-6 than costimulation with TNF and IL-17A. We saw 1.2 ng/mL IL-6 after stimulation with IL-17A alone (100 ng/mL), and 2.6 ng/mL IL-6 after TNF alone (10 ng/mL) (**Figure 6.2A**). In contrast, we saw 49.3 ng/mL of IL-6 in the supernatant after costimulation with TNF (10 ng/mL) and IL-17A (100 ng/mL) – the addition of IL-17A caused a 20-fold increase in IL-6 levels relative to stimulation with TNF alone (**Figure 6.2A**). Matrix Metalloproteinase 3 (MMP-3) protein levels showed a similar synergistic response to TNF and IL-17A (**Figure 6.2A**). These results confirm previous reports that TNF and IL-17A synergistically activate IL-6, MMP-3, and other inflammatory cytokines^{167,169–172,174}.

Figure 6.2. Defining the transcriptional response to TNF and IL-17A in fibroblasts with a high resolution time series and dose response study.

(A) Protein levels of IL-6 and MMP-3 measured by ELISA after stimulating with 16 different combinations of doses: TNF (0, 0.1, 1, 10 ng/mL) and IL-17A (0, 1, 10, 100 ng/mL). Costimulation with TNF and IL-17A results in synergistic IL-6 and MMP-3 protein production. (B) Our study design includes 7 cell lines from patient tissues. We stimulated them with TNF (1 ng/mL) and three doses of IL-17A (0, 1, and 10 ng/mL). We measured gene expression at baseline and 8 time points after stimulation. (C) Principal Component Analysis (PCA) with 636 genes responsive to TNF alone or TNF and IL-17A. Samples fall along the principal curves. 60% of the total variation is captured by the first 3 principal components. (D) Differential expression analysis shows that genes fall into 5 major categories, depending on the response to TNF alone (x-axis) or to TNF and IL-17A (y-axis). (E) Heatmaps with normalized expression for 636 genes, where rows are genes and columns are time points. From left to right, each heatmap shows response to TNF (1 ng/mL) and each of the three doses of IL-17A (0, 1, or 10 ng/mL). (F) Mean gene expression (lines) and SEM (ribbons) over time for genes with characteristic profiles: *CSF3*, *ELF3*, *NR4A2*, *CXCL1*, *CXCL2*, *CXCL3*, *ICAM1*, *MMP1*, *TNFAIP2*, *CXCL10*, *CCL5*, *TNFSF13B*, *GDF5*, *HRCT1*, *MEST*. (G) Gene set enrichment with MSigDB Hallmark gene sets. The x-axis shows signed p-values indicating the strength of enrichment with induced or repressed genes for TNF at 6h versus 0h (left) and TNF and IL-17A (1 or 10 ng/mL) versus TNF (right). (H) We selected the 10 genes in the Interferon Alpha Response gene set with the most significant p-values for TNF and IL-17A (10 ng/mL) versus TNF, and show mean +/- SEM of the means of the 10 genes.

Figure 6.2 (Continued)



IL-17A requires TNF to induce a gene expression response in fibroblasts

Since IL-17A resulted in synergistic protein production of IL-6 and MMP-3, we decided to assay the direct genome-wide transcriptional effects of IL-17A. To do this, we conducted two initial genome-wide transcriptional experiments after stimulation with IL-17A or TNF, alone or in combination.

First, we profiled one synovial fibroblast cell line at 9 time points over 72 hours in 4 different conditions: media only, TNF (1 ng/mL), IL-17A (1 ng/mL), and TNF with IL-17A (1 ng/mL and 1 ng/mL). We used microarrays to assay the transcriptional response (Microarray Data 1, **Figure 6.3**). After stringent quality control, we fit a linear model to each gene to find which genes show differential expression between conditions (Materials and Methods). In response to IL-17A alone, 3 genes (*NFKBIZ*, *TNFAIP6*, and *CXCL1*) were differentially expressed relative to media ($FC > 1.5$ and $FDR < 0.05$, **Figure 6.3D**). On the other hand, 1,208 genes responded to TNF alone (**Figure 6.3E**). In response to TNF and IL-17A, 15 genes were differentially expressed relative to TNF alone (**Figure 6.3F**). We also applied principal components analysis (PCA) ¹⁷⁷ to Microarray Data 1. Along the first two components, samples stimulated with IL-17A alone are most similar to unstimulated samples, and there is little difference between samples stimulated with TNF alone or TNF and IL-17A (**Figure 6.3B,C**). We conclude that that IL-17A is dependent on the presence of TNF to induce significant effects on gene expression, consistent with our ELISA results for protein levels of IL-6 and MMP-3 (**Figure 6.2A**).

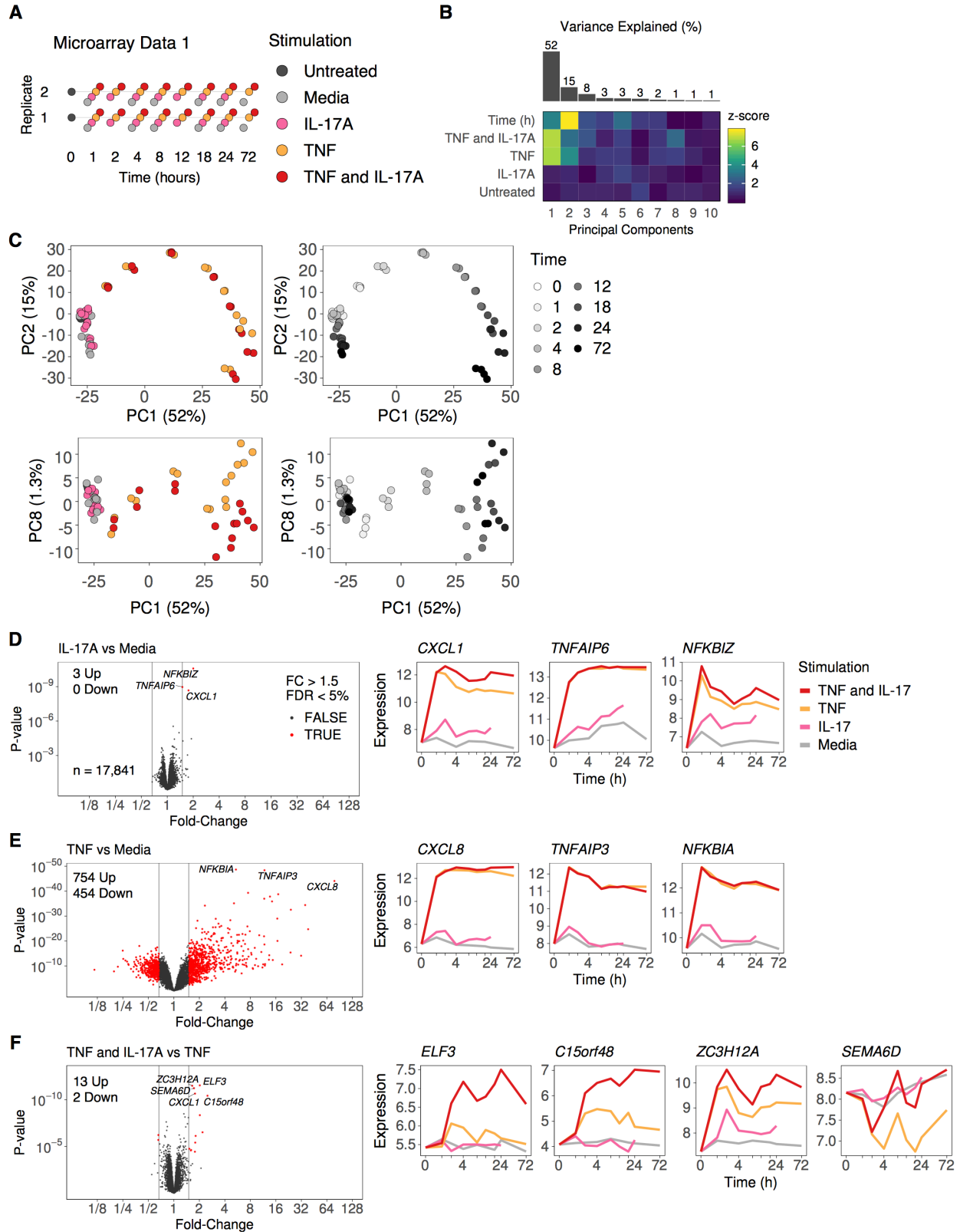
Given the subtle transcriptional difference between costimulation with TNF and IL-17A versus TNF alone, we ran a second experiment where we assayed gene expression response in

four synovial fibroblast cell lines (2 RA and 2 OA) 24 hours after stimulation (Microarray Data 2, **Figure 6.4**). This experiment replicated the differential expression signal between TNF and IL-17A versus TNF from Microarray Data 1 ($r = 0.67$, $P < 10^{-15}$). This gave us confidence that a well-powered transcriptional study could identify the downstream mediators that respond to costimulation with TNF and IL-17A.

Figure 6.3. IL-17A requires TNF to induce gene expression response in fibroblasts.

(A) Experimental design for Microarray Data 1 ($n = 60$): 9 time points, 2 technical replicates, 5 different stimulations (Untreated, Media, IL-17A (1 ng/mL), TNF (1 ng/mL), TNF and IL-17A). We used the GeneChip™ Human Gene 1.0 ST Array ("affy_hugene_1_0_st_v1" on Bioconductor). (B) Summary of principal components analysis (PCA) on 17,841 genes. We see 52, 15, and 8 percent of variation is captured by the first 3 principal components (PCs). We fit a linear model to each PC ($PC \sim Time + Treatment$) where *Time* is a numeric variable and *Treatment* is a factor. Then we transformed the term p-values to normal z-scores and displayed them in a heatmap. We see that PCs 1 and 2 have strongest association with time, and PCs 1 and 8 have the strongest association with TNF and IL-17A costimulation. (C) Samples projected onto PC1 and PC2 (top) or PC1 and PC8 (bottom), colored by stimulation (left) or time (right). Along the first two principal components, stimulation with TNF and IL-17A is similar to TNF alone. Also, stimulation with IL-17A alone is similar to no stimulation at all (Media). (D, E, F) Volcano plots and expression profiles of selected genes. We set a threshold at 0.05 FDR and 1.5 fold-change. At this threshold, 3 genes are induced by IL-17A versus media (D), 754 by TNF versus media (E), and 13 by TNF and IL-17A versus TNF (F). 454 genes are repressed by TNF, and 2 genes repressed by TNF and IL-17A versus TNF. Adjacent panels show gene expression levels over time, colored by stimulation. Lines represent means of 2 observations for each stimulation. The y-axis represents microarray expression intensity on the Log_2 scale.

Figure 6.3 (Continued)



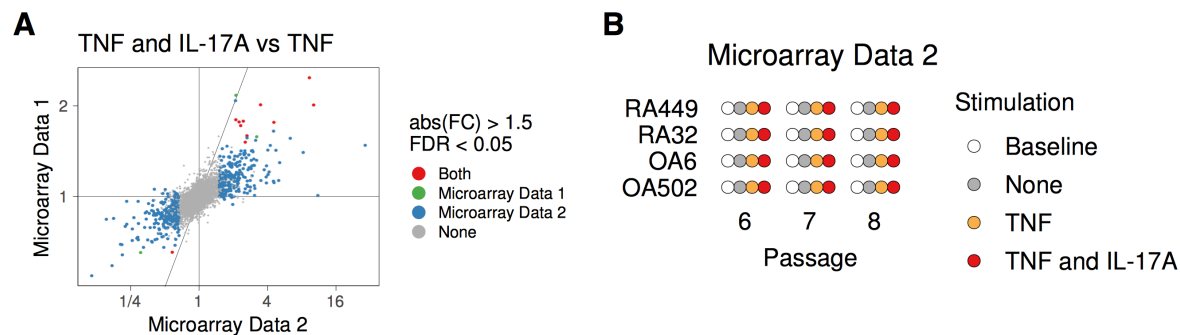


Figure 6.4. Synergy between TNF and IL-17A is reproducible across independent experiments, laboratories, technologies, and cell lines.

(A) We tested differential expression in Microarray Data 1 and Microarray Data 2, and found that the log2 fold-change is correlated (Pearson correlation = 0.67, $p < 2 \times 10^{-16}$). We found 11 genes with significant differential expression in both datasets for the contrast of TNF and IL-17A versus TNF: *ABCA8*, *AGPAT4*, *C15orf48*, *CCL8*, *CFAP69*, *CNKS3*, *CXCL1*, *CXCL2*, *EDNRB*, *ELF3*, *SEMA6D*. (B) Experimental design for Microarray Data 2 (n = 48): 1 time point (24 hours), 3 passages, 4 cell lines, 4 different stimulations (Baseline, None, TNF, TNF and IL-17A). We used the GeneChip™ Human Genome U133 Plus 2.0 Array ("affy_hg_u133_plus_2" on Bioconductor).

Defining the transcriptional response to TNF and IL-17A in fibroblasts with a high resolution time series and dose response study

We used RNA-seq to assay transcriptomes of 7 synovial fibroblast cell lines (4 RA, 3 OA) at 9 time points over 24 hours after stimulation with TNF (1 ng/mL) and IL-17A (0, 1, and 10 ng/mL), resulting in a total of 175 RNA-seq profiles (**Figure 6.2B**, Materials and Methods). By using RA and OA cell lines, we could look for responses consistent across tissue states. By using multiple IL-17A dosages, we could identify genes with expression proportional to IL-17A dose.

By modeling the dynamics of each gene over time, we increased our statistical power to identify differentially expressed genes (Materials and Methods).

Broadly, we found that genes could be subdivided into categories based on their response to TNF and IL-17A. We used linear models to estimate the effect sizes for each gene (see Materials and Methods). Next, we identified the genes with greatest response to (1) TNF at 6h vs 0h and (2) costimulation with TNF and IL-17A vs TNF alone (**Figure 6.2C**). We limited our analysis to the 636 statistically significant genes (absolute FC > 2 and FDR < 0.05) and subdivided them into 5 categories (**Figure 6.2D**). We also applied principal components analysis (PCA) ¹⁷⁷ to these 636 genes (**Figure 6.2E, Figure 6.5**), and found that the first component (explaining 31% of the total variance) identifies genes with strong activation or deactivation over time in response to TNF stimulation. The second (15% of variance) and third (14% of variance) identify genes with expression altered by IL-17A dose.

Transcriptional response reveals genes selectively induced by TNF and IL-17A costimulation or TNF alone

We defined Category C1 as the set of genes induced by costimulation with TNF and IL-17A, but unchanged by TNF alone (**Figure 6.2D,F**). C1 includes 26 genes with dose-dependency on IL-17A such as *CSF3*, encoding granulocyte colony stimulating factor G-CSF, a crucial factor for neutrophil survival and development ^{178,179} (**Figure 6.6**). Interleukin 11 (*IL11*), a cytokine that is elevated in RA and promotes fibroblast invasion ¹⁸⁰, was also transcribed in an IL-17A dose-dependent manner. The transcription factor (TF) with greatest expression response to IL-17A was the epithelium-specific ETS factor encoded by *ELF3*, so we hypothesized that ELF3 might mediate the IL-17A dose-dependent expression of other genes. For this reason, we included it in an siRNA gene silencing experiment (see below). Another TF in Category C1 is the orphan nuclear receptor encoded by *NR4A2*, which has been shown to induce synoviocyte proliferation and invasion ¹⁸¹, and also to directly control expression of *CXCL8* in inflammatory arthritis ¹⁸². Category C1 also includes “POU domain, class 2, transcription factor 2” encoded by *POU2F2*: a TF with a binding site motif in the promoters of *CSF3* and immunoglobulin genes (**Figure 6.6**).

Category C2 has 25 genes induced in response to TNF and further induced by addition of IL-17A (**Figure 6.2D,F**), including members of the CXC subfamily of chemokines *CXCL1*, *CXCL2*, *CXCL3*, *CXCL6*, and *CXCL8*. Notably, *CXCL8* mRNA was 1000-fold more abundant after TNF at 6h relative to 0h, and the other 4 chemokines were 17-fold more abundant on average (**Figure 6.6**). While TNF alone caused increased CXC chemokine transcription, the addition of IL-17A further increased transcription of these five chemokines in a dose-dependent manner (an additional 7-fold and 15-fold for 1 and 10 ng/mL IL-17A, respectively) (**Figure 6.6**).

We note that IL-8 (CXCL8), CXCL1, and CXCL2 are potent neutrophil chemoattractants, complementing the function of G-CSF (encoded by *CSF3*).

