

Generalizable Machine Learning Methods for Network Inference in Systems Biology

by

Gabrielle Rabadam

DISSERTATION

Submitted in partial satisfaction of the requirements for degree of

DOCTOR OF PHILOSOPHY

in

Bioengineering

in the

GRADUATE DIVISION

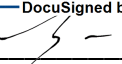
of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

AND

UNIVERSITY OF CALIFORNIA, BERKELEY

Approved:

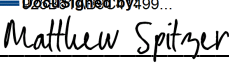
DocuSigned by:

6E99B613A5754DD... Zev Gartner
Chair

DocuSigned by:

Iain Clark

DocuSigned by:

Marina Sirota

DocuSigned by:

4793AE6232BE479... Matthew Spitzer

Committee Members

*To my mother Eleanor Paras Rabadam,
who first made science seem possible.
And to my father, Luisito Rabadam,
who patiently taught me how to count—look, I made it past 10.*

ACKNOWLEDGMENTS

First and foremost, I want to thank my advisor Dr. Zev Gartner for his support. Among several things, I am grateful that he has continuously challenged me to trust my ambition as a scientist just as much as falling back on my pragmatism as an engineer. I am by far a better scientist because of Zev's encouragement to think and dream bigger. Furthermore, I am grateful to Zev for imparting on me the importance of scientific storytelling. It is this capacity for storytelling—the ability to share science with such infectious enthusiasm that people cannot help but want to work with you—that has opened so many doors for me as a member of the Gartner Lab.

In particular, I am grateful to Zev for catalyzing my collaboration with Dr. Jessica Neely and Dr. Marina Sirota. Based on Zev's vote of confidence, Dr. Neely entrusted me with precious patient data that had taken the Department of Pediatric Rheumatology over a decade to acquire. Throughout this project, it was primarily Dr. Neely's recognition and respect of my expertise that empowered me to start seeing myself as a computational biologist. I would also like to thank Dr. Sirota for extending the invitation to join her lab meetings where I could receive feedback on multiple iterations of my computational work and learn from her lab's complementary expertise.

And of course, I would not be here were it not for the Gartner lab, members past and present. Be it through lab meetings, informal conversations, writing revisions, or preparation for conferences, everyone's input has refined my scientific thinking. I specifically want to thank Dr. Lyndsay Murrow who mentored me when I first joined the lab and gave me the space to grow into an independent collaborator. I also want to thank

Dr. Brittany Moser for training me on the MULTiseq protocol and empowering me to do these experiments independently.

At UCSF, I am grateful for Laura Shub, who finally realized after two quarters of classes together as first-years that I was trying to be her friend. What a joy it is to watch your best friend grow into an absolute force to be reckoned with as a scientist, and a privilege to earn her loyalty. I am also indebted to Dr. Scott Nanda who continued to remind me during this hardest year yet, “Just keep pushing. PhD is PhD is PhD.” Thank you both for stoking the fire and ensuring the flames never blew out completely.

I also want to thank my beloved friends across multiple time zones who have continued to cheer me on, check on me, and remind me to take a breath of fresh air: Dustin, Brenna, Lindsay, Selena, Kristina, Emily, and Marina, I owe you all so much. And to my climbing family scattered across the world, thank you for the sing-alongs, family dinners, and much-needed reminder of my own strength.

Finally, to my (very large) family whose faith in me has never wavered, maraming salamat. I am particularly grateful to my Tito Mike and Tita Jean, Tito Nelson and Tita Irma, Tito Raul and Tita Carol for reminding me why CA will always be home. Thank you to my Kuya Carlo and Ate Jeanelle, who showed me just how far a couple of first-gen kids from Sacramento could go. To my niece Elena, your excitement and pure sense of wonder continues to remind me how fun science can be—I am so lucky to be your Ate. To my brother Adrian, thank you for the laughter and for always being willing to talk, watch, live, and breathe basketball. And finally to my parents, Eleanor and Luisito, who continue to believe in me more than I believe in myself, thank you for the freedom to chase my dreams and for instilling in me the work ethic to make them reality. Mahal kita.

SOURCES AND CONTRIBUTIONS

Chapter 2 contains unpublished material from a manuscript in preparation:

Rabadam G, Murrow LM, Gartner ZJ. *deciphR*: A dimensionality reduction framework for inferring networks of cellular coordination in single-cell genomics data. 2024, In Preparation.

GR and LMM developed the computational method and GR wrote the software package *deciphR* implementing the method. GR wrote the manuscript, and all authors revised the manuscript.

Chapters 3 and 4 contains work previously published or currently in press in a peer reviewed journal:

Murrow LM, Weber RJ, Caruso JA, McGinnis CS, Phong K, Gascard P, **Rabadam G**, Borowsky AD, Desai TA, Thomson M, Tlsty T, Gartner ZJ. Mapping hormone-regulated cell-cell interaction networks in the human breast at single-cell resolution. *Cell Syst.* 2022 Aug 17;13(8):644-664.e8. doi:10.1016/j.cels.2022.06.005

GR and LMM developed the computational method used to analyze the cohort data and generate findings. GR and LMM wrote and revised the manuscript sections describing the computational method and its application. All experiments were led by LMM and conducted with other coauthors. Detailed author contributions can be found in Chapter 3, Methods.

Rabadam G, Wibrand C, Flynn E, Hartoularos GC, Sun Y, Ye CJ, Kim S, Gartner ZJ, Sirota M, Neely J. Coordinated immune dysregulation in Juvenile Dermatomyositis revealed by single-cell genomics. JCI Insight. 2024, In Press. In Press Preview at <https://doi.org/10.1172/jci.insight.176963>.

GR performed the analysis of the discovery and validation cohort dataset including integration and quality control, conceptualization and development of the pipeline for the network analyses, and analysis of the second dataset. GR also led writing of the manuscript. Detailed author contributions can be found in Chapter 4, Methods.

*“Show me how you ride
and I’ll show you who you are”
– Dominic Toretto*

GENERALIZABLE MACHINE LEARNING METHODS FOR NETWORK INFERENCE IN SYSTEMS

BIOLOGY

Gabrielle Rabadam

ABSTRACT

Tissues comprise a multiplicity of specialized cell types that must coordinate state changes in order to function collectively. These state changes are orchestrated by coordinated direct cellular interactions and indirect responses to microenvironmental and systemic cues. Consequently, chronic perturbations to this collective behavior can result in disease states that are difficult to reprogram such as autoimmunity and cancer. As such, studying the self-reinforced dynamics of tissue function can benefit from a systems biology approach where the aim is to understand how individual components of biological systems interact to give rise to emergent properties.

The recent growth in the availability of single-cell resolution genomics platforms has further expanded biologists' ability to do this kind of unbiased inquiry. However, despite the increasing ease of generating these high-dimensional datasets, analyzing these data still presents significant computational challenges because of their noise and sparsity, which are further exacerbated on the level of individual cells and genes. As such, there is a need to develop computational methods that enable scientists to extract systems-level biological insight from noisy high dimensional data.

This dissertation introduces DECIPHER, a machine learning framework tailored for network inference in systems biology, with a focus on applications to single-cell RNA

sequencing data. Chapter 2 details the DECIPHER algorithm and its implementation for the R computing environment, `deciphR`, that is designed to reconstruct cell state networks from high-dimensional molecular profiles. Chapter 3 applies DECIPHER to unveil cell-cell interaction networks in the human breast, elucidating how state changes on the cell-level propagate throughout tissue in response to hormonal fluctuations. Chapter 4 extends DECIPHER's application to investigate peripheral immune dysregulation in a rare pediatric autoimmune disease, revealing underlying immune imbalances that persist even in disease remission and potential therapeutic targets. Overall, this dissertation presents a generalizable approach to network inference for systems biology and demonstrates its utility in multiple biological contexts for unravelling cellular coordination in tissue homeostasis and disease.

TABLE OF CONTENTS

Chapter 1 Preface	1
Introduction	2
Review	6
Chapter 2 <i>deciphR: a dimensionality reduction framework for inferring networks of cellular coordination in single-cell genomics data</i>	10
Abstract.....	11
Introduction	12
System and methods	14
Feature extraction of biological activity programs via coinMF	15
Unsupervised network inference of coordinated biological processes	17
Algorithm	20
coinMF Algorithm	21
Discussion	24
Chapter 3 <i>DECIPHER uncovers cell-cell interaction networks in the human breast</i>	25
Abstract.....	26
Introduction	27
Results	29
Person-to-person variability in transcriptional cell state in the human breast.....	29
Inferring shared transcriptional responses and direct cell-to-cell signaling interactions in the human breast	31
ER/PR signaling and the downstream response	40

Discussion	45
Methods	48
Supplement.....	67
 <i>Chapter 4 Coordinated immune dysregulation in Juvenile Dermatomyositis</i>	
<i>revealed by single-cell genomics.....</i>	<i>75</i>
Abstract.....	76
Introduction	77
Results	81
JDM is associated with immunophenotypic differences in B and CD4+ T cell compartments.....	81
SIGLEC-1 expression is a composite measure of the IFN gene signature in JDM	85
Unsupervised network analysis reveals coordinated immune cell states in JDM...	88
JDM CD4+T cells and B cells display persistent alterations in gene expression in both active disease and remission	93
Novel cell states are correlated with IFN gene expression in treatment-naive JDM	97
Regulatory cell death and protein targeting pathways are dysregulated across multiple immune cell populations in JDM	102
JDM-associated signatures identified by DECIPER validated in an independent dataset.....	104
Discussion	108
Methods	115
Supplement.....	126

Chapter 5 Conclusion.....	159
Summary of Advancements	160
Limitations and Assumptions	162
Future Directions	164
References	165

LIST OF FIGURES

Figure 2.1 Overview of DECIPHER method.	14
Figure 2.2 Flowchart depicting algorithm for the dimensionality reduction part of the DECIPHER workflow.	20
Figure 2.3 Flowchart depicting the algorithm for network construction and annotation within the DECIPHER workflow.	23
Figure 3.1 Single-cell transcriptional analysis links biological variables with person-to-person heterogeneity in transcriptional cell state in the premenopausal human breast.....	29
Figure 3.2 UMAP constructed from RNA expression of single cells isolated from breast tissue, colored by major epithelial or stromal cell type.	30
Figure 3.3 Conceptual overview of analysis approach for human breast dataset using DECIPHER.	32
Figure 3.4 Using individual pairwise correlations between cell activities, DECIPHER builds a tissue-level map of the cell-cell interactions present in the healthy human breast and identifies modules of transcriptional states that co-occur across the same sets of samples.....	33
Figure 3.5 DECIPHER identifies cell-cell interaction networks across cell types in the human breast.	35
Figure 3.6 Inferring non-cell-type-specific transcriptional responses in the human breast.....	37
Figure 3.7 DECIPHER infers direct cell-to-cell signaling interactions in the human breast.....	39

Figure 3.8 ER/PR signaling in the human breast.	41
Figure 3.9 ER/PR signaling and the coordinated downstream response confirmed in vivo.	43
Figure 3.10 Activity programs in other epithelial lineages and stromal cell types within the 'ER/PR response' module.	44
Supplemental Figure 3.11 Elbow and knee plots showing standard rank selection metrics calculated for rank sweep of coinMF (left) and PCA (right) run on each major breast cell type in dataset.	71
Supplemental Figure 3.12 Phylogenetic trees of coinMF rank sweep showing final rank selection according to DEW metric.	72
Supplemental Figure 3.13 DEW metric across coinMF rank sweep with final kDEW selection for each cell type.	73
Supplemental Figure 3.14 Network structure at ranks above and below kDEW for each cell type.	74
Figure 4.1 Study design for profiling PBMCs from 27 samples (n=22 JDM, n=5 HC), with an overview of clinical characteristics of study cohort.	81
Figure 4.2 Analysis strategy for CITEseq data from PBMCs.	82
Figure 4.3 Cell types in peripheral blood from patients with JDM and HCs.	83
Figure 4.4 Immunophenotypes in peripheral blood associated with JDM.	84
Figure 4.5 Type I IFN-induced gene and protein expression is associated with disease activity in JDM in CD14+ monocytes.	87
Figure 4.6 DECIPHER learns coordinated biological activity programs through dimensionality reduction.	89

Figure 4.7 DECIPHER reveals network of coordinated biological activity from scRNA-seq data in JDM.	91
Figure 4.8 JDM is associated with a central IFN hub and cell specific gene programs in the B and CD4T compartments.	93
Figure 4.9 Heatmap showing significant differences in expression of selected programs between HC (n=5) and JDM patients (n=22), with columns annotated by p-values of case-control (t-test) and disease activity association (4-group ANOVA).	96
Figure 4.10 Disease activity in JDM is associated with central hub of IFN response in network, correlated with novel dysregulated cell states.	98
Figure 4.11 Heatmap showing significant differences in expression of selected disease activity associated programs between HC (n=5), Inactive JDM (n=6), Active JDM (n=7), and TNJDM patients (n=9).	100
Figure 4.12 Selected network modules colored by FDR of enrichment for indicated gene ontology set (FDR<0.01) or gene loading similarity within Modules 2 and 5.	103
Figure 4.13 JDM-associated signatures identified by DECIPHER can be validated in independent samples.	106
Supplemental Figure 4.14 RNA and surface protein markers used to annotate cell type clusters in UMAP space.	127
Supplemental Figure 4.15 Analysis of reclustered B cells based on ADT measurements alone.	128
Supplemental Figure 4.16 Multi-modal differential analysis for TNJDM and HCs.	129

Supplemental Figure 4.17 Heatmap of top five enriched GO terms per cell type with FDR<0.01.....	130
Supplemental Figure 4.18 Heatmap of differentially expressed genes between TNJDM and HC from all cell types clustered by expression likeliness.....	131
Supplemental Figure 4.19 Gene set enrichment results for GO terms in Module 1 (FDR<0.01).....	132
Supplemental Figure 4.20 Gene set enrichment results for GO terms in Module 2 (FDR<0.01).....	133
Supplemental Figure 4.21 Gene set enrichment results for GO terms in Module 3 (FDR<0.01).....	134
Supplemental Figure 4.22 Gene set enrichment results for GO terms in Module 4 (FDR<0.01).....	135
Supplemental Figure 4.23 Gene set enrichment results for GO terms in Module 5 (FDR<0.01).....	136
Supplemental Figure 4.24 Gene set enrichment results for GO terms in Module 6 (FDR<0.01).....	137
Supplemental Figure 4.25 Mean patient expression of JDM-associated programs (t- test, $p < 0.05$) in B cell (A) and CD4T cell (B) compartments, respectively (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).....	138
Supplemental Figure 4.26 UMAPs of CD4T cells showing expression of NMF program CD4T1 (A) in subcluster corresponding to CD4+ effector cells (B).....	139
Supplemental Figure 4.27 Heatmaps showing top 10 marker genes for selected disease-associated programs for the indicated cell type.....	140

Supplemental Figure 4.28 GEP scores for longitudinal samples collected at an individual's high and low disease activity point.....	141
Supplemental Figure 4.29 Scatter plots showing mean sample expression (n=27) of type I interferon response programs in each corresponding cell type (Pearson).	142
Supplemental Figure 4.30 Expression of program B9 in naive B cells.....	143
Supplemental Figure 4.31 Mean patient expression of disease activity associated programs (4-way ANOVA, $p < 0.05$) in Module 5 (* $p < 0.05$ Post-hoc pairwise Tukey test).....	144
Supplemental Figure 4.32 wnnUMAPs showing normalized expression of GEPs CD4T1 and CD4T10 (A-B) with co-expression of surface protein CCR4 (C).....	145
Supplemental Figure 4.33 Subset of Modules 1, 2, and 5 from original heatmap in Figure 4.6 highlighting the negative correlations.	146
Supplemental Figure 4.34 Proxy GEP metric (AUCell) recapitulates key signatures discovered via DECIPHER in original dataset.....	147
Supplemental Figure 4.35 Relationship between disease activity and proxy scores for GEPs in validation cohort.....	148
Supplemental Figure 4.36 Original multi-modal wnnUMAP, using Leiden clustering at a resolution of 1.4.....	157
Supplemental Figure 4.37 Elbow plots for rank selection for NMF ran on each major cell type, with rank k indicated by vertical line.....	158

LIST OF TABLES

Supplemental Table 3.1 <i>Sample information for breast tissue samples from reduction mammoplasties of 28 individuals.....</i>	67
Supplemental Table 3.2 <i>Sequencing batch info for each sample.....</i>	68
Supplemental Table 4.1 <i>Clinical cohort disease characteristics.</i>	126
Supplemental Table 4.2 <i>ADT sequences and targets for surface protein panel.....</i>	149

LIST OF ABBREVIATIONS

ADT	<i>Antibody Derived Tags</i>
AI	<i>Artificial Intelligence</i>
BSA	<i>Bovine Serum Albumin</i>
CDASI	<i>Cutaneous Disease Area and Severity Index</i>
CITEseq	<i>Cellular Indexing of Transcriptome and Epitope Sequencing</i>
coinMF	<i>Consensus Online Integrative NMF</i>
cSLE	<i>Childhood Systemic Lupus Erythematosus</i>
DECIPHER	<i>Deconstructing Cell-cell Interactions Using Phenotypic Heterogeneity in Single-cell RNA Sequencing</i>
DEW	<i>Depth-encompassing Weight</i>
dsDNA	<i>Double-stranded DNA</i>
ECM	<i>Extra-cellular Matrix</i>
ER	<i>Estrogen Receptor</i>
FACS	<i>Fluorescence-activated Cell Sorting</i>
FBS	<i>Fetal Bovine Serum</i>
FDR	<i>False Discovery Rate</i>
GEP	<i>Gene Expression Program</i>
GSEA	<i>Gene Set Enrichment Analysis</i>
GO	<i>Gene Ontology</i>
HC	<i>Healthy Control</i>
HR+	<i>Hormone Receptor Positive</i>
IFN	<i>Interferon</i>

iNMF	<i>Integrative NMF</i>
JDM	<i>Juvenile Dermatomyositis</i>
KNN	<i>K-nearest Neighbors</i>
LMO	<i>Lipid-modified Oligo</i>
MMT	<i>Manual Muscle Testing</i>
ML	<i>Machine Learning</i>
MSA	<i>Myositis-specific Antibodies</i>
MULTI-seq	<i>Multiplexing Using Lipid-tagged Indices for Single-cell and Single-nuclei Sequencing</i>
NMF	<i>Non-negative Matrix Factorization</i>
NK	<i>Natural Killer Cells</i>
PBMC	<i>Peripheral Blood Mononuclear Cells</i>
PBS	<i>Phosphate-buffered Saline</i>
PCA	<i>Principal Component Analysis</i>
PGA	<i>Physician Global Assessment</i>
PR	<i>Progesterone Receptor</i>
RNA	<i>Ribonucleic Acid</i>
scRNA-seq	<i>Single-cell RNA Sequencing</i>
SNP	<i>Single-nucleotide Polymorphism</i>
TNJDM	<i>Treatment-naive Juvenile Dermatomyositis</i>
UMAP	<i>Uniform Manifold Approximation and Projection</i>
UMI	<i>Unique Molecular Identifier</i>
VAS	<i>Visual Analog Scale</i>

wnnUMAP *Weighted Nearest Neighbors UMAP*
wTO *Weighted Topological Overlap*

CHAPTER 1 PREFACE

Introduction

As dynamic living systems with multiple levels of hierarchical organization, tissues must coordinate state changes in order to perform their biological functions. These state changes occur directly through local cell-to-cell signaling, or indirectly through response to a common microenvironment or systemic signals like hormones. Systems biology, an approach that integrates experimental data, computational analysis, and mathematical modeling, offers a comprehensive framework for studying these complex dynamics of living tissues.^{1,2} At its core, a systems approach aims to elucidate how individual components of biological systems interact to give rise to emergent properties, such as cellular coordination, signaling cascades, and homeostasis despite perturbations.

For example, in the human breast, hormonal fluctuations control cell growth, differentiation, and tissue structure. However, only a small portion of epithelial cells express estrogen and progesterone receptors.³ As such, hormone-induced changes are primarily mediated through a cascade of paracrine signaling from hormone-responsive cells (HR+) to others in the breast. The impact of these changes is profound: cumulative lifetime exposure to cycling hormones is a major modifier of breast cancer risk⁴, and the majority of breast tumors are estrogen-dependent. Thus, cell-cell interactions between HR+ cells and other cell types are key to normal breast morphogenesis

Similarly in the immune system, the dynamic exchange of cellular information between the dispersed network of localized tissue-resident immune microenvironments and central lymphoid organs is critical for mounting an effective immune response and maintaining immune tolerance.^{5,6} For example, during an inflammatory response, a small proportion of cells first respond to stimuli and then drive the widespread signal

propagation to trigger a population-level response.^{7,8} In complex juvenile autoimmune diseases like childhood systemic lupus erythematosus (cSLE) and juvenile dermatomyositis (JDM), multiple immune populations are disrupted resulting in an inflammatory signature and disease-specific autoantibodies.^{9–12} In contrast to monogenic autoinflammatory diseases¹³—which are characterized by innate immune dysregulation and etiologies rooted in individual aberrant pathways—complex juvenile autoimmune diseases are distinguished by the involvement of multiple components of the immune system. However, how these clinically observable disease phenotypes are rooted in immunopathology remains insufficiently understood. Systems-level studies based on single-cell measurements are required to reveal how dysregulated cell populations produce the observed disease signatures.

Single-cell genomics has emerged as a transformative tool^{14,15} for dissecting cellular heterogeneity and uncovering the molecular underpinnings of various biological processes.^{16–18} By profiling individual cells at the transcriptomic,^{8,19} proteomic,^{20,21} and epigenomic levels,^{22,23} single-cell genomics enables researchers to unravel the complexity inherent in biological systems and gain insights into cellular states, transitions, and interactions.^{24–26} Despite these advantages, the analysis of single-cell genomics data presents significant computational challenges due to its high dimensionality, sparsity, and noise.^{27,28} Traditional analytical methods that focus on individual genes often struggle to extract meaningful biological information from such complex datasets.²⁹

As such, there is a need to develop computational methods that both enable scientists to extract biological insight from high dimensional data and address technical challenges of data integration and normalization. Machine learning (ML) techniques, with

their ability to uncover hidden patterns and relationships within large-scale data, have become valuable tools for analyzing and interpreting single-cell genomics data.^{30,31} Even before the advent of next generation sequencing technologies, it has long been appreciated that biology is rife with questions amenable to formulations as classical ML, or more broadly, computing problems. Historical examples range from the immune system's effective pathogen response as a distributed autonomous system in the language of control theory³²⁻³⁴ to the question of sensory processing in visual neurons as a sparse coding problem when borrowing concepts from information theory.³⁵⁻³⁷ Across learning tasks, dimensionality reduction approaches, where the data is decomposed into its relevant basis factors that minimally reconstruct the core features of the data, are central to deriving meaningful insight and reducing computational load.

In this dissertation, I present DECIPHER, a dimensionality reduction framework developed for network inference in systems biology, with a specific focus on inferring cellular coordination and interaction networks from single-cell genomics data: Deconstructing Cell-cell Interactions using Phenotypic Heterogeneity in single-cell RNA sequencing data. The remainder of this introductory chapter is devoted to a brief review on foundational concepts in the fields of Artificial Intelligence (AI) and ML and their utility for dealing with uncertainty in biology, particularly genomics. Specifically, I discuss the intuition behind mathematical formalisms central to the DECIPHER method: matrix factorization, parts-based learning, and graphical representations of Bayesian networks.

Chapter 2 introduces the DECIPHER algorithm and its package implementation deciphR. By leveraging feature extraction and network inference techniques, deciphR facilitates the reconstruction of cell state networks, enabling researchers to infer latent

biological programs from high-dimensional molecular profiles and gain insights into cellular coordination and function. Building upon this foundation, Chapter 3 applies the DECIPHER method to uncover cell-cell interaction networks within the human breast tissue. Using this approach, we explore the dynamic interplay between different cell types within the mammary gland microenvironment in response to cycling hormone levels. In Chapter 4, I present the application of DECIPHER to investigate coordinated immune dysregulation in a rare autoimmune disease, Juvenile Dermatomyositis. By integrating immunophenotyping data and high-dimensional transcriptomic profiling, this chapter reveals the functional immune imbalance underpinning JDM, offering insights into potential therapeutic targets and personalized treatment strategies. Finally, in Chapter 5, I summarize and reflect on the body of work produced throughout my PhD, discuss the limitations and underlying assumptions of the computational methods I have developed, and propose future investigations to further the impact and scope of this work. Overall, this dissertation presents a generalizable approach to network inference in biological systems, leveraging machine learning methods to parse the networks of cellular coordination governing tissue homeostasis and disease. The findings presented herein hold promise for advancing our understanding of self-organization in living tissues, signaling dynamics, and disease pathogenesis, with potential implications for personalized medicine, biomarker discovery, and therapeutic intervention.

Review

“...[W]hat casts the pall over our victory celebration? It is the curse of dimensionality, the malediction that has plagued the scientist from earliest days”

– Richard Bellman

For next generation sequencing genomics data, dimensionality reduction is a standard component of single-cell data processing workflows,^{38–40} although its roots as an ML task are often glossed over in biological applications. As one would infer, dimensionality reduction techniques are used to decrease the complexity of high-dimensional datasets by identifying relevant features. For some applications, dimensionality reduction is sufficient as a ‘feature engineering’⁴¹ step, where it is deployed to reduce the computational load or improve the performance of downstream classification algorithms.^{42–44} However, dimensionality reduction techniques also have been successfully applied as insight-generating methods in biology.^{29,45,46} Matrix factorization, also referred to as matrix decomposition, refers to a broad class of dimensionality reduction techniques where the objective is to learn the basis vectors that recreate the original structure of the data. When matrix factorization is effectively done, i.e. arriving at ‘inherent dimensionality’, these basis vectors capture the relevant features while minimizing technical noise.

Principal component analysis (PCA), one of the pioneering dimensionality reduction techniques, collapses the data into linear components or features ordered by their contributed variation.⁴⁷ Because PCA has a unique convex solution, it has been shown to reconstruct data that can best be parametrized by linear combinations.

However, because PCA optimizes components based on maximum contributed variation, the output basis vectors do not always represent interpretable features of the original data.⁴⁶ For example, Lee and Seung demonstrated that when PCA is used to extract components from a training set of facial images, the basis images are holistic ‘eigenfaces’⁴⁸ that resemble highly distorted transformations of entire faces. Because the statistical properties of PCA encodings require each input image to be represented by a linear combination of all basis features, the individual eigenfaces are combined via a convoluted set of positive and negative coefficients, thereby obscuring the underlying meaning of many basis features. In contrast, when trained on the same set of facial images, non-negative matrix factorization (NMF) learned basis images that represent localized facial features.⁴⁶ Because of NMF’s non-negativity constraint, where basis vectors can only be additively combined rather than combined with arbitrary signs, the algorithm is able to extract features that intuitively represent parts of the input data it was trained on.

Learning these intuitively representative parts is referred to as ‘parts-based learning.’ In addition to the physiologically and cognitively representative advantages of parts-based learning,^{49–51} matrix factorization that emulates parts-based learning is advantageous for analysis of biological signals because the basis features are sparsely encoded.^{52–54} Examples of sparse biological signals amenable to parts-based learning include perturbations to epigenetic activity in cancer⁵⁴ and abnormal signals indicative of sleep disturbances.⁵² For genomics, sparse coding is reflected in the co-regulation of multiple pathways simultaneously, or the inherent low-dimensionality of gene regulation.⁵⁵ This low-dimensionality has been exploited at both the bulk and single-cell genomics level

to do module-level analysis that focuses on cellular processes rather than individual genes.^{56–58}

Naturally, module-level analysis lends itself to graph theory whereby hierarchical relationships can be represented and studied through network structure. Portraying these relationships between components as graphs allows us to bypass the limitations of numerical representations of probability which dictate that in order to make any inference, we must define a joint distribution of all combinations of probabilities.⁵⁹ In his seminal work *Probabilistic Reasoning in Intelligent Systems*, Judea Pearl highlights the inadequacies of formalizing human reasoning in this way:

Human performance shows the opposite pattern of complexity: probabilistic judgments on a small number of propositions...are issued swiftly and reliably, while judging the likelihood of a conjunction of propositions entails much difficulty and hesitancy. This suggests that the elementary building blocks of human knowledge are not entries of a joint-distribution table. Rather, they are low-order marginal and conditional probabilities defined over small clusters of propositions.⁶⁰

This observation asserts that we make sense of uncertainty in conjunction with empirical observations by processing low-dimensional blocks of knowledge, i.e. modules of relevant conditional properties. This reflects scientists' observation of functional co-regulation of biological processes, whereby dynamic function in living systems can be described by clusters of coordinated processes rather than individual pathways.^{29,54,58}

Beyond the psychologically meaningful representations networks allow, marrying probability to graph theory, herein referred to as Bayesian networks, has allowed researchers to parametrize inferred relationships by encoding uncertainty in the structure of networks themselves.⁶¹ Furthermore, by providing researchers the mathematical language with which they can model inferential reasoning, Bayesian networks also enable

scientists to map how new information propagates through the network as a property of the network's topology.⁶² This property of Bayesian networks—now termed 'belief propagation' describing the computational problem of quantifying dynamic inference updates as a linkage parameter of networks—has allowed them to revolutionize the field of 'modern AI.'^{61,63,64} Within the realm of biology and thus at the core of the DECIPHER method, Bayesian networks provide both a conceptual syntax and quantitative calculus to describe how biological functions are dynamically coordinated across lower-dimensional modules of activity.^{57,65,66}

**CHAPTER 2 DECIPHR: A DIMENSIONALITY
REDUCTION FRAMEWORK FOR INFERRING
NETWORKS OF CELLULAR COORDINATION IN
SINGLE-CELL GENOMICS DATA**

Abstract

Single-cell genomics technologies have revolutionized our ability to dissect the heterogeneity and dynamics of cellular populations, providing unprecedented insights into biological processes at the individual cell level. However, the analysis of these data poses significant computational challenges due to their high dimensionality and sparsity. In this chapter, we introduce DECIPHER, a dimensionality reduction framework tailored for inferring networks of cellular coordination from single-cell genomics data. This approach, implemented as an R package `deciphR`, serves as a versatile and interpretable tool for biologists seeking to integrate and gain biological insight into coordinated cell states from high dimensional genomics data.

Introduction

Tissues comprise a multiplicity of specialized cell types that must coordinate state changes in order to function collectively. State changes occur directly through local cell-to-cell signaling, or indirectly through response to a common microenvironment or systemic signals like hormones. We reasoned that coordinated changes in cell state could be detected as covarying gene expression programs between cell types and across large cohorts of heterogeneous single cell RNA sequencing data. When heterogeneity in the cohort arises from biological variation such as age, disease progression, or developmental time, we hypothesized that these correlations encode information about the biological processes coordinated across cells that drive tissue homeostasis and change.

However, detecting correlations between individual genes in RNA expression data is a statistical problem with high type I error, i.e. a high rate of false positives that only increases as the number of pairwise comparisons scales.^{67–69} There is a known tradeoff between minimizing the rate of Type I errors and decreased statistical power,⁷⁰ which is further exacerbated by multiple testing.^{71,72} Thus, the goal of identifying biological coordination in high dimensional gene expression data can be formulated as a multiple hypothesis testing problem. While methods to correct for the influence of multiple testing comparisons is an active area of statistical research,^{68,73–75} we instead sought to reduce the number of statistical comparisons made overall. To do this, we use a matrix factorization approach to reduce the dimensionality of the data from the order of individual genes to the order of co-varying gene sets.

This dimensionality reduction approach takes advantage of two key principles: 1) the previously outlined inverse relationship between Type I error and number of hypotheses tested and 2) the inherent low-dimensionality of eukaryotic gene expression. The latter has long been appreciated as a feature of the human genome^{58,76} whereby multiple genes are coordinated such that functional programming of biological processes can be observed as patterns in up- and down-regulated gene expression. With the advent of next generation sequencing, this low dimensionality has been shown to enable the detection of robust biological programs in noisy single-cell RNA expression data even when samples are sequenced at a shallow read depth.⁵⁵

Given this, we developed the DECIPHER algorithm to detect correlated changes in the expression of ‘biological activity programs,’ or activity programs, with the hypothesis that in this lower dimensional space, a subset of these correlations represent biologically meaningful coordinated cell states that underly tissue homeostasis in dynamic living systems. After learning these activity programs in an unsupervised manner from the data, we then apply network analysis to aid in the biological annotation and interpretation of the dimensionality reduction results. The end output reveals a graph representation that maps how biological processes are coupled in tissue and cooperatively change according to heterogeneous conditions across the cohort such as age, disease progression, or developmental time. We present *deciphR*, a user-friendly package implementing the DECIPHER algorithm in R with integrated visualization functionalities and interoperability with Seurat, a popular single-cell analysis pipeline.³⁹

System and methods

An overview of the DECIPHER algorithm, can be seen in **Figure 2.1** below. Briefly, the workflow consists of two phases. First, the method extracts activity programs from the single-cell expression data using a consensus matrix decomposition. Next, the method applies network analysis to annotate and identify biologically relevant hubs, or ‘modules’, of activity programs. The code for the *deciphR* package will be available on Github at ‘GartnerLab/deciphR’ upon preprint release of this chapter’s corresponding manuscript.

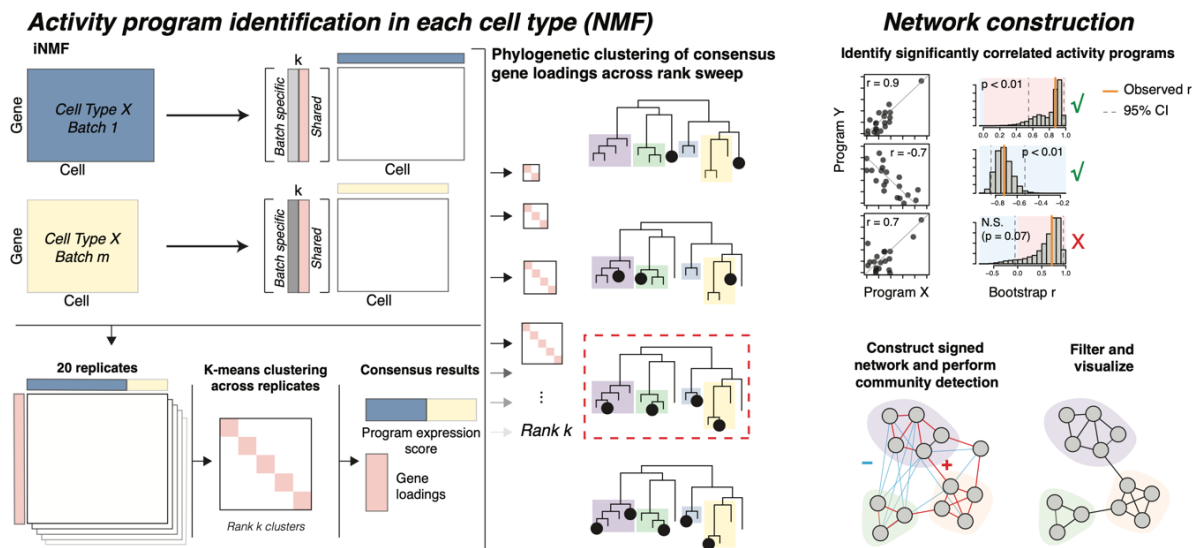


Figure 2.1 Overview of DECIPHER method.

Activity programs, or GEPs as we also refer to them, are identified through consensus NMF. Briefly, a k -sweep of i NMF is run on expression data subset by major cell type and batch, if any, for R replicates (20 by default). Results at each rank k are aggregated across replicates such that consensus matrices are output for gene loadings (W), batch-specific expression loadings (V), and shared cell loadings (H). The optimal rank is selected using the DEW metric calculated from the phylogenetic trees constructed from consensus gene loadings W across the k -sweep. Next, network of correlated cell states is constructed from significantly correlated activity programs from NMF results at optimal k_{DEW} .

Feature extraction of biological activity programs via coinMF

We apply non-negative matrix factorization (NMF) to reduce the dimensionality of the single cell data at a resolution appropriate for identifying activity programs.⁴⁶ In this implementation, NMF seeks to decompose the input matrix of data into two sets of orthogonal vectors: one set of gene loading vectors, W , comprised of weights that quantify how much a given gene contributes to that feature or activity program, and a second set of cell loading vectors, H , comprised of weights that represent how strongly that activity program is expressed in a given cell. As a dimensionality reduction technique, NMF is distinct from principal component analysis in that there is no single solution for the number of patterns or components into which the data is segmented. As such, it is necessary to optimize the parameter 'rank k ' via a 'k-sweep' such that the NMF results capture the relevant biology at an appropriate granularity.

At each rank k , we deploy a consensus NMF approach first described by Kotliar et al. to ensure that the downstream matrix factorization results are robust to multiple random initializations of the algorithm.⁷⁷ We combine this consensus approach with integrative NMF (iNMF)⁷⁸ that takes advantage of online learning⁷⁹ to decompose both batch-specific latent factors and shared latent factors across samples and batches. Hereafter, we refer to this consensus online iNMF approach as 'coinMF.' We chose to take advantage of an implementation of NMF that uses online learning for two reasons: first, improved computational performance over batch learning algorithms by segmenting the learning task into 'mini-batches' and second, the possibility of updating the learned

latent factors as new data is added rather than re-running the NMF procedure on the expanded dataset.⁷⁹

To optimize the choice of rank, we propose a novel metric for parametrizing the matrix decomposition, which we call the Depth Encompassing Weight or ‘DEW.’ First, the consensus gene loadings across all decompositions from the k-sweep are clustered phylogenetically.^{80,81} We then compute the number of phylogenetic subtrees captured at a given rank k, weighted by the depth of the subtrees. We proposed this metric based on two principled assumptions about the optimal rank K: first, that the optimal rank captures the appropriate breadth of biological activity programs, represented as phylogenetic subtrees; second, that meaningful biological features are robust across several ranks k, represented as the depth of these subtrees. Thus, the optimal decomposition of the data extracts features at an appropriate coverage and depth as parametrized by the metric DEW:

$$DEW = \frac{t_{phylo}}{w_{depth}}$$

Previously we reported selection of rank k according to this procedure as simply the saturation of this depth-weighted subtree metric, identifiable as the inflection point on the curve such as below.⁶⁵ However, depending on the underlying structure of the input data, it became evident that plotting this metric often results in ‘jittery’ curves characteristic of discrete data rather than continuous curves, leading to ambiguous manual selection of k_{DEW} . The *deciphR* package release further advances the original method by implementing an automated k-selection procedure based on the *kneedle* algorithm for identifying inflection points in discrete data.⁸² We later demonstrate that the optimized *kneedle*-based k selection procedure identified the appropriate rank k in a real dataset from

a highly heterogeneous patient cohort, in Chapter 4, while still allowing for a range of multiple appropriate factorization ranks as with the original method.