Category C3 genes increase in response to TNF, but they are unaffected by the addition of IL-17A (**Figure 6.2D,F**). This is the largest category with 370 genes, including TNF alpha induced genes *TNFAIP6* and *TNFAIP2*, as well as matrix metalloproteinase 1 (*MMP1*) and bradykinin receptor (*BDKRB1*) (**Figure 6.6**). We tested MSigDB Hallmark gene sets ⁸⁶ for enrichment with TNF responsive genes and found significant enrichments in the gene sets "TNFA signaling via NFKB", "Interferon Alpha Response", and "Interferon Gamma Response" (**Figure 6.2G**).

Category C4 genes are transcribed in response to TNF, but transcription is dampened by the addition of IL-17A. C4 includes 13 genes related to interferon response such as *IRF1*, *IRF9*, *IFI30*, and *CXCL11* (**Figure 6.7**). Gene set enrichment analysis shows that genes in the MSigDB hallmark gene set "Interferon Alpha Response" were induced by TNF, but repressed by TNF and IL-17A (**Figure 6.2G**). Indeed, an increase from 1 ng/mL IL-17A to 10 ng/mL IL-17A causes more repression of these genes (**Figure 6.2H**). This suggests that TNF activates – and IL-17A attenuates – the genes responsive to interferon alpha.

Category C5 has 202 genes repressed by TNF but largely unaffected by IL-17A. Enrichment with MSigDB hallmark pathways reveals that many of these genes are related to cell cycle such as *STMN1*, *CDKN2C*, *CDC20*, *CDKN3*, *HMGB2*, *CDK1*, *CCNB2*, *CDCA8*, *TOP2A* (**Figure 6.7**). While TNF stimulation caused reduced transcription of genes involved in cell division, it caused increased transcription of genes involved in apoptosis such as *SOD2* and *BIRC3*. The genes with greatest reduction after TNF stimulation include growth differentiation factor 5 (*GDF5*), histidine rich carboxyl terminus 1 (*HRCT1*), and mesoderm specific transcript

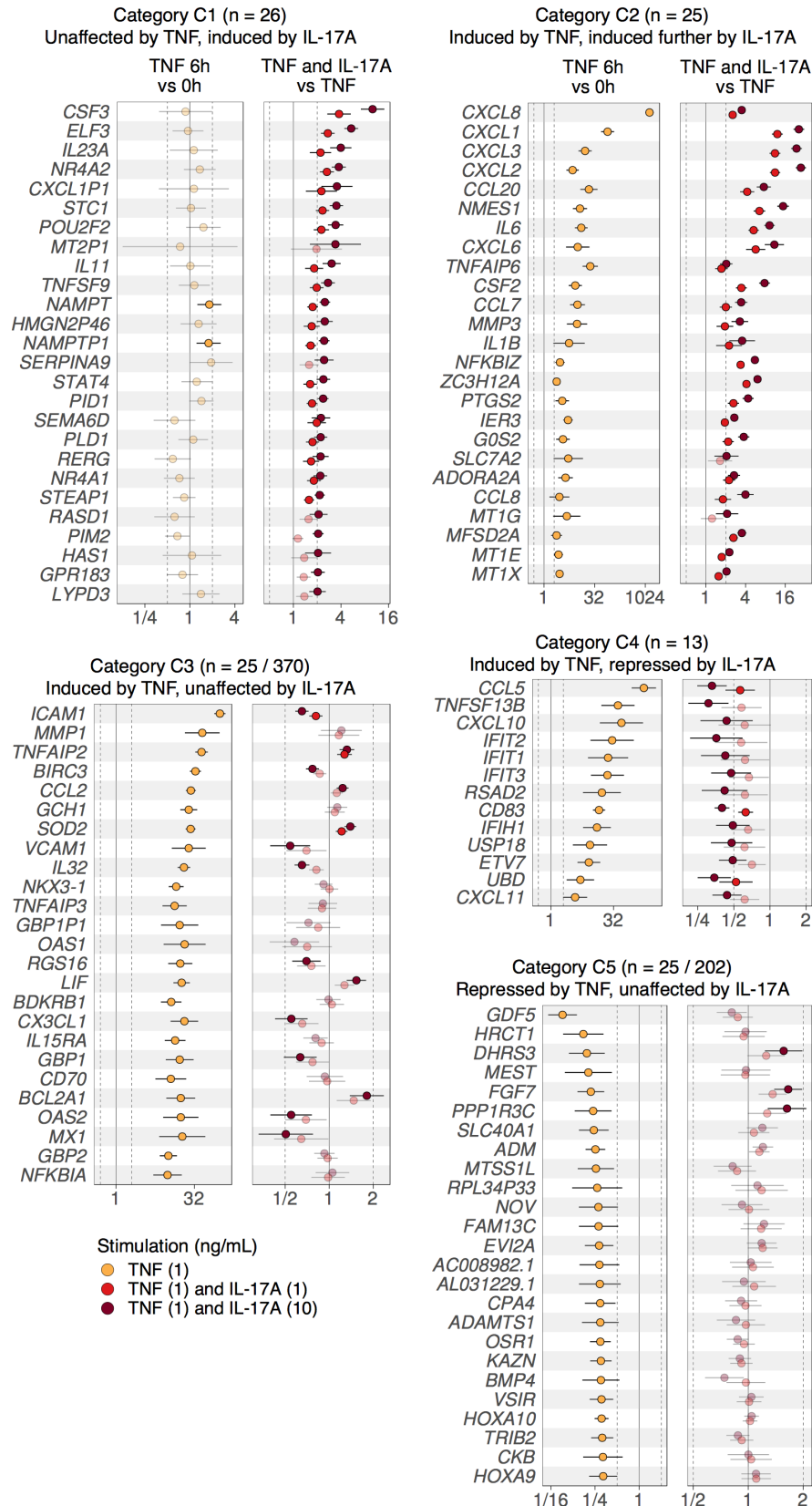
(*MEST*) (**Figure 6.2D**). For example, *GDF5* – a critical factor for normal synovial joint function¹⁸³ and synovial lining hyperplasia¹⁸⁴ – decreased 11-fold (fold change 0.09, 95% CI 0.058-0.14, $P = 3\text{e-}20$) after 6h of TNF.

We conclude that TNF alone is sufficient to activate transcription of the CXC subfamily of neutrophil recruitment chemokines, and that IL-17A synergizes with TNF to increase transcription of these chemokines. Costimulation with both TNF and IL-17A is required for transcription of *CSF3*, while TNF alone is insufficient. Thus, costimulation of synovial fibroblasts with TNF and IL-17A activates a transcriptional program for recruiting neutrophils with CXC chemokines and sustaining neutrophil survival with G-CSF (*CSF3*).

Figure 6.6. Response to TNF alone or the combination of TNF and IL-17A.

Differential gene expression effect sizes and 95% confidence intervals for genes in each of the 5 Categories (C1, C2, C3, C4, C5). Three contrasts are shown: TNF at 6h versus 0h, TNF and IL-17A (1 ng/mL) vs TNF, and TNF and IL-17A (10 ng/mL) vs TNF. Faded color indicates $FDR > 0.05$.

Figure 6.6 (Continued)



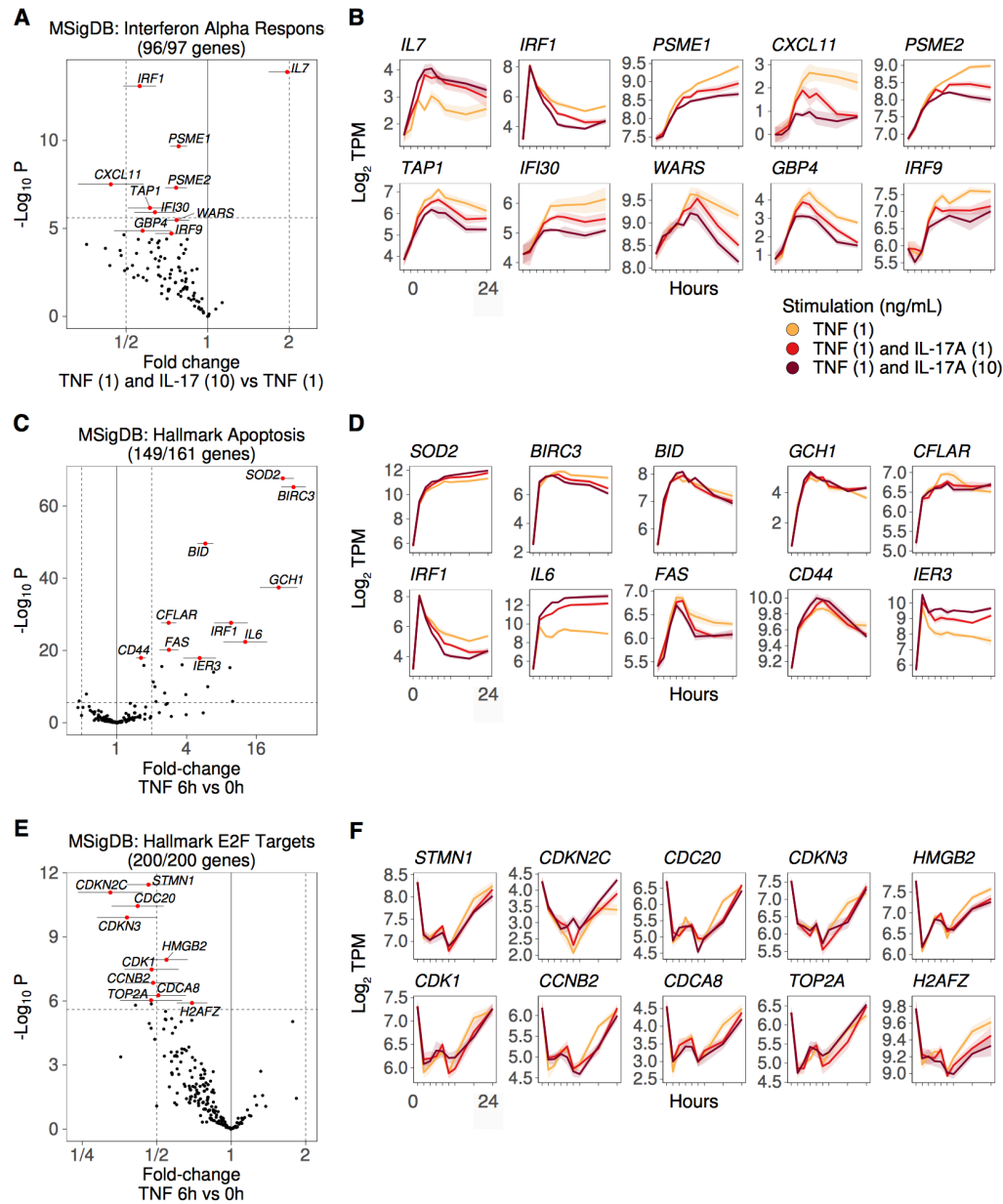


Figure 6.7. Genes induced by TNF in the MSigDB Hallmark gene set called "Interferon Alpha Response" are repressed by IL-17A.

(A,C,E) Volcano plots, showing fold change estimates and p-values for the contrast of TNF and IL-17A (10 ng/mL) versus TNF. The top 10 genes with lowest p-values are highlighted in red. The titles indicate the number of genes measured in RNA-seq Data 1 and the total number of genes in the gene set, e.g. 96 of the 97 genes in the MSigDB Hallmark gene set called "Interferon Alpha Response" are measured in RNA-seq Data 1. (B,D,F) Expression over time, shown as mean $\pm$ SEM of the Log2 transcripts per million (TPM).

Transcription factors underlying the synergy between TNF and IL-17A

We used two strategies to find putative transcription factors that mediate the response to costimulation with TNF and IL-17A. First, we identified TFs with transcription proportional to IL-17A dose (Categories C1 and C2), including *ELF3*, *POU2F2*, *ZC3H12A*. Since *ELF3* expression increased (FC 5.4, 95% CI 4.4-6.6, $P = 2e-37$) after stimulation with IL-17A (10 ng/mL), we decided to study the global transcriptome-wide effects of silencing *ELF3* with siRNA.

Next, we searched for TF binding motifs in the promoters of differentially expressed genes. We used HOMER¹⁸⁵ to perform *de novo* motif enrichment analysis with 26 genes within Categories C1 and C2 that had greater expression for 10 versus 1 ng/mL IL-17A ($\log_2$ FC > 1.5 (FC > 2.83) and FDR < 0.05) (**Methods**). As expected, we found motifs that closely resemble NF- κ B, REL, and JUND binding sites (**Table 6.1**). We also found a motif that might represent the binding site for cut like homeobox 1 (CUX1), TCCGGATCGATC, that is present in the promoters of 4 of the 26 genes ($P = 1e-10$): *CXCL1*, *CXCL2*, *CXCL3*, *PIM2*. When we examined public CUX1 ChIP-seq data from ENCODE for three cell lines (GM12878, K562, MCF-7), we found evidence that CUX1 might bind these motifs. Consistent with our motif analysis and differential expression analysis, there are CUX1 ChIP-seq peaks in the promoters of *CXCL2*, *CXCL3*, *C15orf48* (*NMES1*), and *IL6*, as well as less than 5kb upstream of the promoter of *CXCL6* (**Figure 6.8A**). These results motivated us to assay the transcriptome-wide effects of silencing *CUX1*.

In addition to silencing *ELF3* and *CUX1*, we also decided to silence the genes encoding leukemia inhibitory factor receptor (*LIFR*) and signal transducer and activator of transcription 4

(*STAT4*). Previously, we found that LIFR and STAT4 are key components of an autocrine loop controlling IL-6 production that is specific to fibroblasts and not leukocytes ¹⁶⁴. We also decided to silence *STAT3* because it encodes a known mediator of inflammatory pathways ¹⁸⁶, and we wanted to see how its function might be distinct from STAT4 in the fibroblast response to inflammatory chemokines.

Table 6.1. HOMER *de novo* motifs in gene promoters.

Best Match	P	FG (%)	BG (%)	Consensus	Promoters
NFkB-p65-Rel(RHD)/ThioMac-LPS-Expression(GSE23622)	10 ⁻¹²	16	0.0043	CAGGGAATTTCC	<i>CXCL1, CXCL2, CXCL3, CXCL6</i>
CUX1/MA0754.1/Jaspar	10 ⁻¹⁰	16	0.016	TCCGGATCGATC	<i>CXCL1, CXCL2, CXCL3, PIM2</i>
Nur77(NR)/K562-NR4A1-ChIP-Seq(GSE31363)	10 ⁻¹⁰	16	0.025	ACCTTYMCAWKA	<i>CCL20, CSF2, MMP3, TNFSF13B</i>
REL/MA0101.1/Jaspar	10 ⁻⁹	20	0.11	TCCGGGSTTTCC	<i>CXCL1, CXCL2, CXCL3, NFKBIZ, PTGS2</i>
GLI2/MA0734.1/Jaspar	10 ⁻⁹	16	0.027	GACCGCGCTGGC	<i>C15orf48, CD83, G0S2, IL4I1</i>
Hic1/MA0739.1/Jaspar	10 ⁻⁸	24	0.51	GCGATGGCCC	<i>C15orf48, CCL20, CXCL1, CXCL2, CXCL3, G0S2, NFKBIZ</i>
EWSR1-FLI1/MA0149.1/Jaspar	10 ⁻⁷	12	0.019	GAAGGAAGGCGA	<i>CXCL1, CXCL2, CXCL3</i>
JUND/MA0491.1/Jaspar	10 ⁻⁶	20	0.56	GACTCATCCT	<i>CCL7, CSF2, IL11, MMP3, NFKBIZ</i>

HOMER *de novo* motif analysis of the promoters of 25 genes with statistically significant difference ($\log_2$ FC > 1.5 and FDR < 0.05) for 10 versus 1 ng/mL of IL-17A. For each discovered motif, we show the name of a known motif that is similar to the newly discovered motif, the hypergeometric p-value from HOMER, the percent of matches in the foreground (FG) gene set of 25 genes, the percent of matches in the background (BG) gene set of 39,486 genes, and the list of genes in the foreground that have a match for the motif.

Figure 6.8. ENCODE CUX1 ChIP-seq peaks in promoters of *CXCL2*, *CXCL3*, *C15orf48* (*NMES1*), *IL6*, and near *CXCL6*.