Unsupervised network inference of coordinated biological processes

Using the decomposition at the selected k_{DEW} , we calculate the pseudo-bulk average expression of each program, i.e. the per-sample expression of each activity program, H . Then, the method constructs a correlation matrix from all pair-wise combinations of the filtered activity program matrix of expression, H , with significance determined by a bootstrapping procedure for re-sampling pairs of programs.⁸³

An adjacency matrix is constructed from the statistically significant correlations, and transformed into a force-directed network, where activity programs are represented by nodes and correlations are represented by edges, with only positive significant correlations shown in visualizations for clarity. However, both the sign and magnitude of significant correlations are accounted for in the final structure of the network, such that positively correlated programs are ‘pushed’ closer together and negatively correlated programs are ‘pulled’ further.

To reflect the fundamental observation that biology consists of higher-order organized processes than simple pairs of interacting pathways, we implement a community detection algorithm based on a Constant Potts’ model of modularity⁸⁴ and quantify each node’s connectivity to the rest of the metric, as parametrized by weighted topological overlap (wTO).⁸⁵ A parameter sweep of modularity is performed to optimize

the community detection resolution using the 'leidenalg' package.⁸⁶ The choice of a module formulation for further analysis was based on the assumption that biologically meaningful programs are tightly coupled to other programs and conversely, less relevant programs are isolated and poorly connected to the rest of the network. For applications where one is interested in biological activity programs that appear to turn on and off in isolation, perhaps representing putatively rare phenotypes, DECIPHER is likely not the appropriate tool.

Once the structure of the network and its modules are determined, the network can be annotated on both the gene loading and cell loading level. Marker scores that quantify the strength of contribution of an individual gene to a given activity program can be calculated as the ordinary least squares regression coefficient between each gene's z-scored expression and the expression score of that program in the given cell type.⁷⁷ This enables comparison of each gene's contributions across activity programs, which could not be done with the raw gene loadings due to the non-negativity constraint of the matrix decomposition.⁴⁶ Gene lists of these ranked marker scores can then be used as input to standard gene set enrichment analysis pipelines for biological annotation.

The DECIPHER algorithm builds on these gene set enrichment results by then incorporating the network's module structure to inform data-driven and unsupervised annotation of the network. We quantify the uniqueness of a given gene set to a module, or module enrichment as we call it, as the likelihood P of that gene set being enriched, g , in as many or more neighboring nodes than it currently is within that module, G , if module membership were randomly assigned, M , from the whole network with N nodes:

$$P(g \geq G) \text{ when } \binom{N}{M}$$

Thus, the module enrichment metric can then be used to assign biological processes to each of the network's modules in a principled manner that takes advantage of both the wealth of curated gene annotation databases available^{87,88} and the inherent network structure of the dataset.

Similarly, the metadata describing biological conditions of the samples, such as spatial location or disease activity can be associated specific nodes or modules within the network. Depending on the kind of biological variable, i.e. categorical, discrete numerical, continuous, etc. and the sample distribution in the dataset, various statistical tests can be deployed to quantify the association between a biological condition and program expression. For example, to associate hormone receptor signaling status with components of the network derived from reduction mammoplasty tissue, we tested the association between individual's oral contraceptive use and expression of each activity program in the network.⁶⁵ The *deciphR* package includes several built-in functions to run this biological meta-analysis of the network.

Algorithm

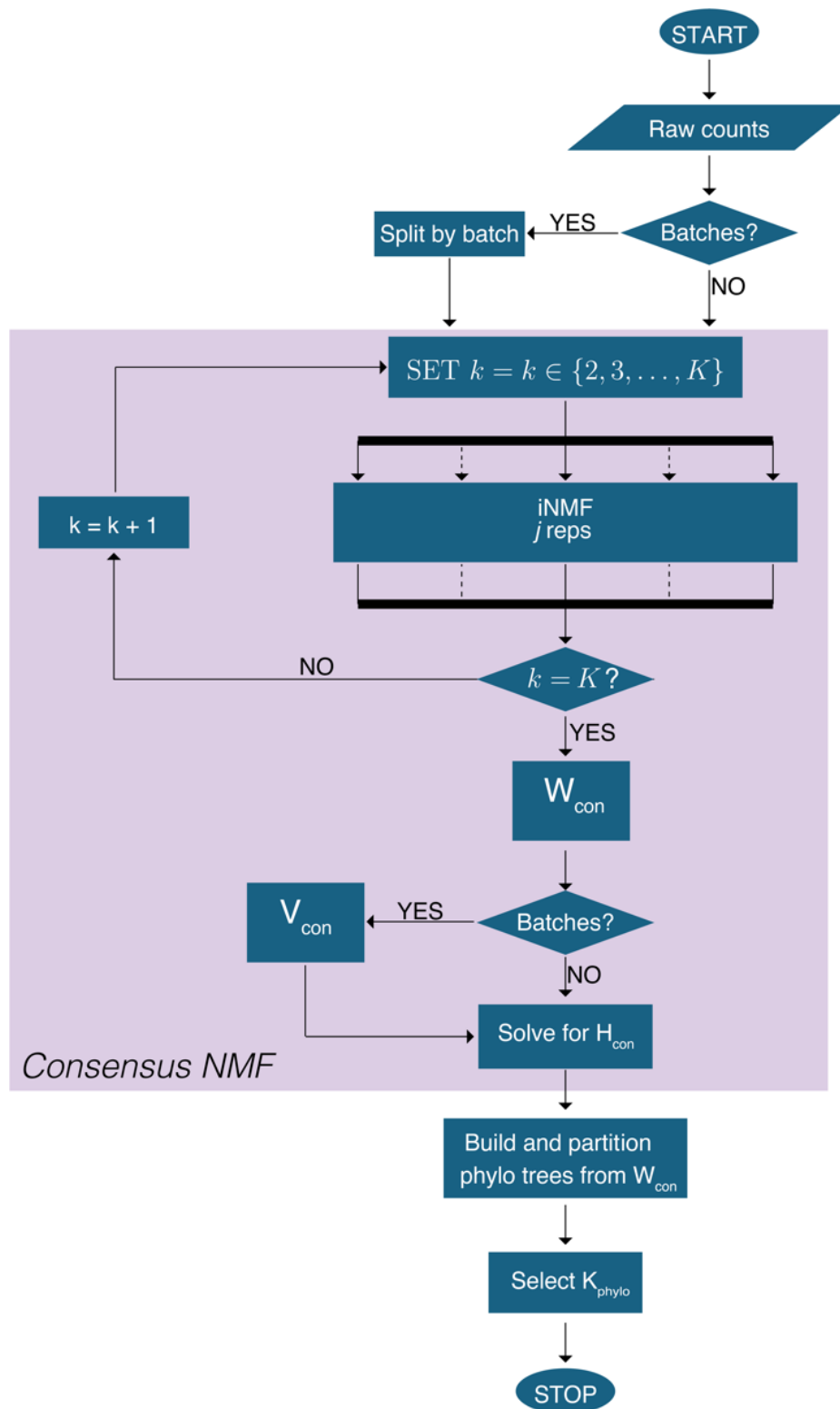


Figure 2.2 Flowchart depicting algorithm for the dimensionality reduction part of the DECIPHER workflow.

coinMF Algorithm

1. Input data: raw counts $\tilde{\mathbf{X}}$
2. If batches, split by batch.
3. Normalize and scale counts.
4. For each $k = k \in \{2, 3, \dots, K\}$:
 - a. At each initialization $r \in R$ replicates of iNMF:

$$\mathbf{W}^{(r)} \times \mathbf{H}^{(r)} \approx \tilde{\mathbf{X}} \text{ for } r = 1 \dots R$$

$$\mathbf{H}^{(r)} = \arg \min_{\mathbf{H} \geq 0} \left\| \begin{pmatrix} \mathbf{W}^{(r)} + \mathbf{V}^{(r)} \\ \sqrt{\lambda} \mathbf{V}^{(r)} \end{pmatrix} \mathbf{H}^{(r)\top} - \begin{pmatrix} \tilde{\mathbf{X}} \\ \mathbf{0}^{m \times n} \end{pmatrix} \right\|_F^2$$

5. Compile consensus gene loadings \mathbf{W} and \mathbf{V} , if batches, at each rank $k \in K$:
 - a. Concatenate L2 normalized shared gene loadings $\mathbf{W}^{m \times k}$, $\tilde{\mathbf{W}}$, and batch-specific loadings $\mathbf{V}^{m \times k}$, $\tilde{\mathbf{V}}$, across R replicates into \mathbf{W} and \mathbf{V} where each row is a component from one of the NMF replicates:

$$\mathbf{W} = \begin{bmatrix} \tilde{\mathbf{W}}^{(1)} \\ \vdots \\ \tilde{\mathbf{W}}^{(r)} \\ \vdots \\ \tilde{\mathbf{W}}^{(R)} \end{bmatrix} \text{ and } \mathbf{V} = \begin{bmatrix} \tilde{\mathbf{V}}^{(1)} \\ \vdots \\ \tilde{\mathbf{V}}^{(r)} \\ \vdots \\ \tilde{\mathbf{V}}^{(R)} \end{bmatrix}$$

- b. Filter components of \mathbf{W} and \mathbf{V} with outlier mean Euclidean distance from their 6 nearest neighbors, with outliers determined by the threshold third quartile plus 1.5 the interquartile range $q_{0.75} + 1.5IQR$.
 - c. Let these filtered components be $\mathbf{W}^{(f)}$ and $\mathbf{V}^{(f)}$. Next, group the rows of $\mathbf{W}^{(f)}$ and $\mathbf{V}^{(f)}$ using k-means clustering, with the same number of clusters set to the given NMF rank k .

- d. Collapse each cluster of shared and batch-specific gene loadings to consensus vectors $\overline{\mathbf{W}}$ and $\overline{\mathbf{V}}$ by taking the median value for each gene j across components κ in that cluster K .

$$\overline{\mathbf{W}}_{j\kappa} = \text{median} \left(\{W_{j\kappa}^{(f)} \text{ for } \kappa \in K\} \right)$$

$$\overline{\mathbf{V}}_{j\kappa} = \text{median} \left(\{V_{j\kappa}^{(f)} \text{ for } \kappa \in K\} \right)$$

with $\widetilde{\mathbf{W}}$ and $\widetilde{\mathbf{V}}$ representing the normalized consensus components.

6. Solve for consensus cell loadings H , at each rank $k \in K$:

The consensus cell loading matrix $\overline{\mathbf{H}}$ is solved for by using one last iteration of NMF initialized with the consensus shared $\overline{\mathbf{W}}$ and batch-specific $\overline{\mathbf{V}}$ gene loading matrices.

$$\overline{\mathbf{H}} = \arg \min_{H \geq 0} \left\| \begin{pmatrix} \overline{\mathbf{W}} + \overline{\mathbf{V}} \\ \sqrt{\lambda} \overline{\mathbf{V}} \end{pmatrix} \overline{\mathbf{H}}^\top - \begin{pmatrix} \widetilde{\mathbf{X}} \\ 0^{m \times n} \end{pmatrix} \right\|_F^2$$

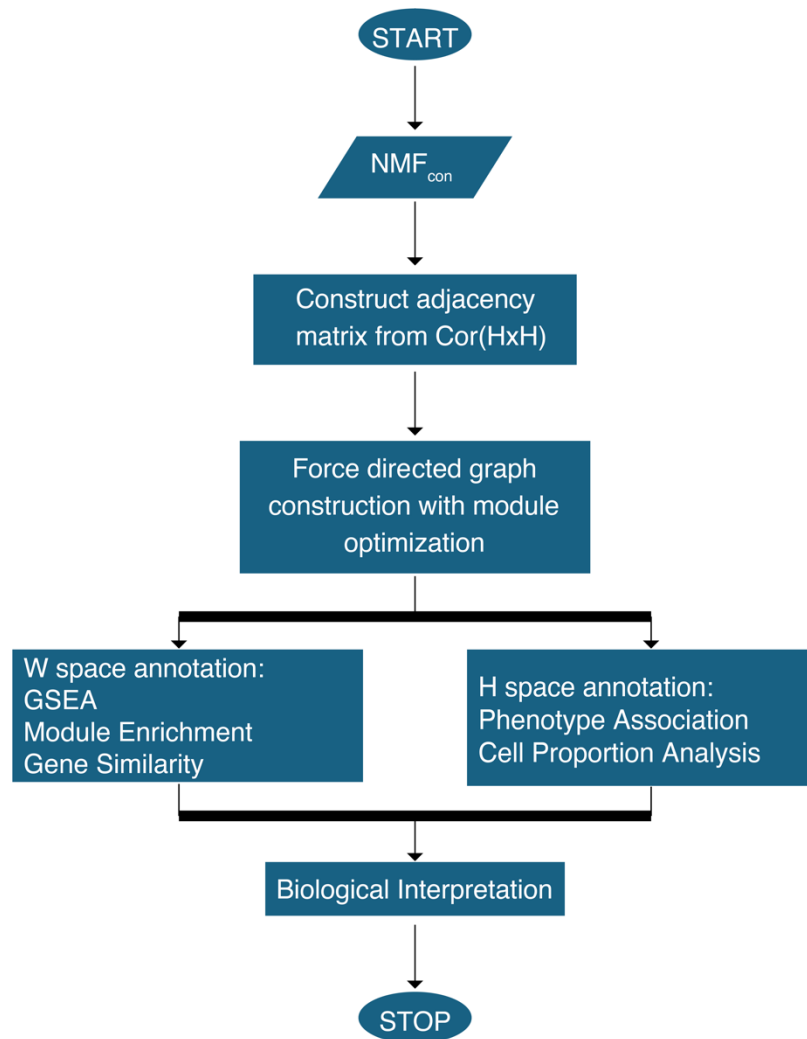


Figure 2.3 Flowchart depicting the algorithm for network construction and annotation within the DECIPHER workflow.

Discussion

Here, we present a method DECIPHER for inferring networks of cellular coordination in single-cell sequencing data. DECIPHER is a method based on dimensionality reduction that takes advantage of fundamental properties of the data: first, the low dimensionality of gene expression owing to patterns of covariation indicative of co-regulated pathways and second, the decreased rate of false positives when the number of statistical comparisons is decreased. We deploy a novel metric, the Depth Encompassing Weight, for optimizing the dimensionality reduction of the data and implement an algorithm to automatically suggest the proper dimensionality parameter. Finally, we show in simulated data that biological activity can be detected via NMF as parametrized by the rank selected according to the DEW metric. Overall, this method provides an interpretable approach to nominating hypotheses from high-dimensional data in a principled manner, thereby addressing a key need in the field for methods that allow researchers to go from big data to biological insight.

CHAPTER 3 DECIPHER UNCOVERS CELL-CELL INTERACTION NETWORKS IN THE HUMAN BREAST

Abstract

The rise and fall of estrogen and progesterone levels across menstrual cycles and during pregnancy regulates postpubertal breast development and modifies cancer risk. How these hormones uniquely impact each cell type in the breast is not well understood, because many of their effects are indirect—acting through paracrine interaction networks. Here, we apply DECIPHER, a computational approach that leverages inter-sample heterogeneity in scRNA-seq data to identify sets of cellular activity programs in multiple cell types that co-occur across samples. Applied to a dataset of 28 healthy reduction mammoplasty tissue samples, DECIPHER identifies a network of correlated activity programs that represent the dynamic tissue-level response of the human breast to changing hormone levels.

Introduction

Coordinated interactions between cells are essential for the development and maintenance of normal tissue function, and dysregulation of cell-cell interactions is a key driver of disease. In the human breast, fluctuations in the levels of estrogen and progesterone with each menstrual cycle and during pregnancy control cell growth, survival, differentiation, and tissue morphology. The impact of these changes is profound: cumulative lifetime exposure to cycling hormones is a major modifier of breast cancer risk⁴, and the majority of breast tumors are estrogen-dependent. However, many of the effects of ovarian hormones within the breast are indirect. The estrogen and progesterone receptors (ER/PR) are expressed in only 10-15% of cells within the epithelium³, and most of the changes that occur in response to hormone receptor activation are mediated by a complex cascade of paracrine signaling from hormone-responsive (HR+) cells to other cell types in the breast. Accordingly, cell-cell interactions between HR+ cells and other cell types are key to normal breast morphogenesis. However, a systems-level understanding of how different cell populations in the breast respond to cycling hormone levels remains unclear.

A key challenge is that the human breast is both heterogeneous across individuals and characterized by a highly dynamic microenvironment. There is a high degree of variability between individuals in terms of epithelial architecture,⁸⁹ cell composition,^{90,91} and hormone-responsiveness,⁹²⁻⁹⁴ and these differences likely impact both normal breast function and breast cancer susceptibility. Within individuals, the menstrual cycle and pregnancy/lactation/involution cycle are major drivers of epithelial remodeling, characterized by alternating periods of epithelial expansion and regression in response

to changing hormone levels.^{89,95–97} Histological analyses of paraffin-embedded human tissue sections have also identified cyclical alterations in epithelial architecture and stromal organization across the menstrual cycle^{98,99} and broad remodeling following weaning.^{100,101} However, little is known about how this underlying heterogeneity impacts cell state and the intercellular signaling networks that control tissue morphogenesis. As it enables unbiased analysis of cell types within the human mammary gland at single-cell resolution, scRNA-seq is particularly well-suited to investigate this problem.

To provide insight into the cellular interactions that regulate breast tissue homeostasis, we applied DECIPHER, a computational approach that leverages inter-sample transcriptional heterogeneity to identify coordinated interaction networks across cell types in scRNA-seq datasets. We applied DECIPHER to a dataset consisting of twenty-eight premenopausal reduction mammoplasty tissue specimens. Based on this approach, we identified a network of coordinated gene activity programs in HR+ cells and other cell types that represent the dynamic tissue-level response of the human breast to changing hormone levels. Using differences in cell-type proportions across samples, we infer a subset of activity programs that depend on direct cell-to-cell signaling and find that these direct interactions primarily comprise signaling from HR+ cells to other cell types. Second, we use DECIPHER to generate new hypotheses about how person-to-person variation at the tissue level is linked to specific biological mechanisms at the cellular level. Overall, these results provide a comprehensive map of the cycling human breast and the dynamic cell-cell interactions that underlie normal breast function and breast cancer risk.

Results

Person-to-person variability in transcriptional cell state in the human breast

To identify inter-individual differences in transcriptional cell state in the human breast, we performed scRNA-seq analysis on 86,136 cells collected from 28 healthy premenopausal donors who underwent reduction mammoplasty surgery (**Figure 3.1, Supplemental Tables 3.1, 3.2**). To obtain an unbiased snapshot of the epithelium and stroma, we collected live (DAPI negative) singlet cells from all samples by fluorescence activated cell sorting (FACS). For a subset of samples, we also collected purified epithelial cells or purified luminal and basal/myoepithelial cells. We used MULTI-seq barcoding and *in silico* genotyping for sample multiplexing to minimize technical variability between samples (Chapter 3, *Methods*).^{19,102}

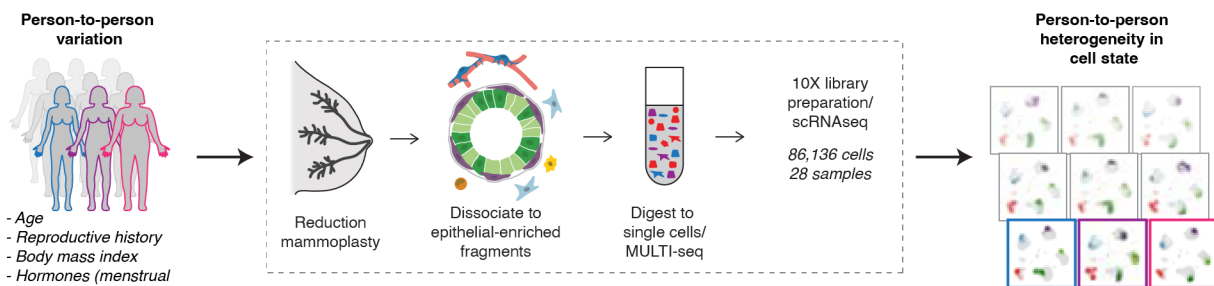


Figure 3.1 Single-cell transcriptional analysis links biological variables with person-to-person heterogeneity in transcriptional cell state in the premenopausal human breast.

Overview of scRNA-seq workflow: Reduction mammoplasty samples were processed to epithelial-enriched tissue fragments, then to single cells, followed by MULTI-seq sample barcoding, library preparation using the 10X Chromium system, and sequencing.

Sorted basal and luminal cell populations were well-resolved by UMAP. Unsupervised clustering identified one basal/myoepithelial cluster, two luminal clusters, and six stromal clusters (**Figure 3.2**). Based on the expression of known markers, the two

luminal clusters were annotated as hormone-responsive (HR+) and secretory luminal cells, and the six stromal clusters were annotated as fibroblasts, vascular endothelial cells, lymphatic endothelial cells (“lymphatic”), smooth muscle cells/pericytes (“vascular accessory”), lymphocytes, and macrophages (**Figure 3.2**).

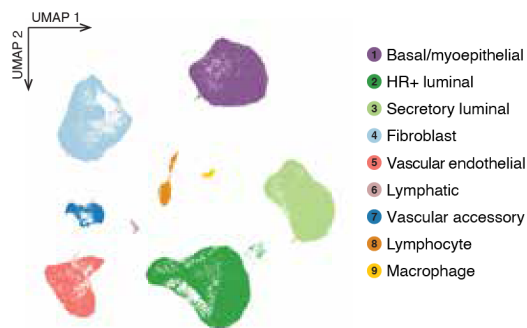


Figure 3.2 UMAP constructed from RNA expression of single cells isolated from breast tissue, colored by major epithelial or stromal cell type.

The luminal populations described here closely match those identified as “hormone-responsive/mature luminal” and “secretory/luminal progenitor” in previous scRNA-seq analyses of the human.^{103,104} Here, we use the nomenclature “hormone-responsive/HR+” and “secretory” to refer to these two luminal cell types. The HR+ cluster was enriched for the hormone receptors ESR1 and PGR, and other known markers such as ANKRD30A.¹⁰⁴ Consistent with previous studies demonstrating variable hormone receptor expression across the menstrual cycle,¹⁰⁵ expression of ESR1 and PGR transcripts were sporadic and often non-overlapping. Within the HR+ luminal cluster, 22% of the cells had detectable levels of ESR1 or PGR, with only 2% of hormone-responsive cells expressing both transcripts.

Inferring shared transcriptional responses and direct cell-to-cell signaling interactions in the human breast

Since estrogen and progesterone are master regulators of breast development, and the levels of these hormones fluctuate across the menstrual cycle, we predicted that ER/PR signaling and the downstream paracrine response would be a major source of transcriptional heterogeneity across samples in our dataset (**Figure 3.3**). Based on random sampling across the menstrual cycle and differences in hormonal contraceptive use, we would expect to identify samples with varying levels of ER/PR activation in hormone-responsive (HR+) luminal cells. If these hormone-responsive cells are signaling to cell types, such as basal cells, we would further expect to see a second activity program in those cells representing the downstream paracrine response. Finally, this “paracrine response” activity program should co-vary with the level of ER/PR activation across different samples (**Figure 3.3**). Thus, we developed DECIPHER based on the hypothesis that inter-sample transcriptional variation contains meaningful information about how the behaviors of different cell types are coordinated at the tissue level, and that transcriptional signatures (“activity programs”) representing interactions between two cell types should correlate across samples.

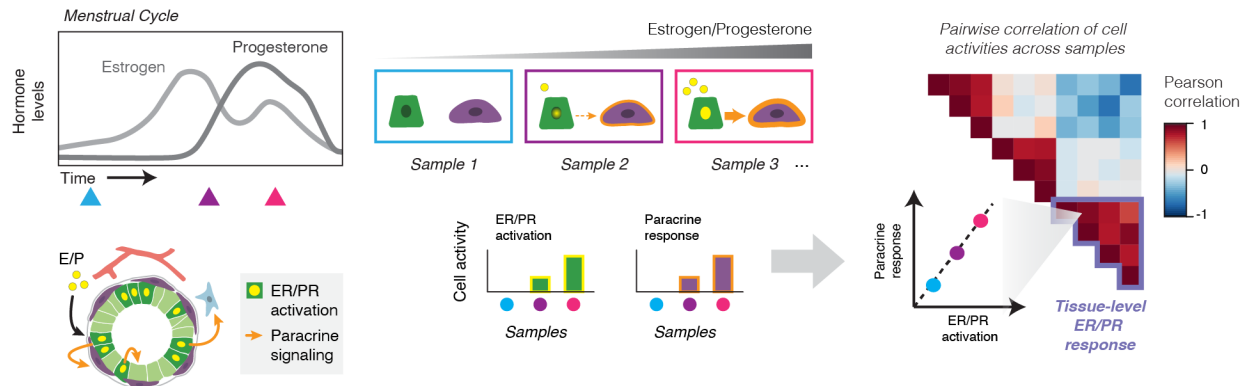


Figure 3.3 Conceptual overview of analysis approach for human breast dataset using DECIPHER. We hypothesized that hormone receptor activation in hormone-responsive (HR+) luminal cells would correlate with transcriptional changes in other cell types, representing the downstream paracrine response. Based on differences in hormone levels due to menstrual cycling (depicted, left) or hormonal contraceptive use, we predicted that gene expression programs representing ER/PR signaling in HR+ luminal cells and the downstream signaling response in other cell types would co-vary across samples.

DECIPHER identifies activity programs within cell types in scRNA-seq data and uses individual pairwise correlations between activity programs to build a higher-order map of coordinated cell-state changes (**Figure 3.4**). Using this network view, DECIPHER identifies modules of activity programs representing transcriptional states that co-occur across the same sets of samples. In downstream analyses, we infer modules that are enriched for direct cell-cell signaling interactions (i.e. modules containing links that depend on the proportion of one cell type across samples), or driven by non-cell-type specific responses to shared microenvironmental signals (i.e. modules containing transcriptionally similar activity programs) (**Figure 3.4**). We define individual activity programs and modules by performing gene set enrichment analysis, which allows us to infer higher-order functional interactions between multiple cell types. Finally, we uncover associations between annotated metadata features and sets of activity programs to infer potential sources of biological variation (**Figure 3.4**).

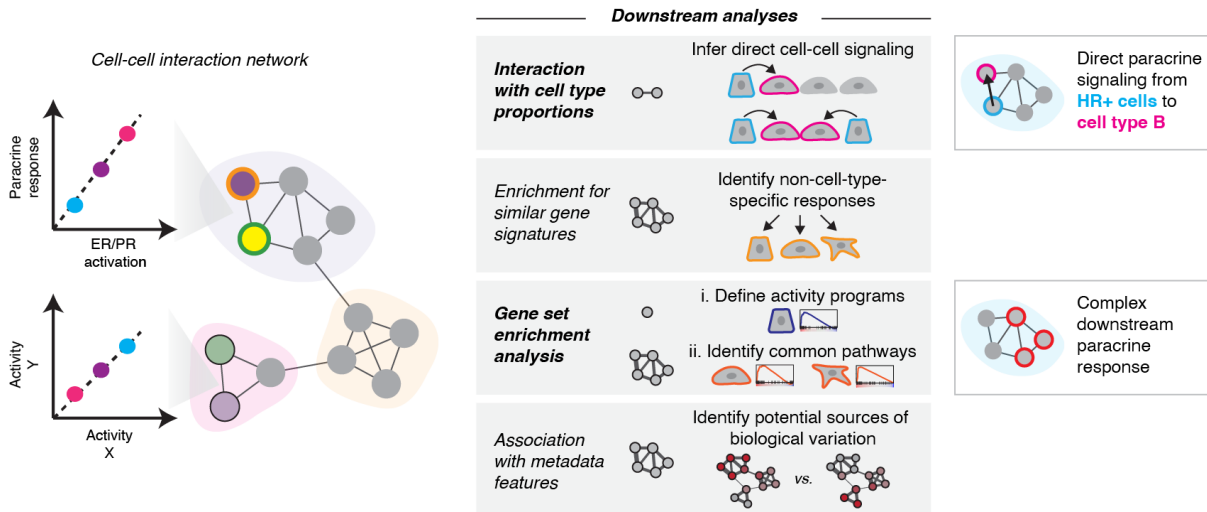


Figure 3.4 Using individual pairwise correlations between cell activities, DECIPHER builds a tissue-level map of the cell-cell interactions present in the healthy human breast and identifies modules of transcriptional states that co-occur across the same sets of samples. In downstream analyses, we uncover modules driven by non-cell-type specific responses to shared signals, or enriched for putative direct cell-cell signaling interactions. We define activity programs using gene set enrichment analysis, identify common pathways enriched across activity programs in a module, and uncover potential sources of biological variation by testing association with annotated metadata features.

To identify activity programs within cell types in the premenopausal breast, we performed non-negative matrix factorization (NMF) on each of the major cell type clusters represented in our dataset.^{77,78,106} To account for batch differences in our dataset, we used integrative NMF,^{78,106} and performed subsequent gene set enrichment analyses on the shared—rather than batch-specific—components of each activity program. This approach successfully corrected for batch differences while retaining sample-to-sample transcriptional variability. As solutions to NMF are non-unique, we adapted a consensus matrix factorization approach⁷⁷ to identify activity programs that were consistent across replicates (**Figure 2.1**, Chapter 3 *Methods*). A key parameter in matrix factorization is number of activity programs found for each cell type (rank, k). None of the three commonly used heuristics for guiding the choice of K identified an obvious “elbow” in our dataset (**Supplemental Figure 3.11**). We therefore deployed a new metric for choosing k , the Depth Encompassing Weight (*DEW*), based on the goal of identifying the greatest number of robust (i.e. consistent across values of k) and unique (i.e. distinct from other

programs at the same k) activity programs. We perform consensus iNMF over a range of ranks, build a phylogenetic tree based on the correlation matrix of gene loadings across all ranks, and partition the phylogenetic tree into subtrees to define distinct sets of programs (**Figures 2.1, Supplemental Figure 3.12**, Chapter 3 *Methods*).⁸¹ After filtering “outlier” activity programs that are expressed in only rare contaminating cells (Chapter 3 *Methods*), we choose the optimum k according to the *DEW* (k_{DEW}) as the point at which increasing the granularity of matrix factorization does not identify activity programs that comprise major new subtrees (**Supplemental Figures 3.11, 3.13**, Chapter 3 *Methods*).

Finally, to build a network map of cell-cell interactions, we quantified the average expression of each cell type-specific activity program for each sample and constructed a weighted network of coordinated activity programs based on the pair-wise Pearson correlation (**Figure 2.1**). To remove spurious correlations driven by outlier samples, we used bootstrap resampling to estimate confidence intervals associated with each correlation coefficient and transformed the resulting Pearson correlation matrix into a signed weighted adjacency matrix by setting all Pearson correlation coefficients with p -values greater than 0.05 to zero. We identified modules of highly correlated gene expression programs using a community detection algorithm for signed graphs.⁸⁴

This approach identified eight major modules comprising highly correlated transcriptional states across cell types in the breast (**Figure 3.5**). Consistent with our goal of choosing the rank k for each cell type that captured the greatest number of unique activity programs (**Supplemental Figure 3.13**), the overall organization of modules into cell-cell interaction networks remained highly robust to the choice of rank at values of $k \geq$

k_{DEW} , whereas the network structure at $k \leq k_{DEW}$ had much sparser connections between modules (Supplemental Figure 3.14).

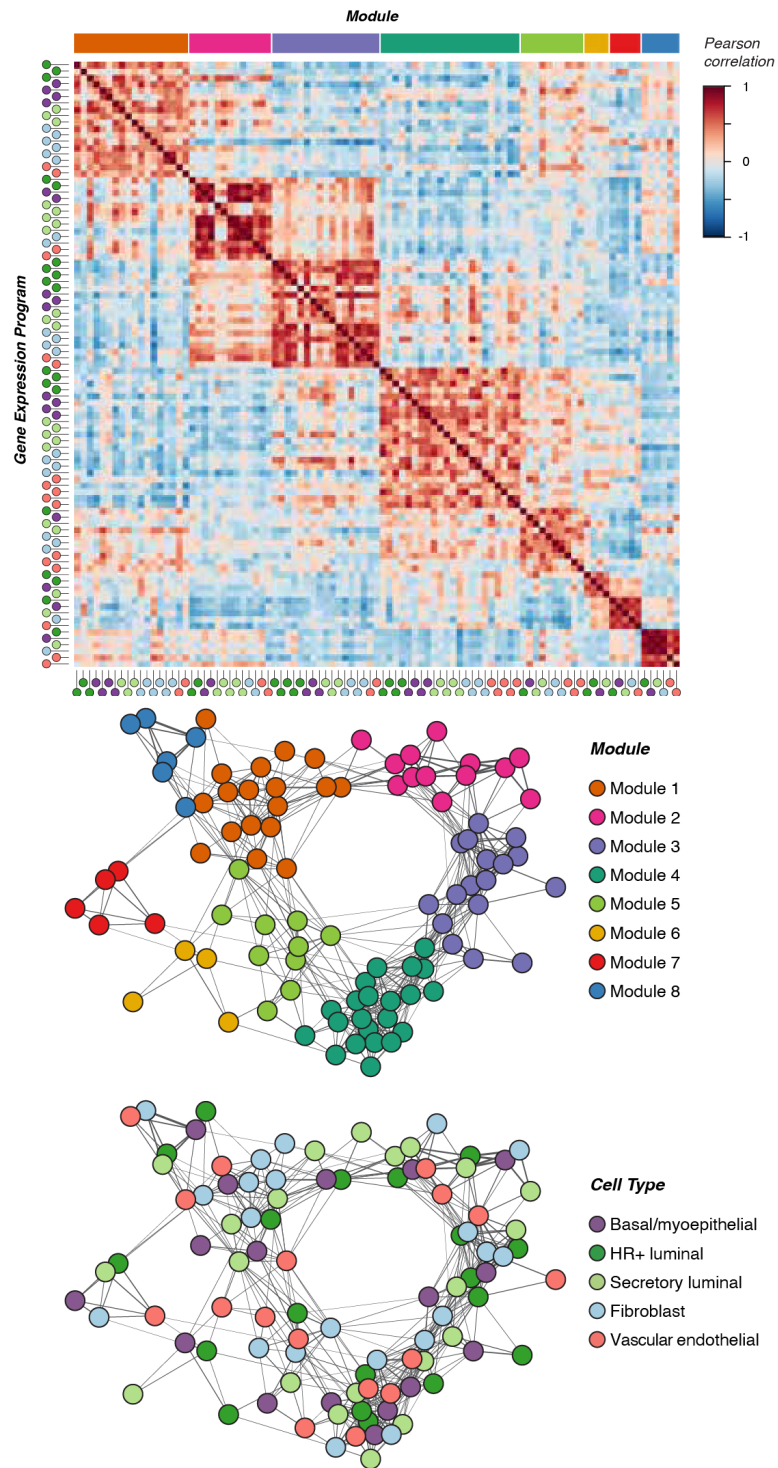


Figure 3.5 DECIPHER identifies cell-cell interaction networks across cell types in the human breast. (Figure caption continued on next page)

(Figure caption continued from previous page)

Top: Heatmap depicting Pearson correlation coefficients between activity programs in the eight major modules identified by DECIPHER. Bottom: Network graphs of correlated activity programs in the human breast. Nodes represent distinct activity programs in the indicated cell types, and edges connect significantly correlated programs (Pearson correlation coefficient > 0, $p < 0.05$). Modules of correlated programs were identified using a Constant Potts Model for community detection.

To interpret this network, we reasoned that the activities of two cell types can be coordinated in a tissue in multiple ways. First, cells can respond to the same microenvironmental cues (or the loss of cues), leading to either cell-type specific or non-cell-type specific transcriptional responses in each cell. Second, cells can engage in direct cell-to-cell signaling, where a transcriptional response in one cell type leads to a change in a second cell type. In the premenopausal breast, fluctuating levels of estrogen and progesterone with each menstrual cycle control cell growth, survival, differentiation, and tissue morphology. As only a subset of cells expresses ER and/or PR, most of these changes are mediated by a complex cascade of paracrine signaling originating in HR+ cells that goes on to affect other cell types. Therefore, we expect the tissue-level response to hormones to lead to at least two types of coordinated interactions in the breast: direct cell-to-cell signaling interactions between HR+ cells and other cell types, and more complex downstream interactions involving cell-type-specific responses to a shared microenvironment—which we predict would involve transcriptionally unique (e.g. cell-type-specific) activity programs that may be enriched for similar biological processes.

To identify non-cell-type specific transcriptional responses—that are unlikely to be directly related to hormone signaling in the breast—we identified modules made up of activity programs with similar gene loadings (**Figure 3.4**). We found that Modules 7 and 8 were highly enriched for activity programs with correlated gene loadings (**Figure 3.6**). Programs in Module 7 primarily consisted of ribosomal transcripts and genes involved in cellular respiration, whereas programs in Module 8 consisted of stress response genes

such as heat shock and chaperone proteins (**Figure 3.6**). We speculate that Module 8 represents an artifact of tissue processing rather than biologically meaningful transcriptional variation, as prior studies have identified a similar signature in dissociated solid tissues.¹⁰⁷ Notably, as DECIPHER describes cells as a combination of activity programs rather than forcing cells into distinct clusters, samples with high expression of “dissociation-related” activity programs may also contain biologically meaningful signals from other programs.