(A) We used UCSC Genome Browser to visualize CUX1 ChIP-seq data (<https://www.encodeproject.org/targets/CUX1-human/>) in three cell lines: MCF-7, K562, GM12878. ChIP-seq tracks show fold change over control and optimal IDR peak calls. We also show H3K27Ac (7 cell lines from ENCODE) and DNase 1 hypersensitivity peak clusters from ENCODE (95 cell types). (B) A 74 bp sequence with motifs for NFkB1, CUX1, and STAT3 is only present in the promoters of *CXCL1*, *CXCL2*, *CXCL3*, and *CXCLIP1*. We used the following human genomic sequence of 74 nucleotides (hg38 chr4:74038713-74038786 strand=+) as a query with NCBI BLAST:

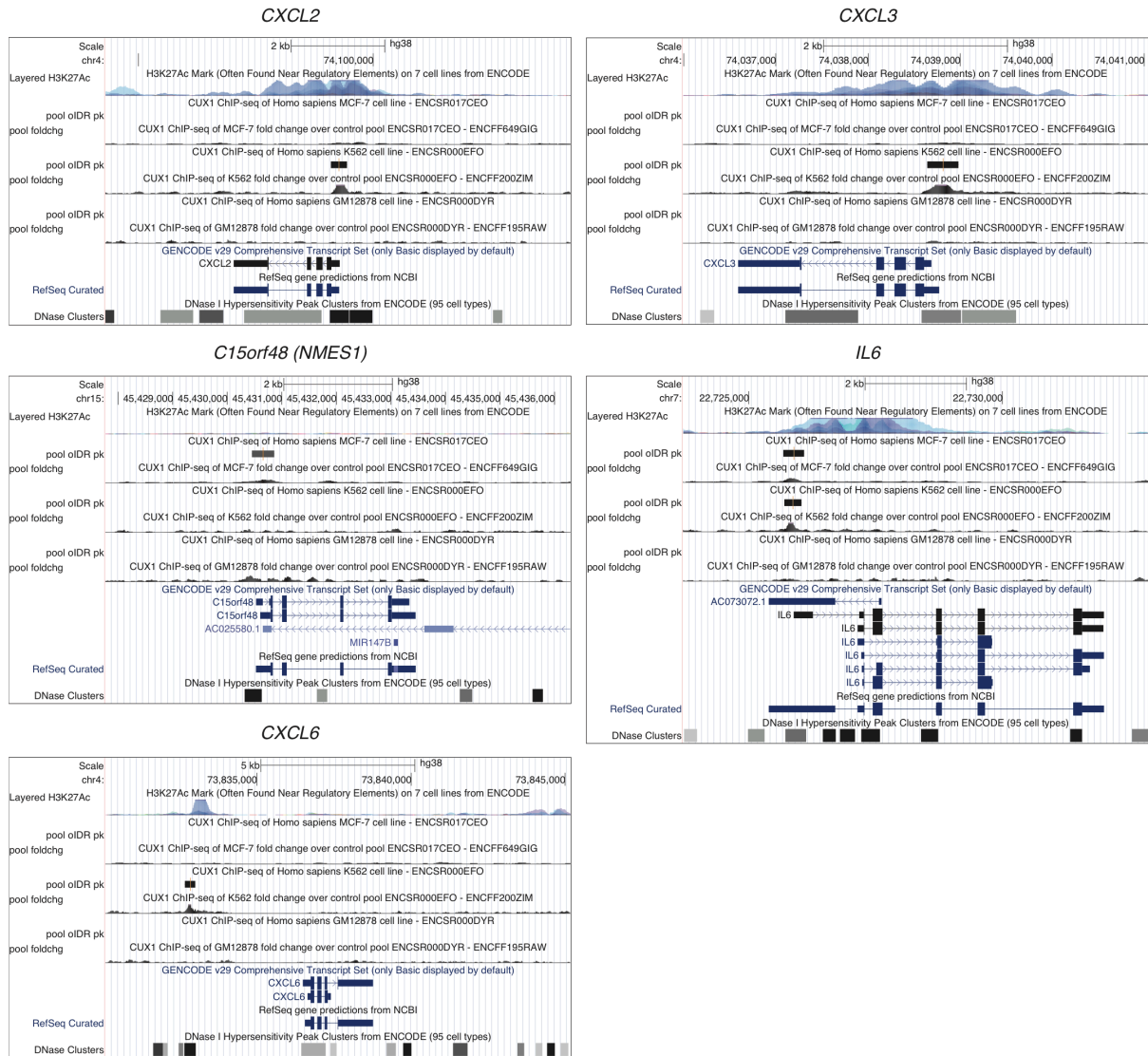
```
GGATCGATCTGGAGCTCCGGAATTTCCCTGGCCCGGCCGCTCCGGGCTTTCCAGTCTCAACCATGCAT
AAAAA
```

We queried against “Human genomic plus transcript (Human G + T)” and found 7 hits. 3 hits were on transcripts (*CXCL3* mRNA, *CXCL2* mRNA, and uncharacterized LOC105377275). 4 hits were on genomic sequences, shown in this figure to be located at the promoters of *CXCL1*, *CXCL2*, *CXCL3*, and pseudogene *CXCLIP1*.

Figure 6.8 (Continued)

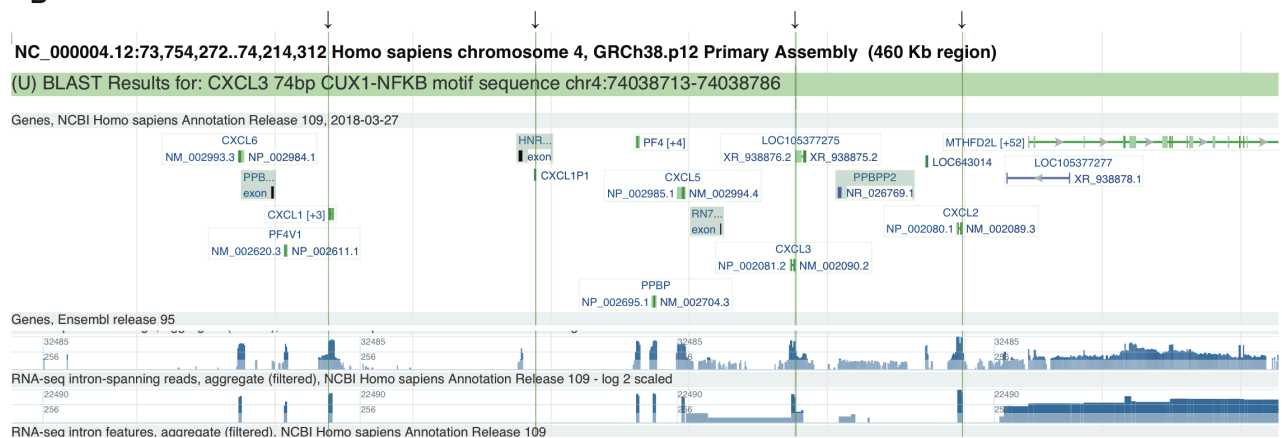
A

CUX1 ChIP-seq (ENCODE)



B

BLAST Hits for 74bp sequence containing CUX1 and NFKB motifs



Transcriptomics time series after silencing putative mediators

We selected five genes encoding putative mediators of IL-17A transcriptional effects (*ELF3*, *CUX1*, *LIFR*, *STAT3*, *STAT4*), silenced them with short interfering RNAs (siRNAs), and assayed global gene expression with RNA-seq. We used qPCR to confirm that each of our selected siRNAs silences its intended target gene (**Figure 6.9A**). After incubating 4 RA synovial fibroblast cell lines with a siRNA for 2 days, we stimulated them with TNF (0.1 ng/mL) alone or TNF and IL-17A (1 ng/mL) and assayed the transcriptomes at four time points (0, 1, 6, and 16 hours) (**Figure 6.10A**). This experimental design let us confirm that these putative mediators are acting downstream of IL-17A, and allows us to directly compare their transcriptional effects.

When we examined genes in Categories C1 and C2, we found two distinct groups of genes with different responses to gene silencing (**Figure 6.10B**). One group of genes was more repressed by silencing *CUX1*, and the other group was more repressed by silencing *LIFR*. Consistent with our motif analysis, silencing *CUX1* repressed Category C2 genes *CXCL1*, *CXCL2*, *CXCL3*, *CXCL6*, *CXCL8* (mean FC 0.60, 95% CI: 0.43-0.84). In contrast, these five genes were not strongly affected by silencing *ELF3*, *LIFR*, *STAT3*, or *STAT4* (mean FC 0.92, 95% CI: 0.66-1.3) (**Figure 6.10B**). We validated this result by using qPCR to measure selected genes in the same cDNA libraries (**Figure 6.9B**). These data suggest at least two independent regulatory pathways for genes induced by costimulation with TNF and IL-17A.

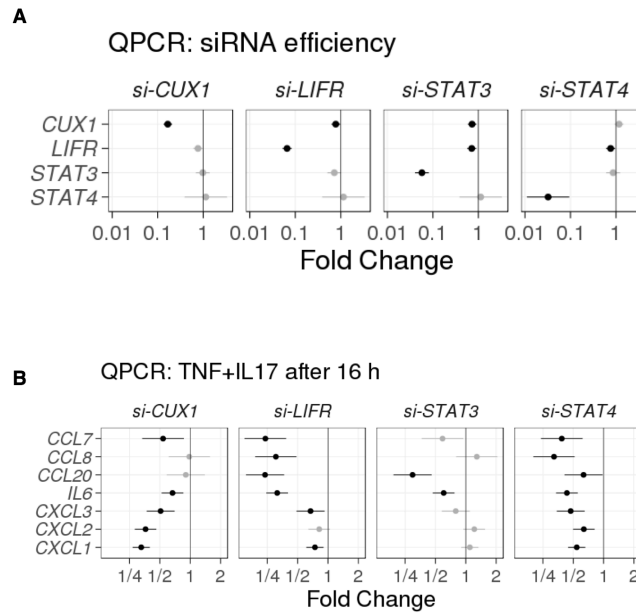


Figure 6.9. Efficiency of silencing each mediator with siRNA, and effects on selected genes.

(A) *CUX1*, *LIFR*, *STAT3*, and *STAT4* were effectively silenced by siRNA. We cultured 4 RA synovial fibroblast cell lines and measured silencing efficiency in each one. For *CUX1*, *LIFR* and *STAT3* siRNA, the silencing efficiency was measured at basal conditions. For *STAT4* siRNA, it was quantified 16 hours after costimulation with TNF and IL-17A. (B) qPCR analysis of the cDNA libraries prepared for RNA-seq. We show results for *CCL7*, *CCL8*, *CCL20*, *IL6*, *CXCL3*, *CXCL2*, *CXCL1*. After incubating 4 cell lines with siRNAs for *CUX1*, *LIFR*, *STAT3*, or *STAT4*, we measured the selected genes with qPCR relative to *GAPDH*. Transcription of *CXCL1*, *CXCL2*, and *CXCL3* is repressed after silencing *CUX1*, but not as much after silencing *LIFR*, *STAT3*, or *STAT4*. Transcription of *CCL7*, *CCL8*, *CCL20*, and *IL6* is repressed after silencing *LIFR*, but not as much after silencing *CUX1*. Dots represent mean fold change relative to control siRNA, and error bars represent 95% confidence intervals. Black color indicates FDR < 0.05.

We downloaded the JASPAR2018 database of genome-wide transcription factor binding motif matches¹⁸⁷ to look more closely at gene promoters, and we found a unique CUX1-NF- κ B nucleotide sequence motif in the promoters of *CXCL1*, *CXCL2*, and *CXCL3* (Figure 6.10C). Next, we ran a genome-wide BLAST search for the 74 bp sequence in the promoter of *CXCL3*

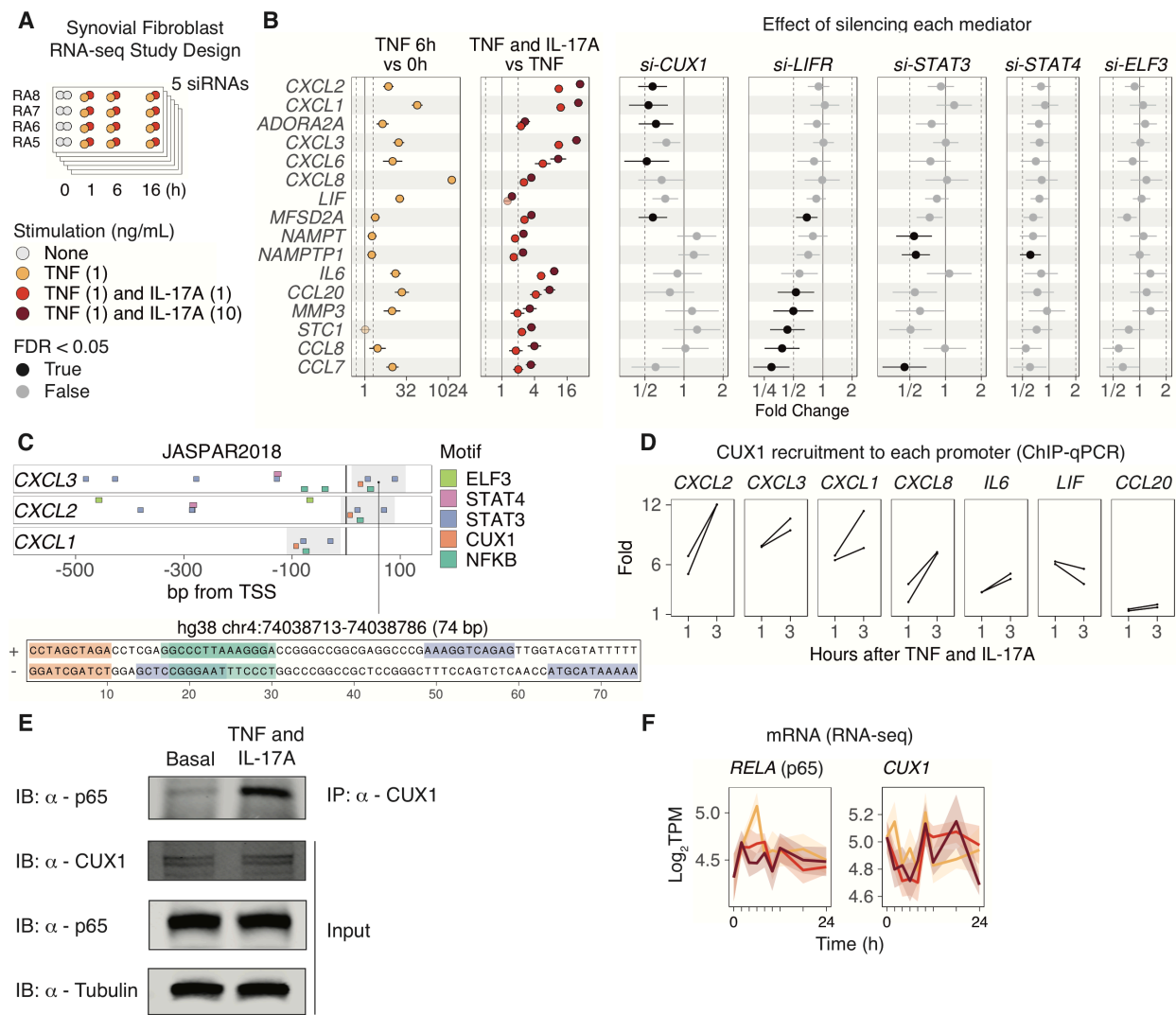
(hg38 chr4:74038713-74038786) containing the CUX1 and NF- κ B motifs. This sequence is present only in four genomic locations: the promoters of *CXCL1*, *CXCL2*, *CXCL3*, and pseudogene *CXCLIP1* (**Figure 6.8B**). Next, we used ChIP-qPCR to test if CUX1 binds selected promoters after stimulation with TNF and IL-17A. Indeed, CUX1 recruitment to the promoters of *CXCL1*, *CXCL2*, *CXCL3* increases more than 10-fold, on average, after 3 hours of stimulation relative to baseline (**Figure 6.10D**). Since CUX1 binds the promoters of genes with the CUX1-NF- κ B motif, we hypothesized that CUX1 might bind NF- κ B. Using an antibody against CUX1, we detected an increased amount of NF- κ B p65 in the pull-down fraction after stimulating fibroblasts with TNF and IL-17A (**Figure 6.10E**). Since protein and mRNA levels of RELA and CUX1 remain relatively constant over time (**Figure 6.10E,F**), our data suggests that CUX1 and NF- κ B form a molecular complex after costimulation with TNF and IL-17A. This complex likely binds the promoters of CXC chemokines and regulates transcription of these genes in synovial fibroblasts.

In addition to the group of genes repressed by silencing *CUX1*, a second group of genes was repressed after silencing *LIFR*. This includes the genes encoding C-C type chemokines encoded by *CCL8*, *CCL20*, as well as *MMP3* (Matrix metalloproteinase-3) and *STC1* (Stanniocalcin-1) (mean FC 0.34, 95% CI: 0.22-0.51, **Figure 6.10B**). These four genes were unaffected after silencing *CUX1* (mean FC 1.0, 95% CI: 0.67-1.6), suggesting they are not mediated by CUX1. CCL8 (MCP-2) is a chemoattractant and activator of monocytes, whereas CCL20 is a chemoattractant for lymphocytes and is known to recruit CCR6⁺ Th17 cells to sites of inflammation¹⁸⁸. Our data suggests that two independent pathways activate in response to costimulation with TNF and IL-17A, and chemoattraction of leukocytes is a primary function of this response in synovial fibroblasts.

Figure 6.10. CUX1 mediates transcription of CXCL1, CXCL2, CXCL3, independent of LIFR, STAT3, STAT4, and ELF3.

(A) Our experimental design included 4 synovial fibroblast cell lines from patient tissues, 4 time points (0, 1, 6 and 16 hours), 2 stimulations (TNF (0.1 ng/mL); TNF (0.1 ng/mL) and IL-17A (1 ng/mL)), and 5 siRNAs (*CUX1*, *LIFR*, *STAT3*, *STAT4*, *ELF3*). (B) Fold change estimates and 95% confidence intervals for 16 select genes that were measured by ChIP-qPCR, or that were responsive to IL-17A and had at least one significant siRNA effect. Left 2 panels show effect estimates from RNA-seq Data 1, and right 5 panels show effect estimates from RNA-seq Data 2. Dark color indicates FDR < 0.05. (C) The promoters of *CXCL1*, *CXCL2*, and *CXCL3* have a unique 74 bp sequence that contains motifs for CUX1, NF- κ B, and STAT3. (D) ChIP-qPCR of the CUX1 protein to 7 gene promoters at 3 time points (0, 1, and 3 hours) after costimulation with TNF and IL-17A. The y-axis shows fold recruitment relative to the first time point at 0 hours, shown for 2 technical replicates. (E) After 4 hours of stimulation with TNF (1 ng/mL) and IL-17A (1 ng/mL), NF- κ B p65 is detected in a pull down of CUX1. (F) The mRNA levels of *RELA* (p65) and *CUX1* are insensitive to IL-17A, and are relatively stable over time.

Figure 6.10 (Continued)



I κ B ζ binds NF- κ B p65 and p100/52, and mediates the response to TNF and IL-17A

Since NF- κ B is a known mediator of inflammatory pathways, and *NFKBIZ* transcriptional response is driven by IL-17A dose (**Figure 6.11A**), we tested the effects of silencing *NFKBIZ*, the gene encoding I κ B ζ . We measured protein levels of IL-6, IL-8 (CXCL8), and MMP-3 after stimulating synovial fibroblasts with TNF alone or with the combination of TNF and IL-17A. Silencing *NFKBIZ* reduced protein levels of IL-6, IL-8, and MMP-3 after TNF and IL-17A, but not after TNF alone (**Figure 6.11B**). Next, we tested if I κ B ζ binds NF- κ B to modulate the transcription of its downstream targets in fibroblasts. Using an antibody against NF- κ B p65, we detected increased levels of I κ B ζ in the pull-down fraction after stimulating fibroblasts with TNF and IL-17A relative to unstimulated fibroblasts (**Figure 6.11C**). Furthermore, using an antibody against NF- κ B p100/52 showed a similar result, indicating that I κ B ζ forms a complex with p65 and p100/52 (**Figure 6.11D**).

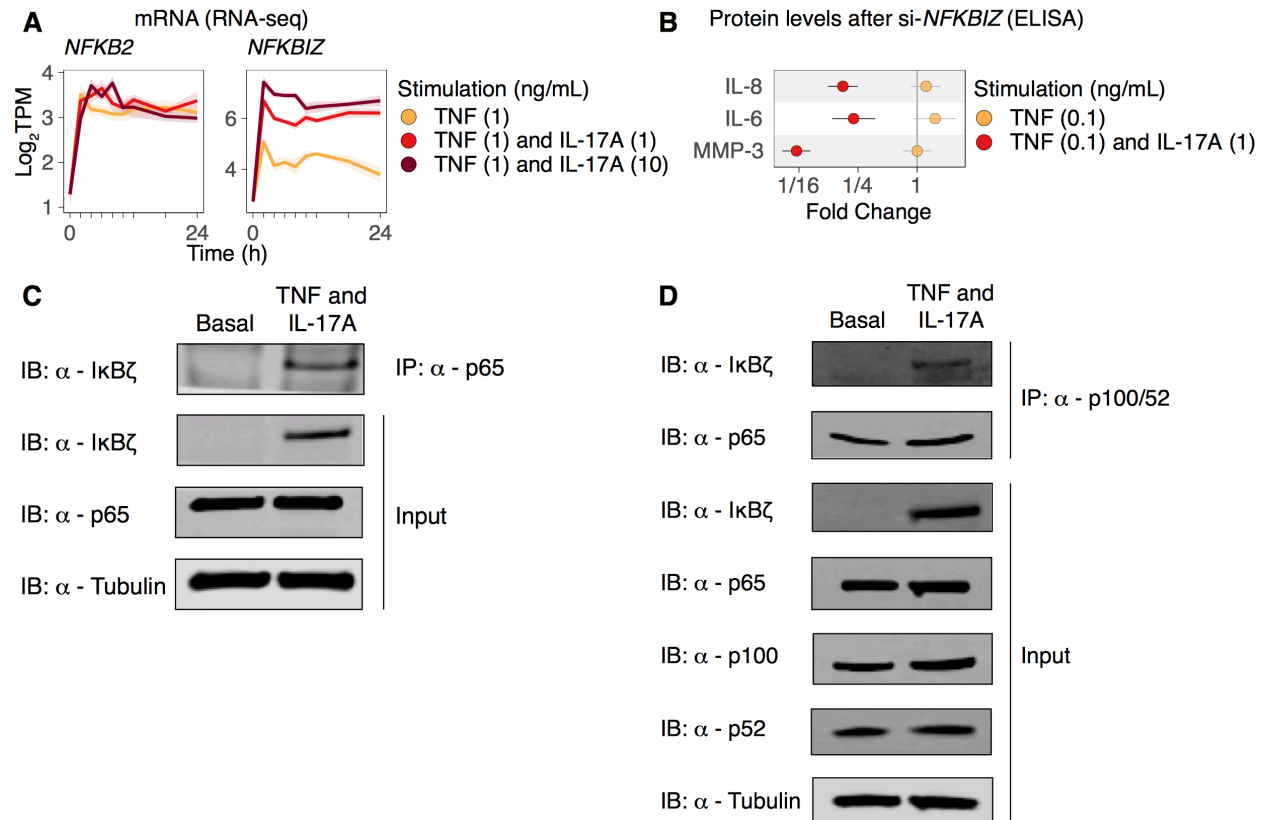


Figure 6.11. IκBζ binds NF-κB and p100/52, and mediates the response to TNF and IL-17A.

(A) Gene expression of *NFKB2* (p100) and *NFKBIZ* (IκBζ) over time and IL-17A dosage in RNA-seq Data 1. (B) Silencing *NFKBIZ* reduces protein levels of MMP-3, IL-6, and IL-8 after costimulation with TNF (0.1 ng/mL) and IL-17A (1 ng/mL), but not after stimulation with TNF alone. (C) After 4 hours of costimulation with TNF (1 ng/mL) and IL-17A (1 ng/mL), IκBζ is detected in a pull down of NF-κB p65. (D) After 4 hours of costimulation with TNF (1 ng/mL) and IL-17A (1 ng/mL), IκBζ is detected in a pull down of NF-κB p100/52.

CUX1 and IκBζ mediate neutrophil and monocyte recruitment

Neutrophils are the most abundant cells in the synovial fluid of inflamed joints in RA^{189,190}, and abundance of synovial macrophages correlates with radiologic score of joint damage¹¹⁰ and

disease activity¹⁹¹. So, we performed *in vitro* assays to test which mediators are required for recruitment of neutrophils and monocytes. Briefly, we stimulated synovial fibroblast cell lines with TNF and IL-17A for 24 hours after silencing *CUX1*, *LIFR*, *STAT3*, *STAT4*, or *NFKBIZ*. Then, we tested the supernatants in a transwell migration experiment (**Figure 6.12A, Methods**).

After silencing *CUX1* or *LIFR*, the supernatant from synovial fibroblasts was less effective at inducing migration of neutrophils (t-test $P = 0.05$ and 0.06 , respectively) (**Figure 6.12B**). The number of migrated cells is proportional to protein level of IL-8 (CXCL8) in the supernatant, and silencing *NFKBIZ* had the greatest effect. While monocyte migration was reduced after silencing of *CUX1* or *NFKBIZ* (t-test $P = 0.04$ and 0.06 , respectively) (**Figure 6.12C**), silencing *STAT4* had the greatest effect on monocyte migration and CCL2 protein level. We conclude that CUX1 and $\text{I}\kappa\text{B}\zeta$ mediate recruitment of neutrophils and monocytes.

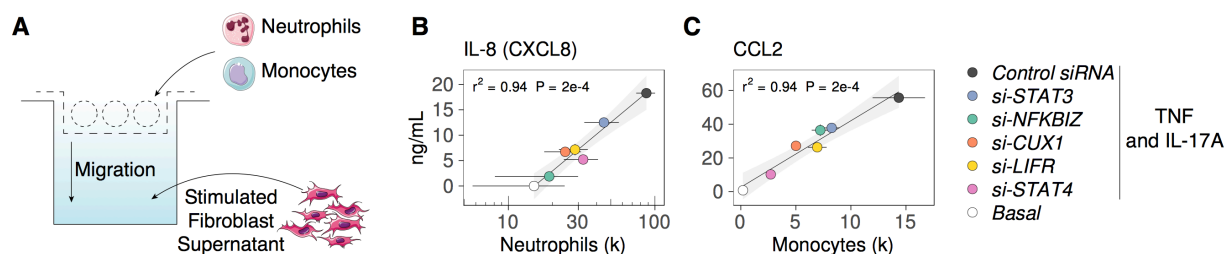


Figure 6.12. CUX1 and $\text{I}\kappa\text{B}\zeta$ mediate neutrophil and monocyte recruitment.

(A) We tested migration and protein levels in supernatant from synovial fibroblasts before stimulation (Basal) and 24 hours after stimulation with TNF and IL-17A. (B) Silencing *CUX1* or *NFKBIZ* ($\text{I}\kappa\text{B}\zeta$) reduces neutrophil recruitment in an *in vitro* migration assay. Neutrophil migration is correlated with protein level of IL-8 in the supernatant measured by ELISA. (C) Silencing *CUX1* or *STAT4* reduces monocyte recruitment in an *in vitro* migration assay. Monocyte migration is correlated with protein level of CCL2 in the supernatant measured by ELISA.

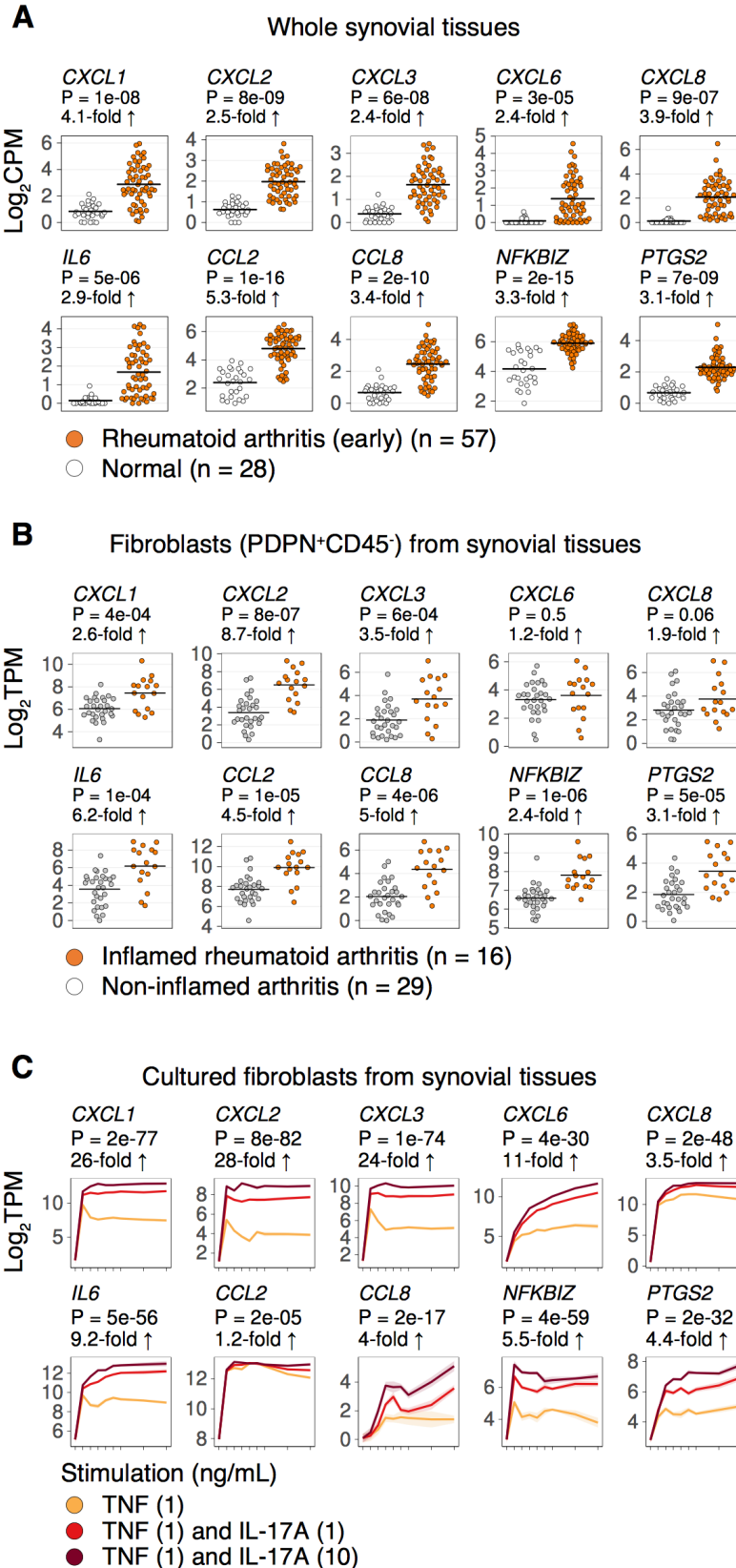
Genes induced by TNF and IL-17A are highly expressed in RA synovial tissues

To test if genes induced by TNF or TNF and IL-17A are highly expressed *in vivo*, we examined public RNA-seq data from two large-scale transcriptomics studies that include inflamed RA synovial tissues^{45,50}. We focused on the 51 genes in Categories C1 or C2 and tested for differential gene expression in each study. First, we compared 57 synovial tissues (all cell types) from early RA versus 28 normal tissues⁴⁵ (**Figure 6.13A**). 14 of 23 genes in Category C1 and 20 of 25 genes in Category C2 were higher in early RA (FDR < 0.05). Since this study gave us a glimpse of the transcriptomic profile of the whole tissue, we next looked only at the contribution from fibroblasts. In the second study, sorted CD45⁺PDPN⁺ fibroblasts were isolated from synovial tissues, and here we compare 16 samples from inflamed RA tissues versus 29 from non-inflamed tissues (lower abundance of infiltrating leukocytes)⁵⁰ (**Figure 6.13B**). 2 of 26 genes in Category C1 (*LYPD3* and *NR4A2*) and 9 of 25 genes in Category C2 (including *CXCL1*, *CXCL2*, *CXCL3*, *NFKBIZ*) were higher in fibroblasts from inflamed RA than non-inflamed arthritis tissues (FDR < 0.05). This suggests that fibroblasts may be an important source of chemokines encoded by *CXCL1*, *CXCL2*, *CXCL3*, *CXCL6*, *CXCL8*, *IL6*, *CCL2*, *CCL8*. This highlights the relevance of CUX1 and IκBζ mediated transcription in inflamed human tissue.

Figure 6.13. Genes induced by TNF and IL-17A are highly expressed in RA synovial tissues.