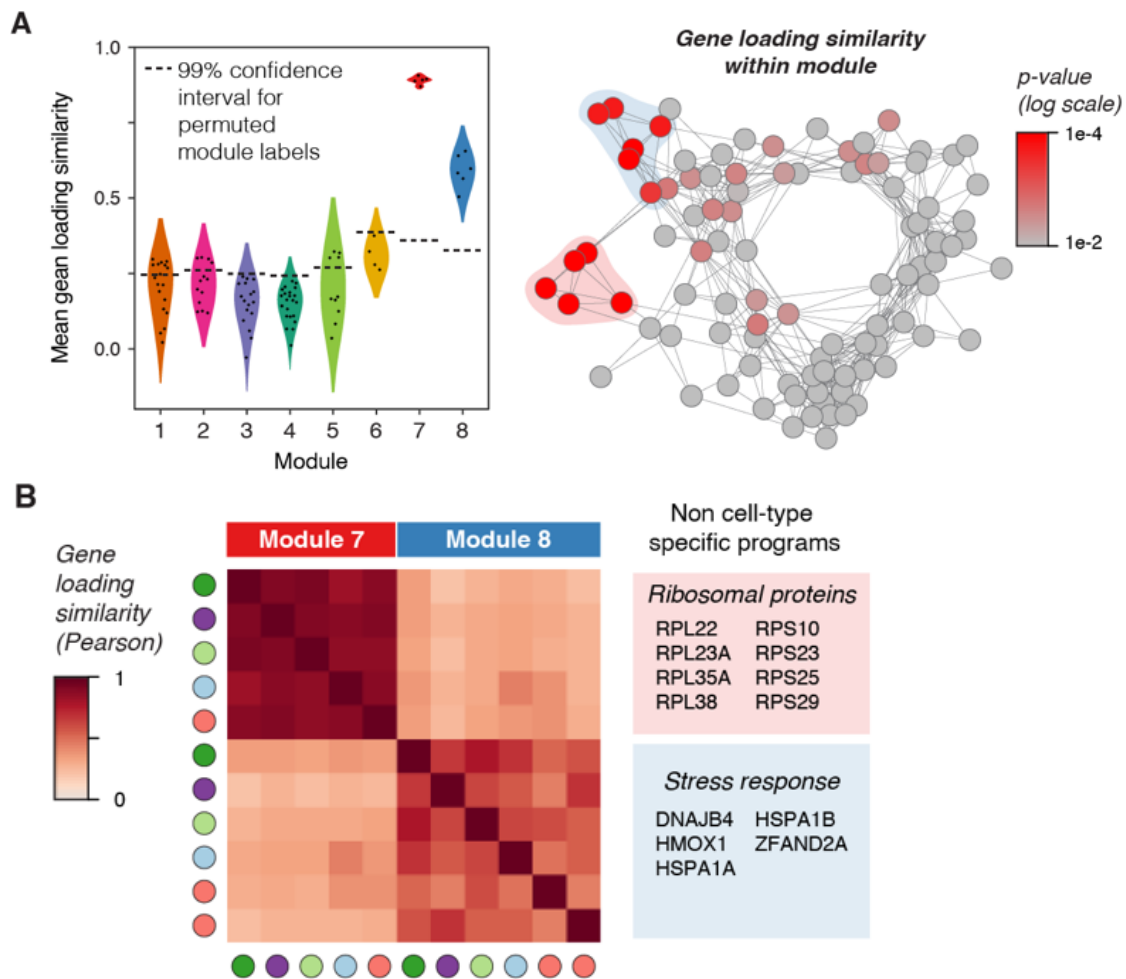


Figure 3.6 Inferring non-cell-type-specific transcriptional responses in the human breast. (A) Left: Violin plot of the mean Pearson correlation between gene loadings for each activity program and all other activity programs in the same module (“gene loading similarity”). The horizontal dashed line represents the 99% confidence interval for permuted module labels. Right: Network graph of activity programs in the human breast, colored by the *p*-value for gene loading similarity for each program (log scale). *P*-values were calculated by permutation testing. (Figure caption continued on next page)

(Figure caption continued from previous page)

(B) Heatmap depicting Pearson correlation coefficients between gene loadings for the indicated activity programs. The colored boxes list the top-loading genes shared by all programs in the indicated modules.

Next, we inferred modules enriched for putative direct cell-cell signaling interactions, by identifying links between two nodes that depended on both the magnitude of activity program expression in a “sender” cell type and the proportion of that sender cell type in the tissue (**Figure 3.4, Figure 3.7**). We reasoned that if one cell type was signaling to another, the activity program representing the transcriptional response in the “receiver” cell type should be sensitive to the proportion of sender cells in the tissue, particularly for direct interactions involving short-range signaling molecules. As the proportion of epithelial versus stromal cells in our samples may be influenced by tissue dissociation, we restricted this analysis to links between epithelial cell types as “sender” cells (HR+ luminal, secretory luminal, or basal cells) and all other cell types as “receivers”. We modeled each pairwise interaction as a linear response to three variables: signaling from a sender cell type (i.e. the mean expression score of an activity program in that cell type), the proportion of the sender cell type in the epithelium, and an interaction term representing the combined effects of signaling and cell proportions (**Figure 3.7**).

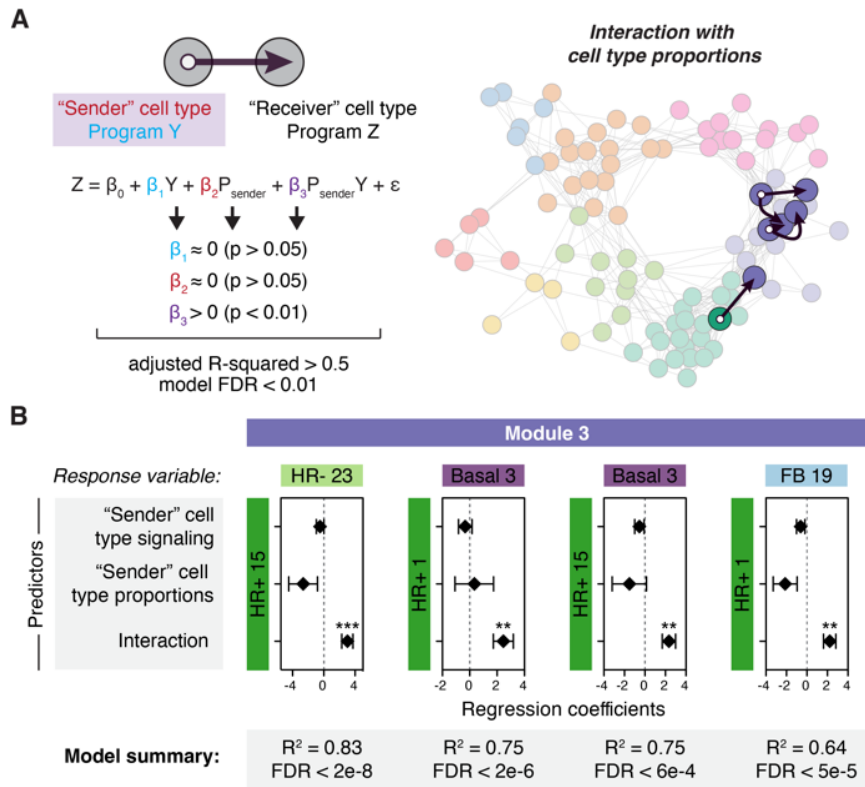


Figure 3.7 DECIPHER infers direct cell-to-cell signaling interactions in the human breast.

(A) Network graph of activity programs in the human breast, with arrows highlighting inferred direct cell-cell interactions. We modeled each pairwise combination of activity programs as a linear response to the mean expression score of an activity program in a “sender” cell type ($\beta_1 Y$), the proportion of the “sender” cell type in the epithelium ($\beta_2 P_{\text{sender}}$), and an interaction term representing the combined effect of both terms ($\beta_3 P_{\text{sender}} Y$). Arrows highlight pairs where only the interaction term is significant, the model describes over 60% of the variation in the response variable, and the FDR-corrected p-value for the overall model is less than 0.01.

(B) Results from multiple linear regression analysis, depicting the four most significant (FDR < 0.01) inferred direct cell-cell interactions. For each pairwise combination, the response variable was modeled in response to three predictors: the expression score in a “sender” cell type (signaling), the proportion of the “sender” cell type, and an interaction term between both predictors.

While this simplified model does not consider the effects of signal amplification, cooperation between signaling pathways, or higher-order interactions between more than two cell types, it identifies a subset of “high-confidence” direct cell-cell interactions that meet a series of simple criteria. We annotated putative direct cell-cell signaling interactions as those where the combined effects of signaling from a sender cell type and the proportion of the sender cell type in a tissue described over 50% of the variation in activity program expression in a second cell type (**Figure 3.7**). Consistent with our prediction about the nature of hormone signaling in the breast, four out of the five of the

high-confidence inferred direct cell-cell interactions (FDR < 0.01) were part of the same module (Module 3), and consisted of a link between HR+ luminal cells as the “sender” cell type and a second “receiver” cell type (**Figure 3.7**).

ER/PR signaling and the downstream response

We next performed gene set enrichment analysis to define activity programs within each module and identify common pathways upregulated across multiple activity programs in a module (Chapter 3 *Methods*, Supplemental Data). We first focused on Module 3 (**Figure 3.8**), as our previous analysis demonstrated that this module was highly enriched for putative direct cell-cell signaling interactions. Since estrogen and progesterone are master regulators of breast development that act via paracrine signaling from HR+ luminal cells to other cell types, we predicted that ER/PR signaling and the paracrine response would represent a major source of direct cell-cell signaling signatures present in our dataset.

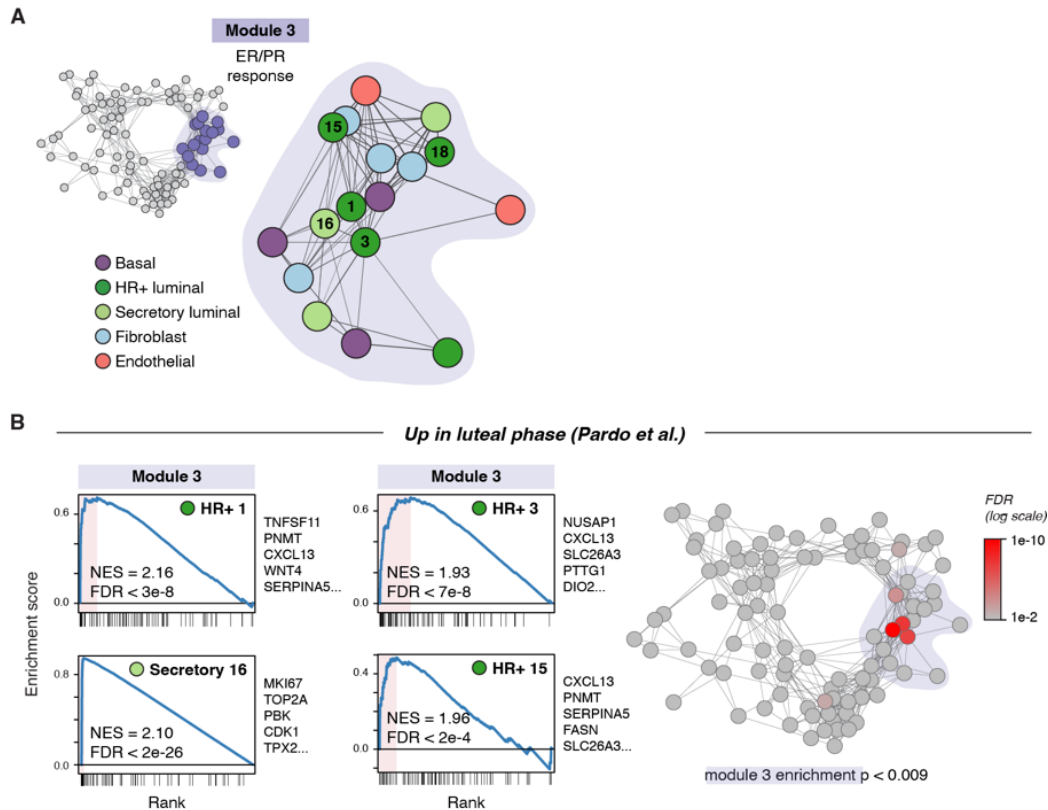


Figure 3.8 ER/PR signaling in the human breast.

(A) Diagram highlighting activity programs in the “ER/PR response” module.

(B) Left: Gene set enrichment analysis of the indicated activity programs in the “ER/PR response” module, showing enrichment of genes upregulated during the luteal phase of the menstrual (Pardo et al., 2014). The top five leading edge genes for each activity program are listed. Right: Network graph of activity programs, colored by the FDR for gene set enrichment of genes upregulated during the luteal phase of the menstrual (Pardo et al., 2014). Overall enrichment of this gene set in the “ER/PR response” module was determined by permutation analysis.

Consistent with this hypothesis, activity programs in Module 3—here annotated as the “ER/PR response” module—were highly enriched for genes previously found to be upregulated during the luteal phase of the menstrual cycle in a bulk RNA sequencing analysis (module enrichment $p < 0.01$; **Figure 3.8**).¹⁰⁸ Notably, activity program 1 in HR+ luminal cells (“ER/PR signaling”) was associated with high expression of the essential PR target genes *WNT4* and *TNFSF11* (RANKL),^{94,109} and enriched for transcripts in the Molecular Signatures Database Hallmark “early estrogen response” ($p < 0.001$) and “late estrogen response” ($p < 0.01$) gene sets (**Figure 3.8**).⁸⁷ Additional canonical hormone-responsive genes including *TFF1*, *AREG*, *PGR*, and *VEGFA* were highly expressed

across multiple activity programs in this module.^{110–113} Consistent with previous work demonstrating that STAT5 acts as a cofactor to mediate signaling downstream of PR activation in the breast, the ER/PR response module was also enriched for genes involved in IL-2/STAT5 signaling (module enrichment $p < 1e-4$). Finally, gene set enrichment analysis identified a subpopulation of proliferative secretory luminal cells within the ER/PR response module (**Figure 3.8**). This “proliferation” activity program (Secretory program 16) was highly enriched for cell-cycle related genes previously found to be upregulated during the luteal phase of the menstrual cycle (**Figure 3.8**).¹⁰⁸

Notably, our analysis also revealed that high levels of ER/PR signaling in HR+ cells (HR+ 1) coincided with the emergence of a second transcriptional state in a distinct subpopulation of HR+ luminal cells (HR+ 18) (**Figure 3.9**). Marker analysis and gene set enrichment analysis demonstrated that HR+ program 18 was characterized by upregulation of a hypoxia gene signature and pro-angiogenic factors such as *VEGFA* and *ANGPTL4*. The identification of this “hypoxia” gene signature is consistent with a previous study using microdialysis of healthy human breast tissue which found that *VEGF* levels increased in the luteal phase of the menstrual cycle.¹¹⁴ As estrogen response elements have been identified in the untranslated regions of *VEGFA*,¹¹¹ our results suggest that this increased expression may be, in part, a direct effect of hormone signaling to a subpopulation of HR+ cells.

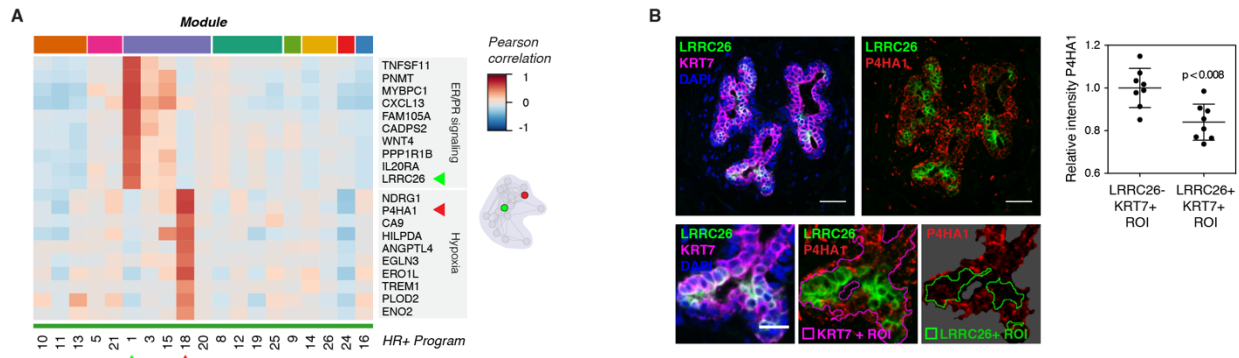


Figure 3.9 ER/PR signaling and the coordinated downstream response confirmed in vivo.

(A) Heatmap of the top 10 marker genes for HR+ 1 and HR+ 18. Results depict the Pearson correlation between the expression score of the indicated activity programs and the normalized expression of the indicated genes across cells.

(B) Immunostaining for LRRC26, P4HA1, and KRT7 and quantification of the relative intensity of P4HA1 signal in LRRC26-/KRT7+ and LRRC26+/KRT7+ regions of interest. Scale bars 20 μ m.

To confirm these results in vivo, we performed marker analysis to identify genes specific to each cluster that could be used for immunohistochemical staining (**Figure 3.9**). We identified *LRRC26* as a marker of the ER/PR signaling activity program HR+ 1 and *P4HA1* as a marker of the hypoxia/pro-angiogenic activity program HR+ 18 (**Figure 3.9**). In intact human tissue sections, we found that LRRC26 staining marked a distinct set of luminal cells from P4HA1 (**Figure 3.9**). Moreover, these two subpopulations co-occurred within the same regions of the breast, demonstrating that they are unlikely to be an artifact of sample processing. Together, these results identify at least two diverging transcriptional states in HR+ cells in samples with high ER/PR signaling, one associated with signaling via RANK ligand and WNT4 to the surrounding epithelium and a second associated with a hypoxia-related/pro-angiogenic transcriptional signature (**Figure 3.9**).

We next expanded our analysis of gene activity programs to other epithelial lineages and stromal cell types in the “ER/PR response” module. Similar to program 18 in HR+ cells, multiple activity programs across other cell types in this module were enriched for transcripts involved in hypoxia and blood vessel remodeling including VEGFA and ANGPTL4 (**Figure 3.10**). The ER/PR response module was also enriched for genes

involved in tissue remodeling, cell migration, and extra-cellular matrix (ECM) organization (Figure 3.10), consistent with previously reported morphological changes observed in the breast epithelium⁹⁸ and alterations in stromal organization and ECM composition^{115,116} across the menstrual cycle.

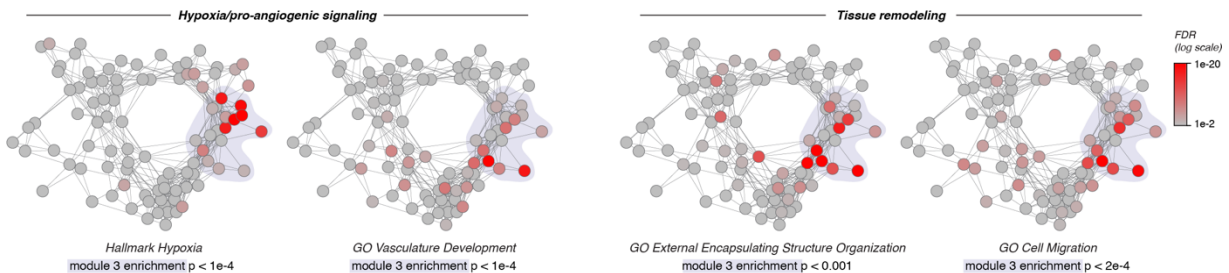


Figure 3.10 Activity programs in other epithelial lineages and stromal cell types within the ‘ER/PR response’ module. Network graph of activity programs, colored by the FDR for enrichment of the indicated gene sets in each activity program (log scale). Overall enrichment of gene sets within module 1 was determined by permutation analysis.

Stromal cell types in this module were characterized by upregulation of ECM remodeling proteins including collagens (COL3A1, COL1A2), the crosslinking enzyme LOXL2, and the cytokine TGFB3. Together, these results demonstrate that DECIPHER-seq identified distinct transcriptional signatures for ER/PR activation in HR+ luminal cells and the downstream paracrine response in other cell types.

Discussion

In this study, we combine single-cell analyses, immunostaining, and computational analysis to deconstruct the major sources of sample-to-sample heterogeneity in the human breast. We apply DECIPHER, a computational pipeline that leverages the inter-sample transcriptional heterogeneity in our dataset to identify coordinated “tissue states” made up of modules of correlated transcriptional programs. A key insight of this approach is that a subset of “high confidence” direct cell-cell interactions can be identified based on their dependence on the proportion of one cell type in the tissue. Because DECIPHER corrects for batch effects while maintaining meaningful biological variation and optimizes both the granularity and robustness of identified activity programs, it has the potential to be flexibly applied to a broad range of preexisting single-cell datasets, or across datasets from multiple sources. While we focus on single-cell transcriptional data in this study, the integrative NMF backbone that underpins activity program identification in DECIPHER has also been applied to multi-omic datasets containing spatial or epigenetic data together with transcriptional information.^{78,106}

Using DECIPHER, we identify a set of highly correlated activity programs representing the *in situ* response to hormone receptor activation in HR+ cells and the effects of downstream paracrine signaling in other cell types, as well other coordinated programs representing the dynamic response to changing hormone levels (e.g. “involution-like”). By providing a systematic approach to explore scRNA-seq datasets at the tissue level, and organizing individual cell types’ transcriptional signatures into higher order modules of cell-cell interactions, we anticipate that DECIPHER will be broadly useful in other contexts. In this study, we used variation across individuals to identify cell-

cell interaction networks in the healthy human breast, but the same approach could be used to study variation across individuals in disease states or in response to drug treatment, to identify spatial variation across different regions or cellular neighborhoods in a tissue, to uncover coordinated transcriptional responses in large-scale drug or genetic perturbation screens, or to understand coordinated changes in cell-cell interactions over time during developmental or disease processes.

Cumulative lifetime hormone exposure is a major determinant of breast cancer risk.⁴ Using DECIPHER, we mapped the coordinated changes in cell state that occur in response to paracrine signaling from HR+ luminal cells. Strikingly, many of these changes closely mimic those seen during the pregnancy/involution cycle that have been linked to a transient increased breast cancer risk following pregnancy.^{100,101,117} First, we identify a proliferative gene signature in secretory luminal cells that is highly correlated with hormone signaling in HR+ luminal cells, consistent with previous studies demonstrating that TNFSF11 (RANKL) and WNT control progesterone-mediated epithelial proliferation.¹¹⁸ Second, we also observe upregulation of hypoxic gene signatures in multiple epithelial and stromal cell types that are highly correlated with hormone signaling in HR+ cells. A previous study identified these same pathways as highly enriched following post-lactational involution in the mouse. More importantly from the perspective of breast cancer risk, this “hypoxia/pro-angiogenic” signature identified breast cancers with increased metastatic activity,¹¹⁹ suggesting that these pathways can be co-opted by cancer cells to support a permissive tumor microenvironment. Thus, we speculate that many of the same mechanisms underlie both the short-term increased breast cancer risk

following pregnancy and the lifetime increased risk due to total number of menstrual cycles.

In summary, by applying DECIPHER to scRNA-seq data from a unique cohort of 28 healthy premenopausal women, we provide a comprehensive, systems-level view of the cellular and transcriptional variation within the human breast, which profoundly affects the response to hormones and may impact breast cancer risk. As the human breast is one of the only human organs that undergoes repeated cycles of morphogenesis and involution, this study serves as a roadmap for deeper interrogation of the cell state changes associated with hormone dynamics. Finally, it provides a foundation for future systems-level studies dissecting how the paracrine communication networks downstream of hormone signaling are altered during ER+/PR+ breast cancer progression.

Methods

Tissue samples and preparation

Reduction mammoplasty tissue samples were obtained from the Cooperative Human Tissue Network (CHTN, Vanderbilt University Medical Center, Nashville, TN) and Kaiser Permanente Northern California (KPNC, Oakland, CA). Tissues were obtained as de-identified samples and all subjects provided written informed consent. When possible, medical reports or other patient data were obtained with personally identifiable information redacted. Use of breast tissue specimens to conduct the studies described above were approved by the UCSF Committee on Human Research under Institutional Review Board protocols 16-18865 and 10-01532. A portion of each sample was fixed in formalin and paraffin-embedded using standard procedures. The remainder was dissociated mechanically and enzymatically to obtain epithelial-enriched tissue fragments. Tissue was minced, followed by enzymatic dissociation with 200 U/mL collagenase type III (Worthington CLS-3, samples RM108 - RM203) or collagenase type II (Worthington CLS-2, samples RM216 - RM314) and 100 U/mL hyaluronidase (Sigma H3506) in RPMI 1640 with HEPES (Corning 10-041-CV) plus 10% (v/v) dialyzed FBS, penicillin, streptomycin, amphotericin B (Lonza 17-836E), and gentamicin (Lonza 17-518) at 37 °C for 16 h. For reduction mammoplasty samples, the cell suspension was centrifuged at 400 x g for 10 min and resuspended in RPMI 1640 plus 10% FBS. Digested tissue fragments enriched for epithelial cells and closely-associated stroma were collected after serial filtration through 150 µm and 40 µm nylon mesh strainers. Following

centrifugation, tissue fragments and filtrate were frozen and maintained at -180 °C until use.

Dissociation to single cells

The day of sorting, epithelial-enriched tissue fragments from the 150 µm fraction, or total banked material for the KTB samples, were thawed and digested to single cells by trituration in 0.05% trypsin for 2 min, followed by trituration in 5 U/mL dispase (Stem Cell Technologies 07913) plus 1 mg/mL DNase I (Stem Cell Technologies 07900) for 2 min. Single-cell suspensions were resuspended in HBSS supplemented with 2% FBS, filtered through a 40 µm cell strainer, and pelleted at 400 x g for 5 min. The pellets were resuspended in 10 mL of complete mammary epithelial growth medium with 2% v/v FBS without GA-1000 (MEGM; Lonza CC-3150). Cells were incubated at 37 °C for 1 h, rotating on a hula mixer, to regenerate surface antigens.

MULTI-seq sample barcoding (Batches 3 and 4)

Single-cell suspensions were pelleted at 400 x g for 5 min and washed once with 10 mL mammary epithelial basal medium (MEBM; Lonza CC-3151). For each sample, one million cells were aliquoted, washed a second time with 200 µL MEBM, and resuspended in 90 µL of a 200 nM solution containing equimolar amounts of anchor lipid-modified oligonucleotides (LMOs) and sample barcode oligonucleotides in phosphate buffered saline (PBS). Following a 5-minute incubation on ice with anchor-LMO/barcode, 10 µL of 2 µM co-anchor LMO in PBS was added to each sample (for a final concentration of 200

nM), and wells were mixed by gentle pipetting and incubated for an additional 5 min on ice. Following incubation, cells were washed twice in 200 μ L PBS with 1% BSA and pooled together into a single 15 mL conical tube containing 10 mL PBS/1% BSA. All subsequent steps were performed on ice.

Sorting for scRNA-seq

Cells were pelleted at 400 x g for 5 min and resuspended in PBS/1% BSA at a concentration of 1 million cells per 100 μ L, and incubated with primary antibodies. Cells were stained with Alexa 488-conjugated anti-CD49f to isolate basal/myoepithelial cells, PE-conjugated anti-EpCAM to isolate luminal epithelial cells, and biotinylated antibodies for lineage markers CD2, CD3, CD16, CD64, CD31, and CD45 to remove hematopoietic (CD16/CD64-positive), endothelial (CD31-positive), and leukocytic (CD2/CD3/CD45-positive) lineages by negative selection (Lin⁻). Sequential incubation with primary antibodies was performed for 30 min on ice in PBS/1% BSA, and cells were washed with cold PBS/1% BSA. Biotinylated primary antibodies were detected with a streptavidin-Brilliant Violet 785 conjugate. After incubation, cells were washed once and resuspended in PBS/1% BSA plus 1 μ g/mL DAPI for live/dead discrimination. Cell sorting was performed on a FACSAria II cell sorter. Live/singlet (DAPI⁻), luminal (DAPI⁻/Lin⁻/CD49f⁻/EpCAM⁺), basal/myoepithelial (DAPI⁻/Lin⁻/CD49f⁺/EpCAM⁻), or total epithelial (pooled luminal and basal/myoepithelial) cells were collected for each sample as specified in **Supplemental Table 3.2** and resuspended in PBS/1% BSA at a concentration of 1000 cells/ μ L. For Batch 4, an aliquot of MULTI-seq barcoded cells were separately stained with biotinylated-CD45/streptavidin-Brilliant Violet 785 to enrich for immune cells, and

sorted CD45⁺ cells were pooled with the Live/singlet fraction as specified in **Supplemental Table 3.2**. Antibodies and dilutions used (μL /million cells) were as follows: FITC-EpCAM (1.5 μL ; BD 550257, clone AD2), APC-CD49f (4 μL ; Stem Cell Technologies 10109, clone VU1D9), Biotin-CD2 (8 μL ; BioLegend 313636, clone GoH3), Biotin-CD3 (8 μL ; BD 55325, clone RPA-2.10), Biotin-CD16 (8 μL ; BD 55338, clone HIT3a), Biotin-CD64 (8 μL ; BD 555526, clone 10.1), Biotin-CD31 (4 μL ; Invitrogen MHCD31154, clone MBC78.2), Biotin-CD45 (1 μL ; BioLegend 304004, clone HI30), BV785-Streptavidin (1 μL ; BioLegend 405249).

scRNA-seq library preparation

cDNA libraries were prepared using the 10X Genomics Single Cell V2 (CG00052 Single Cell 3' Reagent Kit v2: User Guide Rev B) or Single Cell V3 (CG000183 Single Cell 3' Reagent Kit v3: User Guide Rev B) standard workflows as specified in **Supplemental Table 3.2**. Library concentrations were quantified using high sensitivity DNA Bioanalyzer chips (Agilent, 5067-4626) and Qubit dsDNA HS Assay Kit (Thermo Fisher Q32851). Individual libraries were sequenced on a lane of a HiSeq4500 or NovaSeq, as specified in **Supplemental Table 3.2**, for an average of $\sim 150,000$ reads/cell.

Expression library pre-processing

Cell Ranger (10x Genomics) was used to align sequences, filter data and count unique molecular identifiers (UMIs). Data were mapped to the human reference genome GRCh37 (hg19). The resulting sequencing statistics are summarized in Table S2. For

samples run across multiple 10X lanes, the cellranger aggr pipeline (10X Genomics) was used to normalize read depth across droplet microfluidic lanes (see “sort gate” information in **Supplemental Table 3.2**).

Cell calling

For V2 experiments, cell-associated barcodes were defined using Cell Ranger. For V3/MULTI-seq experiments, cells were defined as barcodes associated with ≥ 600 total RNA UMIs and $\leq 20\%$ of reads mapping to mitochondrial genes. We manually selected 600 RNA UMIs and 20% mitochondrial genes to exclude low-quality cell barcodes.

MULTI-seq barcode library pre-processing

Raw barcode FASTQs were converted to barcode UMI count matrices as described previously.¹⁹ Briefly, FASTQs were parsed to discard reads where: 1) the first 16 bases of read 1 did not match a list of cell barcodes generated as described above, and 2) the first 8 bases of read 2 did not align with any reference barcode with less than 1 mismatch. Duplicated UMIs, defined as reads with the same cell barcode where bases 17-28 (V3 chemistry) of read 2 exactly matched, were removed to produce a final barcode UMI count matrix.

Sample demultiplexing

Barcode UMI count matrices were used to classify cells using the MULTI-seq classification suite.¹⁹ In Batch 3, sample RM192 was poorly labeled for the lane of cells from the epithelial cell sort gate. Therefore, to reduce spurious doublet calls in this dataset, we manually set UMI counts which were <10 for this barcode to zero. For all experiments, raw barcode reads were log₂-transformed and mean-centered, the top and bottom 0.1% of values for each barcode were excluded, and a probability density function (PDF) was constructed for each barcode. Next, all local maxima were computed for each PDF, and the negative and positive maxima were selected. To define a threshold between these two maxima, we iterated across 0.02-quantile increments and chose the quantile maximizing the number of singlet classifications, defined as cells surpassing the threshold for a single barcode. Multiplets were defined as cells surpassing two or more thresholds, and unlabeled cells were defined as cells surpassing zero thresholds. Unclassified cells were removed and the procedure was repeated until all remaining cells were classified.

To classify cells that were identified as unlabeled by MULTI-seq, we used the SoupOrCell pipeline¹⁰² to assign cells to different individuals based on single nucleotide polymorphisms (SNPs). For each dataset, we set the number of clusters (k) to the total number of samples in that experiment. To avoid local minima, SoupOrCell restarts clustering multiple times and takes the solution that minimizes the loss function. For Batch 3, we chose the number of restarts that produced less than a 1.5% misclassification rate between MULTI-seq and SoupOrCell singlet sample classifications (Live/singlet: 30 restarts/1.2% mismatch rate; Epithelial: 75 restarts/1.5% mismatch rate). SoupOrCell classification performed more poorly across parameters for Batch 4 (Live/singlet plus

CD45+: 50 restarts/8.1% mismatch rate, 75 restarts/4.8% mismatch rate; Epithelial: 50 restarts/8.6% mismatch rate, 75 restarts/14.9% mismatch rate, 100 restarts/4.1% mismatch rate). Therefore, for these datasets we used sample classifications that were consistent across two restarts (Live/singlet plus CD45+: consistent calls across 50 and 75 restarts/0.4% overall mismatch rate; Epithelial: consistent calls across 50 and 100 restarts/1% overall mismatch rate) to identify high-confidence singlets.

Quality control, dataset integration, and cell type identification using Seurat

Cell type identification was performed using the Seurat package (version 3.1.5) in R.⁴⁰ To identify and remove doublets formed from cells from the same sample that would not be identified by MULTI-seq or SoupOrCell, we filtered each lane to remove cells with greater than 20% of reads mapping to mitochondrial genes and ran DoubletFinder (version 2.0) on each data subset,¹²⁰ using parameters identified by the 'paramSweep_v3' function. Aggregated data for singlet cells for each batch was filtered to remove cells that had fewer than 200 genes and genes that appeared in fewer than 3 cells. Cells with a Z score of 4 or greater for the total number of genes expressed were presumed to be doublets and removed from analysis. The remaining cells were log transformed and scaled to a total of 1e4 molecules per cell, and the top 2000 most variable genes based on variance stabilizing transformation were identified for each batch.³⁸ Data from all four batches were integrated using the standard workflow and default parameters from Seurat v3.⁴⁰ This data integration workflow identifies pairwise correspondences between cells

across datasets and uses these anchors to transform datasets into a shared expression space. Following dataset integration, the resulting batch-corrected expression matrix was scaled, and PCA was performed using the identified integration genes. The top 28 statistically significant PCs as determined by visual inspection of elbow plots were used as an input for UMAP visualization and k-nearest neighbor (KNN) modularity optimization-based clustering using Seurat's 'FindNeighbors' and 'FindClusters' functions.

Quantification of sample-to-sample heterogeneity: cluster entropy and similarity scores/alignment

Cluster entropy: To measure how well-mixed cells from different samples were across cell type clusters, we quantified the normalized relative cluster entropy for our dataset, weighted by cluster size.¹²¹ A cluster entropy value of 1 represents complete intermixing of samples across clusters.

Similarity scores/alignment: To measure transcriptional variation in cell state within cell types between cells from the same versus different batches and/or samples, we measured the pairwise alignment between each sample/batch,³⁹ where batches consisted of sets of samples processed on the same day (**Supplemental Table 3.2**). This "similarity score" examines the local neighborhood of each cell in a particular sample/batch, asks how many of its k nearest neighbors (in PC or iNMF space) belong to a second sample/batch, and averages this over all cells. We chose k to be 1% of the total number of cells within a cluster. The result was normalized by the expected number of cells from each sample/batch. Notably, for repeat measurements, samples run across

multiple batches were highly similar. We calculated the pairwise similarity score between each sample/batch using the first 14 principal components for each cell type. We calculated the pairwise similarity score between each sample/batch using all iNMF components for each cell type (at k_{DEW} , see text below for optimization of k).

PCA of individual cell types

To perform principal component analysis on individual cell types, we subset out each cluster from the integrated dataset and repeated the standard workflow from Seurat v3 to identify integration genes specific to this cell type. The resulting batch-corrected expression matrices were scaled, and PCA was performed using the identified integration genes.

Activity program identification in each cell type

(consensus iNMF)

To identify gene expression signatures, or “activity programs”, within individual cell types, we subset raw counts data from each of the five most abundant cell type clusters (HR+ luminal cells, secretory luminal cells, basal/myoepithelial cells, fibroblasts, and endothelial cells) and performed matrix factorization. We chose to perform matrix factorization independently on each cell type rather than on the combined dataset, as preliminary analyses demonstrated that the number of gene programs identified for each cell type was highly dependent on the relative sizes of each cluster in the combined

dataset. To correct for batch differences between samples run on different days, we used the LIGER package in R to perform integrative NMF (iNMF),^{78,106} and performed subsequent gene set enrichment analyses on shared, rather than batch-specific, gene loadings for each activity program. Importantly, activity program expression in cells from the same sample run across different batches was more similar than program expression in cells from different samples processed in the same batch, demonstrating that this approach successfully corrected for batch differences while retaining sample-to-sample transcriptional variability. To avoid identification of gene signatures dominated by highly-expressed transcripts, we normalized the raw counts matrix for each cell based on its total expression, multiplied by a scale factor of $1e4$, and log-transformed and scaled the result without centering. The resulting datasets (one for each cell type) were decomposed using the 'online_iNMF' function from LIGER.⁷⁸ Online iNMF uses an online learning algorithm to iteratively cycle through the data in small mini-batches, greatly increasing convergence times for large datasets. We performed 10 complete passes ('max.epochs' parameter) through each dataset, and chose the mini-batch size ('miniBatch_size') by rounding down to the nearest 500 from the smallest batch size in that cell type (HR+ luminal cells: 1000, Secretory luminal cells: 2000, Basal cells: 500, Fibroblasts: 500, Endothelial cells: 500).

Since solutions to NMF are non-unique, we adapted a consensus matrix factorization approach from Kotliar et al. to identify activity programs that were consistent across multiple replicates.⁷⁷ For each cell type, we ran 20 replicates of iNMF on the same normalized dataset with the same choice of rank K , starting from different random seeds. We row normalized the resulting 20 shared gene loading matrices (W , each of dimension

$K_{programs} \times N_{genes}$) to have an L2 norm of one. Following normalization, we combined the shared gene loading matrices from each matrix into a $20K_{programs} \times N_{genes}$ dimensional matrix, where each row represents the gene loading from one activity program in one replicate. Next, we filtered out programs with a high mean Euclidean distance from their 6 nearest neighbors (30% of replicates), using the third quartile plus 1.5 times the interquartile range ($q_{0.75} + 1.5 \cdot IQR$) as an outlier threshold. After filtering outlier programs, we grouped the rows of the resulting matrix using k-means clustering, with the number of clusters set to the chosen iNMF rank K . Next, we collapsed each group of shared gene loadings to a single consensus vector by taking the median value for each gene across activity programs in that cluster, to produce a final $K_{programs} \times N_{cells}$ consensus program matrix, \mathbf{W} . We performed the same row normalization on the batch-specific gene loading matrices, filtered programs identified as outliers in the shared gene loading matrix, and collapsed groups of batch gene loadings into a consensus vector by taking the median value for each gene across programs in that cluster to produce consensus batch matrices \mathbf{V}_{batch} , each of dimension $K_{programs} \times N_{genes}$. Finally, we solve for the consensus cell expression score matrix \mathbf{H} ($X_{cells} \times K_{programs}$), by using non-negative least squares initialized with the consensus shared (\mathbf{W}) and batch-specific (\mathbf{V}_{batch}) gene loading matrices.

A key parameter in matrix factorization is the choice of rank K . This parameter determines the granularity of identified activity programs. Three commonly used heuristics for guiding the optimum choice of K are: 1) minimizing the Frobenius reconstruction error of the final solution,⁷⁷ 2) maximizing the median Kullback-Leibler (KL) divergence of activity program loadings across cells relative to a uniform distribution,¹⁰⁶

and 3) estimating the “dimensionality” of the dataset via elbow plot of the proportion of variance explained across principal components.⁷⁷ We propose a new metric for choosing an optimum K , based on the goal of identifying the greatest number of activity programs that are robust (i.e. consistent across multiple choices of K) and unique (i.e. distinct from other programs at a particular choice of K). First, we perform consensus iNMF as described above over a range of ranks, with the sweep range guided by the heuristics described above. Here, we chose a range of 2 to 40 for all cell types. Next, we use the ‘fastme.bal’ function in the ‘ape’ R package to build a balanced minimum evolution phylogenetic tree based on the correlation matrix of the gene loadings for activity programs across all ranks.⁸⁰ For each cell type, we partitioned the resulting phylogenetic tree into clusters using an empirical distance threshold to define distinct groups of activity programs.⁸¹ To identify partitions, we first artificially rooted each tree by taking the median of the activity programs at $K = 2$. Next, we identified clusters by performing a depth-first search starting from this artificial root, stopping at sub-trees where the median value of the pairwise patristic distance between all programs in that sub-tree was below an empirically determined threshold of 0.3. To filter out “outlier” activity programs that represent rare contaminating cells (e.g. a “fibroblast-like” gene signature in HR+ luminal cells), we calculated the maximum expression score for each activity program divided by the mean expression score for the next 50 highest-scoring cells, and removed programs where this ratio was greater than 5. We also removed subtrees with fewer than 5 total activity programs. Finally, we plotted the number of subtrees identified at each K (excluding outlier programs), weighted by the total number of programs in each subtree. We choose the optimum K (K_{opt}) as the saturation point in this curve, representing the

point at which increasing the granularity of matrix factorization does not identify activity programs that comprise major new subtrees.

Network clustering of correlated activity programs

To identify sets of activity programs that co-varied across samples, we first decomposed each cell type into a set of distinct gene expression signatures, or “activity programs”, using consensus iNMF with k_{DEW} chosen for each cell type as described above. We then quantified the average expression of each gene program in each sample and constructed a weighted network of coordinated gene expression programs based on the pair-wise Pearson correlations between gene programs. To account for correlations driven by outlier samples, we used bias-corrected and accelerated bootstrap resampling to estimate confidence intervals associated with each correlation coefficient. The resulting Pearson correlation matrix was transformed into a weighted adjacency matrix by setting all Pearson correlation coefficients with p-values greater than 0.05 (based on the null hypothesis $r = 0$) to zero. We identified modules of highly correlated gene expression programs using a Constant Potts Model for community detection in signed graphs in the ‘leidenalg’ package in python.⁸⁴ We ran this algorithm at a range of resolutions from 0.001 to 0.4 and chose the resolution that maximized overall modularity. To filter out isolated links and modules, we calculated the signed weighted topological overlap (wTO) between activity programs in each module⁸⁵ and filtered nodes with low wTO and modules containing fewer than four nodes. In contrast to Pearson correlation values which consider each pair of nodes in isolation, wTO is based on the similarity of two activity programs’ correlation values with all other programs in the network. We calculated the

mean wTO between each node and all other nodes in the same module, and compared this to the value calculated for nodes in randomly selected modules of equal size. We determined p-values for each node's mean wTO by determining the fraction of permutation trials where the mean wTO of nodes from "random" modules was greater than the mean wTO of nodes from tested modules, and removed nodes where $p > 0.01$. Community detection results remained unchanged after this filtering step. For visualization, we use positive edges to create a force-directed layout.

Identification of similar gene loadings across activity programs

To identify transcriptionally similar activity programs representing non-cell-type specific responses, we calculated the Pearson correlation of gene loadings between activity programs using pairwise complete observations (i.e. excluding genes that are not expressed in either cell type). We defined each node's "mean gene loading similarity" as the mean correlation between the tested node and all other nodes in the same module. To determine p-values for each node's gene loading similarity, we compared this value to that calculated for nodes in randomly selected modules of equal size. The reported p-values represent the fraction of permutation trials where the mean gene loading similarity for nodes from "random" modules was greater than the mean gene loading similarity for nodes in tested modules.

Inferring direct cell-cell interactions

To infer modules enriched for putative direct cell-cell signaling interactions, we identified links between nodes that depended on both the magnitude of activity program expression in a “sender” cell type and the proportion of that “sender” cell type in the tissue. Since the proportion of epithelial versus stromal cells in our samples was highly dependent on tissue dissociation conditions, we restricted this analysis to links between epithelial cell types as “sender” cells (HR+ luminal, secretory luminal, or basal cells) and other cell types as “receivers”. We modeled activity program expression in the “receiver” cell type as a linear response to three predictors: activity program expression Y in the “sender” cell type (i.e. “signaling” from that cell type), the proportion P_{sender} of the “sender” cell type in the epithelium, and an interaction term representing the combined effects of signaling and cell proportions (Signaling \times Proportions). For links between two epithelial cell types, we tested both directions as “sender” versus “receiver” nodes. To infer high-confidence direct cell-cell signaling interactions, we identified pairwise combinations of activity programs where a) the individual effects of Y and P_{sender} were not significant ($p > 0.05$), b) there was a positive interaction effect between Y and P_{sender} (Signaling \times Proportions; $p < 0.01$ and $\beta > 0$), c) the adjusted R-squared for the overall model was at least 0.5, and d) the false discovery rate-corrected p-value for the overall model was less than 0.05.

Gene set enrichment analysis

To identify marker genes statistically associated with each gene program, we used ordinary least squares regression of each gene's normalized (z-scored) expression against the activity program expression score for each program in each cell type, after filtering genes not expressed in that cell type.⁷⁷ This results in a vector of regression coefficients representing the strength of the relationship between a cell's expression score for a particular activity program and its scaled expression of each gene. The resulting ranked gene lists (Supplemental Data) were analyzed by gene set enrichment analysis, using the 'fgsea' package in R.¹²²

Enrichment of gene sets within modules

To identify gene sets enriched across activity programs in a module, we first calculated the false discovery rate (FDR) for each gene set in each node. We performed false discovery rate correction for Hallmark and GO Biological Process gene sets separately, as many of the pathways in each database are highly related. For all gene sets enriched across at least 5 activity programs in our network, we calculated the number of activity programs in each module that were significantly enriched for each gene set (FDR < 0.01), and compared this value to randomly selected modules of equal size. We determined p-values for enrichment of gene sets in each module by determining the fraction of permutation trials where the number of significantly enriched nodes from "random" modules was greater than number of significantly enriched nodes from tested modules.

Fluorescent Immunohistochemistry

For immunofluorescent staining, formalin-fixed paraffin-embedded tissue sections were deparaffinized and rehydrated using standard methods. Endogenous peroxides were blocked using 3% hydrogen peroxide in PBS, and antigen retrieval was performed in 0.1 M citrate buffer pH 6.0. Sections were blocked for 5 min at room temperature using Lab Vision Ultra-V block (Thermo TA-125-UB) and rinsed with TNT wash buffer (1X Tris-buffered saline with 5 mM Tris-HCl and 0.5% TWEEN-20). Primary antibody incubations were performed for 1 hour at room temperature or overnight at 4°C. Sections were washed three times for 5 min each with TNT wash buffer, incubated with Lab Vision UltraVision LP Detection System HRP Polymer (Thermo Fisher TL-060-HL) for 15 min at room temperature, washed, and incubated with one of three colors of tyramide signal amplification (TSA) reagent at a 1:50 dilution. After TSA, antibody complexes were removed by boiling in citrate buffer, followed by blocking and incubation with additional primary antibodies as above. Finally, sections were rinsed with deionized water and mounted using Vectashield HardSet Mounting Media with DAPI (Vector H-1400). Immunofluorescence was analyzed by spinning disk confocal microscopy using a Zeiss Cell Observer Z1 equipped with a Yokagawa spinning disk and running Zeiss Zen Software.

Antibodies, TSA reagents, and dilutions used are as follows: p63 (1:2000; CST 13109, clone D2K8X), KRT7 (1:4000; Abcam AB68459, clone EPR1619Y), KRT23 (1:2000; Abcam AB156569, clone EPR10943), ER (1:4000; Thermo Scientific RMM-9101-S, clone SP1), PR (1:3000; CST 8757, clone D8Q2J), TCF7 (1:2000; CST 2203,

clone C63D9), P4HA1 (1:9000; Thermo PA5-55353), LRRC26 (1:2000; Thermo PA5-63285), FITC-TSA (2 min; Perkin Elmer NEL701A001KT), Cy3-TSA (3 min; Perkin Elmer NEL744001KT), Cy5-TSA (7 min; Perkin Elmer NEL745E001KT).

Data and code availability

Single-cell RNAseq data have been deposited at the Gene Expression Omnibus (GEO: [GSE198732](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE198732)). Processed data and code is available at <https://github.com/lmurrow/DECIPHER-seq>. Any additional information required to reanalyze the data or reproduce the figures in this study is available from the corresponding authors upon request.

Author Contributions

LMM, RJW, and ZJG conceived the project. LMM, JAC, RJW, CSM, and KP performed the sequencing experiments. CSM generated aligned reads and barcode matrices. CSM and LMM performed sample demultiplexing. PG coordinated sample acquisition and provided critical guidance for sample selection. LMM and JAC performed fluorescent immunohistochemistry and RNA-FISH experiments. LMM and JAC performed flow cytometry experiments. LMM and JAC performed histopathology on tissue sections. ADB performed histopathological tissue analysis. LMM analyzed and visualized the data. LMM and GR wrote and tested the code used in data analysis. MT provided guidance in data analyses and computational approaches. TT and ADB provided guidance in human breast biology. TT, MT, and ZJG provided critical resources. TAD, MT, TT, and ZJG

supervised the project. LMM and ZJG wrote the manuscript. All authors reviewed and edited the manuscript.

Supplement

Supplemental Table 3.1 Sample information for breast tissue samples from reduction mammoplasties of 28 individuals.

Sample information					
Sample ID	Source	Age	BMI	Parity	HC use
RM142	KPNC	26	unknown	0	none
RM166	CHTN	24	33	unknown	none
RM169	CHTN	22	unknown	unknown	combined
RM172	CHTN	20	37.5	1	progestin
RM176	KPNC	21	unknown	1	none
RM181	CHTN	19	unknown	unknown	combined
RM183	CHTN	20	unknown	unknown	none
RM192	KPNC	27	unknown	1	none
RM193	KPNC	27	unknown	1	none
RM198	KPNC	23	unknown	0	none
RM203	KPNC	35	unknown	1	none
RM216	CHTN	22	23.7	0	none
RM222	KPNC	19	unknown	0	progestin
RM231	CHTN	22	23.5	0	combined
RM234	CHTN	32	38.6	0	none
RM248	CHTN	19	25.5	0	combined
RM249	CHTN	23	41	1	progestin
RM253	CHTN	39	35.1	1	none
RM261	CHTN	32	39.7	2	combined
RM263	CHTN	24	27.3	0	none
RM264	CHTN	37	31.5	3	none
RM272	CHTN	23	26	0	none
RM273	CHTN	24	26.2	0	none
RM274	CHTN	22	unknown	3	combined
RM278	CHTN	19	31.8	0	combined
RM282	CHTN	36	44.3	2	none
RM288	CHTN	24	42.4	unknown	combined
RM307	CHTN	19	unknown	unknown	combined

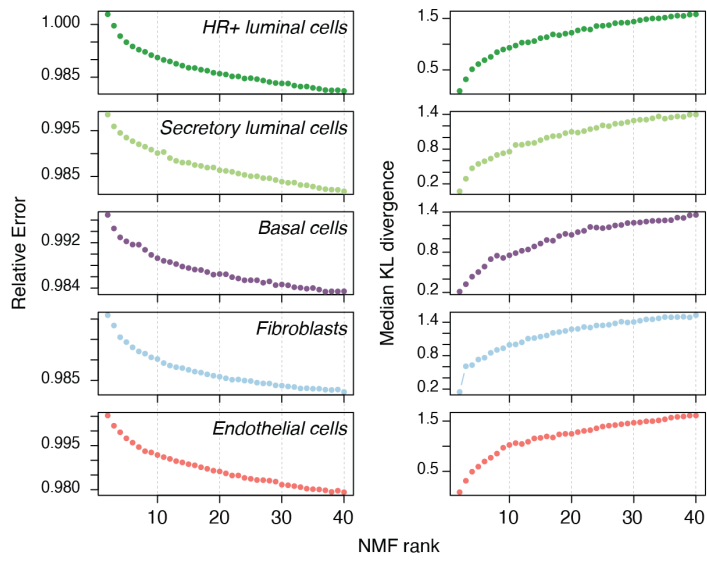
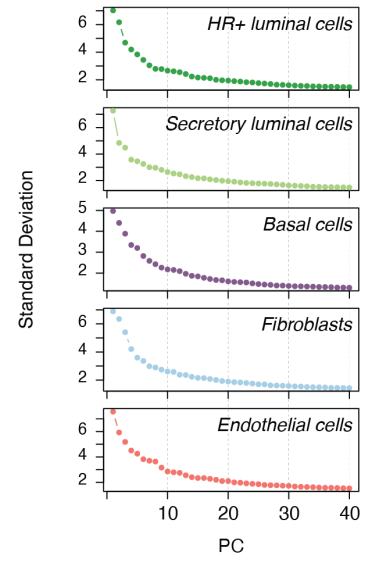
Supplemental Table 3.2 Sequencing batch info for each sample.

Sample ID	Sort gate	10X chemistry	Batch	Seq	# reads	% reads mapped	Median genes per cell	# cells post QC
RM264	Live singlet	V2	1	HiSeq 4000	665,013,964	58.1%	1,552	4,190
	Luminal				343,078,479	63.2%	1,207	2,200
	Basal				332,878,058	62.0%	1,469	2,747
RM272	Live singlet	V2	1	HiSeq 4000	656,991,798	54.6%	1,319	3,502
	Luminal				337,334,049	55.2%	1,351	4,205
	Basal				337,758,232	54.8%	601	5,003
RM273	Live singlet	V2	1	HiSeq 4000	338,792,766	65.8%	2,590	1,783
	Luminal				353,798,722	66.2%	2,772	2,482
	Basal				335,833,078	62.6%	1,891	5,126
RM282	Live singlet	V2	1	HiSeq 4000	314,274,078	53.6%	2,359	2,923
	Luminal				330,627,133	54.7%	2,888	2,744
	Basal				317,563,455	51.0%	2,095	3,101
RM263	Live singlet	V2	1	HiSeq 4000	331,865,524	60.7%	2,765	2,653
RM222	Live singlet	V2	2	HiSeq 4000	346,749,732	61.5%	2,914	802
RM234	Live singlet	V2	2	HiSeq 4000	335,184,746	61.2%	2,188	2,109
RM248	Live singlet	V2	2	HiSeq 4000	333,160,029	59.8%	2,602	2,004
RM249	Live singlet	V2	2	HiSeq 4000	329,434,014	63.3%	3,036	1,547
RM142	Live singlet	V3/MULTI-seq	3	Nova Seq	1,665,766,405	59.6%	2,882	1,350
RM166								481
RM176								180
RM183								510
RM192								867
RM193								768
RM198								1,317
RM203								1,166
RM216								741
RM253								698

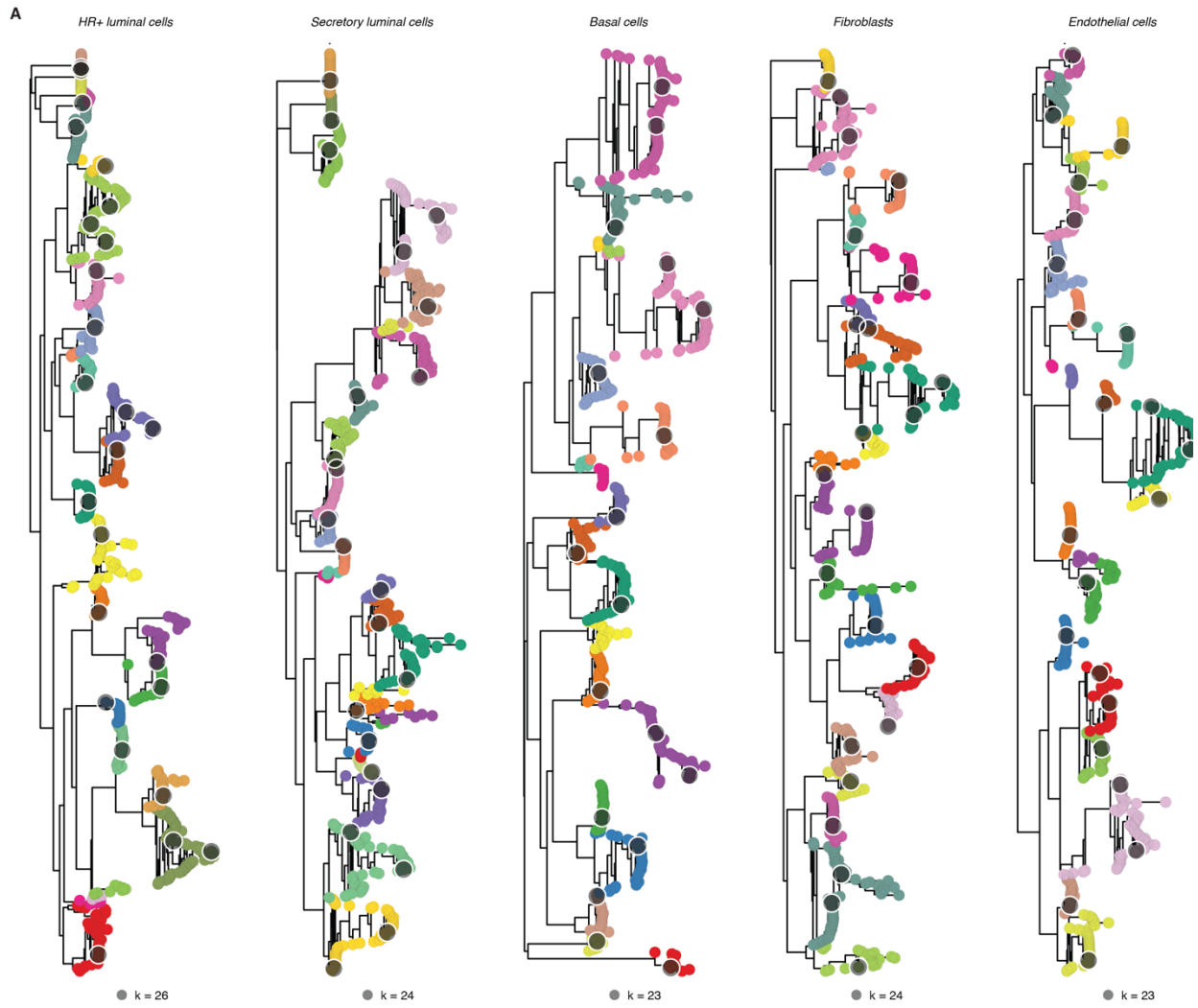
Sample ID	Sort gate	10X chemistry	Batch	Seq	# reads	% reads mapped	Median genes per cell	# cells post QC
RM272								1,563

RM142	Epithelial	V3/MULTI-seq	3	Nova Seq	1,553,893,597	59.6%	3,223	1,665
RM166								175
RM176								81
RM183								507
RM192								831
RM193								645
RM198								824
RM203								532
RM216								723
RM253								408
RM272								1,249
RM169								Pooled: Live singlet (90%) / CD45+ immune (10%)
RM172	119							
RM181	1,192							
RM198	1,377							
RM231	1,477							
RM261	854							
RM263	1,704							
RM272	1,596							
RM273	793							
RM274	828							
RM278	837							
RM282	2,202							
RM288	1,505							
RM307	795							
RM169	Epithelial	V3/MULTI-seq	4	Nova Seq	1,591,437,578	55.0%	2,934	617
RM172								72
RM181								322
RM198								1,008
RM231								678
RM261								336
RM263								822
RM272								1,554
RM273								827

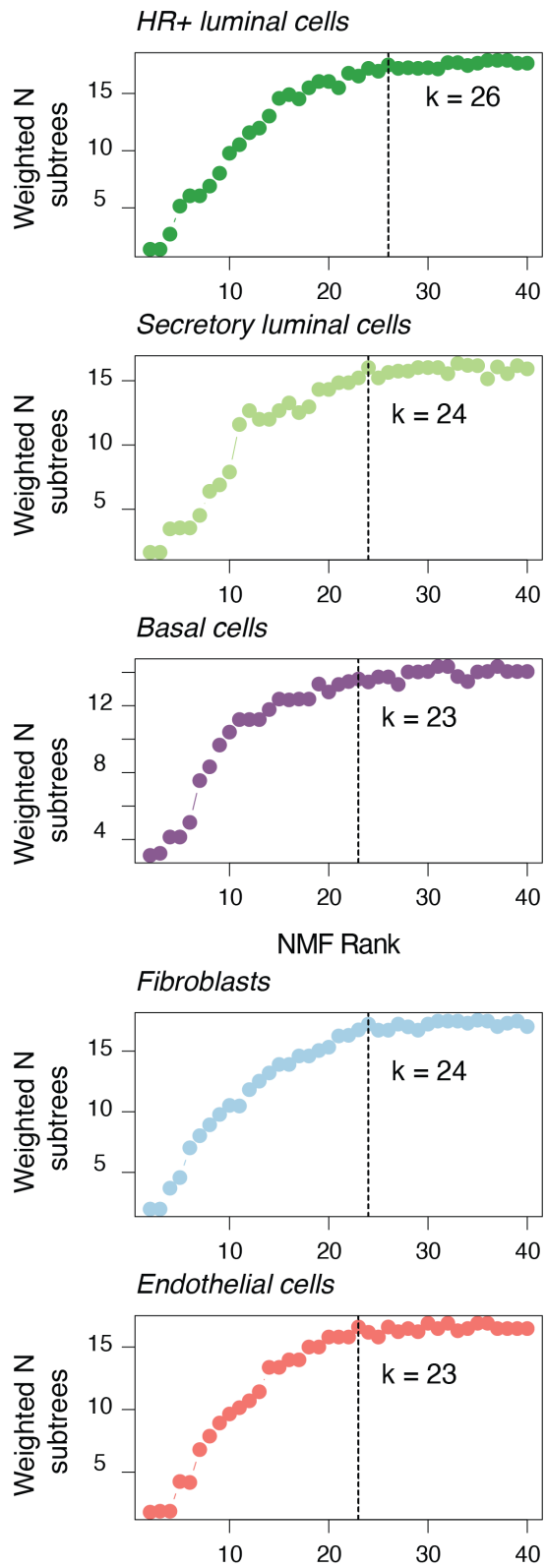
Sample ID	Sort gate	10X chemistry	Batch	Seq	# reads	% reads mapped	Median genes per cell	# cells post QC
RM274								287
RM278								1,147
RM282	Epithelial	V3/MULTI-seq	4	Nova Seq	1,591,437,578	55.0%	2,934	3,055
RM288								1,092
RM307								657

A**B**

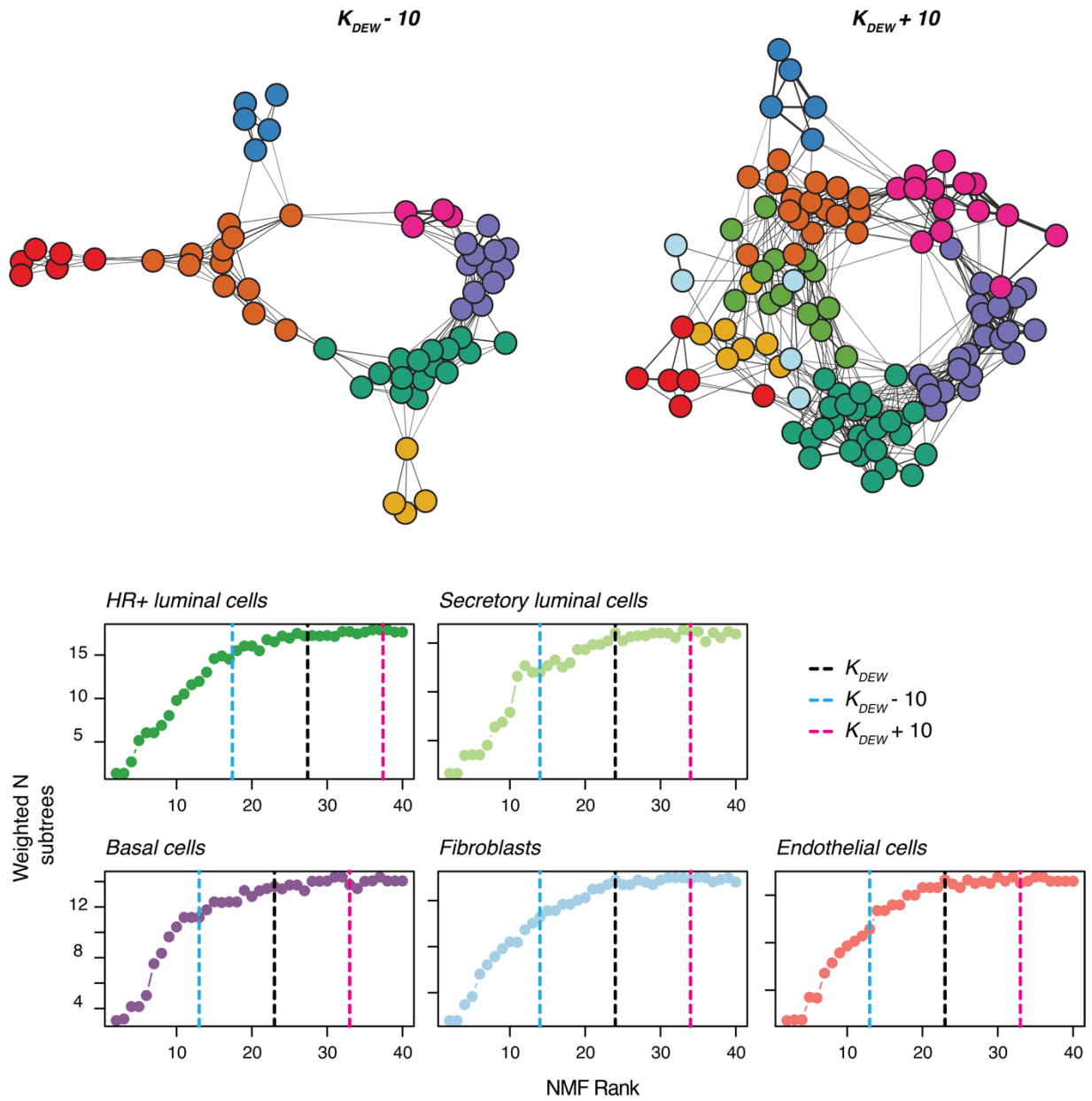
Supplemental Figure 3.11 Elbow and knee plots showing standard rank selection metrics calculated for rank sweep of coinMF (left) and PCA (right) run on each major breast cell type in dataset.



Supplemental Figure 3.12 Phylogenetic trees of *coinMF* rank sweep showing final rank selection according to DEW metric.



Supplemental Figure 3.13 DEW metric across coinMF rank sweep with final k_{DEW} selection for each cell type.



Supplemental Figure 3.14 Network structure at ranks above and below k_{DEW} for each cell type.

**CHAPTER 4 COORDINATED IMMUNE
DYSREGULATION IN JUVENILE
DERMATOMYOSITIS REVEALED BY SINGLE-
CELL GENOMICS**

Abstract

Juvenile Dermatomyositis (JDM) is one of several childhood-onset autoimmune disorders characterized by a type I interferon response and autoantibodies. Treatment options are limited due to incomplete understanding of how the disease emerges from dysregulated cell states across the immune system. We therefore investigated the blood of JDM patients at different stages of disease activity using single-cell transcriptomics paired with surface protein expression. By immunophenotyping peripheral blood mononuclear cells, we observed skewing of the B cell compartment towards an immature naive state as a hallmark of JDM at diagnosis. Furthermore, we find that these changes in B cells are paralleled by T cell signatures suggestive of Th2-mediated inflammation that persist despite disease quiescence. We applied network analysis to reveal that hyperactivation of the type I interferon response in all immune populations is coordinated with previously masked cell states including dysfunctional protein processing in CD4⁺ T cells and regulation of cell death programming in NK, CD8⁺ T cells and gdT cells. Together, these findings unveil the coordinated immune dysregulation underpinning JDM and provide novel insight into strategies for restoring balance in immune function.

Introduction

Juvenile Dermatomyositis (JDM) is part of a broad group of childhood-onset autoimmune conditions characterized by a type I interferon (IFN) gene signature and specific autoantibodies ranging from related systemic conditions such as systemic lupus erythematosus (SLE) to endocrine-specific disorders such as type I diabetes.^{9–12} Despite a shared IFN signature, JDM is associated with pathognomonic rashes and proximal muscle weakness resulting in distinct clinical phenotypes. The etiology of JDM is not fully understood but studies have shown that JDM is autoimmune-mediated and associated with a combination of genetic and environmental risk factors.¹²³ While mortality is low with corticosteroid treatment, long-term patient follow-up studies have reported that 60-70% of patients have cumulative tissue damage^{124,125} with the risk of damage increasing almost linearly for each year after diagnosis.¹²⁶ This finding highlights the importance of early disease intervention and the need for a personalized approach to disease management to improve upon these outcomes.

Clinical management of JDM currently relies on compiled empirical metrics such as physician global visual analog scale (VAS) of disease activity and muscle strength quantified via the childhood myositis assessment scale (CMAS) or manual muscle testing (MMT).¹²⁷ However, how these clinically observable phenotypes are rooted in disease immunopathology remains insufficiently understood. The presence of myositis-specific antibodies (MSA) that correspond to distinct clinical phenotypes and recent work showing that MSAs may be pathogenic^{128,129} suggest the involvement of B cells.¹³⁰ The expansion of naïve B cells in JDM has been highlighted by three independent studies using flow cytometry, mass cytometry, and single-cell RNA sequencing, respectively.^{131–133} The

adaptive arm of the immune system is further implicated in disease pathogenesis by several large immunophenotyping studies that demonstrated the expansion of extra-follicular Th2 memory cells and central memory B cells.^{134,135} Additionally, the innate immune system has emerged as a contributor in JDM. Inflamed muscles of JDM patients exhibited the presence of plasmacytoid dendritic cells and macrophage-secreted proteins,^{136,137} while similarly, biopsies of JDM and adult DM skin lesions showed an increase in CD14⁺ and CD68⁺ macrophages.^{138,139} In peripheral blood, NK cells were found to be both dysfunctional and hyperproliferative in JDM.^{132,140} Together, these findings highlight the involvement of both the adaptive and innate immune compartments in JDM in blood and disease-affected tissues. However, it also raises the question of whether the cause of JDM lies in a single cell type or is a manifestation of broadly dysregulated cellular interactions across the immune system.

Systems-level studies based on single-cell measurements are required to reveal how dysregulated cell populations act individually or cooperatively to produce the observed inflammation. Accordingly, several groups have turned to next generation sequencing as it enables unbiased profiling of tissues at single-cell resolution. In the first single-cell study of peripheral blood of JDM patients, we previously described a pan-cell-type IFN gene signature over-expressed in treatment-naive JDM that was most strongly correlated with disease activity in cytotoxic cell types.¹³³ This signature has since been independently identified in the peripheral blood of treatment-naive patients.¹⁴¹ However, these studies have utilized small cohorts and lack pediatric controls, in part due to the rarity of JDM in the human population. Thus, it has been challenging to determine which of these cell populations are specific to JDM compared to healthy children, how these

disease-specific dysregulated cell states are coordinated with one another, and which of these states cooperatively change in response to treatment.

In this study, we addressed this challenge by profiling JDM across several stages of disease activity using multiplexed Cellular Indexing of Transcriptomes and Epitopes by sequencing (CITEseq)²⁰ of peripheral blood mononuclear cells (PBMCs) from 15 JDM patients, totaling 22 samples, and 5 healthy controls (HC). Compositional analysis of immune populations identified a disease activity-associated imbalance of naive and mature lymphocytes, corroborated by distinct immunophenotypes in treatment-naive disease. To move beyond the identification of disease-associated cell populations and towards an understanding of immune-scale dysregulation in JDM, we applied a recently developed computational method DECIPHER⁶⁵ to infer networks of coordinated cell states from large cohorts of single-cell data. Importantly, this unsupervised method takes advantage of the biological heterogeneity in the entire dataset, improving upon previous work that relied on pairwise comparisons of subsetted disease groups. Among other signatures previously masked by traditional single-cell analysis, this approach revealed specific co-occurring cell states in CD4+ T and B cell populations suggestive of extra-follicular responses. A subset of these CD4+ T signatures implicate disruption of protein targeting and immune tolerance processes; notably, these cell states persist even in patients in remission off medication. Furthermore, we show that the ubiquitously hyperactive type I IFN response in disease is paralleled by impaired cell death processes in cytotoxic immune cells, highlighting the functional imbalance across immune compartments that typifies this complex autoimmune disease. Translationally, this

broadened understanding of the underlying immune dysregulation in disease can inform precision treatment strategies for JDM.

Results

JDM is associated with immunophenotypic differences in B and CD4+ T cell compartments

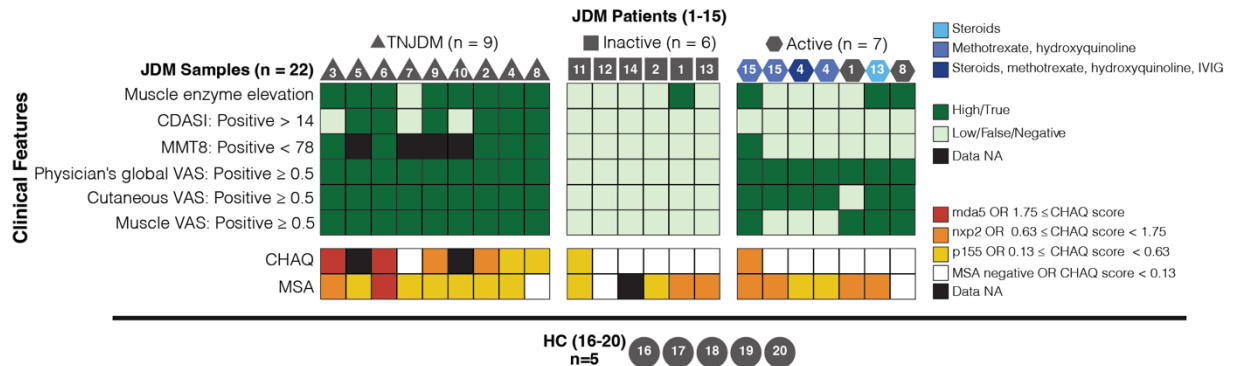


Figure 4.1 Study design for profiling PBMCs from 27 samples ($n=22$ JDM, $n=5$ HC), with an overview of clinical characteristics of study cohort.

Individuals are labelled by the donor ID used throughout the paper (JDM 1-15, HC 16-20). Longitudinal samples were collected from the following donors: JDM1, JDM2, JDM4, JDM8, JDM13, and JDM15. Icon shapes denote disease activity group and shades of blue denote medication regimen.

To gather a dataset with appropriate controls and limited confounding, patients were selected according to disease activity and medication status (**Figure 4.1**, **Supplemental Table 4.1**). Of the JDM patients, serial samples were collected from 5 individuals totaling 22 samples from 15 patients. To minimize confounding by immune suppression, the study included 9 treatment-naive samples as well as 6 samples from patients with inactive disease off medication. CITEseq was performed on PBMCs to generate single-cell libraries (**Figure 4.2**). Surface protein expression was measured using antibody-derived tags (ADT).