Transcript abundances from (A) RNA-seq data for whole synovial tissues or (B) RNA-seq data for sorted synovial fibroblasts (CD45⁻PDPN⁺) from synovial tissues. Each dot represents one sample. Fold-changes and t-test p-values. (C) Transcript abundances from RNA-seq Data 1, showing abundance over time and IL-17A dose. Fold-changes and t-test p-values for TNF (1 ng/mL) and IL-17A (10 ng/mL) versus TNF (1 ng/mL).

Figure 6.13 (Continued)



Conclusion

We provide a genome-wide view of transcriptional responses over time to the combination of TNF and IL-17A and to TNF alone. Since IL-17A by itself has weak transcriptional effects, we used time series transcriptomics after stimulation with TNF and three dosages of IL-17A (0, 1, 10 ng/mL) across seven human donor derived cell lines. Multiple dosages of IL-17A allowed us to confidently identify genes with transcriptional abundance proportional to IL-17A dosage. We defined categories of genes by looking at their transcriptional responses to TNF alone and costimulation with TNF and IL-17A. For example, many genes known to be transcriptionally induced by interferon alpha and interferon gamma are induced by TNF alone and repressed by the addition of IL-17A. By querying public RNA-seq data, we also showed that many genes induced by TNF and IL-17A are more highly expressed in inflamed RA synovial tissues, and in particular in synovial fibroblasts from RA joint tissues (**Figure 6.13**).

TNF is known to induce fibroblasts to produce many inflammatory cytokines, and IL-17A has been shown to synergize with TNF to further induce key cytokines such as IL-6^{169–172}, neutrophil attracting chemokines^{162,173}, and matrix metalloproteinases^{167,174}. Two transcription factors, NFκB and C/EBP, control synergistic production of IL-6^{167,171} in response to the combination of TNF and IL-17A.

Among the list of IL-17A dose-responsive genes, we saw known IL-17A mediators with dose-responsive transcription such as *NFKBIZ*¹⁹² and *ZC3H12A*¹⁹³. We used motif enrichment analysis to discover a unique CUX1-NF-κB motif and predicted that CUX1 mediates transcription of CXC chemokines and that CUX1 might form a complex with NFκB. We validated this prediction by showing that silencing CUX1 with siRNA represses transcription of

the genes with the CUX1-NF- κ B motif (*CXCL1*, *CXCL2*, *CXCL3*). Moreover, CUX1 binds NF- κ B p65 after costimulation with TNF and IL-17A. We found increased CUX1 binding to the promoters of *CXCL1*, *CXCL2*, *CXCL3* after stimulation with TNF and IL-17A, suggesting that a CUX1 NF- κ B complex directly regulates transcription of these chemokines. Our data suggests that CUX1 mediated transcription is independent of LIFR, STAT3, STAT4, and ELF3, because silencing these mediators did not repress transcription of *CXCL1*, *CXCL2*, and *CXCL3*.

CUX1 is a homeobox transcription factor with ubiquitous expression across mammals and essential roles in development and homeostasis¹⁹⁴, but its role in inflammation is not well understood. Previously, in tumor associated macrophages, CUX1 was found to bind NF- κ B p65 and suppress chemokines *CXCL10* and *CCL5* transcription¹⁹⁵. Here, in fibroblasts, we find that the CUX1 NF- κ B p65 complex induces chemokine transcription, suggesting cell type-specific regulation. In future studies, we might expect to find that additional homeobox transcription factors work with NF- κ B to regulate inflammation, with different cell-type specific effects.

Previously, I κ B ζ , a unique member of the I κ B family with a distinct role as a coactivator of NF- κ B, was shown to be induced in macrophages after stimulation with a number of different toll-like receptor (TLR) ligands and IL-1 β , but not after stimulation with TNF alone¹⁹⁶. I κ B ζ regulates transcription of a specific set of genes such as *Csf3* in mouse embryonic fibroblasts¹⁹⁷. Here, we found that I κ B ζ is one of the critical mediators for the synergistic response to TNF and IL-17A in fibroblasts. By silencing *NFKBIZ*, the gene encoding I κ B ζ , production of IL6 and MMP3 was reduced after costimulation with TNF and IL-17A, but not after stimulation with TNF alone. We also found that synovial fibroblasts require I κ B ζ for transcription of *CXCL8*, for production of the encoded protein IL-8, and for recruitment of neutrophils *in vitro*. Thus, I κ B ζ has a critical role in driving inflammation after activation of fibroblasts.

IL-17A is an important cytokine for pathogenesis in autoimmune and inflammatory diseases ^{198,199} such as psoriasis, psoriatic arthritis, ankylosing spondylitis, rheumatoid arthritis (RA) ¹⁷⁵, inflammatory bowel disease (IBD), and multiple sclerosis ²⁰⁰. Two antibodies targeting IL-17A have shown excellent results for treatment of psoriasis: secukinumab is approved for psoriasis, psoriatic arthritis, and ankylosing spondylitis ²⁰¹, and ixekizumab is approved for psoriasis ²⁰². One might have expected similar results for RA, since IL-17A knockout suppresses arthritis in a collagen-induced arthritis mouse model ²⁰³, and mouse models have shown IL-17A involvement in all stages of rheumatoid arthritis ¹⁶¹. However, clinical trials of IL-17A inhibitors with RA patients have shown mixed results ²⁰⁴. There are several possible explanations for this discrepancy. One possibility we might expect is that the heterogeneity in clinical response to IL-17A inhibition could be partially explained by genetic variation in the components of regulatory pathways downstream of TNF and IL-17A.

Materials and Methods

Cell Lines and Reagents. Human synovial fibroblasts were isolated as previously described ⁹⁴ from tissues discarded after synovectomy or joint replacement surgery. Each line was derived from a unique donor and used experimentally between passages 5 and 8. Cells were maintained in Dulbecco's modified Eagle's medium (DMEM, Gibco) supplemented with 10% fetal bovine serum (FBS; Gemini), 2 mM L-glutamine, 50 μ M 2-mercaptoethanol, antibiotics (penicillin and streptomycin), and essential and nonessential amino acids (Life Technologies). The following antibodies were used: anti-CUX1 (EMD Millipore); anti-p100/p52, anti-I κ B ζ (Cell Signaling Technology); anti-p65 (Santa-Cruz); anti- β -tubulin (Sigma); anti-CD45, anti-CD11b, anti-

CD66b, (BioLegend); anti-CD14 (eBioscience). Other reagents were purchased from the following vendors: IL-6, IL-8, CCL2 and MMP3 ELISA kits, TNF, IL-17A (R&D Systems). All siRNAs (Silencer Select) were purchased from Life Technologies.

Immunoprecipitation. Total cell lysates were collected by washing cells once with cold PBS followed by addition of lysate buffer (50mM HEPES pH 7.5, 5% glycerol, 100 mM NaCl, 0.25% TritonX-100, supplemented with a protease inhibitor cocktail (Roche) and phosphatase inhibitors (sodium orthovanadate, sodium fluoride, and beta-glycerol phosphate). Cells were lysed for 30 minutes on ice followed by centrifugation at 15000rpm for 15 minutes at 4°C. Immunoprecipitation was done overnight at 4°C using Protein A resins together with a rabbit antibody against the relevant target.

Quantitative Real-Time PCR. mRNA samples were extracted from cells using an RNeasy Micro Kit (Qiagen). cDNA synthesis was carried out using the QuantiTect Reverse Transcription Kit (Qiagen). qPCR reactions were performed in duplicate using the Brilliant III Ultra-Fast SYBR reagent (Agilent). Relative transcription level was calculated by using the $\Delta\Delta C_t$ method with GAPDH as the normalization control. Fold induction was calculated by dividing the normalized mRNA at a certain time point with that at time 0hr. Fold change is the ratio of the normalized mRNA from cells expressing a specific siRNA vs. Ctrl siRNA.

Chromatin Immunoprecipitation Assays. Fibroblasts were stimulated with TNF (1 ng/mL) + IL-17A (1 ng/mL) for different durations of time as indicated followed by fixation with formaldehyde (1% final concentration) for 10 minutes. Cells were lysed in swelling buffer for 15

minutes (25mM HEPES pH 7.8, 1.5 mM MgCl₂, 10mM KCl, 0.1% NP-40, 1mM DTT, protease inhibitor cocktail) then in sonication buffer (50mM HEPES pH 7.8, 140mM NaCl, 1mM EDTA, 0.1% SDS, 1% TritonX-100 supplemented with protease inhibitor cocktail). Samples were sonicated for 2 minutes with 20-second pulse intervals and 1 minute off at each interval at 4°C using a Qsonica sonicator set at 28% output. Chromatin was immunoprecipitated using Dynabead Protein A resins (Life Technologies) together with a rabbit antibody against the relevant target at 4°C overnight. DNA was purified using the phenol/chloroform precipitation method. The amount of DNA precipitated was quantified using qPCR. Fold recruitment was calculated by dividing the amount of ChIP-ed DNA at indicated time point (normalized with input DNA) with that at 0hr (normalized with input DNA).

Western Blotting. Total cell lysates were collected by washing cells once with cold PBS followed by addition of lysate buffer (50mM HEPES pH 7.5, 10% glycerol, 100 mM NaCl, 0.1% SDS, 1% NP-40, supplemented with protease inhibitors and phosphatase inhibitors sodium orthovanadate, sodium fluoride, and beta-glycerol phosphate). Cells were lysed for 30 minutes on ice followed by centrifugation at 15000rpm for 15 minutes at 4°C. Protein concentration was measured by the microBCA kit (Pierce). Equal amounts of total protein (~20 µg per lane) were separated on an 8% SDS-PAGE gel. Proteins were transferred onto a PVDF membrane and blocked with 5% BSA in PBS and probed with primary antibodies overnight at 4°C, followed by secondary antibodies conjugated with IRDye 680 or 800 (Rockland). Membranes were scanned with a Li-COR Odyssey scanner.

Leukocyte Migration. Fibroblasts were transfected with a control siRNA, or siRNA against a gene by reverse transfection using the RNAiMax reagent (Life Technologies) on day 1. The following day, cells were switched to 0.5% serum media and after 24 hours they were stimulated with TNF (0.5 ng/mL) + IL-17A (0.5ng/mL) for another 24 hours before supernatants were collected and added to the bottom well of a 24-well plate. Neutrophils (PMNs) purified from blood using the neutrophil isolation kit (Stem Cell Technologies) and peripheral blood mononuclear cells (PBMC) containing monocytes purified from blood using Ficoll isolation technique were then introduced to the top well (Corning, 3µm pores insert) at the concentration of 1 million 1×10^6 cells per mL for PMNs and 2×10^6 cells per mL for PBMC. After 75 mins (PMN migration) or 3 hours (monocyte migration), cells from the bottom well were collected and the number of neutrophils, monocytes recruited toward the bottom well were quantified by flow cytometry using (anti-CD66b and anti-CD11b for neutrophils), (anti-CD14 and anti-CD45 for monocytes), and cell numbers were normalized using counting beads (Spherotech).

RNA-seq Expression Profiling. Fibroblasts were plated on day 1 at 50,000 cells per well in 24-well plates in 10% FBS containing media. Cells were serum-starved on day 2 by changing to 1% FBS-containing media. Cells were either left unstimulated or stimulated with TNF (1 ng/mL) or TNF (1 ng/mL) + IL-17A (1 ng/mL) for various durations of time. Samples were collected and RNA was extracted using the RNeasy Micro Kit (Qiagen). RNA-seq libraries were prepared with the Smart-Seq2 protocol.

siRNA Silencing. Fibroblasts were transfected with an siRNA by reverse transfection at 30 nM using the RNAi Max reagent (Life Technologies) in 10% FBS containing media. Cells were then

switched to serum-starving media containing 1% FBS on day 2. Cells were stimulated as indicated on day 3. Efficiency of siRNA silencing was assessed by qPCR (**Figure 6.9**).

Data Visualization. All of the figures were created with R (<https://www.R-project.org>) and many supporting packages, including: ggplot2, ggrepel, patchwork, RColorBrewer, seriation, viridis.

Microarray Data Analysis. We normalized the data from 60 microarrays with the robust multi-array averaging (RMA) algorithm²⁰⁵ from the oligo R package²⁰⁶, downloaded from Bioconductor²⁰⁷. To assess quality of the microarray data, we looked at the Normalized Unscaled Standard-Errors (NUSE) and Relative Log-Expression (RLE), and decided to keep all microarrays in our analysis.

We used the BioMart R package^{208,209} (from Bioconductor) to download gene names and identifiers for all of the oligonucleotide probe identifiers that match the array name "affy_hugene_1_0_st_v1". We averaged all of the probes for each gene.

We assess the quality of expression data for each array as follows. For each array, we computed the median M RMA expression value across all genes. For each gene, we computed the percent P of microarrays with expression greater than M . We defined a set of 7,891 genes with $P > 90\%$ and call them "common genes". In other words, the common genes had high expression in most microarrays. For each microarray, we computed how many of the common genes had expression greater than M . This is the percent of common genes detected for each microarray. All of the 60 microarrays detected 97% or more of the common genes, so we decided to keep all of them in our analysis.

We used the limma R package ¹⁰⁴ to fit a linear model to each gene and test for differential gene expression across contrasts of interest. For Microarray Data 1, we fit the following model:

$$GeneExpression \sim TimeFactor + TreatmentFactor$$

TimeFactor represents 7 indicator variables, one for each time point (we excluded $t = 0$). *TreatmentFactor* represents 3 indicator variables, one for each treatment: (1) IL-17A, (2) TNF, and (3) TNF and IL-17A. We illustrate the experimental design in **Figure 6.3A**.

For Microarray Data 2, we fit the following model:

$$GeneExpression \sim TreatmentFactor$$

TreatmentFactor represents 3 indicator variables, one for each treatment after 24 hours: (1) media, (2) TNF, and (3) TNF and IL-17A. We illustrate the experimental design in **Figure 6.4B**.

RNA-seq Data Analysis. We used kallisto (version 0.43.1) ¹⁰³ to quantify transcripts per million (TPM) for all of the transcripts reported in Ensembl release 89 ²¹⁰. To get TPM for each gene, we added the TPM values for all of the transcripts that belong to a gene.

We assessed the quality of each RNA-seq sample with a similar approach to the microarray data analysis. For each gene, we computed the percent P of 175 samples with expression greater than 0. We defined a set of 12,087 genes with $P > 90\%$ and call these the "common genes". These are genes for which expression was detected in most samples. For each

sample, we computed how many of the common genes had expression greater than 0. This is the percent of common genes detected for each sample. All of the 175 samples detected 98% or more of the common genes, so we kept all of them in our analysis.

We used the limma R package ¹⁰⁴ to fit a linear model to each gene and test for differential gene expression across contrasts of interest. For RNA-seq Data 1, we fit the following model:

$$GeneExpression \sim TimeFactor + DoseFactor$$

TimeFactor represents 8 indicator variables: one for each time point after stimulation with TNF. *DoseFactor* represents 2 indicator variables: one for TNF and IL-17A (1 ng/mL), and one for TNF and IL-17A (10 ng/mL).

We used the CAMERA algorithm from the LIMMA package to compute gene set enrichment p-values with gene sets from MSigDB ⁸⁶.

We used HOMER ¹⁸⁵ to find transcription factor binding motifs that are overrepresented in promoter sequences. First, we selected 26 genes with at least 1.5 log₂ fold-change (2.83 fold-change) greater expression for 10 versus 1 ng/mL IL-17A at 5% false-discovery rate (FDR). We ran the findMotifs.pl script with the Ensembl gene identifiers for these 26 genes as input to find motifs overrepresented in their promoters relative to the background of all other genes. HOMER accepted 25 of the 26 genes for analysis.

We assessed the power gained in RNA-seq Data 1 by comparing results from analysis of the full data with 9 time points (0, 2, 4, 6, 8, 10, 12, 18, 24 hours) and results from analysis of a subset of the data with 2 time points (0, 24 hours). In the full analysis and the subset analysis, we

fit two linear models (reduced model and full model) and computed the F statistic to test how many genes have a significantly better fit with the full model at $FDR < 0.05$.

Reduced model: $GeneExpression \sim TimeFactor$

Full model: $GeneExpression \sim TimeFactor + DoseFactor$

We found 253 genes were significantly better fit by the full model than the reduced model when we used 9 time points. In contrast, we found 144 genes when we used 2 time points.

Chapter 7

Conclusion

Summary of results

We created SNSPEA, a fast C++ implementation of a robust algorithm for integrating GWAS results with any quantitative or binary gene matrix to answer the question of which cell type or gene set in the matrix is significantly associated with the trait from the GWAS. We showed that the statistical procedure is well calibrated, with appropriate false positive rates (type I error) (**Figure 3.3**) and is sensitive to detect true positive associations such as the genetic association for HDL cholesterol and gene expression in the liver (**Figure 3.6**).