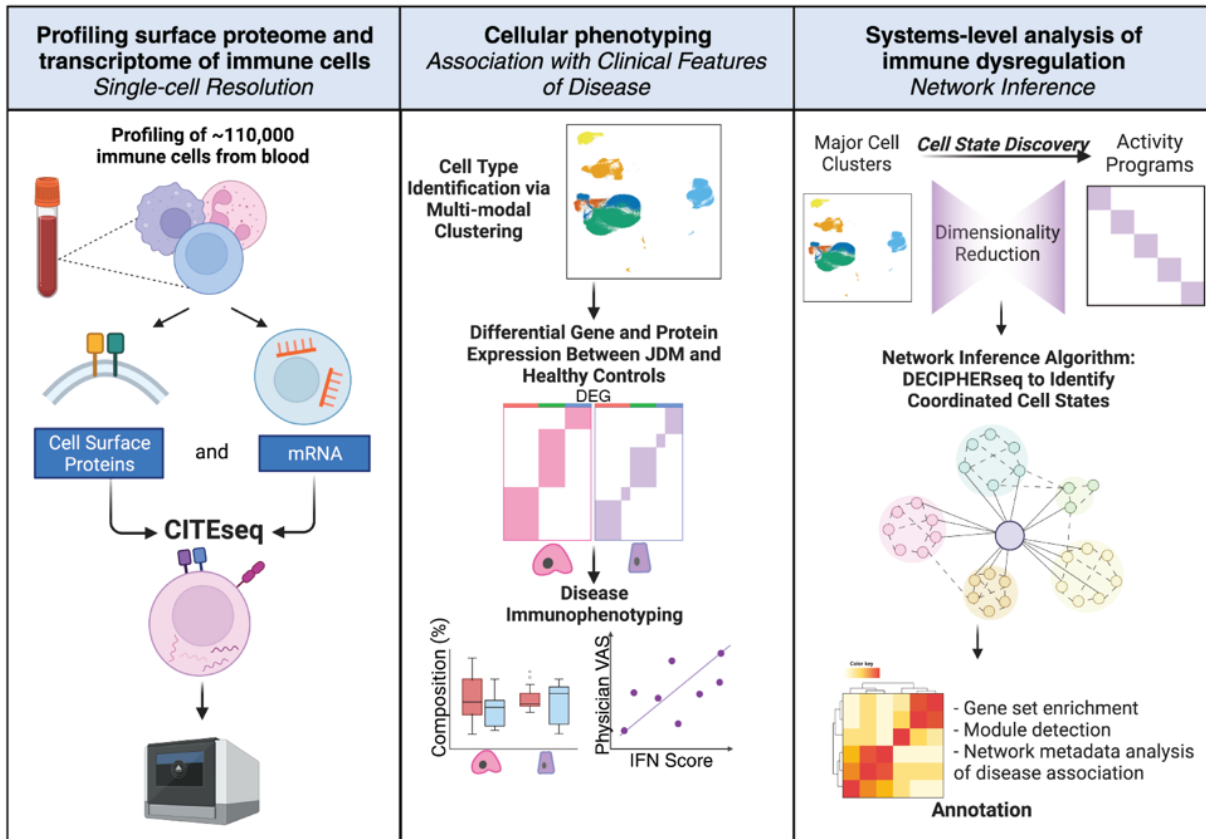


Figure 4.2 Analysis strategy for CITEseq data from PBMCs.

Following pre-processing steps, we identified 29 clusters, which comprised 21 distinct immune cell populations across 105,827 cells (**Figure 4.3**). Clusters were annotated using canonical RNA (**Figure 4.3**) and protein markers (**Supplemental Figure 4.14**) within all major mononuclear immune cell compartments.

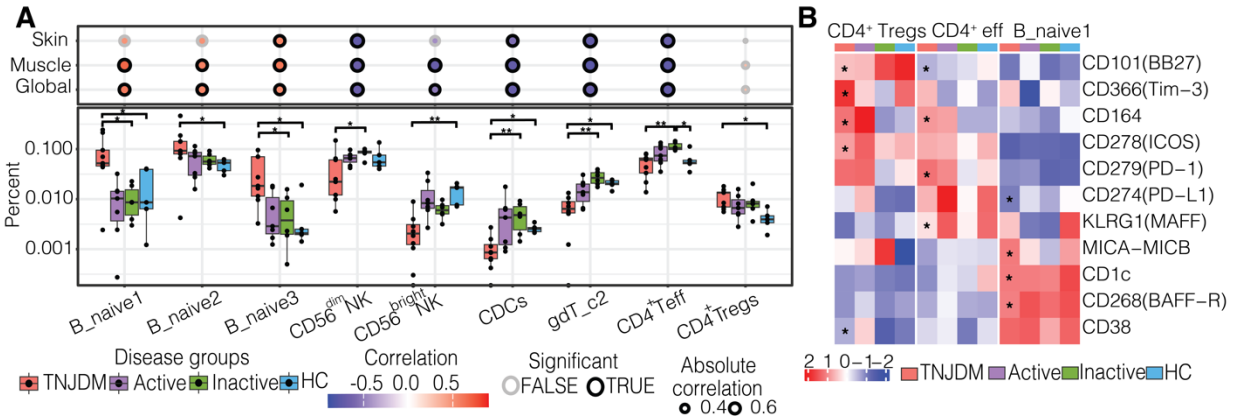


Figure 4.4 Immunophenotypes in peripheral blood associated with JDM. (A) Boxplot shows cell type proportion by disease group, using Kruskal-Wallis test with Dunn’s post hoc test comparing TNJDM to HC, TNJDM to inactive JDM, and inactive to HC (Holm $p_{adj} < 0.05$). The dotplot above shows the Spearman correlation between corresponding cell type proportion in boxplot and Physician Global VAS, where the size of the dot indicates the correlation, the color indicates the direction of the correlation, and the border weight indicates significance ($p_{adj} < 0.05$). (B) Heatmap with selected ADT protein markers. Asterisks mark significant comparisons between TNJDM and HC per cell type with an absolute LFC > 0.5 and $p_{adj} < 0.05$.

Compared to healthy controls and patients with inactive disease, treatment-naive patients had higher proportions of multiple naive B cell populations, including B_naive1 (IgM⁺IgD⁺CD38⁺CD24⁺CD10⁺) corresponding to an immature naive B population, B_naive2 (IgM⁺IgD⁺CD38^{lo}CD24^{lo}), and B_naive3 (IgM⁺IgD⁺CD38⁺CD24⁺), and the proportion of these populations positively correlated with multiple disease activity measures ($p < 0.05$, Spearman) (**Figure 4.4, Supplemental Figure 4.15**). The proportion of B_mem cells, characterized by *TNFRSF13B* (encoding TACI) expression, negatively correlated with the muscle VAS score ($p < 0.05$, Spearman). The immature naive B population had higher expression of CD38 (both RNA and protein) and *MZB1*, two genes essential for plasma cell differentiation, than all other B cell clusters.^{142,143}

Given the observed imbalance of lymphocytes in treatment-naive JDM, we next sought to immunophenotype B cell and CD4⁺ T cell subsets in JDM at the proteomic level to gain molecular insight into cell states (**Figure 4.4**). Differential protein analysis of

immature naive B cells comparing treatment-naive JDM to HC identified increased expression of MICA-MICB and decreased expression of CD1C, BAFF-R and PD-L1 (**Figure 4.4**). Within the CD4⁺ T compartment, CD4⁺Tregs from TNJDM had higher expression of Tim-3, ICOS, CD164 and CD38 and down-regulation of CD101 a molecule which decreases pro-inflammatory T cell responses.¹⁴⁴ CD4⁺Teff in patients with treatment-naive JDM had higher surface expression of CD164 and PD-1 and down regulation of KLRG1, an inhibitory molecule (**Figure 4.4**). The over-expression of PD-1 on the cell surface suggested that peripheral T helper cells might be present in JDM.^{145,146} However, while ICOS expression was higher (Benjamini-Hochberg (BH) $p < 0.05$), no difference was found in surface expression of CXCR5 between CD45RO^{hi}PD-1^{hi}CD4⁺ T cells and CD45RO^{lo}PD-1^{lo}CD4⁺ T cells, and these cells were not significantly expanded in JDM (**Figure 4.4**). Taken together, these compositional and immunophenotyping observations add to the growing body of work showing that JDM in the treatment-naive state is characterized by relative imbalances of peripheral naive and mature lymphocyte states,^{133,141} reduced innate immune populations (22) and distinct CD4⁺ T and B cell immunophenotypes.^{134,135}

SIGLEC-1 expression is a composite measure of the IFN gene signature in JDM

We next compared gene and protein expression between treatment-naive JDM and HC samples in all cell types based on the hypothesis that certain cell types may not be altered in composition but may be functionally altered at the molecular level.

Monocytes displayed the highest number of differentially expressed genes (n = 211) and proteins (n = 19) in this pairwise analysis including CD169 (SIGLEC-1), CD107a (LAMP-1), and CD164 (**Supplemental Figure 4.16**). SIGLEC-1 is a monocyte-restricted IFN-induced protein that was recently identified as a potential biomarker in JDM.¹⁴⁷ Both CD107a and CD164 are cell adhesion molecules involved in trafficking of activated mononuclear cells and adhesion to vascular endothelium.¹⁴⁸

A common finding across all cell types when comparing treatment-naïve JDM and HC samples was overexpression of genes enriched in Type I IFN processes, which was previously reported in bulk expression data^{149,150} and confirmed in single-cell studies (**Supplemental Figure 4.17**).^{133,141} Using an IFN gene score derived from the transcriptional data (**Supplemental Figure 4.18**), we plotted the average score per patient per cell type and applied hierarchical clustering and observed variation among individuals and cell types (**Figure 4.5**). This approach did not detect IFN gene expression to persist beyond the treatment-naïve state and two patients within the treatment-naïve group had negligible IFN gene signature as quantified by this method. This heterogeneity of the IFN gene signature was, in part, explained by disease activity level (**Figure 4.5**), as a bulk IFN gene score significantly correlated with disease activity (R=0.69, Spearman). However, the remaining unexplained heterogeneity of this IFN score suggests additional biological sources of disease activity in patients with JDM and exemplifies a limitation of utilizing gene scores identified through pairwise comparisons between subsets of the data.

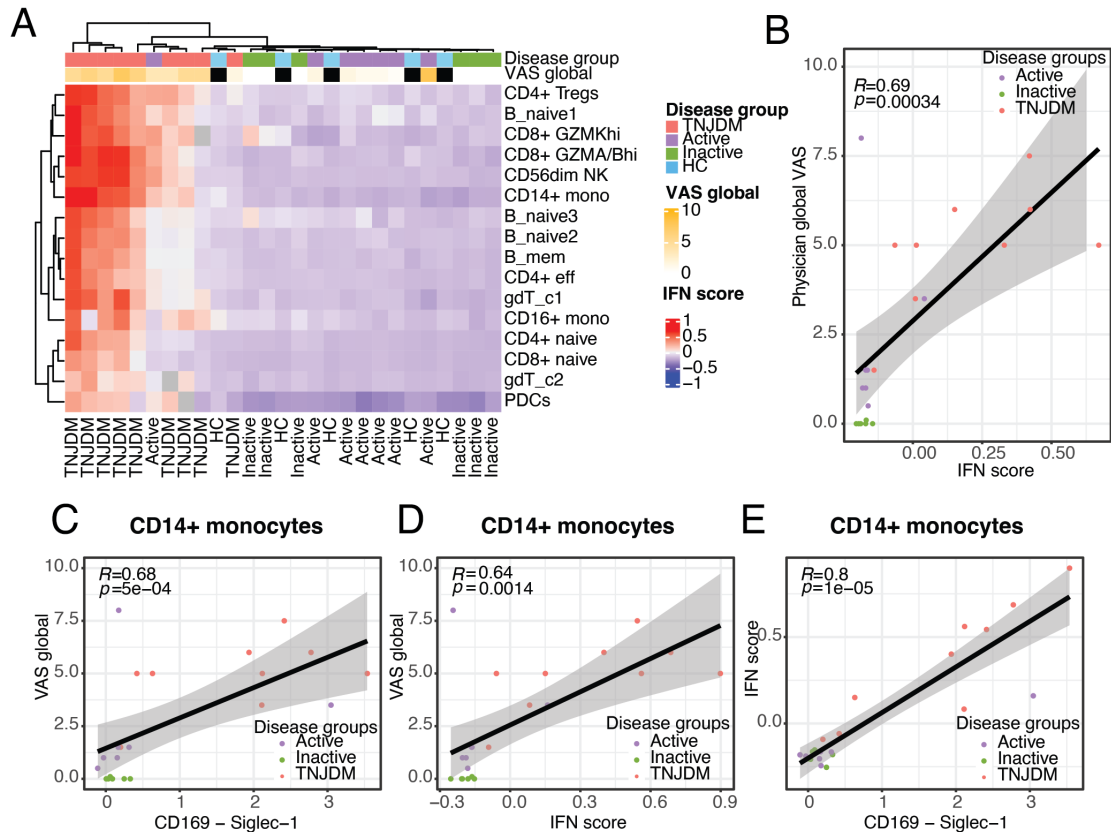


Figure 4.5 Type I IFN-induced gene and protein expression is associated with disease activity in JDM in CD14+ monocytes.

(A) Heatmap of average IFN score per cell type and sample. Hierarchical clustering was performed using Euclidean distance and the complete clustering method. IFN score was calculated based on average expression of IFN module across all cells per sample.

(B) Spearman correlation between IFN score and Physician Global VAS colored by disease group.

(C) Scatter plot showing Spearman correlation between CD169 (SIGLEC-1) expression in CD14+ monocytes and Physician Global VAS.

(D) Scatter plot showing Spearman correlation between IFN score and Physician Global VAS. (E) Scatter plot showing Spearman correlation between CD169 expression and IFN score in CD14+ monocytes.

Given that SIGLEC-1 is a type I IFN-induced protein, we next wanted to determine if patterns of type I IFN stimulated gene expression were reflected at the protein level, as protein biomarkers are more amenable for clinical lab-based testing. SIGLEC-1 expression in CD14⁺ monocytes correlated with disease activity to a similar degree as the IFN gene signature (**Figure 4.5**), and SIGLEC-1 expression was itself highly correlated with the IFN gene signature in monocytes (**Figure 4.5**). This suggests that SIGLEC-1 expression in CD14⁺ monocytes is a representative composite measure of the IFN gene signature in JDM. These results underscore the potential of SIGLEC-1 as a biomarker of

IFN responses in JDM that may be useful for stratifying disease severity and tracking disease activity.

Unsupervised network analysis reveals coordinated immune cell states in JDM

We next turned to a systems level approach to better understand the coordination of immune cell gene programs in JDM relative to healthy controls and in relation to disease activity level. This approach also overcomes limitations of differential expression analysis which relies on pairwise comparisons of subsets of data. We applied an unsupervised network inference method, DECIPHER, to the 6 major cell types annotated in the dataset: B cells, CD4T, CD8T, NK cells, gdT cells, and myeloid cells (**Figure 4.6**). DECIPHER relies on non-negative matrix factorization (NMF)^{46,77,106} to first break the dataset down into gene sets that represent distinct states of biological activity, or ‘activity programs’, and then constructs a network of gene expression programs (GEPs) based on how expression of the programs covaries across patient samples (**Figure 4.6**). After outlier filtering, NMF identified 76 activity programs (**Figure 4.6**).

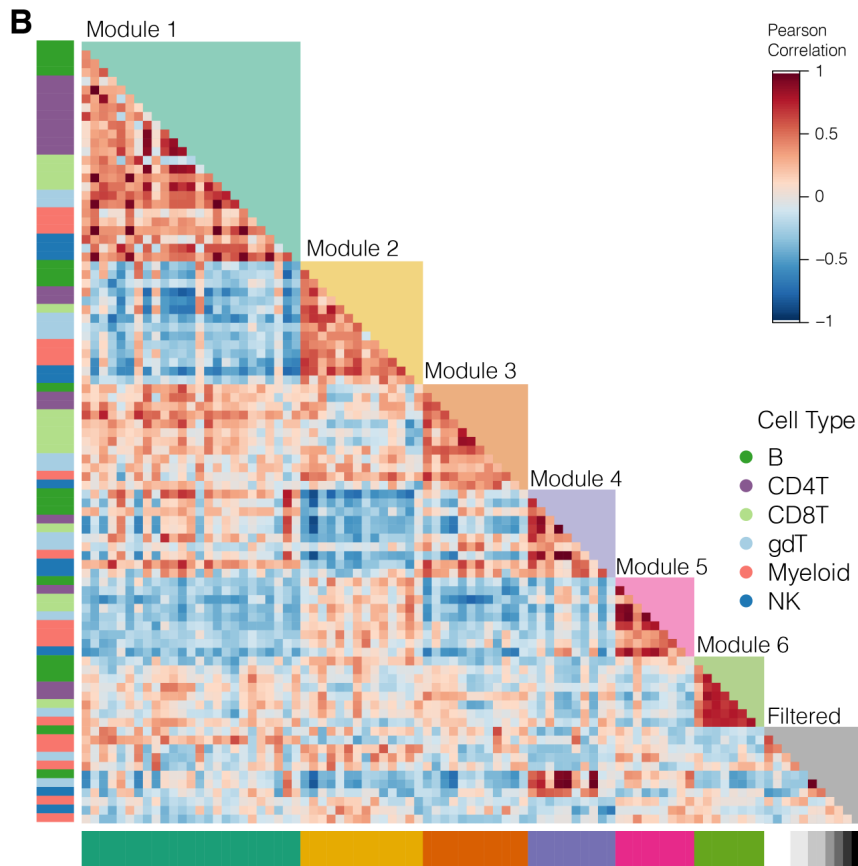
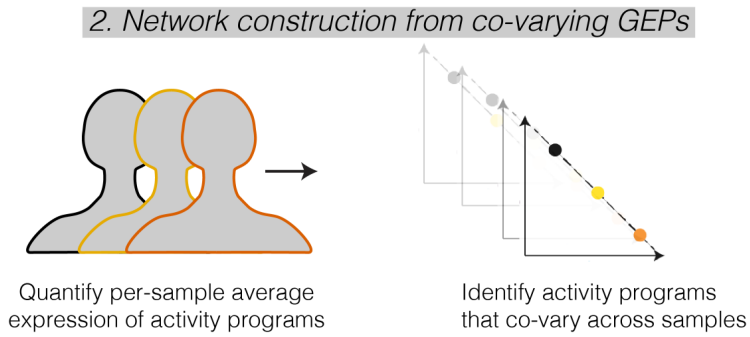
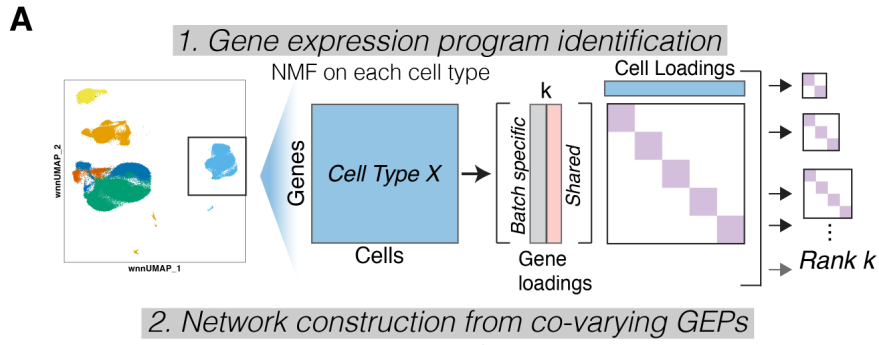


Figure 4.6 DECIPHER learns coordinated biological activity programs through dimensionality reduction.

(A) Overview of the DECIPHER workflow.

(B) Heatmap showing 6 major clusters of GEPs identified by DECIPHER (Pearson). GEPs are clustered into modules, with isolated GEPs filtered out (greyscale).

Next, a force-directed network graph from the correlation matrix of activity programs was constructed where each node represents a program, and each edge represents a statistically significant positive correlation between two nodes. Correlations between programs are accounted for in the network visualization such that further apart nodes can be interpreted as being negatively correlated and closer nodes can be interpreted as being positively correlated programs (**Figure 4.7**). Using DECIPHER's community detection algorithm, we identified 6 hubs of inter-connected activity programs or 'modules.' All modules contained multiple cell types, highlighting that many biological processes in JDM are coordinated across several immune cell types (**Figure 4.6, Figure 4.7**). We annotated each node using gene set enrichment analysis^{151,152} of gene ontology terms (GO)⁸⁸ on each program's ranked marker gene list (**Supplemental Figures 4.19-24**).

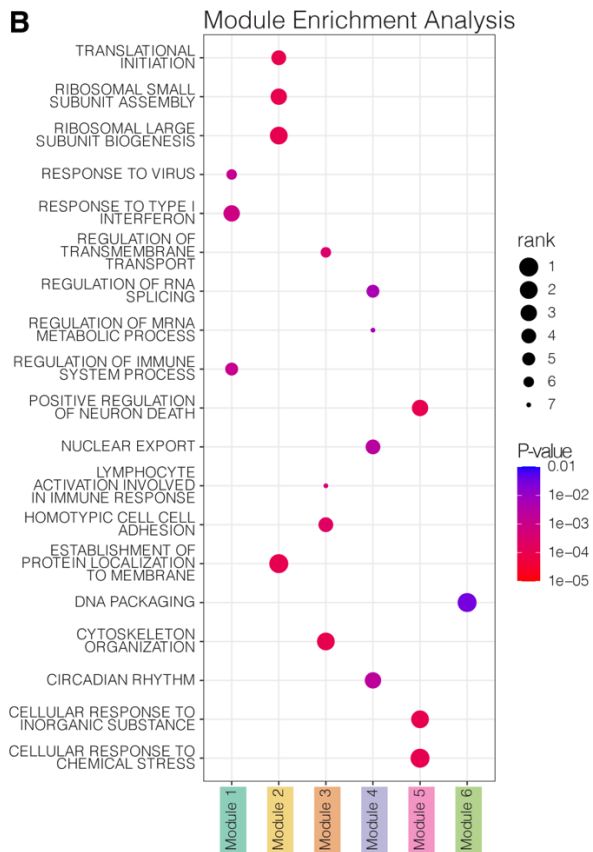
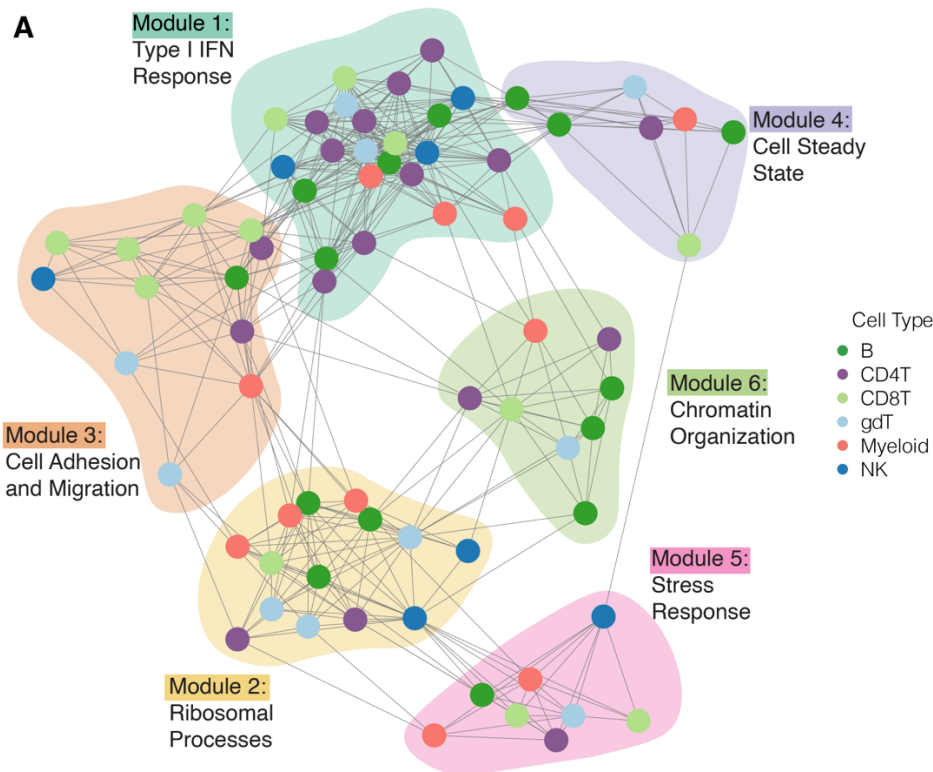


Figure 4.7 DECIPHER reveals network of coordinated biological activity from scRNA-seq data in JDM. (Figure caption continued on next page)

(Figure caption continued from previous page)

(A) Force-directed network constructed from correlated GEPs in PBMCs from JDM patients and healthy controls. Nodes represent programs in the given cell types and edges represent positive significant correlations (Pearson, $p < 0.05$). (B) Dotplot showing selected gene sets found to be enriched within specific modules compared to the rest of the network. Color corresponds to module enrichment p value and size corresponds to a set's rank in list of significantly enriched gene sets for that given module ordered by ascending module enrichment p -value. All gene sets shown fall in the top 10 terms for their respective modules (total gene sets: 626).

DECIPHER's module enrichment analysis identified consensus biological themes for each module in an unsupervised manner (**Figure 4.7**). Module 1 was enriched for Type I IFN response programs including gene sets such as 'Response to Virus'. Module 2 consisted of programs enriched for protein assembly genes used in ribosomal processes including 'Translational Initiation'. Module 3 was comprised of mostly lymphocyte programs and was significantly enriched for gene sets related to cell adhesion and migration. Module 4 represented cells' steady state processes as it was enriched for gene sets like 'Circadian Rhythm'. Module 5 was annotated as a Stress Response module because it was enriched for 'Regulation of Cell Death' and 'Cellular Response to Chemical Stress'. Module 6 contained very few gene sets that were unique to the module, as it consisted of programs enriched for programs intrinsic to eukaryotic cells like 'DNA Packaging.'

JDM CD4+T cells and B cells display persistent alterations in gene expression in both active disease and remission

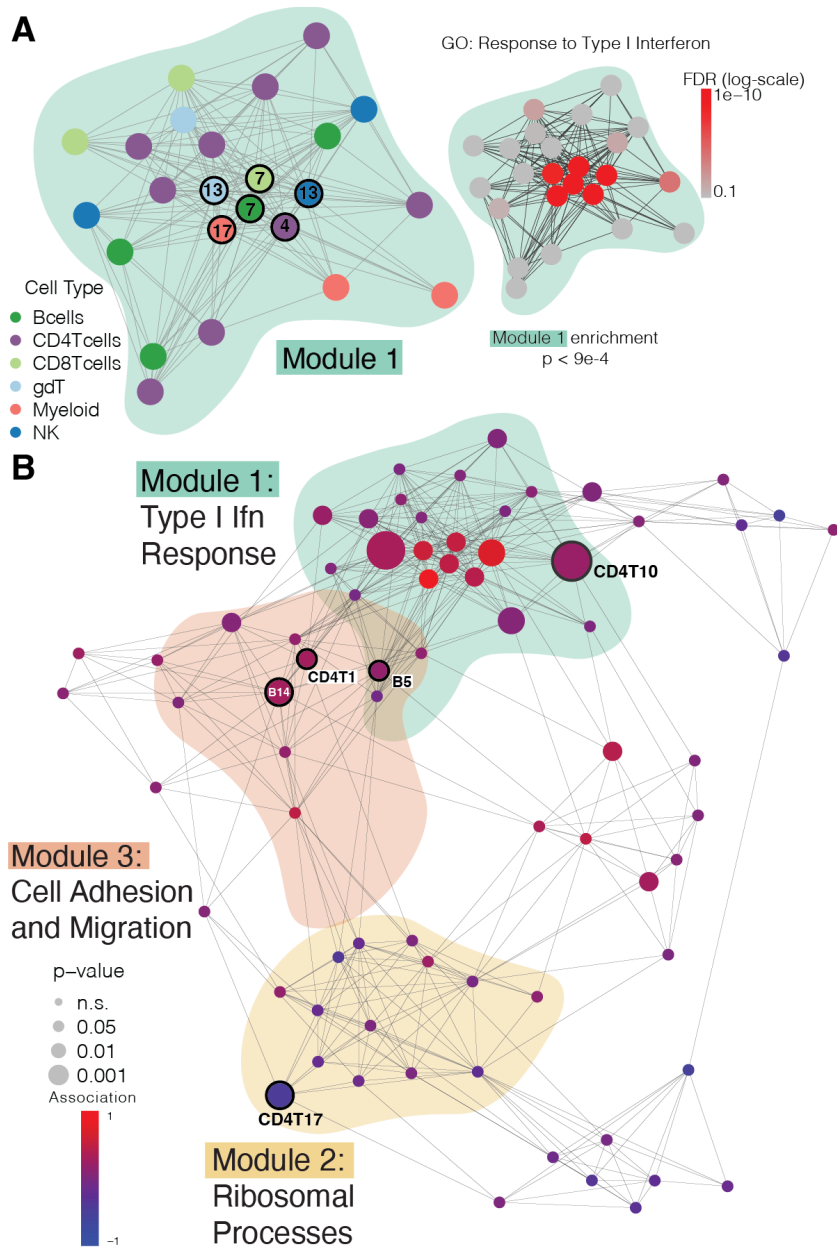


Figure 4.8 JDM is associated with a central IFN hub and cell specific gene programs in the B and CD4T compartments.
(Figure caption continued on next page)

(Figure caption continued from previous page)

(A) Zoomed in graph of Module 1. GSEA results for Response to Type I IFN GO term shown with each node colored according to FDR. Adjusted p-value of module enrichment is also shown.

(B) Network graph showing case-control analysis of each program's expression, with node size scaled according to p-value and colored according to strength of the association between disease status and program expression (t-test).

Next, we aimed to interpret the annotated network in the context of GEPs associated with JDM compared to healthy control patients irrespective of disease activity. In the annotated network, we first focused on Module 1, which was enriched in type I IFN responses and many programs in this module were increased in TN-JDM, as expected (**Figure 4.8**). All 6 major cell types expressed an IFN gene program which were highly correlated to one another, as shown by the closely connected hub at the center of Module 1 (**Figure 4.8**). This IFN hub was associated with JDM as compared to HC patients (t-test, $p < 0.05$) (**Figure 4.8**). IFN modules identified by NMF were highly expressed in all treatment-naïve patients as well as some patients with active disease, inactive disease, and a healthy control patient (**Figure 4.9**), in contrast to the signature of IFN gene expression previously detected by differential gene expression in **Figure 4.5**. This highlights the strength of this method to reflect the low-dimensional space of gene expression where measurement of many genes working together may be needed to detect underlying biological processes more accurately.^{55,58,76}

We next identified additional coordinated gene programs in Modules 1-3 expressed more highly in all JDM patients compared to HCs ($p < 0.05$), irrespective of disease activity. These JDM associated programs included B cell (5 and 14) and CD4T (1, 10, 17) programs and their expression persisted even in patients with inactive JDM who previously achieved remission off medication (**Figure 4.9, Supplemental Figure 4.25**). JDM patients more highly expressed two B cell programs: B5 in Module 1 was enriched in mRNA metabolic processing, RNA splicing, chromatin organization and

modification, and cell cycle regulation, and B14 in Module 3 was enriched in chromatin remodeling and cytoskeletal organization (**Supplemental Figures 4.19, 4.21**). These enriched biological processes suggest that a subpopulation of B cells is more transcriptionally active and undergoing epigenetic regulation in JDM relative to healthy controls.

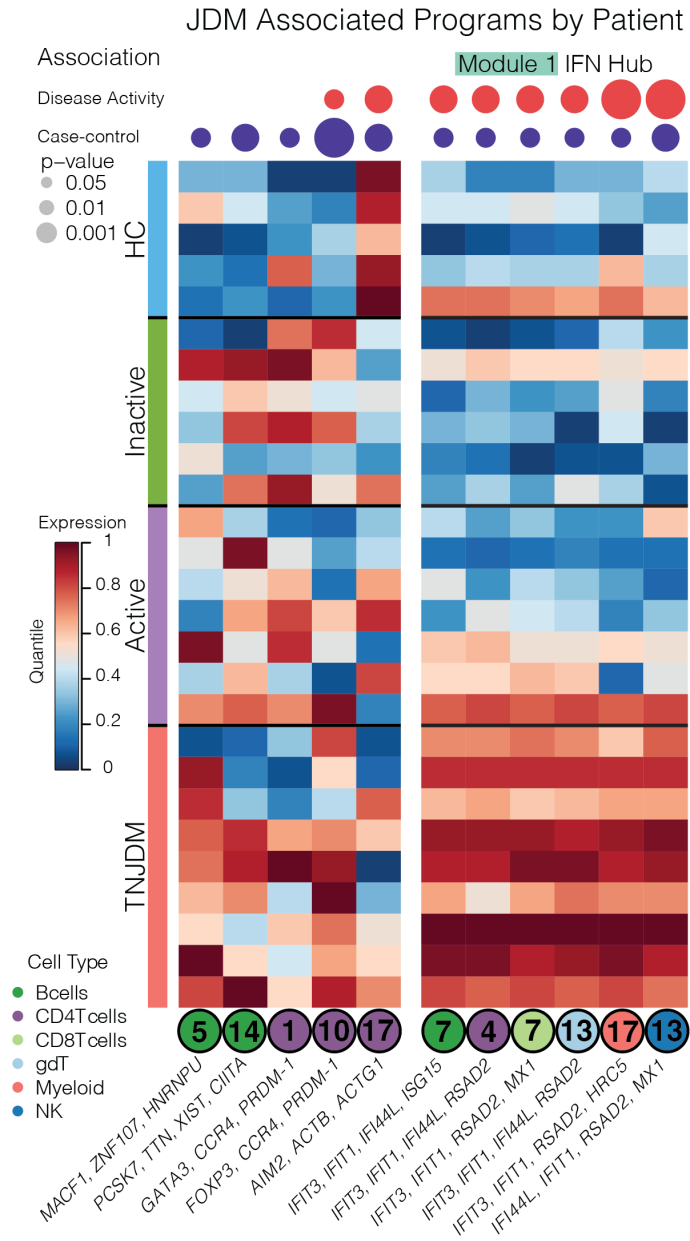


Figure 4.9 Heatmap showing significant differences in expression of selected programs between HC (n=5) and JDM patients (n=22), with columns annotated by p-values of case-control (t-test) and disease activity association (4-group ANOVA).

In Module 3, correlated to B14, CD4T1 (enriched in cell migration, adhesion, activation, and secretion) was expressed more highly in JDM and in the region of the UMAP corresponding to CD4⁺Teff cells (**Supplemental Figure 4.26**). This CD4T1 program expressed by CD4⁺Teff cells contained genes (*GATA3*, *CCR4*, *PRDM1*) that indicate possible skewing towards a Th2 subset while expression of *PRDM1* (Blimp-1) suggests participation in extra-follicular reactions (**Figure 4.9**). Th2 CD4⁺ T cells were previously found to be expanded in JDM and associated with extra-follicular B-T cell help.^{134,135}

We observed similar expression of Th2 genes (*GATA3*, *CCR4*, *PRDM1*) in CD4T10, a Treg program (*FOXP3*, *IKZF2*, *IL2RA*) expressed more highly in JDM (**Figure 4.9**). CD4T1 and CD4T10 included genes for costimulatory molecules OX40 (*TNFRSF4*) and GITR (*TNFRSF18*), both of which have been described to promote survival and proliferation of CD4⁺ T effector cells and have been targets of autoimmune disease therapeutics.^{153–157} Notably, CD4T17 (*AIM2*, *ACTB*, *ACTG1*, *NCF1*, *ID3*, *SOX4*) was negatively associated with JDM and expression was significantly decreased in nearly all patients (**Figure 4.9**, **Supplemental Figure 4.25**, **4.27**). This program was enriched in protein targeting to the membrane and endoplasmic reticulum and included several genes important in T cell regulation. *NCF1* has been found to be a critical regulator of T cell tolerance in a collagen-induced arthritis mouse model¹⁵⁸ and co-expression of *ID3* and *SOX4* transcription factors has been identified as a mechanism of CAR-T cell exhaustion and dysfunction.¹⁵⁹ Together, these results suggest multiple mechanisms by which

CD4+T cell dysfunction may occur in JDM including participation in extra-follicular reactions, expression of co-stimulatory molecules, and down-regulation of genes important in mediating tolerance and exhaustion.

Novel cell states are correlated with IFN gene expression in treatment-naive JDM

We next wanted to identify modules and gene programs associated with stages of disease activity in JDM (HC, Inactive and Active JDM, and treatment-naive JDM). To do so, we performed a 4 group ANOVA on each program in the network and post-hoc pairwise analysis using the Tukey test. We identified programs in Module 1, 2 and 5 that were significantly associated with disease activity (ANOVA $p < 0.05$) (**Figure 4.10**). We confirmed that these biological programs moved in the direction expected within most patients with longitudinal assessments (**Supplemental Figure 4.28**). The IFN gene programs were also significantly overexpressed in treatment-naive JDM patients, as expected (**Figure 4.9**). Notably, expression of the central Module 1 IFN hub GEPs in all six major cell types more strongly correlated to the clinically evaluated PGA than the pseudobulk IFN gene score derived from pairwise DEG analysis (**Supplemental Figure 4.29**), underscoring the utility of a dimensionality reduction approach in uncovering clinically relevant gene signatures.