We made an R package called ggrepel that automatically positions text labels on a figure to identify each data point. The package implements a physical simulation with attraction and repulsion between objects. We show that this is an effective way to layout text labels on a figure. The package has is very flexible and can accommodate many different because it extends the popular ggplot2 package by Wickham *et al* ²¹¹.

We showed how R and the shiny package can be used to share functional genomics data such as RNA-seq on the web. To increase accessibility to this data and accelerate the process of discovery, we developed multiple different pages to view different data sets with bulk RNA-seq, single-cell RNA-seq, and mass cytometry data (**Figure 2.5** and **Figure 2.6**). As one example, we showed how investigators can use one of these pages to answer the question of which cell types and cell subtypes in the synovial tissue are likely sources of the inflammatory cytokine IL-6 (**Figure 2.7**).

We identified subsets of fibroblasts in fresh human synovium with distinct protein surface markers, transcriptomic profiles, and cytokine outputs⁴³. PDPN⁺CD34⁻THY1⁺ fibroblasts are expanded in RA tissues, and these cells might represent a pathogenic subset of fibroblasts in

the synovial tissue. These fibroblasts are mostly found around blood vessels in RA synovium, and their expression profile reveals potential pathogenic roles in lymphocyte recruitment and osteoclastogenesis. This discovery was refined in a comprehensive study of synovial tissue cell populations with single-cell RNA-seq, mass cytometry, bulk RNA-seq, and flow cytometry applied to 51 synovial tissue samples from patients with rheumatoid arthritis (RA) or osteoarthritis (OA) ⁵⁰. There, we found that the PDPN⁺CD34⁻THY1⁺HLA-DR⁺ fibroblasts were overabundant in RA tissues relative to OA (36% versus 2% of all fibroblasts measured by mass cytometry) ⁵⁰.

We provide high resolution, genome-wide transcriptional data for the synovial fibroblast response to the combination of TNF and IL-17A and to TNF alone, and we categorized genes by their transcriptional responses to TNF alone and costimulation with TNF and IL-17A. For example, many genes known to be transcriptionally induced by interferon alpha and interferon gamma are induced by TNF alone but repressed by the addition of IL-17A (**Figure 6.2**). The chemokine transcription in response to TNF and IL-17A is relevant *in vivo*, as evidenced by querying public RNA-seq data (**Figure 6.13**). Inflamed RA synovial tissues, and in particular in synovial fibroblasts from RA joint tissues, have higher expression of many of the genes induced by TNF and IL-17A.

We predicted and validated that CUX1 is a key mediator of CXC chemokine production in fibroblasts. By analyzing the presence of nucleotide sequence motifs in gene promoters, we nominated CUX1 as a putative regulator of the response to TNF and IL-17A (**Figure 6.10**). We found supporting evidence in the form of ChIP-seq read pileups at the promoters of *CXCL2* and *CXCL3*. We demonstrated that CUX1 binds these promoters with ChIP-qPCR. Silencing *CUX1*

caused a reduction in the transcription of the predicted CXC chemokines (**Figure 6.10**), as well as a reduction in the number of migrating neutrophils *in vitro* (**Figure 6.12**).

Assaying transcriptomic response to drugs will advance precision medicine

In the last few years, we've seen wide adoption of RNA-seq and other functional genomics data such as ATAC-seq. Several studies have assayed basal gene expression profiles in many tissues, and some studies have assayed gene expression across large samples of individuals. In the future, we will see studies focusing on the cellular response to perturbations across large samples of individuals. To maintain a healthy organism, all of its cells must respond to the multitude of stimuli around them in a way that leads towards health. Since the transcriptional profile of a cell depends on its context, studying drug responses is critical to understand how cells act in the context of disease and in the context of therapeutic treatment with drugs.

Several studies have assayed transcriptomics across large samples of individuals. For example, Lappalainen *et al* assayed mRNA and miRNA from lymphoblastoid cell lines of 462 individuals from the 1000 Genomes Project²¹². This project clearly demonstrated the existence of many genetic variants that regulate gene expression across the genome, and allowed the authors to infer putative causal variants for many genomic loci previously associated with risk for developing some disease. However, an obvious limitation of this study is that all expression was measured in a lymphoblastoid cell line, so it is not possible to extrapolate from this cellular context to the multitude of cellular contexts that are different across tissues in the organism. The Genotype-Tissue Expression (GTEx) project went further by assaying mRNA from more than 50 different tissues across hundreds of individuals²¹³. The authors report that, as the sample size for

a given tissue increases, so does their statistical power to identify tissue-specific expression quantitative trait loci (eQTLs)²¹⁴. Tissue-specific eQTLs are genetic variants that are only operational in the context of a specific tissue. Now that we have data on tissue-specific gene expression profiles, the next step forward is to collect data on cellular response to perturbations.

The Milieu Intérieur Consortium has already generated data on the transcription of 560 immune genes (measured with NanoString gene expression arrays) in response to infection with 6 different pathogens in a cohort of 1,000 healthy individuals²¹⁵. This study discovered many eQTLs that are specific to the cellular response to just one of the six pathogens. Further, the data helps to establish quantitative boundaries or distributions of what gene expression looks like in healthy individuals in response to common pathogens. More studies like these with larger sample sizes, whole genome sequencing, and genome-wide transcriptomics will help to understand how human tissues respond and regulate their responses to perturbagens.

Some pharmaceutical companies are starting to making use of genetic data to understand which patients may expect the greatest benefits from existing therapies, and to understand the mechanisms inside the cells by which therapies exert their functions. Unfortunately, genetic association studies on a binary trait such as disease cases (true) and healthy controls (false) are not likely to discover the variants that influence the response to a therapy. This is because the genetic effects that confer some risk to develop a disease are not guaranteed to be identical with the variants that affect cellular response to perturbagens such as pharmaceutical therapies. To use genetics in this context, it is necessary to assay the cellular response to a given therapy directly and then perform a genome-wide association study with that cellular response. Functional genomics data such as RNA-seq offers a promising link between the genetic variants and the phenotypic outcome of therapeutic response. By studying the transcriptional response to

therapeutic agents across a large number of individuals, we will begin to appreciate the distribution of different responses to these agents as well as the genetic and non-genetic mediators. When the technologies needed to execute these studies are eventually developed, optimized, and scaled up, we might open new opportunities for precision medicine, where each patient is treated differently based on data collected from their own body.

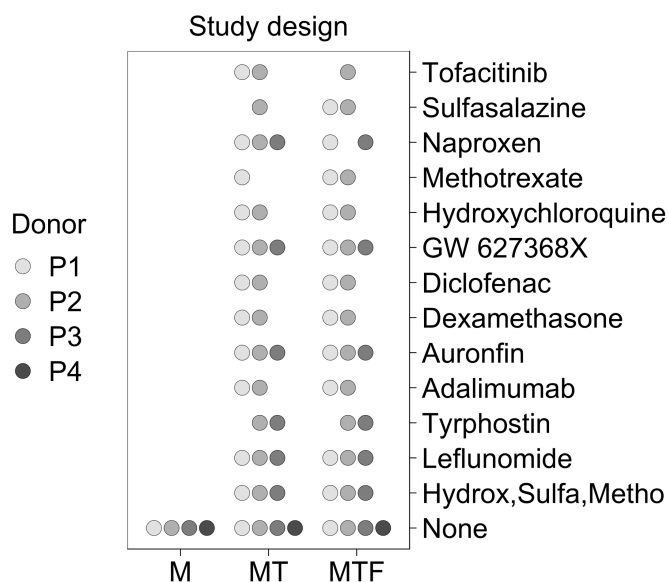


Figure 7.1 Macrophage stimulation study design by Kuo et al 2019.

Each dot represents an RNA-seq sample. Unstimulated macrophages (M), macrophages stimulated with TNF (MT), and co-cultured macrophages and fibroblasts stimulated with TNF (MTF). Stimulated macrophages were treated with 13 different drugs.

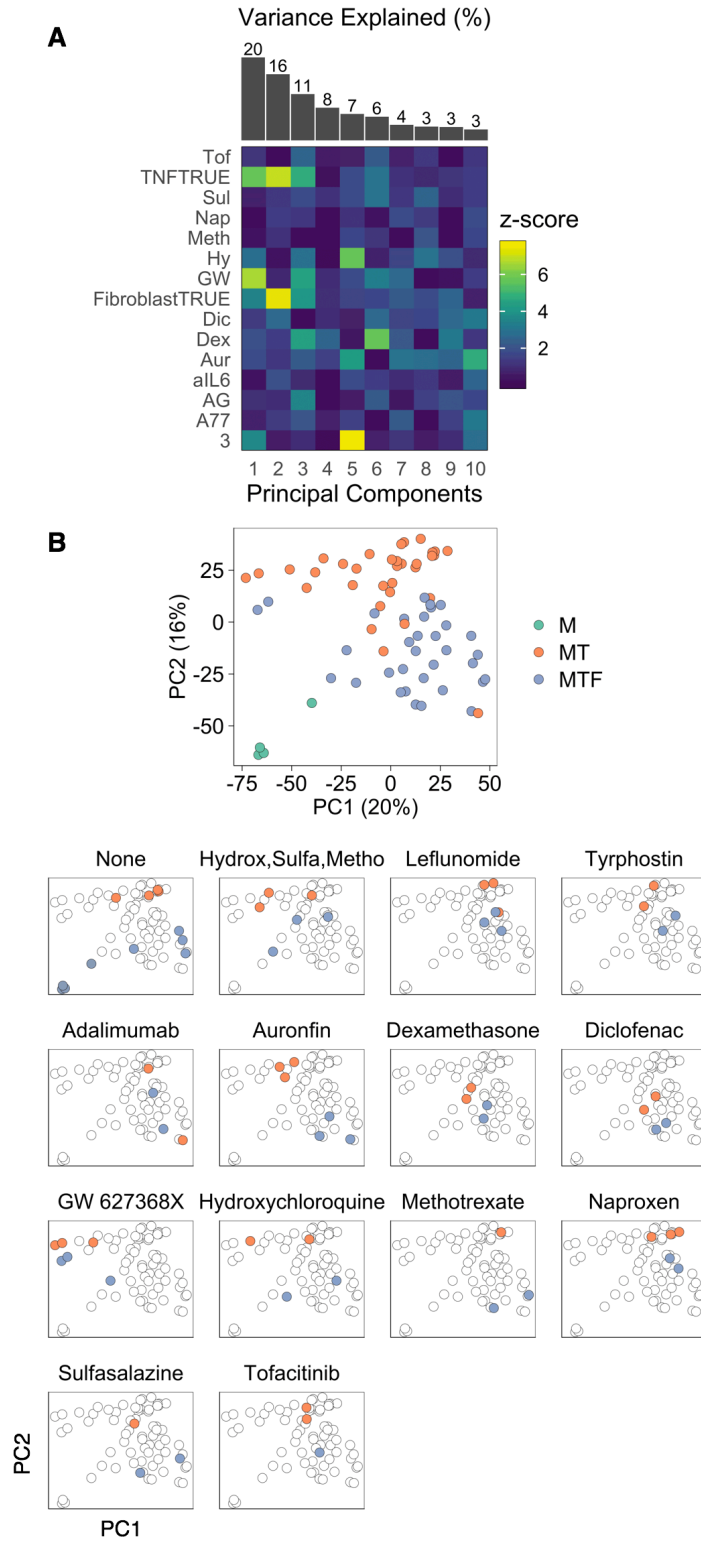
One example of a promising study design is by Kuo *et al*, where the authors assay transcriptomes in an *ex vivo* assay to test the effects of medications on synovial tissue (**Figure 7.1**)²¹⁶. In this work, the authors assayed the transcriptomes of macrophages by themselves or in a coculture with synovial fibroblasts. They show that synovial fibroblast invasiveness and

neutrophil chemoattraction, mediated through the EGF receptor (EGFR), is activated by macrophages producing epidermal growth factor (EGF) ligands. In addition, the fibroblasts also affect the macrophages: the cytokine TNF has different effects on macrophages alone than macrophages cocultured with fibroblasts.

Figure 7.2 Principal components analysis of RNA-seq data from Kuo et al. 2019.

(A) Variance explained by each principal component. Linear model coefficient p-values converted to z-scores for the model $PC_i \sim TNF + Fibroblast + Drug$. (B) Samples projected onto principal components 1 and 2, colored by condition. The 14 panels show the positions of samples treated with each drug. Notably, the samples treated with the EGFR inhibitor (GW 627368X) are closest to the unstimulated macrophages (M) along the first principal component.

Figure 7.2 (Continued)



Kuo *et al* also report the effects of commonly used medications on *ex vivo* tissue explants²¹⁶. For example, they claim that the COX inhibitor (non-steroidal anti-inflammatory drug (NSAID)) naproxen is not effective because it inhibits prostaglandin PGE2. Since PGE2 suppresses the interferon response, NSAIDs can actually enhance the inflammatory interferon response in macrophages. On the other hand, when the authors tested an EGFR inhibitor developed for cancer treatment, they found that the macrophage-induced fibroblast response could be reduced, indicating that the interaction between these two cell types might be reduced by the EGFR inhibitor (**Figure 7.2**). This result could suggest a new treatment direction for RA patients that might not involve broadly suppressing the immune system. This study serves as a proof of concept for how assays of cocultures, organoids, or biopsies with multiple cell types can deliver fine-grained insights about cell type interactions in the synovial tissue. Further, this study shows that we can learn much more from biopsy tissues by perturbing them with medications, cocultures, cytokines, or pathogens than we can learn from resting or basal state transcriptomes. As synovial biopsies become more common in clinical practice^{50,217,218}, the type of study done by Kuo *et al* will be feasible in larger samples across many patients. In the future, by utilizing functional genomics technologies such as RNA-seq on real patient tissues that have been perturbed with drugs, we will learn a great deal about how the drugs work, how genetic variants influence drug response, and – most importantly – which therapy will be most effective for a given patient.

References

1. Ko, S. D., Page, R. C. & Narayanan, A. S. Fibroblast heterogeneity and prostaglandin regulation of subpopulations. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 3429–3432 (1977).
2. Sorrell, J. M. & Caplan, A. I. Fibroblasts—a diverse population at the center of it all. *Int. Rev. Cell Mol. Biol.* **276**, 161–214 (2009).
3. Borrello, M. A. & Phipps, R. P. Differential Thy-1 expression by splenic fibroblasts defines functionally distinct subsets. *Cell. Immunol.* **173**, 198–206 (1996).
4. Fries, K. M. *et al.* Evidence of fibroblast heterogeneity and the role of fibroblast subpopulations in fibrosis. *Clin. Immunol. Immunopathol.* **72**, 283–292 (1994).
5. Rege, T. A. & Hagood, J. S. Thy-1 as a regulator of cell-cell and cell-matrix interactions in axon regeneration, apoptosis, adhesion, migration, cancer, and fibrosis. *FASEB J.* **20**, 1045–1054 (2006).
6. Koumas, L., King, A. E., Critchley, H. O., Kelly, R. W. & Phipps, R. P. Fibroblast heterogeneity: existence of functionally distinct Thy 1(+) and Thy 1(-) human female reproductive tract fibroblasts. *Am. J. Pathol.* **159**, 925–935 (2001).
7. Malmström, E. *et al.* The importance of fibroblasts in remodelling of the human uterine cervix during pregnancy and parturition. *Mol. Hum. Reprod.* **13**, 333–341 (2007).
8. Koumas, L., Smith, T. J. & Phipps, R. P. Fibroblast subsets in the human orbit: Thy-1+ and Thy-1-subpopulations exhibit distinct phenotypes. *Eur. J. Immunol.* **32**, 477–485 (2002).
9. Dudas, J., Mansuroglu, T., Batusic, D., Saile, B. & Ramadori, G. Thy-1 is an in vivo and in vitro marker of liver myofibroblasts. *Cell Tissue Res.* **329**, 503–514 (2007).
10. Zhao, H. & Peehl, D. M. Tumor-promoting phenotype of CD90hi prostate cancer-associated fibroblasts. *Prostate* **69**, 991–1000 (2009).
11. Ospelt, C. & Frank-Bertoncelj, M. Why location matters - site-specific factors in rheumatic diseases. *Nat. Rev. Rheumatol.* **13**, 433–442 (2017).
12. Frank-Bertoncelj, M. *et al.* Epigenetically-driven anatomical diversity of synovial fibroblasts guides joint-specific fibroblast functions. *Nat. Commun.* **8**, 14852 (2017).
13. Ai, R. *et al.* Joint-specific DNA methylation and transcriptome signatures in rheumatoid arthritis identify distinct pathogenic processes. *Nat. Commun.* **7**, 11849 (2016).