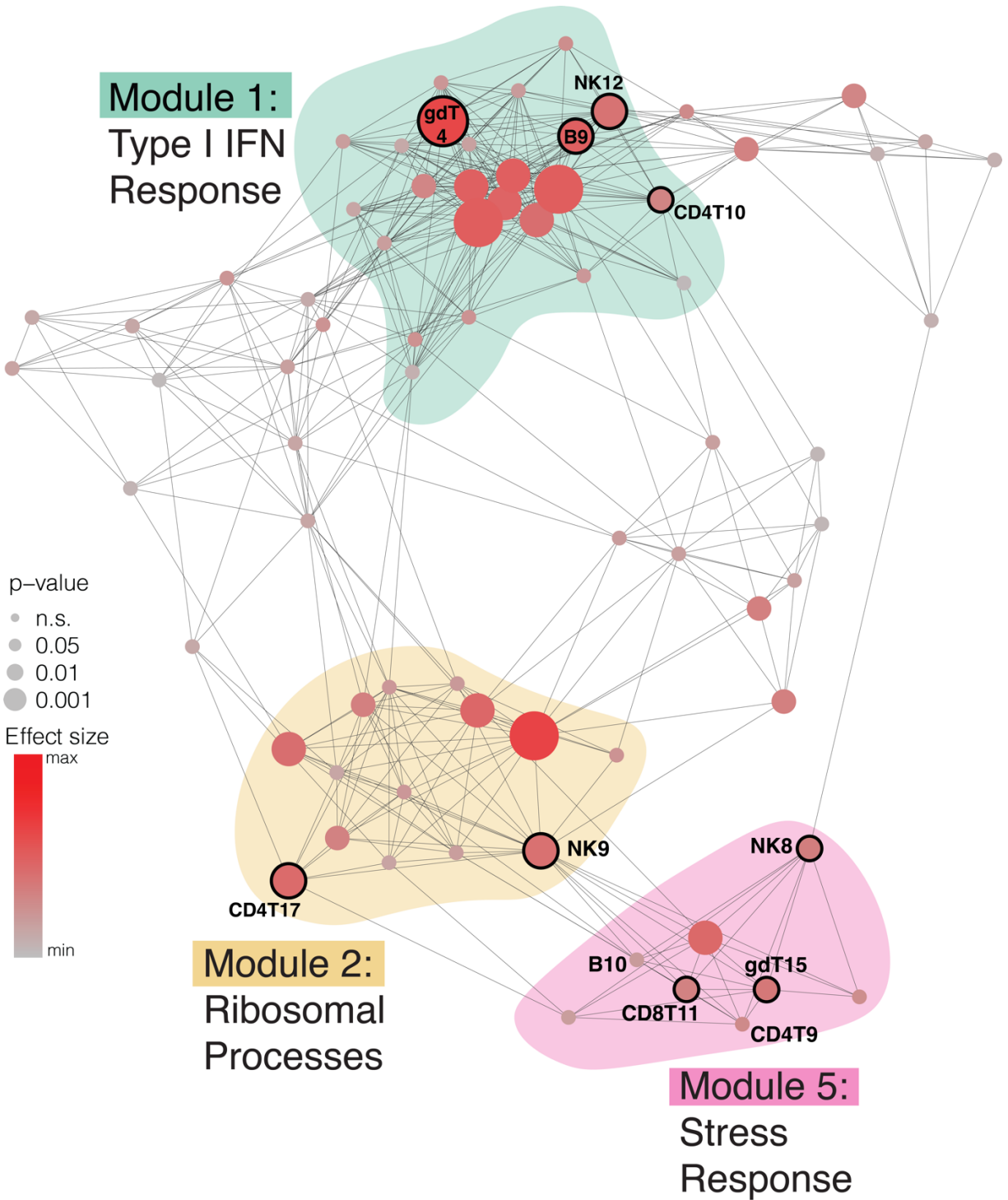


Figure 4.10 Disease activity in JDM is associated with central hub of IFN response in network, correlated with novel dysregulated cell states. Network graph showing results of 4-group ANOVA of each program's expression, with node size scaled according to p-value and colored according to strength of the association between disease status and program expression.

By isolating these IFN GEPs in each cell type, we were able to determine disease activity-associated programs correlated with the IFN hub, some of which corroborate previous findings (**Figure 4.10, Figure 4.11**). This approach identified B9, an immature naive B cell program (*CD24, CD38, MME*), to be significantly associated with disease activity (**Figure 4.11**). This gene program shared several top markers (*TCL1A, SOX4, NEIL1*) with the immature B cell population that was previously found to be expanded in treatment-naive JDM (**Supplemental Figure 4.27**).^{133,141} Notably, expression of this activated immature B cell program could be attributed to the B_naive1 cluster that we observed to be increased in treatment-naive JDM during the compositional analysis (**Supplemental Figure 4.30**). Similarly correlated with the IFN hub, NK12 was associated with treatment-naive JDM compared to active and inactive disease (**Figure 4.11, Supplemental Figure 4.31**). NK12 (*MKI67, HIST1H1B*) was enriched for gene sets related to cell proliferation and epigenetic regulation, confirming findings that a subset of NK cells in JDM are highly activated and proliferative.^{132,140}

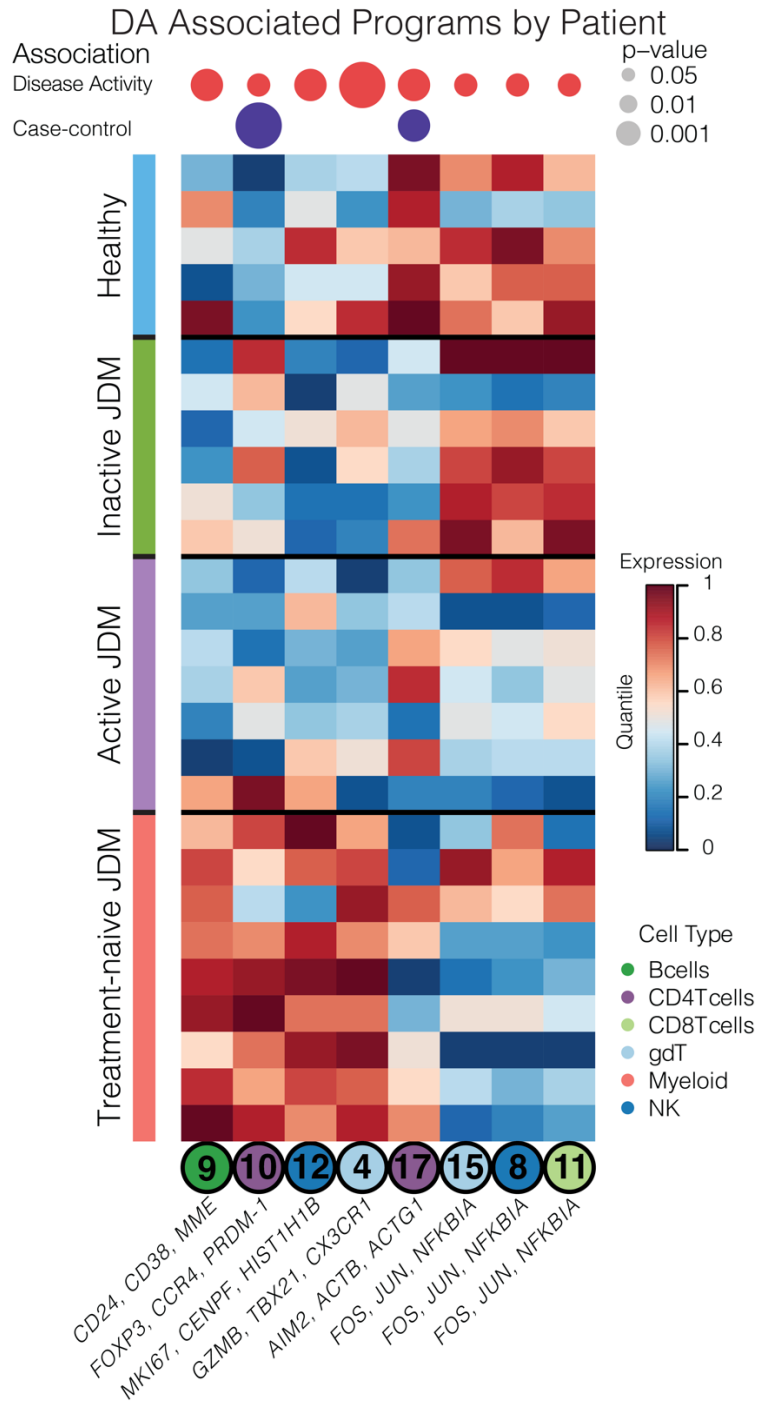


Figure 4.11 Heatmap showing significant differences in expression of selected disease activity associated programs between HC (n=5), Inactive JDM (n=6), Active JDM (n=7), and TNJDM patients (n=9). Columns are annotated by p-values of case-control t-test and disease activity association (4-group ANOVA).

We next focused our attention on the other disease activity-associated programs that DECIPHER identified as correlated with the Module 1 IFN hub. Importantly, these disease activity associations were only revealed in the lower dimensional space of gene

sets identified by NMF rather than the noisier space of differential expression of individual genes. We annotated CD4T10, also significant in our case control analysis, as a proliferative Treg program (*FOXP3*, *IL2RA*, *MKI67*) which expressed genes implicated in extra-follicular B-T interactions (*PRDM1*) and genes associated with Th-2 mediated inflammation (*GATA3*, *CCR4*, **Figure 4.11**).¹⁶⁰ Notably, CD4T10 included the marker *CCR4*, a chemokine receptor highly expressed in Tregs that are preferentially recruited to skin under inflammatory conditions.¹⁶¹ Expression of both CD4T1 and CD4T10 co-localized with surface protein expression of *CCR4* in the UMAP as well (**Supplemental Figure 4.32**), highlighting the advantage of this multi-modal sequencing approach in identifying functional markers of transcriptomic signatures.

This network approach also identified the program gdT4, a cytotoxic Th1 polarized gdT program (*GZMB*, *CX3CR1*, *TBX21*) that was correlated with the central IFN hub and was significantly increased in treatment-naive patients compared to both active and inactive JDM and HC ($p < 0.05$, Tukey, **Figure 4.11**). High expression of *TRGC1* and *TBX21*, encoding the transcription factor T-bet responsible for regulating IFNG expression, specifically identified cells expressing this program as Th1-like TCRVd1 gdT cells (**Figure 4.11**).^{162,163} A similar subpopulation of gdT cells was found to be increased in synovial fluid and blood of juvenile idiopathic arthritis patients, which expressed IFN γ and TNF to the same degree as CD4⁺ T cells.¹⁶⁴ This suggests this subpopulation of gdT cells may reflect an important inflammatory cell state specific to treatment-naive disease that is potentially up- or downstream of the Type I IFN response broadly upregulated across immune cell populations in JDM.

Regulatory cell death and protein targeting pathways are dysregulated across multiple immune cell populations in JDM

The novel disease activity programs that were highly expressed in treatment-naive JDM were components of Module 1 which was enriched for Type I IFN and its associated immune processes. The network-wide ANOVA analysis also revealed disease activity-associated programs in Module 2 and Module 5 (**Figure 4.10**), which were significantly anti-correlated with Module 1 (**Supplemental Figure 4.33**) and expression was decreased in treatment-naive JDM patients compared to healthy controls and other JDM patients (**Figure 4.11**). Module 2 was significantly enriched for gene ontology terms 'ribosome assembly' and 'translational initiation' while Module 5 was enriched for terms 'regulation of cell death' and 'cellular response to chemical stress' (module enrichment $p < 0.005$) (**Figure 4.12**). The disease-associated programs within these modules were expressed significantly lower in treatment-naive JDM, suggesting dysfunction of cellular processes that underpin ribosomal activity and cell death regulatory processes at disease onset (**Figure 4.11, Supplemental Figure 4.31**).

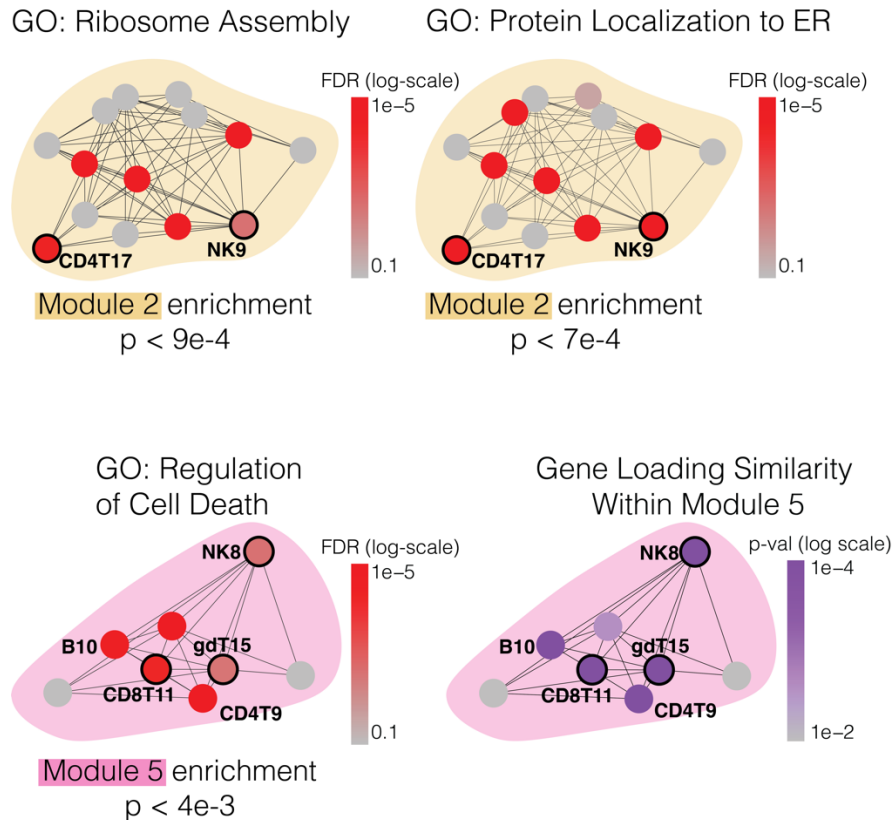


Figure 4.12 Selected network modules colored by FDR of enrichment for indicated gene ontology set ($FDR < 0.01$) or gene loading similarity within Modules 2 and 5.

Notably, disease activity-associated programs CD8T11, NK8 and gdT15 (*FOS*, *JUN*, *DUSP1*, *NR4A2*, *GADD45B*) in Module 5 share a common gene signature (**Figure 4.11**) and are each individually enriched in ‘regulation of cell death’ and ‘regulation of cell cycle’ (**Figure 4.12**, **Supplemental Figure 4.23**). We quantified the overlap in gene expression between activity programs by Fisher’s exact test and confirmed the high gene loading similarity between programs in Stress Response Module 5 (**Figure 4.12**). All three of these programs were expressed at lower levels in active and treatment-naive JDM and negatively correlated with activated disease-associated cell signatures identified in Module 1 (**Supplemental Figure 4.31**). The gene loading similarity analysis revealed that the programs CD4T9 and B10 also share top marker genes (*FOS*, *JUN*, *DUSP1*, *NR4A2*,

GADD45B) (**Figure 4.12**), however, these two programs were not associated with disease activity status. This suggests that regulatory mechanisms of cell death may be uniquely disrupted in circulating cytotoxic cell populations in patients with active disease.

In Module 2, CD4T17 and NK9 were enriched in several gene sets related to protein processing such as protein targeting to the ER (**Supplemental Figure 4.20**). Interestingly, the CD4T17 program was also characterized by high expression of several genes encoding members of the actin protein family (*ACTB*, *ACTG1*). Given the crucial role actin filaments play in antigen recognition during the formation of the immune synapse, dysfunction in components of that protein machinery could have significant effects on the immune system. Among other disease activity-associated programs, differential expression of CD4T17 between HCs and JDM patients persisted even in patients who achieved remission off medication (**Figure 4.11**). Taken together, disease activity-associated programs in Modules 2 and 5 highlight shared cellular processes that may be under-active in JDM, providing new insights into potential cellular mechanisms that accompany the known signature of overactive IFN-response in JDM.

JDM-associated signatures identified by DECIPHER validated in an independent dataset

We next investigated whether these JDM-associated signatures could be identified in an independent set of samples. Using DECIPHER's marker quantification method, we subset the genes in each JDM-associated GEP that contributed the most to that program. With each gene list as input, we calculated a proxy GEP metric that quantified rank-based

expression of each program as the enrichment of that subset of genes in each cell.¹⁶⁵ This proxy NMF GEP metric recovered signatures identified by NMF in the original dataset (**Supplemental Figure 4.34**).

Using this proxy NMF GEP method, we validated key signatures in an independent set of CITEseq data from 5 JDM samples and 2 HC (**Figure 4.13**). Importantly, these samples were obtained from patients who had active disease and 4 of 5 were being treated with medication. We compared GEP expression between cases and controls (t-test, $p < 0.05$), which identified significantly higher expression of CD4T1 (CD4+Teff program, *CCR4*, *PRDM-1*, *GATA3*) and CD4T10 (Treg program, *FOXP3*, *CCR4*, *PRDM-1*, *GATA3*) in JDM and lower expression of CD4T17 (*AIM2*, *ACTB*, *ACTG1*), replicating the original results in the initial cohort (**Figure 4.13**). In the B cell compartment, B5 and B14 trended toward increased expression in patients with JDM. However, a single individual had low expression in both programs indicating heterogeneity of expression of these B cell signatures (**Figure 4.13**).

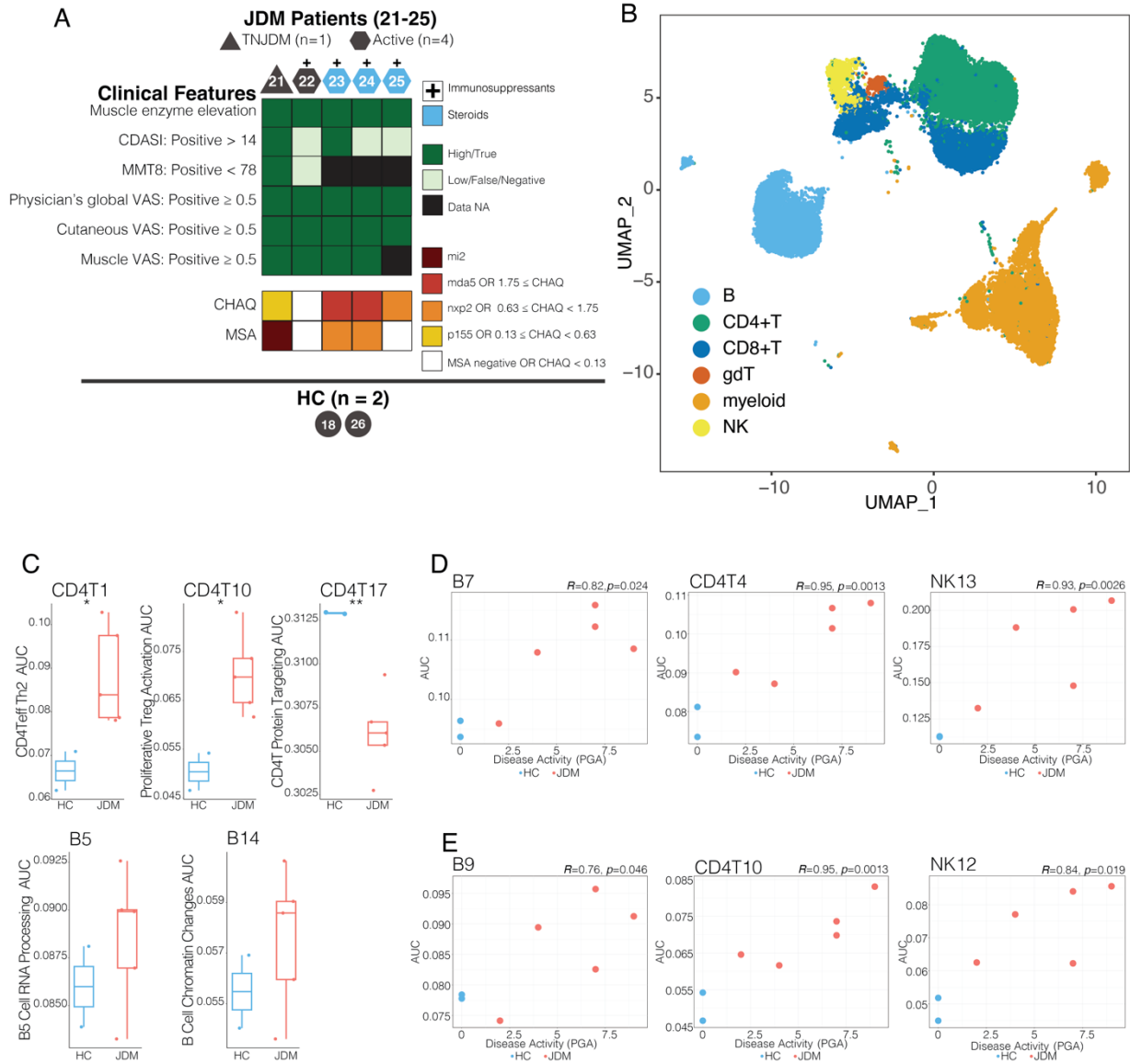


Figure 4.13 JDM-associated signatures identified by DECIPHER can be validated in independent samples.

(A) Clinical characteristics of validation cohort. HC18 was included in original cohort but an independent sample was collected and analyzed for this dataset. Individuals are labelled by the donor ID used throughout the manuscript. Immunosuppressants denoted as '+' for patients JDM22-25 were as follows: (JDM22 - methotrexate), (JDM23 - IVIG, cytoxan), (JDM24 - hydroxychloroquine, MMF, IVIG, tofacitinib), (JDM25 - methotrexate, IVIG).

(B) UMAP of scRNA-seq data from validation cohort PBMC samples, colored by six major cell types corresponding to labels used in original cohort.

(C) Boxplots of case-control comparisons (HC=2, JDM=5) for selected programs queried in validation dataset using AUCcell (t-test, * $p < 0.05$, ** $p < 0.01$).

(D-E) Scatterplots correlating disease activity (PGA) with AUCcell scores for selected IFN programs (D) and selected disease activity programs (E) in validation dataset (Spearman).

To validate disease-activity associated programs, we correlated GEP expression of disease activity associated programs with the PGA score. It was infeasible to validate gdT programs due to low cell numbers. IFN GEPs were most strongly correlated with

disease activity in CD4T cells, B cells and NK cells (**Figure 4.13**) and trended toward a positive correlation in myeloid cells, but there was no significant association in CD8T cells (**Supplemental Figure 4.35**). Additionally, B9 (immature B cell signature, *CD38*, *CD24*, *MME*), CD4T10 (Treg program, *FOXP3*, *CCR4*, *PRDM-1*, *GATA3*), and NK12 (proliferative activated NK cell program, *MKI67*, *CENPF*, *HISTH1HB*) strongly correlated with disease activity (**Figure 4.13**). A subset of the key signatures identified in the original cohort with DECIPHER trended with disease activity but did not significantly correlate with PGA in the validation cohort (**Supplemental Figure 4.35**). Together, these results demonstrate the robustness of many signatures quantified with a less precise method even in an independent dataset with fewer samples. Cell death regulatory signatures previously identified to be negatively correlated with IFN signaling in cytotoxic populations may be more strongly associated with a treatment-naïve state or more heterogeneous across healthy individuals such that significant differences in expression could not be identified in this smaller cohort.

Discussion

Multiple components of the adaptive and innate immune compartments have been implicated in the pathogenesis of JDM consistent with its categorization as a complex autoimmune condition. However, previous studies have been unable to uncover how multiple disease-associated cell states are coordinated to produce inflammation. Here, in the largest single-cell study of JDM to date, we provide an unbiased, comprehensive picture of immune dysregulation in peripheral blood, including a subset of aberrant signatures that persist despite disease quiescence in individuals off medication. Through traditional analyses, we first show that immune dysregulation in JDM manifests at the level of compositional imbalance of immune populations and that these compositional changes are correlated to clinical metrics of disease activity. Next, we identify distinct disease-associated molecular signatures of lymphocyte and myeloid subsets through multi-modal differential analysis and demonstrate that of these markers, surface expression of SIGLEC-1 in CD14⁺ monocytes is a composite metric of disease activity and reflects the type I IFN response in JDM.^{150,166–168} Using DECIPHER to deconvolve disease-associated programs beyond the broad type I IFN response, we uncover novel CD4⁺ T cell states that persist in JDM despite disease remission coordinated with down-regulated cell death processes in cytotoxic immune populations. Together, these findings generate new hypotheses for disease etiology.

Within the B cell compartment, we observe skewing toward an immature state in treatment-naïve disease and observe the distinct transcriptomic and proteomic signature of immature naive B cells consistent with what we and others previously reported.^{133,141} Given that autoantibodies are thought to play a role in disease pathogenesis, this skewing

of the B cell compartment would seem counterintuitive. However, given recent findings emphasizing the importance of extra-follicular B cell differentiation pathways through which autoreactive “activated naive” B cells are precursors to antibody-secreting cells, we hypothesize that this skewing may be suggestive of extra-follicular reactions in JDM.^{145,146,169} In fact, the expanded immature naive population had higher expression of CD38 and MZB1, genes important for plasma cell differentiation, than all other B cell clusters. The overall low expression of CD27 and CXCR5 across all B cells made it difficult for to conclude if this population matches the double negative B cell population associated with SLE.¹⁶⁹ However, recent immunophenotyping work in a large cohort of JDM patients that found simultaneous expansion of CXCR5- central memory B cells and Th2 cells provides support for further investigation into extra-follicular B-T cell help in JDM.¹³⁴ Alternatively, this skewing could represent more mature B cells homing to tissues as has been described in antisynthetase syndrome.^{170,171} Functional work to support or negate the extrafollicular pathway in JDM will be critical to determine if this is a targetable pathway therapeutically.

Accompanying these immunophenotypic changes in B cells, we observe complementary dysregulation in the T cell compartment that lends further support to the hypothesis of extra-follicular interactions in JDM. In populations of peripheral blood *FOXP3*⁺ Tregs and CD4⁺ effector T cells, we identify a shared disease associated signature comprised of genes suggestive of Th2 activation (*GATA3*, *CCR4*), involved in promotion of survival and proliferation (*GITR*, *OX40*), and associated with extra-follicular T cell responses (*PRDM1*). Notably, this signature persisted in disease remission in patients off medication. Likewise, we identify a CD4⁺ T signature, decreased in all JDM

patients regardless of medication status and disease activity, containing genes crucial for tolerance (*NFC1*) and regulation (*ID3*, *SOX4*). This cell state could represent an inflammatory signature or a compensatory mechanism of CD4⁺ T cells in more long-standing disease. These findings are consistent with previous work, which identified skewing of CD4⁺ T cells toward a Th2 phenotype in JDM and showed in vitro that peripheral Th2 cells were efficient in helping B cells, including stimulating antibody production.¹³⁵ A previous study also showed that tertiary lymphoid structures are present in muscle of new-onset JDM, further supporting a role for extra-follicular reactions in JDM.¹³⁶ Future work using paired blood and tissue with spatial information to immunophenotype interacting cells within these structures would further justify investigating therapeutic strategies that prevent homing of CD4⁺ T and B cells to sites of inflammation in tissue.

While other studies have reported that Tregs in JDM have diminished suppressive capacity raising the possibility of Treg exhaustion,¹⁷² the results in this study show that the expanded population of peripheral Tregs in blood are proliferative and activated (*MKI67*, *IL2RA*, *IRF4*), taking on an effector phenotype. Likewise, the shared signature with CD4⁺ Teff cells suggests these Tregs are coopting the transcriptional machinery of effector T cells as has been described by others.^{173–175} In this dataset, JDM Tregs also upregulate transcriptomic and proteomic expression of CCR4—paralleled by increased expression of CCR4 in CD4⁺ effector T cells—which is preferentially expressed in Tregs recruited to the skin.¹⁶¹ Thus, we speculate that this expanded population of Tregs in JDM could represent a peripheral response to site-specific Th2-mediated inflammation in disease-affected tissue. Alternatively, these Tregs could be functioning in a reparative

manner at sites of tissue damage. Future functional studies of peripheral blood and tissue-specific Tregs, particularly investigation of the influence of type I IFN on Treg suppressive capacity in JDM, would provide mechanistic insight on this population's role in disease pathogenesis. The potential translational impact of investigating Treg dysfunction is corroborated by active development of multiple therapeutics targeting Tregs for autoimmune diseases.^{176,177}

More broadly, we show that unsupervised approaches such as DECIPHER can be used to consolidate disparate findings into a systems-level understanding of how interactions among cell states could manifest in disease. Here, our network analysis revealed that a module of hyper-activated IFN response across cell types is coordinated with dysfunction in ribosomal biogenesis, protein processing, and the regulation of cell death that is also shared across many cell types. This model contextualizes recent work that has identified ribosomal dysfunction in NK cells as a disease signature in JDM but also raises the possibility that defective translational machinery is not unique to that cell population.^{132,140} Given that Type I IFN directly promotes the activation and proliferation of NK cells,^{178–180} we speculate that NK cells in JDM are unable to properly translate cytolytic protein machinery required for effector function in response to IFN signaling, potentially perpetuating the IFN response. Similarly, the shared program between CD8T, gdT, and NK cells that describes regulation of cell death and cellular stress response suggests a common dysfunction across cytotoxic cell populations in JDM. Given the importance of cytotoxic cells in clearing cellular debris including autoantigenic neutrophil extracellular traps shown to be pathogenic in JDM,¹²⁹ dysfunctional cytotoxic populations

could result in accumulation of such debris thereby triggering an autoimmune response mediated by lymphocytes.

Finally, the observation that type I IFN responses increase with clinical metrics of disease activity adds to the growing body of work suggesting that disease activity in JDM correlates with this transcriptional signature.^{133,166,181} However, given the time and cost, it remains infeasible to use transcriptomic sequencing as a lab-based clinical diagnostic tool. Our data points to surface expression of SIGLEC-1 in monocytes as a composite measure of the IFN gene signature in JDM and disease activity. Together with a recent independent study of JDM, we provide external validation that SIGLEC-1 is a suitable biomarker for disease monitoring to pursue in larger immunophenotyping studies given the lower cost and ease of implementing screening by flow cytometry.¹⁴⁷ Importantly, we show that SIGLEC-1 directly reflects the IFN gene signature using paired gene and protein expression measurements, strengthening support for its use as a biomarker. Further study of this biomarker, and the role of SIGLEC-1 in disease, is an important step toward precision care of JDM.

These findings should be interpreted in the context of the study's limitations. First, despite being the largest single-cell study in JDM to date, sample numbers are still limited such that the study lacks statistical power to quantify the contribution of MSA status to disease heterogeneity. Furthermore, a majority of patients in the treatment-naïve group are TIF1y+, which could introduce a bias to disease-activity related programs, though it remains unknown whether MSA status is associated with distinct biological mechanisms. Some patients in the “active” disease group had relatively low disease activity, which may have prevented us from identifying more associations with disease activity. Additionally,

this study lacked data from matched JDM skin and muscle which would have enabled insight into how dysregulated cell states in blood might influence local microenvironments in disease-affected tissue. Although profiling blood limits the mechanistic insight compared to skin or muscle, it is a more suitable sample type for biomarker discovery, particularly in a pediatric disease that requires longitudinal monitoring, and future comparison to tissue data will enable us to identify populations in peripheral blood with tissue correlates. Lastly, the DECIPHER algorithm relies on a k selection procedure to accurately decompose the data. As a dimensionality reduction technique, NMF is distinct from principal component analysis in that there is no single solution for the number of patterns or components into which the data is segmented. As such, it is necessary to optimize the parameter 'rank K' such that the NMF results capture the relevant biology at an appropriate granularity. We addressed this limitation of NMF by using the phylogenetic clustering-based k-selection method described by Murrow et al. where the authors demonstrated that saturation of this metric reflects the appropriate granularity of biological programs such that results are robust across multiple choices of rank K.⁶⁵

In summary, using CITEseq to profile compositional and functional imbalance of peripheral blood immune cells and the relationship to disease activity, we provide a comprehensive map of the coordinated immune dysregulation underlying JDM. We identify persistent transcriptional changes in B and CD4+ T cells associated with JDM that persist even in patients in remission off medication and reveal novel cell states associated with the IFN signature that generate new hypotheses for the role of extra-follicular interactions in disease pathogenesis, drawing parallels to other autoimmune diseases. Importantly, these findings pose a new paradigm to how we approach JDM treatment.

The dysregulation of processes simultaneously with hyperactivation of other cell states necessitates that we identify therapeutic strategies that restore balance to the dynamic interactions between immune populations rather than simply turning off a set of pathways. Taken together, our work sets the stage for improving clinical management of JDM by providing a foundation for systems-level inquiry into the cellular basis of this disease. More broadly, application of a similar analytical strategy could provide insight into the immunologic basis of other childhood-onset autoimmune diseases characterized by a type I IFN gene signature.

Methods

Sex as a biological variant

This study contained samples from human males and females. Sex was not considered as a biological variable in downstream analyses.

Study Cohort & Sample Processing

Patients were recruited to the Juvenile Myositis Precision Medicine Biorepository between 2018 and 2021 and underwent informed consent. The diagnosis of JDM was per clinician judgement, however, all patients included in this study met EULAR/ACR classification criteria for “definite” juvenile idiopathic inflammatory myopathy based on typical skin manifestations of either Gottron’s and or heliotrope rashes.¹⁸² This study was approved by the UCSF IRB. Clinical data was collected by study investigators and recorded in a secure REDCap database. Treatment-naive JDM was defined as a new diagnosis of JDM as deemed by the treating clinician with no systemic immune suppressant use in the prior 4 weeks. Inactive JDM was defined as normal CK, MMT8 \geq 78 and Physician Global VAS score $<$ 0.5 to reflect PRINTO clinically inactive disease¹⁸³ definitions but with some modifications based on the data available. Active disease was defined as Physician Global VAS score \geq 0.5, and all patients in this category were taking immune suppressive medications. Longitudinal samples from n=6 patients with JDM were included separated by at least 4 months in time and accompanied by a change in disease activity. Measures of disease activity, including the Cutaneous Disease Area and Severity Index (CDASI) were collected at study visits.¹⁸⁴ Healthy controls were enrolled who had no prior

autoimmunity, no known or suspected genetic disorders, immunodeficiency, active cancer, or history of organ or bone marrow transplantation, no infection or antibiotics in the prior 4 weeks, no chronic systemic immunomodulatory medication use and no vaccinations in the prior 6 weeks. Peripheral blood samples were collected at each study visit and processed by the Pediatric Clinical Research Core Sample Processing Lab. PBMCs were collected in SepMate tubes (n=9) using Ficoll separation or CPT tubes (n=18), isolated per manufacturer's guidelines, and cryopreserved in liquid nitrogen.

CITE-seq of human PBMCs

Our experimental protocol followed protocol from our previous study¹³³ with certain modifications to account for confounding time-related and batch effects. Note these experiments were carried out using early access kits from BD Genomics before the implementation of commercially-available single-cell protein/RNA assays (e.g. Feature Barcoding, 10x Genomics; BD Abseq, BD Genomics, **Supplemental Table 4.2**), and researchers are recommended to use those newer solutions for any follow-up studies as the techniques and reagents have been refined. PBMCs from 27 distinct samples were gently thawed in a 37°C water bath and re-suspended using a pipette set to 1 mL. Cell counts and viability were determined using a Cellometer Vision (Nexcelcom) with AOPI staining (Nexcelcom cat. CS2-0106-5ML). Cells were multiplexed into four pools: one “cross pool” with all samples that consisted of only one time point and three pools consisting of longitudinal samples. Longitudinal samples from the same individual were assigned to separate pools to enable genetic demultiplexing. After pooling, cells were resuspended in 90 µl of 1% BSA in PBS and Fc blocked with 10 µl Human TruStain FcX

(Biolegend cat. 422302) for 10 minutes on ice then stained on ice for 45 minutes with a pool of 268 antibodies in 100 μ l, for a final staining volume of 200 μ l. Antibodies were pooled on ice with 2.2 μ l per antibody per 1×10^6 cells (BD Genomics). Cells were quenched with 2 ml 1% BSA in PBS and spun at 350xg for 5 minutes and further washed two more times with 2 ml of 1% BSA in PBS. After the final wash, cells were resuspended in 100 μ l and strained through a 40 μ M filter (SP Bel-Art cat. H13680-0040). Each longitudinal pool was split across two 10X lanes while the “cross pool” was split across six 10X lanes (6 wells total, 5×10^5 cells/well). The 10x Chromium was run and post-GEM RT and cleanup were done according to manufacturer’s protocol (10X Genomics 3’ Kit V3). Starting at cDNA amplification, modifications to the protocol were made: 1 μ l of 2 μ M additive primer (BD Genomics, beta kit) specific to the antibodies tags was added to the amplification mixture. During the 0.6X SPRIselect (Beckman Coulter, B23318) isolation of the post-cDNA amplification reaction cleanup, the supernatant fraction was retained for ADT library generation. Subsequent library preparation of the cDNA SPRI-select pellet was done exactly according to protocol, using unique SI PCR Primers (10X Genomics). For the ADT supernatant fraction, a 1.8X SPRI was done to isolate ADTs from other non-specifically amplified sequences, followed by sample index PCR. Sample index PCR for the ADTs was done using the cycling conditions as outlined in the standard protocol (15 cycles) but using unique SI-PCR Primers such that all libraries could be mixed and sequenced together. Subsequent SPRI selection was performed, and all libraries were quantified and analyzed via Qubit 2.0 (Fisher) and Bioanalyzer (Agilent), respectively, for quality control. We sequenced the libraries on 2 lanes of a NovaSeq S4 (Illumina), aligned using CellRanger (10X Genomics) to generate feature barcode matrices.

Sequencing data pre-processing and integration

Data was demultiplexed using genotypes with demuxlet⁸ and doublets were filtered using DoubletFinder.¹²⁰ Next, the data were filtered to remove genes with < 3 cells. Additional filters were applied to the cells, removing cells with greater than 5000 ADT counts to avoid antibody aggregates and with >60% ribosomal or >15% mitochondrial DNA (mtDNA) reads. For the ADT data, cells were additionally filtered to remove those with fewer than 70 antibodies detected, and with any antibody isotype control measurements greater than 50. To remove background ambient RNA signal, we ran SoupX separately on each of the six RNA libraries and then merged them.¹⁸⁵ Aggregated data was log-normalized and scaled, regressing out percent mtDNA, percent ribosomal DNA, and cell cycle (S, G2M).³⁹ Data was then integrated with Harmony, with 20 max iterations and 30 max iterations per cluster.¹⁸⁶

DSB was run on all six ADT libraries individually, using default parameters except for more stringent quantile clipping (0.01, 0.99).¹⁸⁷ The background distribution of empty droplets was defined as suggested in the DSB vignette. Isotype controls were then removed from the dataset, and RPCA was used to integrate the DSB-normalized ADT data across libraries. Following RPCA, the data was re-scaled and cell cycle scores (S, G2M genes) and the number of ADT counts and features were regressed out. The harmonized RNA and RPCA corrected ADT were combined using Weighted Nearest Neighbors, with default parameters except for $\text{prune.SNN} = 1/20$. Leiden clustering was run on the resulting graph (method = igraph), at a 1.4 resolution.⁸⁶ Two clusters were removed with low to no expression of ADT and the object was reclustered with the same

parameters. The Seurat function 'FindAllMarkers' was used to identify the top 5 markers per cluster.

We removed an additional 3 clusters: 2 were small clusters with a transcriptomic profile consistent with doublets (original Leiden clusters 26 and 29, **Supplemental Figure 4.36**), and 1 diffusely expressed cluster (original Leiden cluster 19, **Supplemental Figure 4.36**). We further sub-clustered 3 clusters that expressed genes representative of more than one cell type: original Leiden clusters 16, 17 and 23. Sub-clustering was performed using Seurat's 'FindSubCluster' function using the lowest possible resolution to divide the population into two clusters. Based on minimal transcriptional differences between them, original Leiden clusters 1, 5, 9, 11 and 15 were merged into a single CD4⁺T naïve population, clusters 3 and 10 into a single naive CD8⁺T population, and cluster 7 and part of the subsetting cluster 23 CD56^{dim} NK population. Due to interpersonal heterogeneity in monocytes, all CD14⁺ monocyte clusters were merged into one CD14⁺ monocyte population.^{188,189}

While annotating, we discovered that the FOXP3-signature normally attributed to Tregs was only present in a subset of the cluster and FindSubCluster did not appropriately isolate the *FOXP3*⁺ cells. We therefore subsetted the cluster and re-ran 'FindVariableFeatures', 'ScaleData', 'RunPCA', 'FindNeighbours', 'FindClusters' with the Louvain algorithm and a resolution of 0.8, and 'RunUMAP'. This enabled us to subset a smaller group of cells with a statistically significant expression of *FOXP3* compared to other clusters using FindAllMarkers, which we hence annotated T regulatory cells. Annotation of the remaining clusters was performed using both canonical gene and protein markers. One B cell population consisted almost solely of cells from two donors.

This was annotated as B_naive4, and was not used in downstream analysis, but included in UMAPs.

Cell type proportion analysis

Cell type proportion was calculated as the proportion of each cell type for each individual and was compared for: treatment-naive JDM compared to HC, treatment-naive JDM compared to inactive JDM and inactive JDM compared to HC using Kruskal-Wallis test with Dunn's post-test. To determine the association between cell abundance and disease activity, the Spearman correlation coefficient between cell type proportion and physician global VAS scores was calculate and p values were adjusted using BH.

Differential gene and protein expression analysis

The DGE and DPE analysis were completed using DESeq2. Size factors were set using the function 'computeSumFactors' from the scran package. We used the default settings for single cell data, namely test='LRT', useT = T, minmu = 1e-6, fitType = 'glmGamPoi', and minReplicatesForReplace = Inf in the 'DESeq' function. Batch was included as a co-variate using the 'reduced' argument. We filtered genes and proteins that were not expressed in at least 5% of cells and analyzed only cell types where there were at least 100 cells in each group. We used cutoffs of $|LFC| \geq 1$ for genes, $|LFC| \geq 0.5$ for proteins, and BH $p < 0.05$. Over-representation analysis was performed on up- and downregulated genes per cell type using the clusterProfiler package with GOBP as reference and adjusted $p < 0.05$. For the PD1/CD45R0-subanalysis, we compared groups using

Seurat's FindMarkers with test.use = 'MAST', latent.vars = 'well', |LFC| ≥ 0.5, and BH p<0.05.

Identification of global IFN signature

We created a list of IFN genes by excluding cell types with less than 100 cells in either HC or treatment-naive JDM and then collected genes differentially expressed in at least 2 cell types. The average gene expression was calculated using Seurat's 'AverageExpression' function. Expression was averaged per sample for each cell type. The expression was visualized using dittoSeq's¹⁹⁰ 'dittoHeatmap' with default, unsupervised clustering settings of both rows and columns, and the dendrograms ordered using the dendsort package.¹⁹¹ The clustering organized the genes into 7 distinct modules, where Module 1 consisted exclusively of IFN-related genes. Average Module 1 scores for each cell type were then calculated using Seurat's 'AddModuleScore' with default settings. Correlations between disease activity and IFN score was calculated using Spearman correlation and visualized using ggplot2.¹⁹²

Network inference from RNA data using DECIPHER

We applied NMF to the raw RNA count data as implemented in the DECIPHER method with default parameters.⁶⁵ The main output of NMF is a set of two orthogonal vectors: gene loadings that represent how much a given gene contributes to that activity program, and cell loadings that represent how strongly that program is expressed in a given cell. The NMF rank, k, was chosen using the weighted subtrees metric based on phylogenetic

clustering, as described by Murrow et al.⁶⁵ The final choices of rank k for each cell type were $k_B=17$, $k_{CD4T}=17$, $k_{CD8T}=14$, $k_{gdT}=13$, $k_{Myeloid}=17$, $k_{NK}=11$ according to the saturation point in the elbow plots (**Supplemental Figure 4.37**). Network clustering was performed on the per-sample averaged program scores with default parameters as described by Murrow et al. The corresponding gene loading vectors for each GEP were analyzed as described by Kotliar et al. to quantify the strength of an individual gene's contribution to that program, referred to as 'marker gene scores'.⁷⁷ GSEA was performed on the resultant ranked gene lists using the `fgsea`¹²² package in R with GO and Hallmark gene sets. Module themes were assigned by calculating module enrichment p-values using the 'Get_enrichment_pvals' function in DECIPHER with default parameters. Module and gene set enrichment results were visualized using ClusterProfiler.¹⁹³

Validating signatures in independent data

We performed CITEseq using the same protocol as described above with PBMCs from 5 patients with JDM and 2 healthy pediatric controls. We used the same steps for data processing with the exception that we used CellBender rather than SoupX for ambient RNA removal and clustered cells using the RNA measurements only. To derive proxy scores for GEP in the independent dataset, we ranked the gene lists comprising each program by marker score, which quantifies how strongly a single gene contributes to that GEP. Using the top 5% of genes by marker score, that list was used as input for the rank-based gene subset enrichment method AUCell.¹⁹⁴ Pseudobulk proxy GEP scores were calculated as the per-patient mean expression in the same way for the original NMF programs. Case vs control comparisons were done using t-tests between JDM patients

and HCs. Given the patients per group (HC = 2, Active JDM = 4, TN JDM = 1) in this validation cohort, we could not repeat the ANOVA comparisons used for disease activity association in the original dataset. Instead, proxy GEP scores were correlated to physician-assessed scores of disease activity (VAS global) using the Spearman correlation.

Statistics

All statistical analyses and visualization of results were performed using open-sourced R (version 4.2.3). Pairwise comparisons of cell proportions between patient groups were performed using a Kruskal-Wallis test with post-hoc Dunn comparison, with p-values adjusted for multiple comparisons by Holm correction. Significance of Pearson correlations between GEPs used for network construction was calculated using bootstrapping as implemented in DECIPHER. Analyses of disease association with GEPs was performed using two-tailed unpaired t-test or ANOVA with post-hoc Tukey test. False discovery rates for GSEA annotation and module enrichment across programs were calculated and corrected at the cutoff FDR < 0.01 as described by Murrow et al. Gene loading similarity was calculated as the Pearson correlation between gene loadings for each activity program and all other activity programs in the same module with p-values calculated by permutation testing. Correlation methods used in specific figures are described in the corresponding legends and in Methods, and significance for statistical tests was set at the threshold $P < 0.05$.

Study approval

This study was approved by the UCSF IRB #17-24003. Written informed consent to participate in this study was provided by the participant or the participants' legal guardian depending on the age of the participant. Assent was obtained when appropriate.

Data availability

The datasets presented in this study are deposited in the CZ CELLxGENE Discover resource as 'CITEseq of JDM PBMCs'

([https://cellxgene.cziscience.com/collections/c672834e-c3e3-49cb-81a5-](https://cellxgene.cziscience.com/collections/c672834e-c3e3-49cb-81a5-4c844be4a975)

[4c844be4a975](https://cellxgene.cziscience.com/collections/c672834e-c3e3-49cb-81a5-4c844be4a975)). The code used for this analysis will be made publicly available on Github at "GartnerLab/jdm_crosslong" upon manuscript acceptance. Values for all data points in graphs are reported in the Supporting Data Values file uploaded as part of Supplemental Data.

Author Contributions

JN and SK recruited patients for the study. GCH, YS, and JN performed the experiments. GCH performed the alignment and sample demultiplexing. EF, GR, and CW performed the integration and quality control. CW performed the annotations and the cell proportion and differential analysis. GR conceptualized and wrote the pipeline for the network analyses; EF ran the pipeline. GR, CM, and GKF performed analysis of the second dataset. ZG, MS, and JN provided essential feedback on the analysis strategy and statistical methods. GR and JN wrote the manuscript; MS and ZG revised the manuscript.

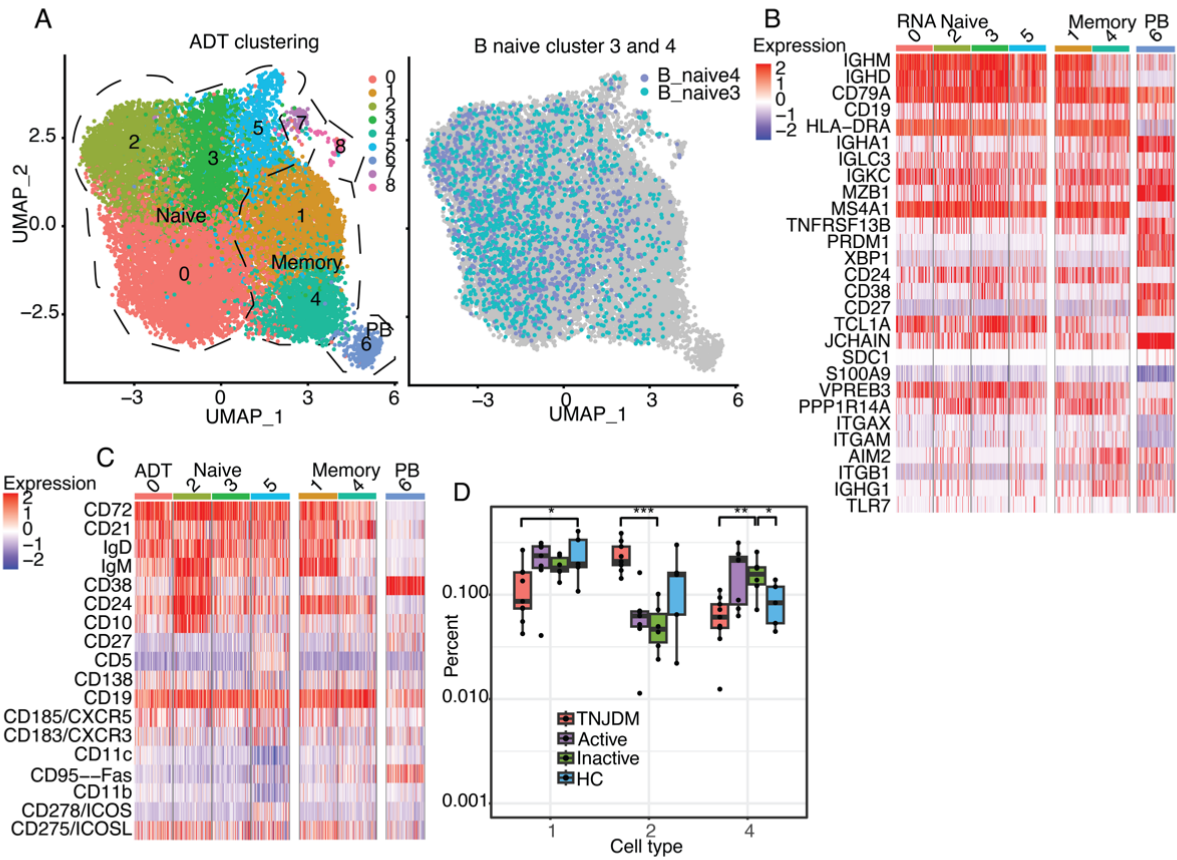
JN, MS, and SK conceptualized the study. JN led the project administration and acquired funding for the study. All authors reviewed and edited the manuscript.

Supplement

Supplemental Table 4.1 Clinical cohort disease characteristics.

	TNJDM (N=9)	Active (N=7)	Inactive (N=6)	HC (N=5)
Sex				
Female	3 (33.3%)	4 (57.1%)	5 (83.3%)	2 (40.0%)
Male	6 (66.7%)	3 (42.9%)	1 (16.7%)	3 (60.0%)
Age (years)				
Median [min-max]	7.0 [2.0-15]	13 [5.0-18]	16 [5.0-18]	3.0 [1.0-16]
Prednisolone treatment — no. (%)	0 (0%)	1 (14.3%)	0 (0%)	
Methyl prednisolone treatment — no. (%)	0 (0%)	1 (14.3%)	0 (0%)	
Methotrexate treatment — no. (%)	0 (0%)	4 (57.1%)	0 (0%)	
HQL treatment — no. (%)	0 (0%)	4 (57.1%)	0 (0%)	
IVIg treatment — no. (%)	0 (0%)	1 (14.3%)	0 (0%)	
Physician Global VAS				
Median [min-max]	5.0 [1.5-7.5]	1.5 [0.50-8.0]	0 [0-0.10]	
Muscle VAS				
Median [min-max]	4.0 [0.50-8.5]	0.50 [0-3.5]	0 [0-0]	
Cutaneous VAS				
Median [min-max]	4.0 [1.5-7.0]	1.0 [0-7.0]	0 [0-0.20]	
Patient/Parent Global VAS				
Median [min-max]	5.0 [2.0-8.2]	2.0 [0-8.0]	0 [0-0.40]	
Missing	2 (22.2%)	0 (0%)	0 (0%)	
CDASIACT				
Median [min-max]	16 [4.0-35]	2.0 [0-8.0]	0 [0-1.0]	
CHAQ-score				
Median [min-max]	1.0 [0-2.4]	0 [0-1.3]	0 [0-0.38]	
Missing	2 (22.2%)	0 (0%)	0 (0%)	
MMT8-score				
Median [min-max]	70 [59-76]	80 [67-80]	80 [79-80]	
Missing	4 (44.4%)	0 (0%)	0 (0%)	
MSA				
MDA5	1 (11.1%)	0 (0%)	0 (0%)	
NEG	1 (11.1%)	1 (14.3%)	1 (16.7%)	
NXP2	1 (11.1%)	4 (57.1%)	2 (33.3%)	
TIF1y	6 (66.7%)	2 (28.6%)	2 (33.3%)	
UNK	0 (0%)	0 (0%)	1 (16.7%)	
Muscle enzyme elevation Present — no. (%)	8 (88.9%)	3 (42.9%)	1 (16.7%)	

Overview of clinical characteristics from the 15 patients with JDM and 5 healthy controls, totalling 27 samples. Some patients had longitudinal sampling at different disease stages. Myositis specific antibodies were sent to Oklahoma Myositis Research Foundation for testing. TNJDM = treatment-naive JDM, Active = Active JDM defined by Physician global visual analog score (VAS) ≥ 0.5 and on medication, Inactive = Inactive JDM defined by physician global VAS < 0.5 and off medication, HC = healthy control, HQL = Hydroxychloroquine, VAS = Visual Analogue Scale, CDASIACT = Cutaneous Dermatomyositis Disease Area and Severity Index Activity Score, CHAQ = Childhood Health Assessment Questionnaire, MMT = Manual Muscle Testing, MSA = Myositis Specific Antibodies.



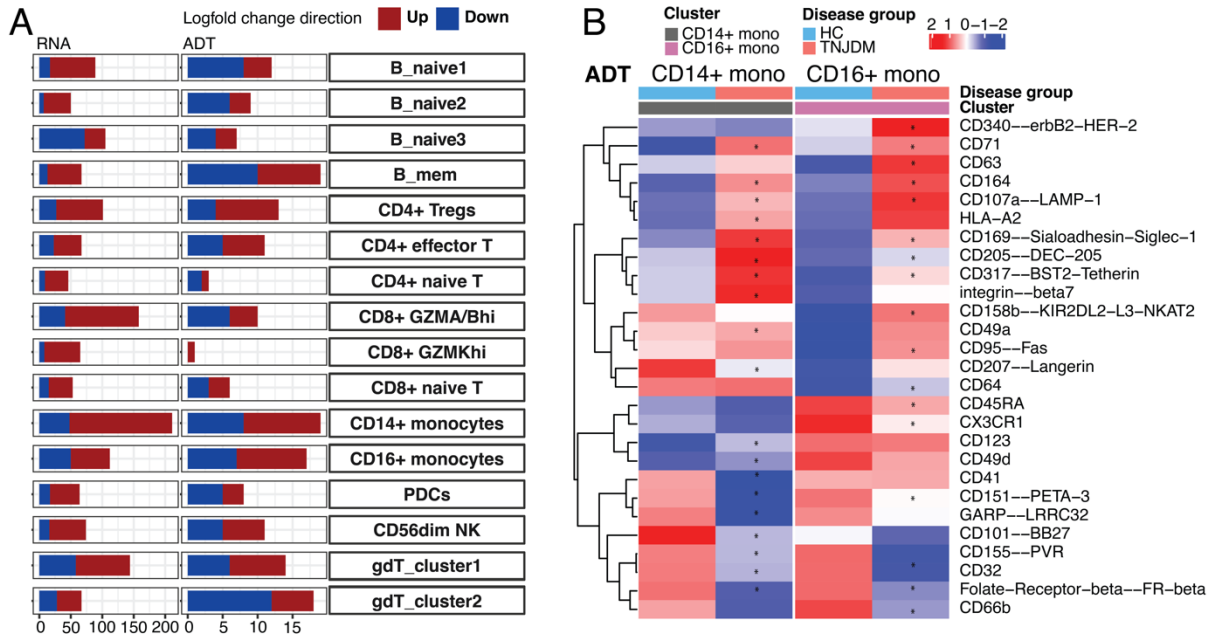
Supplemental Figure 4.15 Analysis of reclustered B cells based on ADT measurements alone.

This re-clustering was completed on the DSB-corrected ADT assay using the Seurat workflow. FindClusters was run using a resolution of 0.4 with the Leiden algorithm. FindUMAP was run using default settings.

(A) UMAP of subsetting B cells where Clusters 0, 3 and 5 corresponded to naïve B cells, cluster 2 to an immature naïve B cell population akin to “B_naive1” in the first analysis, and clusters 1 and 4 to IgM+IgD+ memory and IgM-IgD- memory B cells, respectively, defined by TNFRSF13B (TACI) expression. Cluster 6 consisted of plasmablasts. Clusters 7 and 8 contained few cells with a high expression of platelet and red blood cell-specific genes and were excluded from further analysis. B_naive3 and B_naive4, clusters driven by RNA signatures, did not form specific clusters and patient-specific clustering was resolved.

(B-C) Canonical RNA (B) and ADT (C) markers for reclustered B cells.

(D) Compositional analysis using these ADT clusters verified an increase in the proportion of immature naïve B cells (IgM+IgD+CD24+CD38+CD10+) from cluster 2 in TNJDM as compared to HC. This analysis also found a significant decrease in proportion of memory B clusters 1 and 4, in TNJDM as well as an increase in the proportion of IgM-IgD-memory B cells (cluster 4) in inactive JDM.

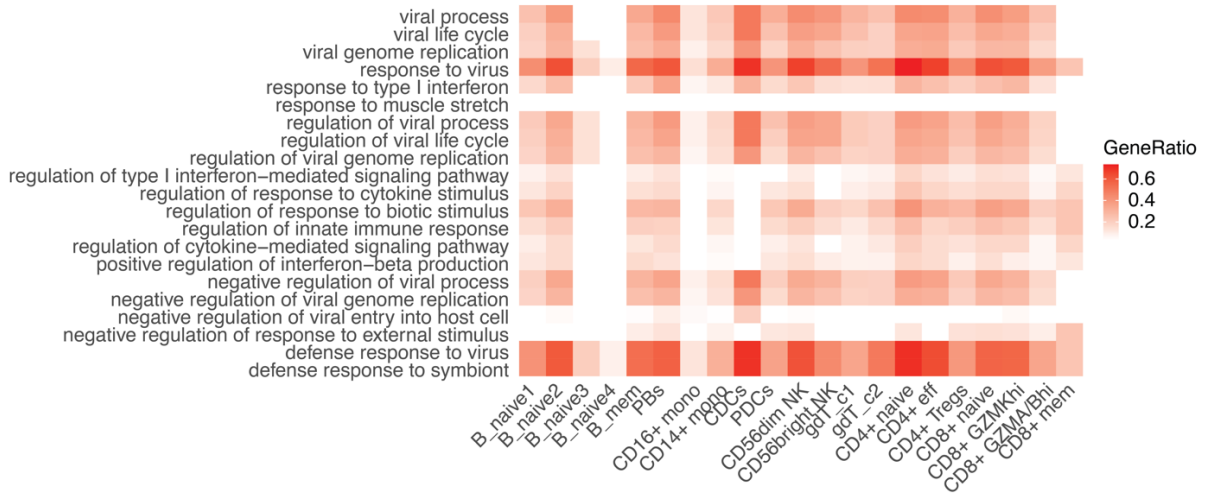


Supplemental Figure 4.16 Multi-modal differential analysis for TNJDM and HCs.

(A) Differential analysis between treatment-naive JDM and HCs for each cell type.

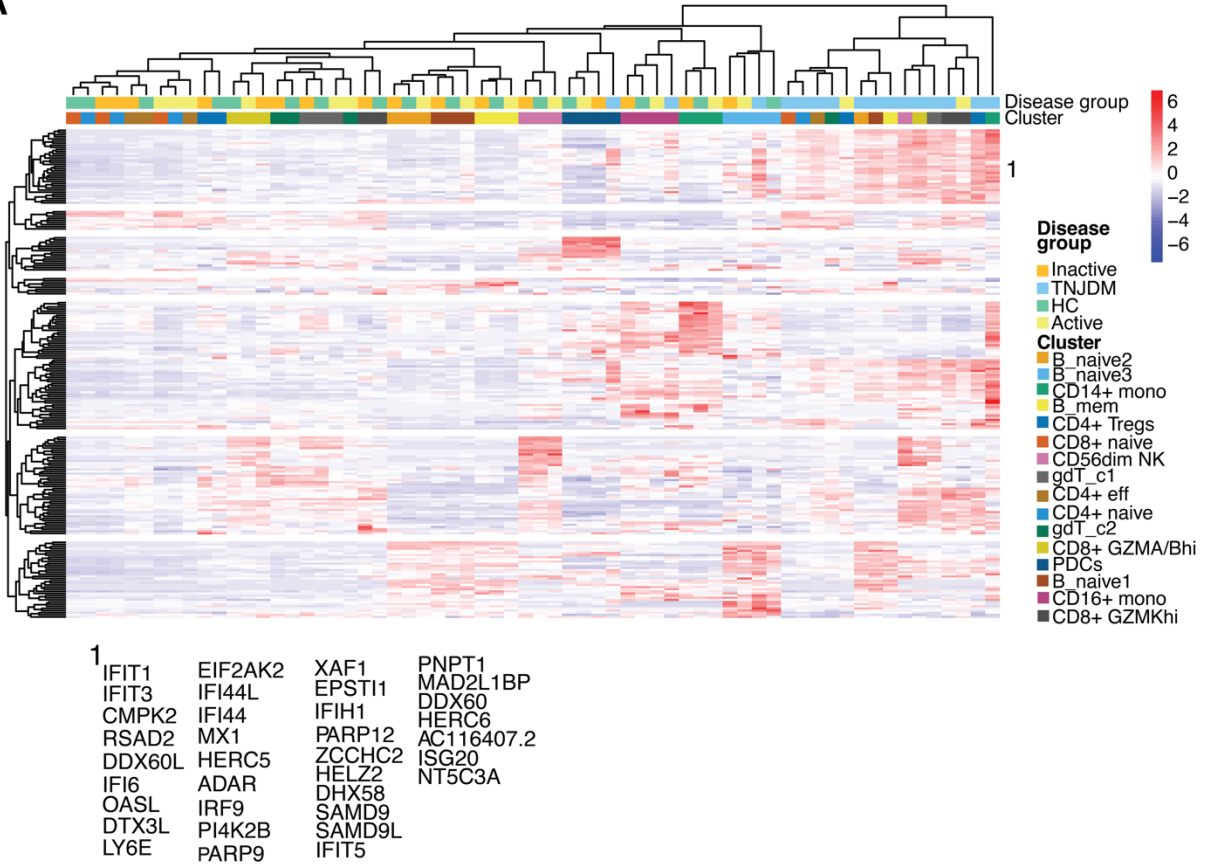
(B) Differential analysis (TNJDM vs HC) of surface protein expression for monocytes. Asterisks indicate significant differential expression (BH-adjusted $p < 0.05$).

A



Supplemental Figure 4.17 Heatmap of top five enriched GO terms per cell type with FDR<0.01.

A

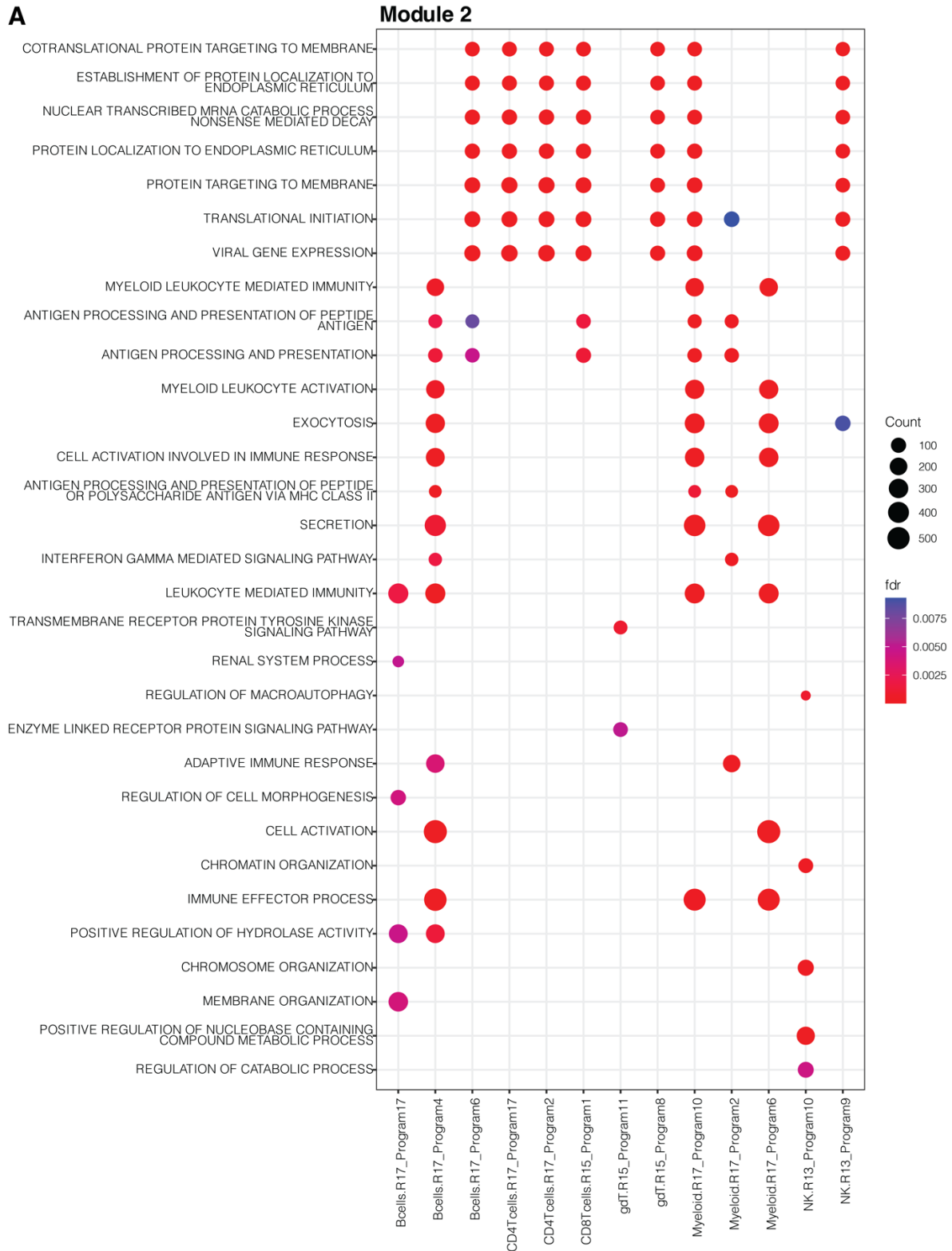


Supplemental Figure 4.18 Heatmap of differentially expressed genes between TNJDM and HC from all cell types clustered by expression likeliness. The genes from Cluster 1 were used to calculate the IFN score.

A



Supplemental Figure 4.19 Gene set enrichment results for GO terms in Module 1 (FDR<0.01).



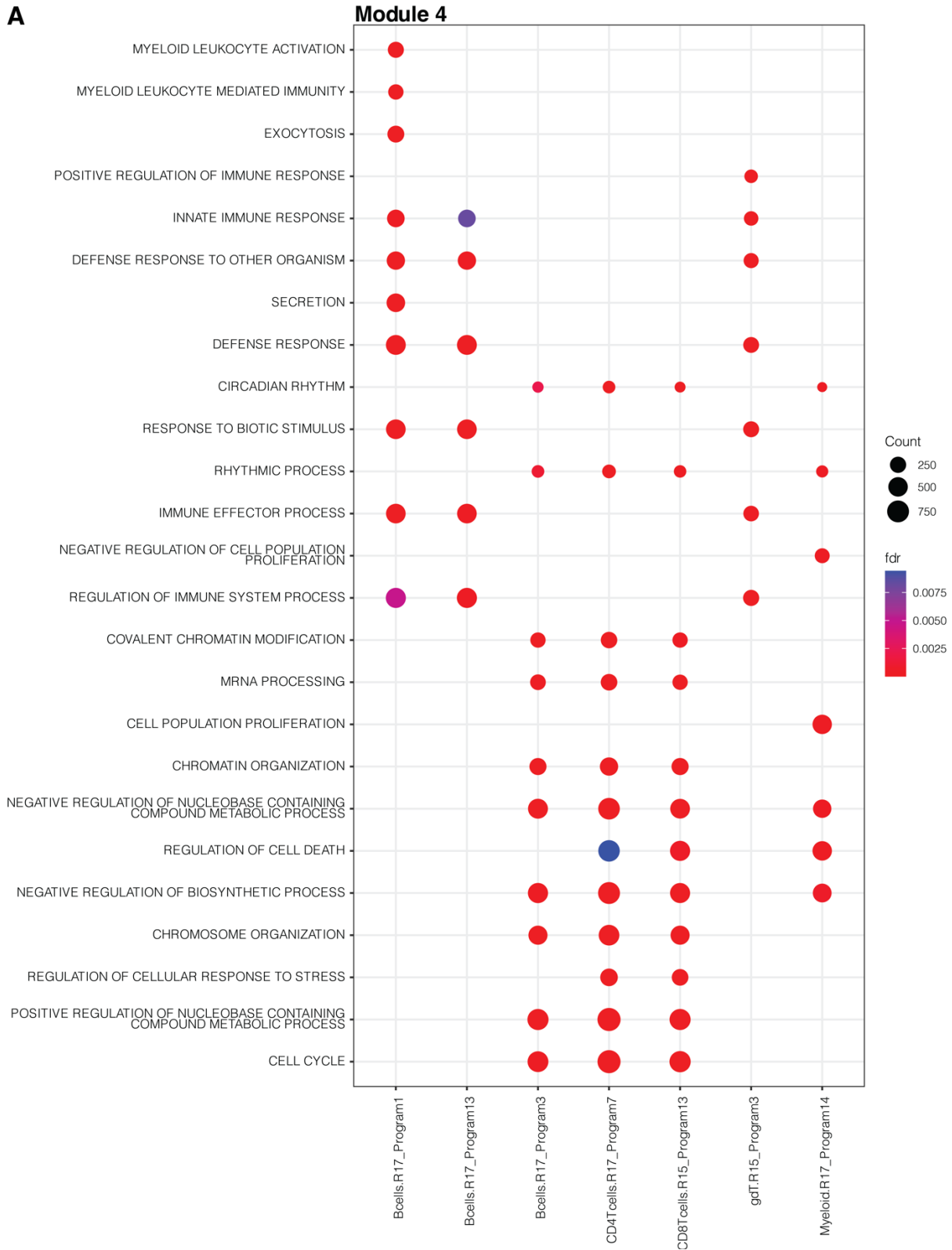
Supplemental Figure 4.20 Gene set enrichment results for GO terms in Module 2 (FDR<0.01).

A

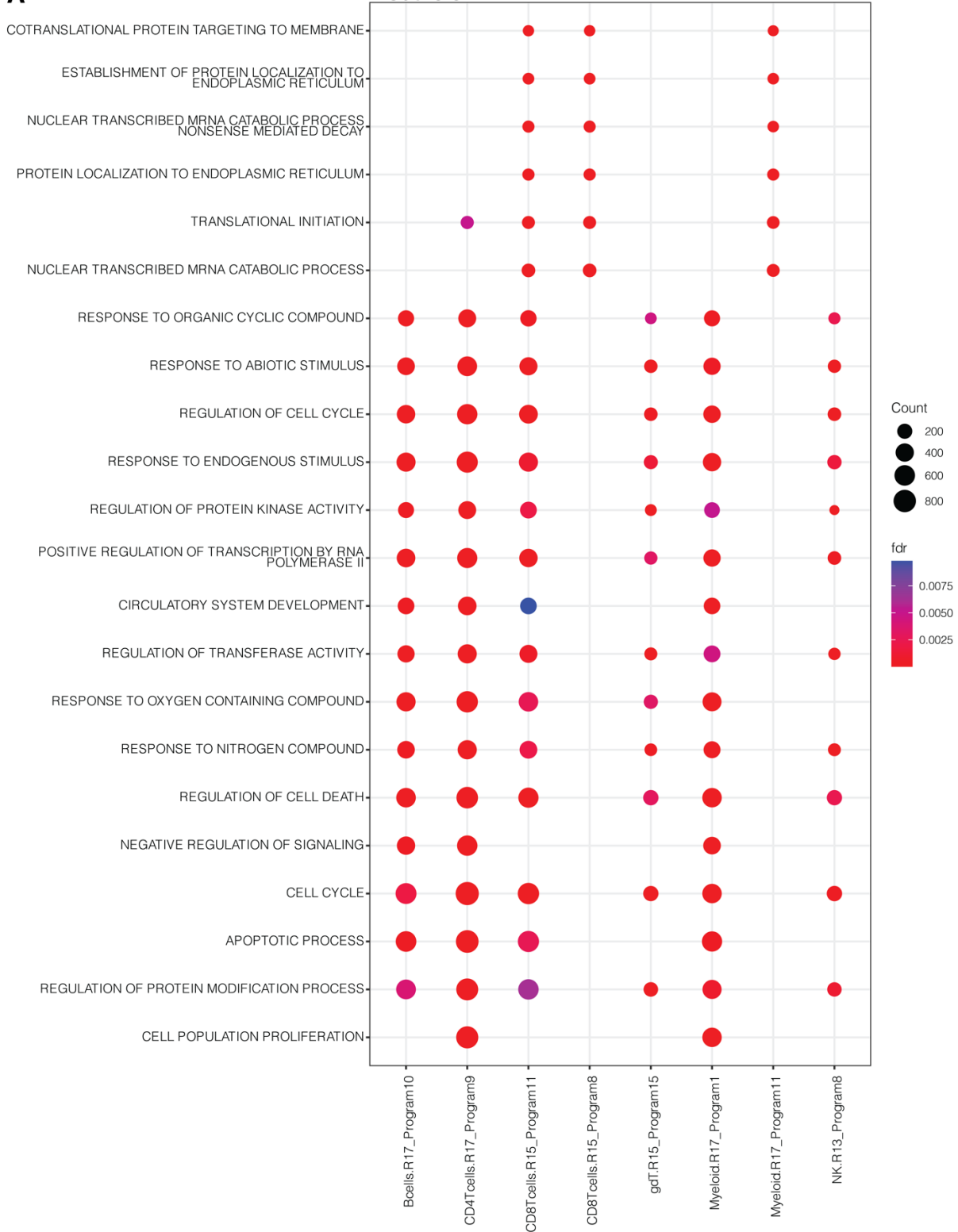
Module 3



Supplemental Figure 4.21 Gene set enrichment results for GO terms in Module 3 (FDR<0.01).

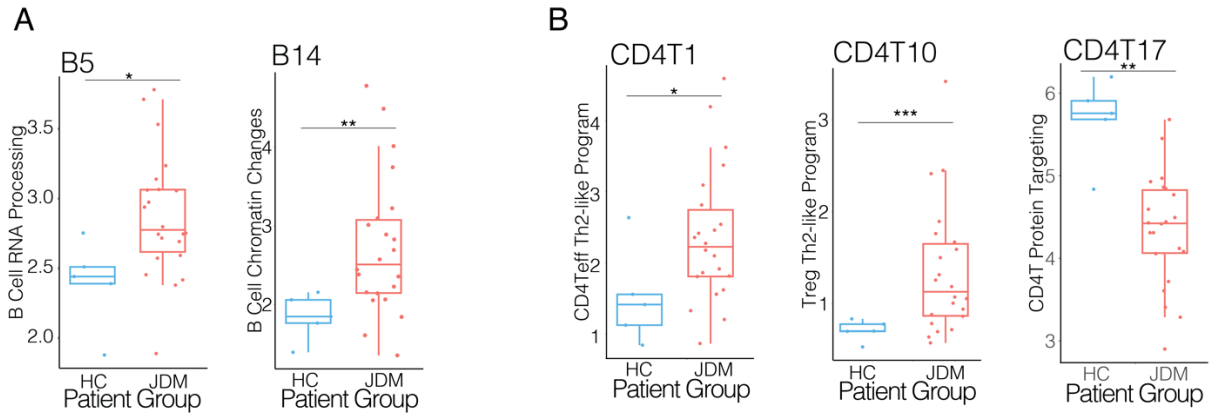
A

Supplemental Figure 4.22 Gene set enrichment results for GO terms in Module 4 (FDR<0.01).