14. 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 (2008).
15. van de Sande, M. G. & Baeten, D. L. Immunopathology of synovitis: from histology to molecular pathways. *Rheumatol.* **55**, 599–606 (2016).
16. Bottini, N. & Firestein, G. S. Duality of fibroblast-like synoviocytes in RA: passive responders and imprinted aggressors. *Nat. Rev. Rheumatol.* **9**, 24–33 (2013).
17. Müller-Ladner, U., Pap, T., Gay, R. E., Neidhart, M. & Gay, S. Mechanisms of disease: the molecular and cellular basis of joint destruction in rheumatoid arthritis. *Nat. Clin. Pr. Rheumatol.* **1**, 102–110 (2005).
18. Sabat, R. & Wolk, K. in *Psoriasis* 28–48 (John Wiley & Sons, Ltd, 2014). doi:10.1002/9781118661796.ch4
19. Ospelt, C. *et al.* Toll-like receptors in rheumatoid arthritis joint destruction mediated by two distinct pathways. *Ann. Rheum. Dis.* **63 Suppl 2**, ii90–ii91 (2004).
20. Buckley, C. D. Why do leucocytes accumulate within chronically inflamed joints? *Rheumatol.* **42**, 1433–1444 (2003).
21. West, N. R. *et al.* Oncostatin M drives intestinal inflammation and predicts response to tumor necrosis factor–neutralizing therapy in patients with inflammatory bowel disease. *Nat. Med.* **23**, 579–589 (2017).
22. de Souza, H. S. P. & Fiocchi, C. Immunopathogenesis of IBD: current state of the art. *Nat. Rev. Gastroenterol. Hepatol.* **13**, 13–27 (2016).
23. Böhm, B. B., Freund, I., Krause, K., Kinne, R. W. & Burkhardt, H. ADAM15 adds to apoptosis resistance of synovial fibroblasts by modulating focal adhesion kinase signaling. *Arthritis Rheum.* **65**, 2826–2834 (2013).
24. Whitaker, J. W. *et al.* An imprinted rheumatoid arthritis methylome signature reflects pathogenic phenotype. *Genome Med.* **5**, 40 (2013).
25. Bartok, B., Hammaker, D. & Firestein, G. S. Phosphoinositide 3-kinase δ regulates migration and invasion of synoviocytes in rheumatoid arthritis. *J. Immunol.* **192**, 2063–2070 (2014).
26. Boots, A. M., Wimmers-Bertens, A. J. & Rijnders, A. W. Antigen-presenting capacity of rheumatoid synovial fibroblasts. *Immunology* **82**, 268–274 (1994).

27. Hutton, A. J. *et al.* Human Lung Fibroblasts Present Bacterial Antigens to Autologous Lung Th Cells. *J. Immunol.* **198**, 110–118 (2017).
28. Tran, C. N. *et al.* Presentation of arthritogenic peptide to antigen-specific T cells by fibroblast-like synoviocytes. *Arthritis Rheum.* **56**, 1497–1506 (2007).
29. You, S. *et al.* Identification of key regulators for the migration and invasion of rheumatoid synoviocytes through a systems approach. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 550–555 (2014).
30. Pohlers, D. *et al.* Constitutive upregulation of the transforming growth factor-beta pathway in rheumatoid arthritis synovial fibroblasts. *Arthritis Res. Ther.* **9**, R59 (2007).
31. Orr, C. *et al.* Synovial tissue research: a state-of-the-art review. *Nat. Rev. Rheumatol.* **13**, 463–475 (2017).
32. Woetzel, D. *et al.* Identification of rheumatoid arthritis and osteoarthritis patients by transcriptome-based rule set generation. *Arthritis Res. Ther.* **16**, R84 (2014).
33. Ungethuem, U. *et al.* Molecular signatures and new candidates to target the pathogenesis of rheumatoid arthritis. *Physiol. Genomics* **42A**, 267–282 (2010).
34. Ducreux, J. *et al.* Global molecular effects of tocilizumab therapy in rheumatoid arthritis synovium. *Arthritis Rheumatol* **66**, 15–23 (2014).
35. Marigorta, U. M. *et al.* Transcriptional risk scores link GWAS to eQTLs and predict complications in Crohn's disease. *Nat. Genet.* (2017). doi:10.1038/ng.3936
36. Regev, A. *et al.* The Human Cell Atlas. *bioRxiv* (2017). doi:10.1101/121202
37. Malhotra, D. *et al.* Transcriptional profiling of stroma from inflamed and resting lymph nodes defines immunological hallmarks. *Nat. Immunol.* **13**, 499–510 (2012).
38. Goetz, J. J. & Trimarchi, J. M. Transcriptome sequencing of single cells with Smart-Seq. *Nat. Biotechnol.* **30**, 763–765 (2012).
39. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331–338 (2017).
40. van der Wijst, M. G. P., Brugge, H., de Vries, D. H. & Franke, L. H. Single-cell RNA sequencing reveals cell-type specific cis-eQTLs in peripheral blood mononuclear cells. *bioRxiv* (2017). doi:10.1101/177568

41. Stephenson, W. *et al.* Single-cell RNA-seq of rheumatoid arthritis synovial tissue using low-cost microfluidic instrumentation. *Nat. Commun.* **9**, 791 (2018).
42. Zheng, C. *et al.* Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing. *Cell* **169**, 1342–1356.e16 (2017).
43. Mizoguchi, F. *et al.* Functionally distinct disease-associated fibroblast subsets in rheumatoid arthritis. *Nat. Commun.* **9**, (2018).
44. Mizoguchi, F. *et al.* Functionally distinct disease-associated fibroblast subsets in rheumatoid arthritis. *Nat. Commun.* **9**, 789 (2018).
45. Guo, Y. *et al.* CD40L-Dependent Pathway Is Active at Various Stages of Rheumatoid Arthritis Disease Progression. *J. Immunol.* **198**, 4490–4501 (2017).
46. Koumas, L., Smith, T. J., Feldon, S., Blumberg, N. & Phipps, R. P. Thy-1 expression in human fibroblast subsets defines myofibroblastic or lipofibroblastic phenotypes. *Am. J. Pathol.* **163**, 1291–1300 (2003).
47. Ruderman, E. M., Mandelin, A. M. & Perlman, H. R. Willie Sutton Was Right: It's Time to Turn to the Synovium to Drive Rheumatoid Arthritis Therapy. *J. Rheumatol.* **43**, 2089–2091 (2016).
48. Gutierrez-Arcelus, M. *et al.* Lymphocyte innateness defined by transcriptional states reflects a balance between proliferation and effector functions. *Nat. Commun.* **10**, (2019).
49. McInnes, L. & Healy, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018).
50. Zhang, F. *et al.* Defining Inflammatory Cell States in Rheumatoid Arthritis Joint Synovial Tissues by Integrating Single-cell Transcriptomics and Mass Cytometry. *bioRxiv* (2018). doi:10.1101/351130
51. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 85 (2008).
52. Yaspan, B. L. & Veatch, O. J. Strategies for pathway analysis from GWAS data. *Curr. Protoc. Hum. Genet.* **Chapter 1**, Unit1.20 (2011).
53. Elbers, C. C. *et al.* Using genome-wide pathway analysis to unravel the etiology of complex diseases. *Genet. Epidemiol.* **33**, 419–431 (2009).
54. Allen, H. L. *et al.* Hundreds of variants clustered in genomic loci and biological pathways

- affect human height. *Nature* **467**, 832–838 (2010).
55. Wang, X. *et al.* Association of polymorphisms in the Chr18q11.2 locus with tuberculosis in Chinese population. *Hum. Genet.* **132**, 691–695 (2013).
 56. Raychaudhuri, S. Mapping Rare and Common Causal Alleles for Complex Human Diseases. *Cell* **147**, 57–69 (2011).
 57. 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 (2011).
 58. Weng, L. *et al.* SNP-based pathway enrichment analysis for genome-wide association studies. *BMC Bioinformatics* **12**, 99 (2011).
 59. Holden, M., Deng, S., Wojnowski, L. & Kulle, B. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics* **24**, 2784–2785 (2008).
 60. Rossin, E. J. *et al.* Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology. *PLoS Genet.* **7**, e1001273 (2011).
 61. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome. *Science (80-.).* **310**, 321–324 (2005).
 62. Consortium, T. 1000 G. P. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
 63. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome. *Science (80-.).* **310**, 321–324 (2005).
 64. Allen, H. L. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
 65. Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 6062–6067 (2004).
 66. van der Harst, P. *et al.* Seventy-five genetic loci influencing the human red blood cell. *Nature* **492**, 369–375 (2012).
 67. Su, A. I. *et al.* Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl.*

- Acad. Sci.* **99**, 4465–4470 (2002).
68. Rossin, E. J. *et al.* Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology. *PLoS Genet* **7**, e1001273 (2011).
 69. Project, G. *et al.* An integrated map of genetic variation. **135**, 0–9 (2012).
 70. 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 (2011).
 71. Johnson, N. L., Kemp, A. W. & Kotz, S. *Univariate discrete distributions*. (Wiley, 2005).
 72. Phipson, B. & Smyth, G. K. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Stat. Appl. Genet. Mol. Biol.* **9**, (2010).
 73. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
 74. Rockey, D. C., Bell, P. D. & Hill, J. A. Fibrosis--a common pathway to organ injury and failure. *N. Engl. J. Med.* **372**, 1138–1149 (2015).
 75. Lee, D. M. *et al.* Cadherin-11 in synovial lining formation and pathology in arthritis. *Science (80-.)*. **315**, 1006–1010 (2007).
 76. Turner, J. D. & Filer, A. The role of the synovial fibroblast in rheumatoid arthritis pathogenesis. *Curr. Opin. Rheumatol.* **27**, 175–182 (2015).
 77. Fletcher, A. L., Acton, S. E. & Knoblich, K. Lymph node fibroblastic reticular cells in health and disease. *Nat. Rev. Immunol.* **15**, 350–361 (2015).
 78. Turley, S. J., Cremasco, V. & Astarita, J. L. Immunological hallmarks of stromal cells in the tumour microenvironment. *Nat. Rev. Immunol.* **15**, 669–682 (2015).
 79. Xue, J. *et al.* Transcriptome-based network analysis reveals a spectrum model of human macrophage activation. *Immunity* **40**, 274–288 (2014).
 80. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science (80-.)*. **344**, 1396–1401 (2014).
 81. Miyamoto, D. T. *et al.* {RNA-Seq} of single prostate {CTCs} implicates noncanonical Wnt signaling in antiandrogen resistance. *Science (80-.)*. **349**, 1351–1356 (2015).

82. Gibofsky, A. Epidemiology, pathophysiology, and diagnosis of rheumatoid arthritis: A Synopsis. *Am. J. Manag. Care* **20**, S128--35 (2014).
83. Firestein, G. S. Evolving concepts of rheumatoid arthritis. *Nature* **423**, 356–361 (2003).
84. Kharchenko, P. V, Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–2 (2014).
85. 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 (2005).
86. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst.* **1**, 417–425 (2015).
87. Danks, L. *et al.* RANKL expressed on synovial fibroblasts is primarily responsible for bone erosions during joint inflammation. *Ann. Rheum. Dis.* **75**, 1187–1195 (2016).
88. Guerrini, M. M. *et al.* Inhibition of the {TNF} Family Cytokine {RANKL} Prevents Autoimmune Inflammation in the Central Nervous System. *Immunity* **43**, 1174–1185 (2015).
89. Kobayashi, Y. The role of chemokines in neutrophil biology. *Front. Biosci.* **13**, 2400–7 (2008).
90. Vukicevic, S. & Grgurevic, L. BMP-6 and mesenchymal stem cell differentiation. *Cytokine Growth Factor Rev.* **20**, 441–448 (2009).
91. Izquierdo, E. *et al.* Synovial fibroblast hyperplasia in rheumatoid arthritis: clinicopathologic correlations and partial reversal by anti-tumor necrosis factor therapy. *Arthritis Rheum.* **63**, 2575–2583 (2011).
92. Chang, S. K. *et al.* Cadherin-11 regulates fibroblast inflammation. *Proc. Natl. Acad. Sci.* **108**, 8402–8407 (2011).
93. Noss, E. H., Chang, S. K., Watts, G. F. M. & Brenner, M. B. Modulation of matrix metalloproteinase production by rheumatoid arthritis synovial fibroblasts after cadherin 11 engagement. *Arthritis Rheum.* **63**, 3768–3778 (2011).
94. Kiener, H. P., Lee, D. M., Agarwal, S. K. & Brenner, M. B. Cadherin-11 induces rheumatoid arthritis fibroblast-like synoviocytes to form lining layers in vitro. *Am. J. Pathol.* **168**, 1486–1499 (2006).

95. Fletcher, A. L. *et al.* Reproducible isolation of lymph node stromal cells reveals site-dependent differences in fibroblastic reticular cells. *Front. Immunol.* **2**, 35 (2011).
96. Ekwall, A.-K. H. *et al.* The tumour-associated glycoprotein podoplanin is expressed in fibroblast-like synoviocytes of the hyperplastic synovial lining layer in rheumatoid arthritis. *Arthritis Res. Ther.* **13**, R40 (2011).
97. Valencia, X. *et al.* Cadherin-11 provides specific cellular adhesion between fibroblast-like synoviocytes. *J. Exp. Med.* **200**, 1673–1679 (2004).
98. Bauer, S. *et al.* Fibroblast activation protein is expressed by rheumatoid myofibroblast-like synoviocytes. *Arthritis Res. Ther.* **8**, R171 (2006).
99. Miyake, K. *et al.* Inflammatory cytokine-induced expression of vasohibin-1 by rheumatoid synovial fibroblasts. *Acta Med. Okayama* **63**, 349–358 (2009).
100. Middleton, J. *et al.* A comparative study of endothelial cell markers expressed in chronically inflamed human tissues: MECA-79, Duffy antigen receptor for chemokines, von Willebrand factor, CD31, CD34, CD105 and CD146. *J. Pathol.* **206**, 260–268 (2005).
101. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
102. Trombetta, J. J. *et al.* in *Current Protocols in Molecular Biology* 4.22.1–4.22.17 (John Wiley & Sons, Inc., 2014). doi:10.1002/0471142727.mb0422s107
103. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic {RNA-seq} quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
104. Ritchie, M. E. *et al.* limma powers differential expression analyses for {RNA-sequencing} and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
105. McInnes, I. B. & Schett, G. The Pathogenesis of Rheumatoid Arthritis. *N. Engl. J. Med.* **365**, 2205–2219 (2011).
106. Wolfe, F. *et al.* The mortality of rheumatoid arthritis. *Arthritis Rheum.* **37**, 481–494 (1994).
107. 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 (1998).
108. Gizinski, A. M. & Fox, D. A. T cell subsets and their role in the pathogenesis of rheumatic disease. *Curr. Opin. Rheumatol.* **26**, 204–210 (2014).

109. Reparón-Schuijt, C. C. *et al.* Secretion of anti-citrulline-containing peptide antibody by B lymphocytes in rheumatoid arthritis. *Arthritis Rheum.* (2001). doi:10.1002/1529-0131(200101)44:1<41::AID-ANR6>3.0.CO;2-0
110. Mulherin, D., Fitzgerald, O. & Bresnihan, B. Synovial tissue macrophage populations and articular damage in rheumatoid arthritis. *Arthritis Rheum.* **39**, 115–124 (1996).
111. Kinne, R. W., Bräuer, R., Stuhlmüller, B., Palombo-Kinne, E. & Burmester, G. R. Macrophages in rheumatoid arthritis. *Arthritis Res.* **2**, 189–202 (2000).
112. Müller-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).
113. Pap, T., Müller-Ladner, U., Gay, R. E. & Gay, S. Fibroblast biology. Role of synovial fibroblasts in the pathogenesis of rheumatoid arthritis. *Arthritis Res.* **2**, 361–367 (2000).
114. Dennis Jr, G. *et al.* Synovial phenotypes in rheumatoid arthritis correlate with response to biologic therapeutics. *Arthritis Res. Ther.* **16**, R90 (2014).
115. Orange, D. E. *et al.* Machine learning integration of rheumatoid arthritis synovial histology and RNAseq data identifies three disease subtypes. *Arthritis Rheumatol* (2018).
116. Lindberg, J. *et al.* Variability in synovial inflammation in rheumatoid arthritis investigated by microarray technology. *Arthritis Res. Ther.* **8**, R47 (2006).
117. Stephenson, W. *et al.* Single-Cell RNA-Seq Of Rheumatoid Arthritis Synovial Tissue Using Low Cost Microfluidic Instrumentation. *bioRxiv* (2017). doi:10.1101/140848
118. Papalexis, E. & Satija, R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.* (2017). doi:10.1038/nri.2017.76
119. Schelker, M. *et al.* Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nat. Commun.* **8**, 2032 (2017).
120. Rao, D. A. *et al.* Pathologically expanded peripheral T helper cell subset drives B cells in rheumatoid arthritis. *Nature* **542**, (2017).
121. Fonseka, C. Y. *et al.* Mixed Effects Association of Single Cells Identifies an Expanded Th1-Skewed Cytotoxic Effector CD4⁺ T Cell Subset in Rheumatoid Arthritis. (2017). doi:10.1101/172403
122. Villani, A.-C. *et al.* Single-cell RNA-seq reveals new types of human blood dendritic

- cells, monocytes, and progenitors. *Science* (80-.). **356**, (2017).
123. Donlin, L. T. *et al.* Methods for high-dimensional analysis of cells dissociated from cryopreserved synovial tissue. *Arthritis Res. Ther.* **20**, 139 (2018).
 124. Becher, B. *et al.* High-dimensional analysis of the murine myeloid cell system. *Nat. Immunol.* **15**, 1181–1189 (2014).
 125. Maesschalck, R. De, De Maesschalck, R., Jouan-Rimbaud, D. & Massart, D. L. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems* **50**, 1–18 (2000).
 126. Krenn, V. *et al.* Grading of chronic synovitis—a histopathological grading system for molecular and diagnostic pathology. *Pathol. Res. Pr.* **198**, 317–325 (2002).
 127. Maaten, L. Van Der & Hinton, G. Visualizing Data using t-SNE. **9**, 2579–2605 (2008).
 128. 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 (2009).
 129. 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 (2015).
 130. Pillai, S. Now you know your ABCs. *Blood* **118**, 1187–1188 (2011).
 131. 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 (2016).
 132. 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 (2018).
 133. Pitzalis, C., Kelly, S. & Humby, F. New learnings on the pathophysiology of RA from synovial biopsies. *Curr. Opin. Rheumatol.* **25**, 334–344 (2013).
 134. Filer, A. The fibroblast as a therapeutic target in rheumatoid arthritis. *Curr. Opin. Pharmacol.* **13**, 413–419 (2013).
 135. Westra, H.-J. *et al.* Fine-mapping and functional studies highlight potential causal variants for rheumatoid arthritis and type 1 diabetes. *Nat. Genet.* (2018). doi:10.1038/s41588-018-0216-7
 136. Al-Mayouf, S. M. *et al.* Loss-of-function variant in DNASE1L3 causes a familial form of

- systemic lupus erythematosus. *Nature Genetics* **43**, 1186–1188 (2011).
137. Snelling, S. J. B. *et al.* Dickkopf-3 is upregulated in osteoarthritis and has a chondroprotective role. *Osteoarthritis and Cartilage* **24**, 883–891 (2016).
 138. Lee, E. B. *et al.* Tofacitinib versus methotrexate in rheumatoid arthritis. *N. Engl. J. Med.* **370**, 2377–2386 (2014).
 139. 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. *The Journal of Immunology* **189**, 3508–3520 (2012).
 140. Frara, N. *et al.* Transgenic Expression of Osteoactivin/gpnmmb Enhances Bone Formation In Vivo and Osteoprogenitor Differentiation Ex Vivo. *Journal of Cellular Physiology* **231**, 72–83 (2016).
 141. 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.e6 (2018).
 142. Smolen, J. S. How well can we compare different biologic agents for RA? *Nature Reviews Rheumatology* **6**, 247–248 (2010).
 143. McInnes, I. B. *et al.* The role of interleukin–15 in T–cell migration and activation in rheumatoid arthritis. *Nature Medicine* **2**, 175–182 (1996).
 144. McInnes, I. B. & Liew, F. Y. Cytokine networks—towards new therapies for rheumatoid arthritis. *Nature Clinical Practice Rheumatology* **1**, 31–39 (2005).
 145. Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* (2017). doi:10.1093/biostatistics/kxx053
 146. Parkhomenko, E., Tritchler, D. & Beyene, J. Sparse Canonical Correlation Analysis with Application to Genomic Data Integration. *Statistical Applications in Genetics and Molecular Biology* **8**, 1–34 (2009).
 147. 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 (2009).
 148. Hashimshony, T. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).

149. Bendall, S. C. *et al.* Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* (80-.). **332**, 687–696 (2011).
150. Bjornson, Z. B., Nolan, G. P. & Fantl, W. J. Single-cell mass cytometry for analysis of immune system functional states. *Current Opinion in Immunology* **25**, 484–494 (2013).
151. Peterson, V. M. *et al.* Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.* (2017). doi:10.1038/nbt.3973
152. Finck, R. *et al.* Normalization of mass cytometry data with bead standards. *Cytometry Part A* **83A**, 483–494 (2013).
153. Gonzalez, I., Déjean, S., Martin, P. & Baccini, A. CCA: AnRPackage to Extend Canonical Correlation Analysis. *Journal of Statistical Software* **23**, (2008).
154. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **21**, 3940–3941 (2005).
155. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
156. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417–425 (2015).
157. Reynolds, A. P., Richards, G., de la Iglesia, B. & Rayward-Smith, V. J. Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms. *J. Math. Model. Algorithms* **5**, 475–504 (2006).
158. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
159. Chabaud, M. *et al.* Human interleukin-17: A T cell-derived proinflammatory cytokine produced by the rheumatoid synovium. *Arthritis Rheum.* **42**, 963–970 (1999).
160. Monin, L. & Gaffen, S. L. Interleukin 17 Family Cytokines: Signaling Mechanisms, Biological Activities, and Therapeutic Implications. *Cold Spring Harb. Perspect. Biol.* **10**, (2018).
161. Benedetti, G. & Miossec, P. Interleukin 17 contributes to the chronicity of inflammatory diseases such as rheumatoid arthritis. *Eur. J. Immunol.* (2014). doi:10.1002/eji.201344184
162. Zrioual, S. *et al.* Genome-wide comparison between IL-17A- and IL-17F-induced effects in human rheumatoid arthritis synoviocytes. *J. Immunol.* **182**, 3112–20 (2009).

163. Hirota, K. *et al.* Autoimmune Th17 Cells Induced Synovial Stromal and Innate Lymphoid Cell Secretion of the Cytokine GM-CSF to Initiate and Augment Autoimmune Arthritis. *Immunity* (2018). doi:10.1016/j.immuni.2018.04.009
164. Nguyen, H. N. *et al.* Autocrine Loop Involving IL-6 Family Member LIF, LIF Receptor, and STAT4 Drives Sustained Fibroblast Production of Inflammatory Mediators. *Immunity* **46**, 220–232 (2017).
165. Smolen, J. S. & Aletaha, D. Rheumatoid arthritis therapy reappraisal: strategies, opportunities and challenges. *Nat. Rev. Rheumatol.* **11**, 276–289 (2015).
166. Kirkham, B. W. *et al.* Synovial membrane cytokine expression is predictive of joint damage progression in rheumatoid arthritis: a two-year prospective study (the DAMAGE study cohort). *Arthritis Rheum.* **54**, 1122–1131 (2006).
167. Moran, E. M. *et al.* Human rheumatoid arthritis tissue production of IL-17A drives matrix and cartilage degradation: synergy with tumour necrosis factor-alpha, Oncostatin M and response to biologic therapies. *Arthritis Res. Ther.* **11**, R113 (2009).
168. Kotake, S. *et al.* IL-17 in synovial fluids from patients with rheumatoid arthritis is a potent stimulator of osteoclastogenesis. *J. Clin. Invest.* **103**, 1345–1352 (1999).
169. Chabaud, M., Page, G. & Miossec, P. Enhancing effect of IL-1, IL-17, and TNF-alpha on macrophage inflammatory protein-3alpha production in rheumatoid arthritis: regulation by soluble receptors and Th2 cytokines. *J. Immunol.* **167**, 6015–6020 (2001).
170. Fossiez, F. *et al.* T cell interleukin-17 induces stromal cells to produce proinflammatory and hematopoietic cytokines. *J. Exp. Med.* **183**, 2593–2603 (1996).
171. Ruddy, M. J., Wong, G. C., Liu, X. K., Yamamoto, H. & others. Functional cooperation between interleukin-17 and tumor necrosis factor- α is mediated by CCAAT/enhancer-binding protein family members. *J. Biol.* (2004).
172. Shen, F., Ruddy, M. J., Plamondon, P. & Gaffen, S. L. Cytokines link osteoblasts and inflammation: microarray analysis of interleukin-17- and TNF-alpha-induced genes in bone cells. *J. Leukoc. Biol.* **77**, 388–399 (2005).
173. Ermann, J., Staton, T., Glickman, J. N., de Waal Malefyt, R. & Glimcher, L. H. Nod/Ripk2 signaling in dendritic cells activates IL-17A-secreting innate lymphoid cells and drives colitis in T-bet^{-/-}.Rag2^{-/-} (TRUC) mice. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E2559-66 (2014).
174. Koshy, P. J. *et al.* Interleukin 17 induces cartilage collagen breakdown: novel synergistic

- effects in combination with proinflammatory cytokines. *Ann. Rheum. Dis.* **61**, 704–713 (2002).
175. Lubberts, E. The IL-23-IL-17 axis in inflammatory arthritis. *Nat. Rev. Rheumatol.* **11**, 415–429 (2015).
 176. Zhang, Q., Lenardo, M. J. & Baltimore, D. 30 Years of NF- κ B: A Blossoming of Relevance to Human Pathobiology. *Cell* **168**, 37–57 (2017).
 177. Raychaudhuri, S., Stuart, J. M. & Altman, R. B. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac. Symp. Biocomput.* **38**, 455–66 (2000).
 178. Panopoulos, A. D. & Watowich, S. S. Granulocyte colony-stimulating factor: molecular mechanisms of action during steady state and ‘emergency’ hematopoiesis. *Cytokine* **42**, 277–288 (2008).
 179. Parsonage, G. *et al.* Prolonged, granulocyte--macrophage colony-stimulating factor-dependent, neutrophil survival following rheumatoid synovial fibroblast activation by IL-17 and TNF α . *Arthritis Res. Ther.* **10**, R47 (2008).
 180. Elshabrawy, H. A. *et al.* IL-11 facilitates a novel connection between RA joint fibroblasts and endothelial cells. *Angiogenesis* **21**, 215–228 (2018).
 181. Mix, K. S. *et al.* Orphan nuclear receptor NR4A2 induces synoviocyte proliferation, invasion, and matrix metalloproteinase 13 transcription. *Arthritis Rheum.* **64**, 2126–2136 (2012).
 182. Aherne, C. M. *et al.* Identification of NR4A2 as a transcriptional activator of IL-8 expression in human inflammatory arthritis. *Mol. Immunol.* **46**, 3345–3357 (2009).
 183. Parrish, W. R. *et al.* Intra-articular therapy with recombinant human GDF5 arrests disease progression and stimulates cartilage repair in the rat medial meniscus transection (MMT) model of osteoarthritis. *Osteoarthr. Cartil.* **25**, 554–560 (2017).
 184. Roelofs, A. J. *et al.* Joint morphogenetic cells in the adult mammalian synovium. *Nat. Commun.* **8**, 15040 (2017).
 185. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
 186. Hillmer, E. J., Zhang, H., Li, H. S. & Watowich, S. S. STAT3 signaling in immunity.

187. Khan, A. *et al.* JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gkx1126
188. Hirota, K. *et al.* Preferential recruitment of CCR6-expressing Th17 cells to inflamed joints via CCL20 in rheumatoid arthritis and its animal model. *J. Exp. Med.* **204**, 2803–2812 (2007).
189. Hollingsworth, J. W., Siegel, E. R. & Creasey, W. A. Granulocyte survival in synovial exudate of patients with rheumatoid arthritis and other inflammatory joint diseases. *Yale J. Biol. Med.* **39**, 289–296 (1967).
190. Zvaifler, N. J. The immunopathology of joint inflammation in rheumatoid arthritis. *Adv. Immunol.* **16**, 265–336 (1973).
191. Haringman, J. J. *et al.* Synovial tissue macrophages: a sensitive biomarker for response to treatment in patients with rheumatoid arthritis. *Ann. Rheum. Dis.* **64**, 834–838 (2005).
192. Muromoto, R. *et al.* IL-17A plays a central role in the expression of psoriasis signature genes through the induction of I κ B- ζ in keratinocytes. *Int. Immunol.* **28**, 443–452 (2016).
193. Koga, T. *et al.* Post-transcriptional regulation of IL-6 production by Zc3h12a in fibroblast-like synovial cells. *Clin. Exp. Rheumatol.* **29**, 906–912 (2011).
194. Hulea, L. & Nepveu, A. CUX1 transcription factors: from biochemical activities and cell-based assays to mouse models and human diseases. *Gene* **497**, 18–26 (2012).
195. Kühnemuth, B. *et al.* CUX1 modulates polarization of tumor-associated macrophages by antagonizing NF- κ B signaling. *Oncogene* **34**, 177–187 (2013).
196. Yamamoto, M. *et al.* Regulation of Toll/IL-1-receptor-mediated gene expression by the inducible nuclear protein IkappaBzeta. *Nature* **430**, 218–222 (2004).
197. Yamazaki, S. *et al.* Gene-specific requirement of a nuclear protein, IkappaB-zeta, for promoter association of inflammatory transcription regulators. *J. Biol. Chem.* **283**, 32404–32411 (2008).
198. Onishi, R. M. & Gaffen, S. L. Interleukin-17 and its target genes: mechanisms of interleukin-17 function in disease. *Immunology* **129**, 311–321 (2010).
199. Miossec, P. Update on interleukin-17: a role in the pathogenesis of inflammatory arthritis

- and implication for clinical practice. *RMD Open* **3**, e000284 (2017).
200. Hu, Y., Shen, F., Crellin, N. K. & others. The IL-17 pathway as a major therapeutic target in autoimmune diseases. *Ann. N. Y. Acad. Sci.* (2011).
 201. Langley, R. G. *et al.* Secukinumab in plaque psoriasis--results of two phase 3 trials. *N. Engl. J. Med.* **371**, 326–338 (2014).
 202. Gordon, K. B. *et al.* Phase 3 Trials of Ixekizumab in Moderate-to-Severe Plaque Psoriasis. *N. Engl. J. Med.* **375**, 345–356 (2016).
 203. Chu, C.-Q., Swart, D., Alcorn, D., Tocker, J. & Elkon, K. B. Interferon- γ regulates susceptibility to collagen-induced arthritis through suppression of interleukin-17. *Arthritis Rheum.* **56**, 1145–1151 (2007).
 204. Miossec, P. & Kolls, J. K. Targeting IL-17 and TH17 cells in chronic inflammation. *Nat. Rev. Drug Discov.* **11**, 763–776 (2012).
 205. Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).
 206. Carvalho, B. S. & Irizarry, R. A. A framework for oligonucleotide microarray preprocessing. *Bioinformatics* **26**, 2363–2367 (2010).
 207. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12**, 115–121 (2015).
 208. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).
 209. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).
 210. Aken, B. L. *et al.* The Ensembl gene annotation system. *Database* **2016**, (2016).
 211. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer, 2016).
 212. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* (2013). doi:10.1038/nature12531
 213. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–

585 (2013).

214. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* (2017). doi:10.1038/nature24277
215. Piasecka, B. *et al.* Distinctive roles of age, sex, and genetics in shaping transcriptional variation of human immune responses to microbial challenges. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E488–E497 (2018).
216. Kuo, D. *et al.* HBEGF+ macrophages identified in rheumatoid arthritis promote joint tissue invasiveness and are reshaped differentially by medications. *bioRxiv* (2019). doi:10.1101/525758
217. Mandelin 2nd, A. M. *et al.* Transcriptional Profiling of Synovial Macrophages Using Minimally Invasive Ultrasound-Guided Synovial Biopsies in Rheumatoid Arthritis. *Arthritis Rheumatol* **70**, 841–854 (2018).
218. Kelly, S. *et al.* Ultrasound-guided synovial biopsy: a safe, well-tolerated and reliable technique for obtaining high-quality synovial tissue from both large and small joints in early arthritis patients. *Ann. Rheum. Dis.* **74**, 611–617 (2015).