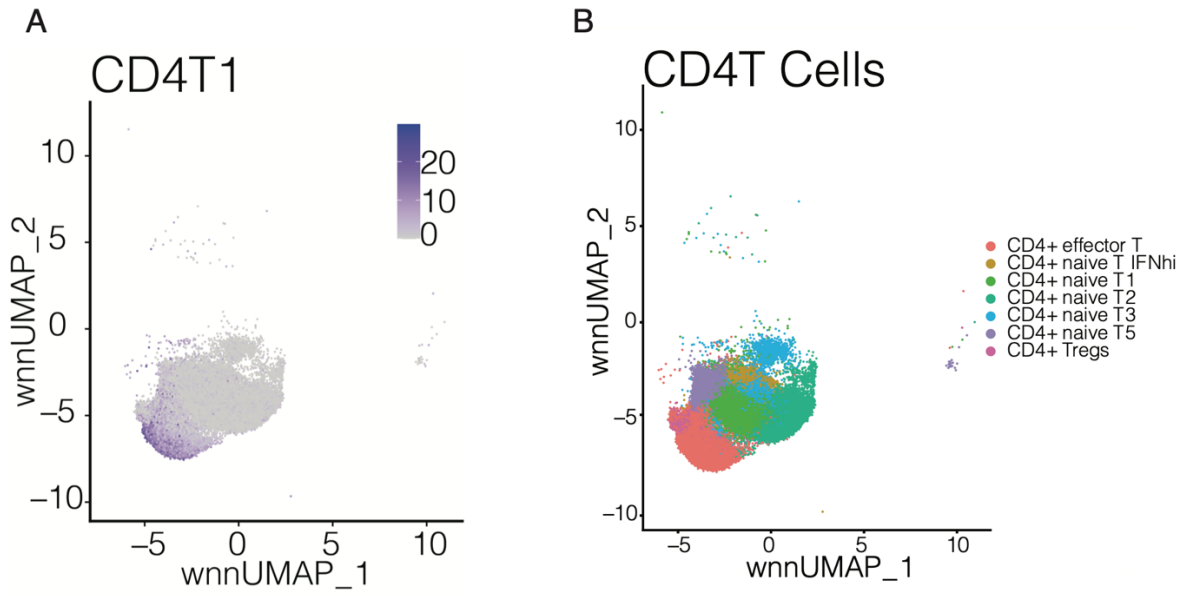
A**Module 5****Supplemental Figure 4.23** Gene set enrichment results for GO terms in Module 5 (FDR<0.01).



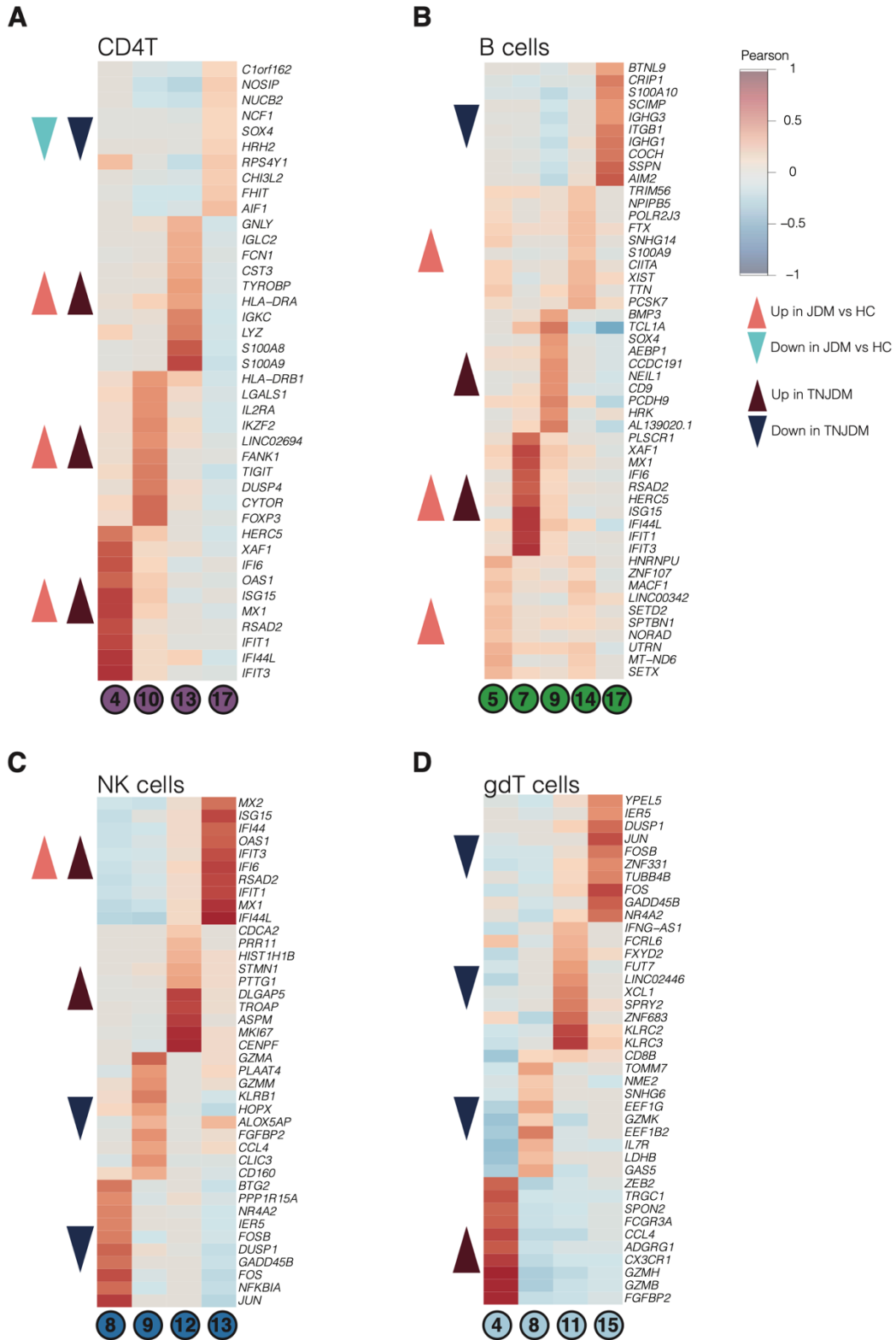
Supplemental Figure 4.24 Gene set enrichment results for GO terms in Module 6 (FDR<0.01).



Supplemental Figure 4.25 Mean patient expression of JDM-associated programs (t-test, $p < 0.05$) in B cell (A) and CD4T cell (B) compartments, respectively (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

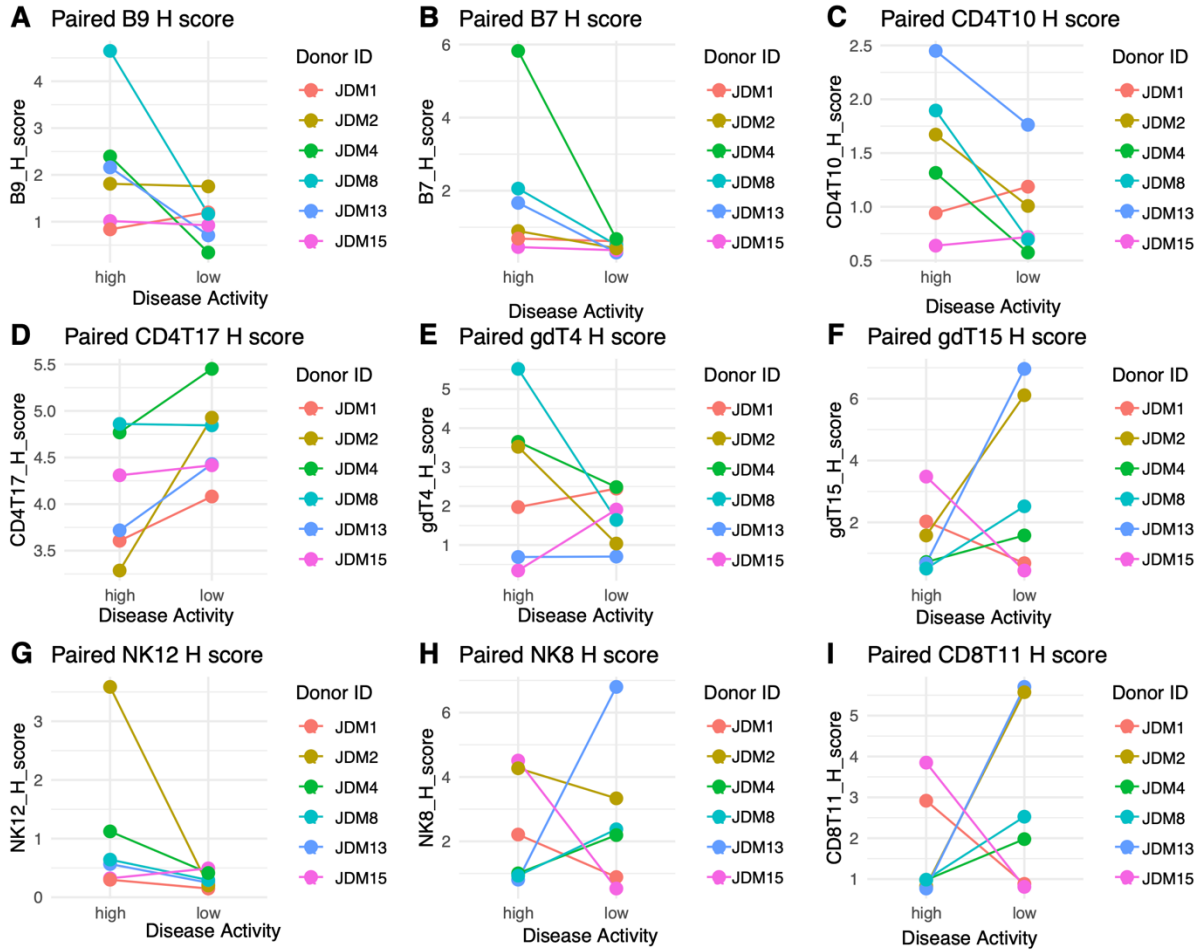


Supplemental Figure 4.26 UMAPs of CD4T cells showing expression of NMF program CD4T1 (A) in subcluster corresponding to CD4+ effector cells (B).



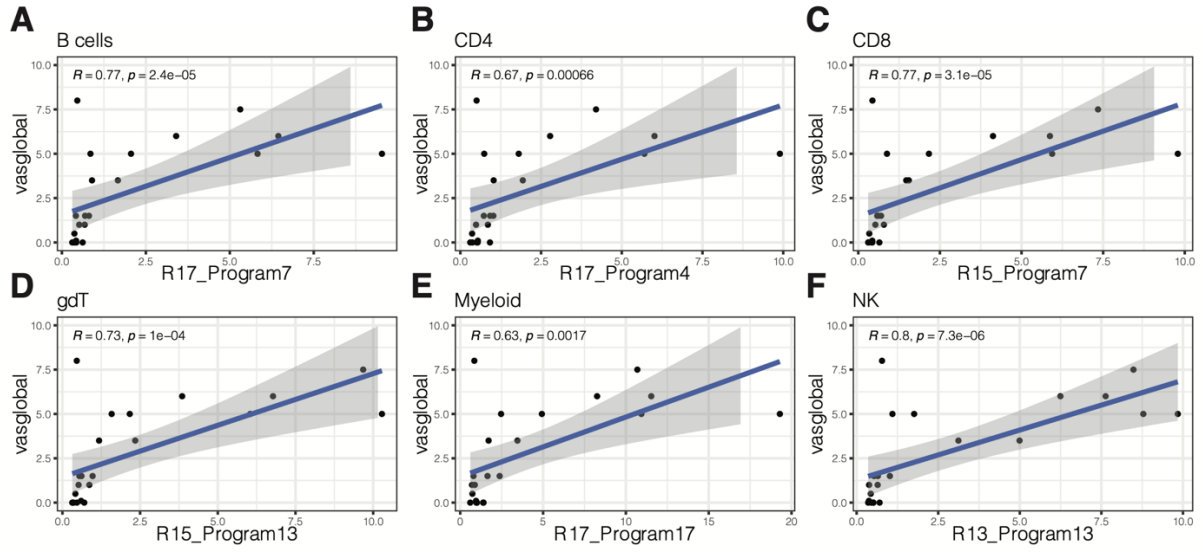
Supplemental Figure 4.27 Heatmaps showing top 10 marker genes for selected disease-associated programs for the indicated cell type.

Colored according to Pearson correlation between gene expression and program expression in the indicated cell type. Arrows indicate whether a given program is expressed higher or lower in a specific subset of patients.

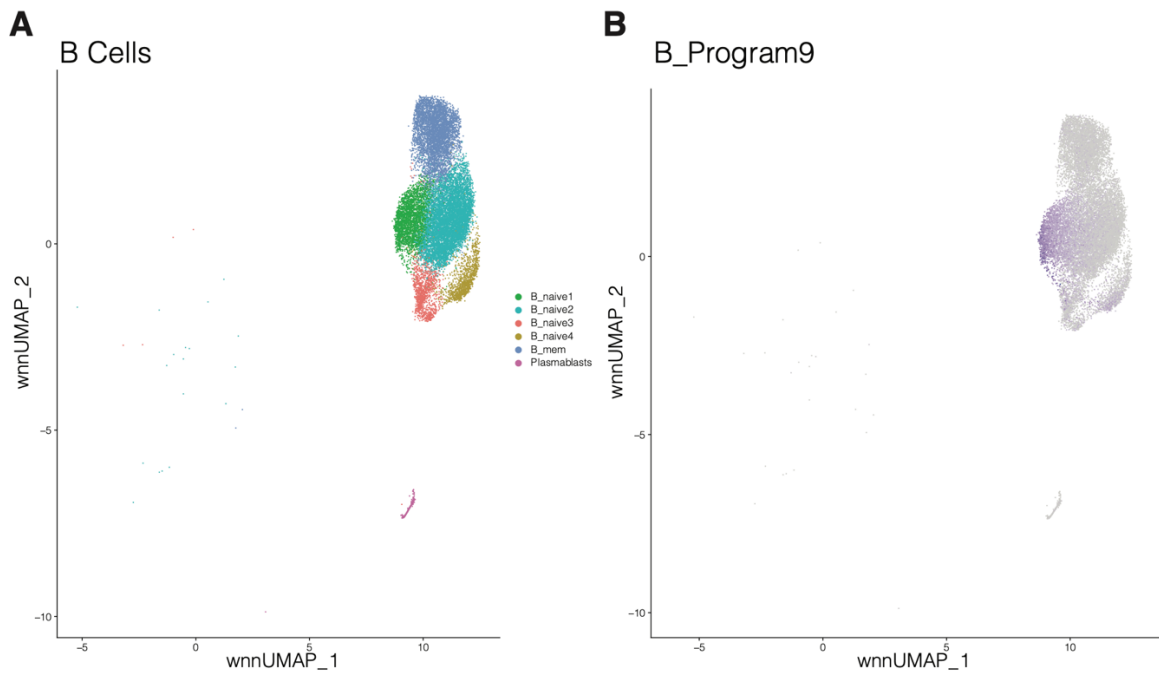


Supplemental Figure 4.28 GEP scores for longitudinal samples collected at an individual's high and low disease activity point.

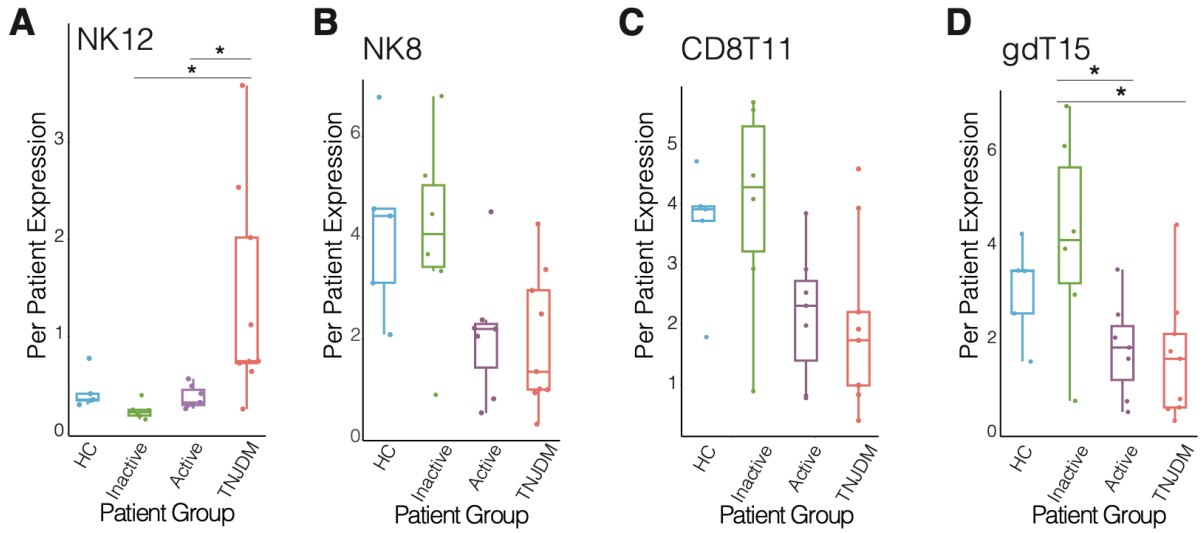
Individuals are labelled using the same Donor ID used throughout the paper. Changes in expression within individuals trended, but in this subset of patients with longitudinal samples, there was insufficient statistical power to quantify statistical significance given disease activity heterogeneity across these 6 individuals.



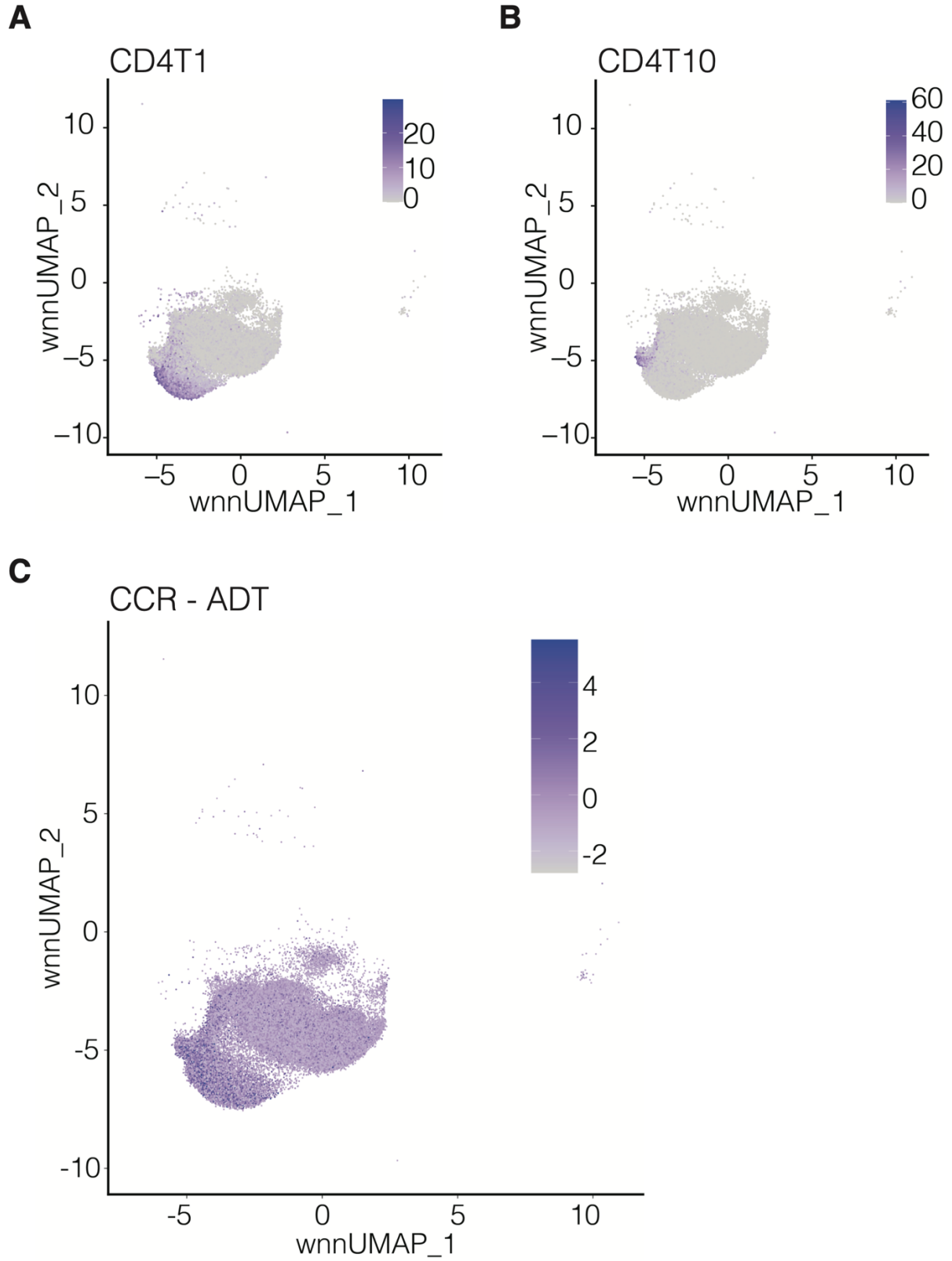
Supplemental Figure 4.29 Scatter plots showing mean sample expression ($n=27$) of type I interferon response programs in each corresponding cell type (Pearson).



Supplemental Figure 4.30 Expression of program B9 in naive B cells.
 (A) wnnUMAP of B cell subsets
 (B) wnnUMAP showing expression of program B9 in B cells.

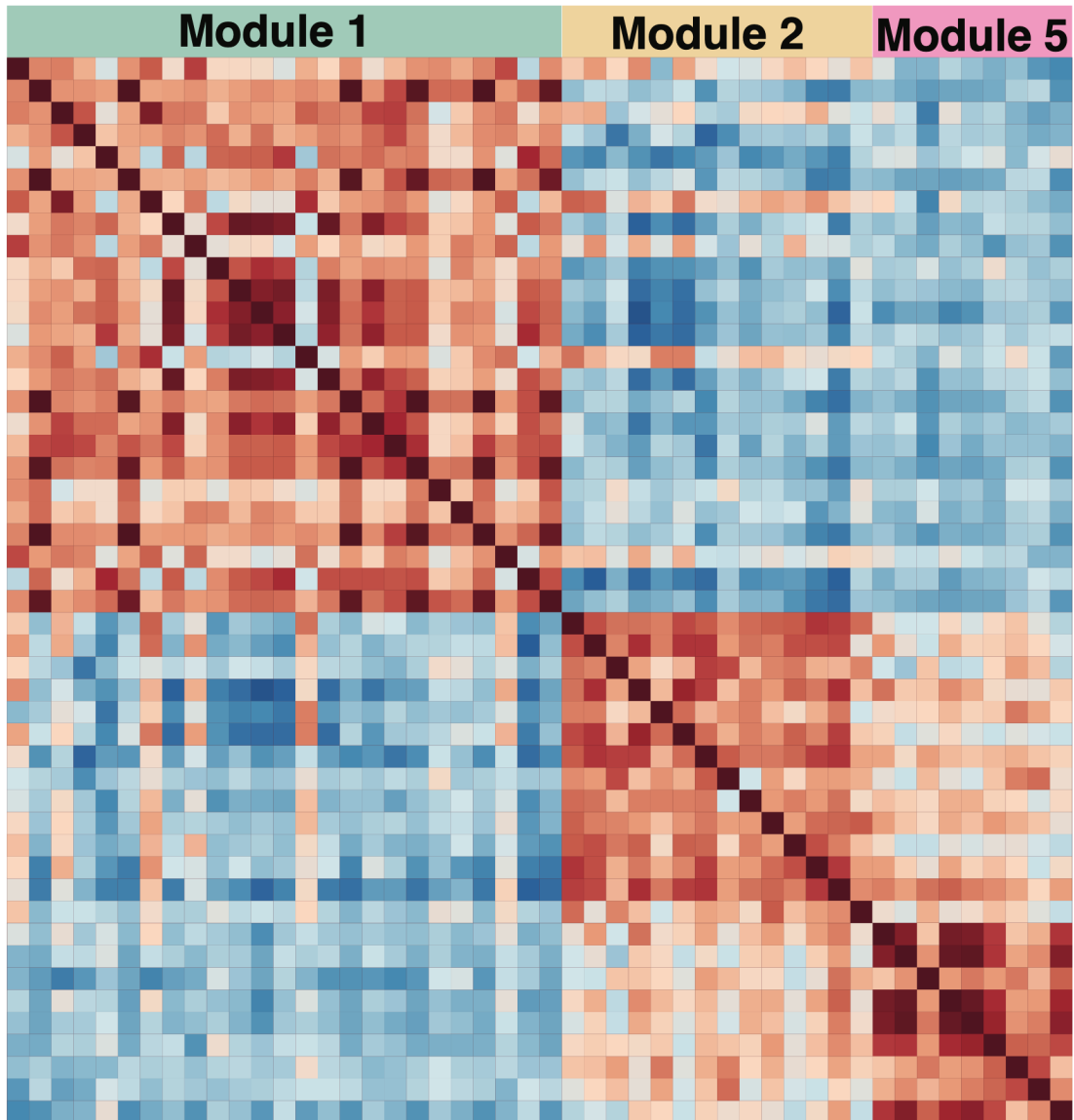


Supplemental Figure 4.31 Mean patient expression of disease activity associated programs (4-way ANOVA, $p < 0.05$) in Module 5 (* $p < 0.05$ Post-hoc pairwise Tukey test).

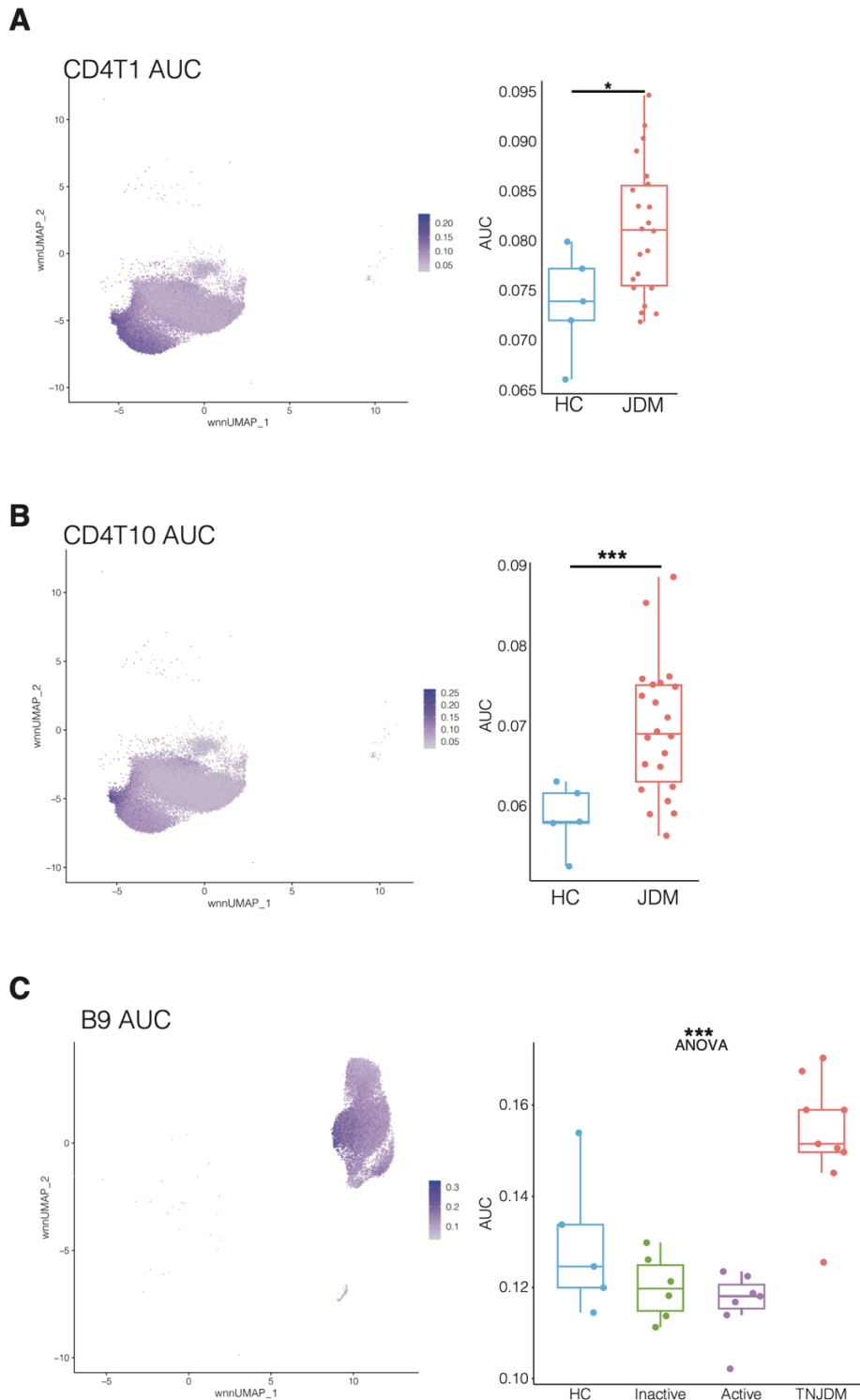


Supplemental Figure 4.32 wnnUMAPs showing normalized expression of GEPs CD4T1 and CD4T10 (A-B) with co-expression of surface protein CCR4 (C).

A

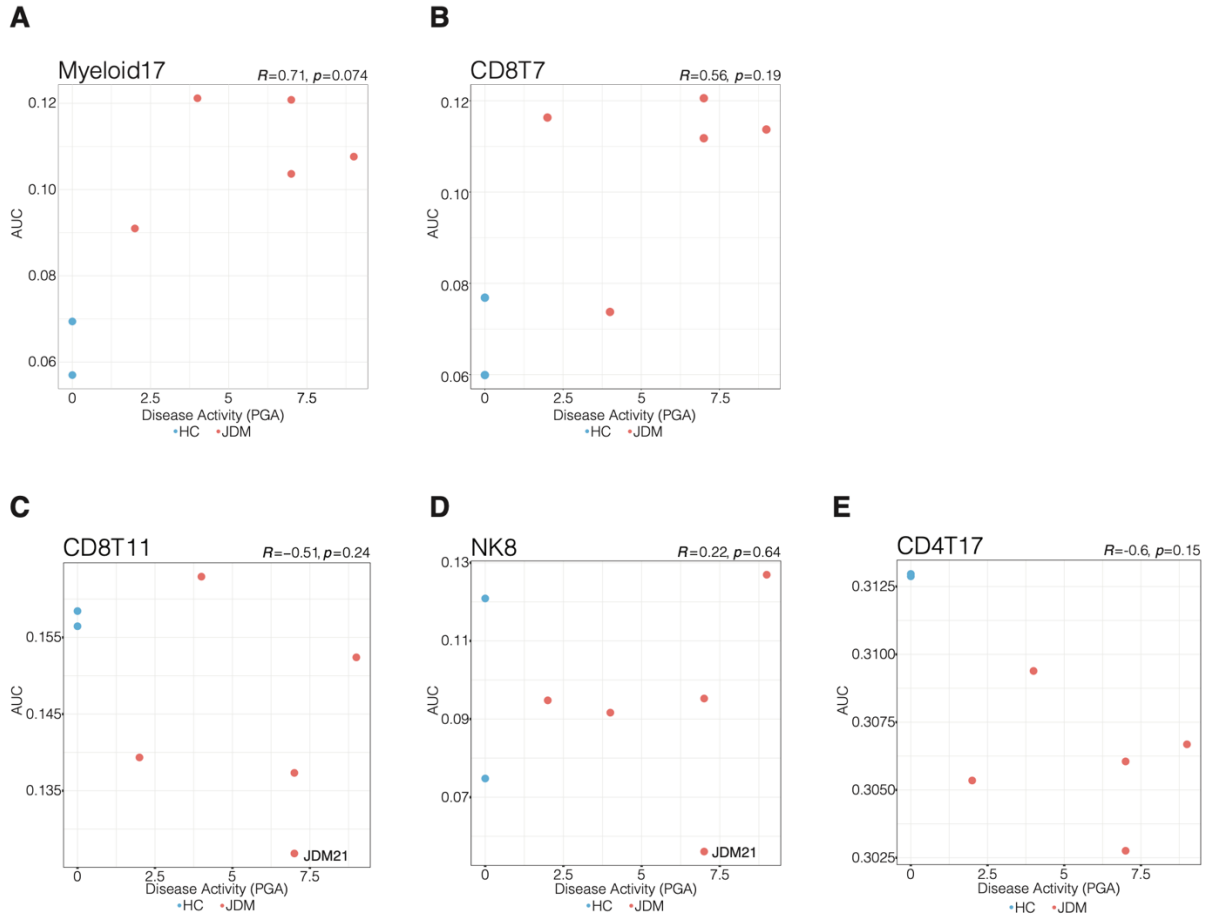


Supplemental Figure 4.33 Subset of Modules 1, 2, and 5 from original heatmap in Figure 4.6 highlighting the negative correlations.



Supplemental Figure 4.34 Proxy GEP metric (AUCcell) recapitulates key signatures discovered via DECIPHER in original dataset.

(A-B) Single-cell expression of proxy GEP metric calculated using AUCcell in the original dataset and quantification of proxy program expression for each patient comparing HC (n=5) to JDM (n=22) (t-test: * $p < 0.05$, *** $p < 0.001$). (C) Single-cell expression of proxy GEP metric calculated using AUCcell in the original dataset and quantification of proxy program expression for each patient comparing HC (n=5), Inactive JDM (n=6), Active JDM (n=7), and TN JDM (n=9) (4-group ANOVA *** $p < 0.001$).



Supplemental Figure 4.35 Relationship between disease activity and proxy scores for GEPs in validation cohort. (A-B) Scatterplots correlating disease activity (PGA) with AUCcell scores for proxy IFN programs, Myeloid17 and CD8T7, in independent dataset (Spearman). (C-E) Scatterplots correlating disease activity (PGA) with AUCcell scores for proxy CD8T11, NK8, and CD4T17 programs in independent dataset (Spearman). The cell death regulatory programs previously found to be lowest in active and treatment-naïve JDM, CD8T11 and NK8, were not significantly correlated in the independent dataset, though the single treatment-naïve patient (JDM21) exhibited the lowest scores for both programs. CD4T17 exhibited a negative trend with disease activity that did not meet the threshold for significance

Supplemental Table 4.2 ADT sequences and targets for surface protein panel.

DNA_ID	Clone	Target	barcode
A0005	2D10	anti-human CD80	ACGAATCAATCTGTG
A0006	IT2.2	anti-human CD86	GTCTTTGTCAGTGCA
A0007	29E.2A3	anti-human CD274 (B7-H1, PD-L1)	GTTGTCCGACAATAC
A0008	24F.10C12	anti-human CD273 (B7-DC, PD-L2)	TCAACGCTTGGCTAG
A0010	DCN.70	anti-human CD276 (B7-H3)	GACTGGGAGGGTATT
A0016	9M1-3	anti-human Galectin-9	ACTCACTGGAGTCTC
A0020	122	anti-human CD270 (HVEM, TR2)	TGATAGAAACAGACC
A0021	11C3.1	anti-human CD252 (OX40L)	TTTAGTGATCCGACT
A0022	5F4	anti-human CD137L (4-1BB Ligand)	ATTCGCCTTACGCAA
A0023	SKII.4	anti-human CD155 (PVR)	ATCACATCGTTGCCA
A0024	TX31	anti-human CD112 (Nectin-2)	AACCTTCCGTCTAAG
A0026	CC2C6	anti-human CD47	GCATTCTGTACACCTA
A0029	BJ40	anti-human CD48	CTACGACGTAGAAGA
A0031	5C3	anti-human CD40	CTCAGATGGAGTATG
A0032	24-31	anti-human CD154	GCTAGATAGATGCAA
A0033	HI186	anti-human CD52	CTTTGTACGAGCAAA
A0034	UCHT1	anti-human CD3	CTCATTGTAACCTCT
A0046	SK1	anti-human CD8	GCGCAACTTGATGAT
A0047	5.1H11	anti-human CD56 (NCAM)	TCCTTTCCTGATAGG
A0048	2D1	anti-human CD45	TCCCTTGCGATTTAC
A0050	HIB19	anti-human CD19	CTGGGCAATTAACCTCG
A0052	P67.6	anti-human CD33	TAACTCAGGGCCTAT
A0053	S-HCL-3	anti-human CD11c	TACGCCTATAACTTG
A0054	581	anti-human CD34	GCAGAAATCTCCCTT
A0058	W6/32	anti-human HLA-A,B,C	TATGCGAGGCTTATC
A0060	5E10	anti-human CD90 (Thy1)	GCATTGTACGATTCA
A0061	104D2	anti-human CD117 (c-kit)	AGACTAATAGCTGAC
A0062	HI10a	anti-human CD10	CAGCCATTCATTAGG
A0063	HI100	anti-human CD45RA	TCAATCCTTCCGCTT
A0064	6H6	anti-human CD123	CTTCACTCTGTCAGG
A0066	CD7-6B7	anti-human CD7	TGGATTCCCGGACTT
A0068	43A3	anti-human CD105	ATCGTCGAGAGCTAG
A0069	RCR-401	anti-human CD201 (EPCR)	GTTTCCTTGACCAAG
A0071	L291H4	anti-human CD194 (CCR4)	AGCTTACCTGCACGA
A0073	IM7	anti-mouse/human CD44	TGGCTTCAGGTCCTA
A0081	M5E2	anti-human CD14	TCTCAGACCTCCGTA

DNA_ID	Clone	Target	barcode
A0083	3G8	anti-human CD16	AAGTTCACTCTTTGC
A0085	BC96	anti-human CD25	TTTGTCCTGTACGCC
A0087	UCHL1	anti-human CD45RO	CTCCGAATCATGTTG
A0088	EH12.2H7	anti-human CD279 (PD-1)	ACAGCGCCGTATTTA
A0089	A15153G	anti-human TIGIT (VSTM3)	TTGCTTACCGCCAGA
A0090	MOPC-21	Mouse IgG1, κ isotype Ctrl	GCCGGACGACATTAA
A0091	MOPC-173	Mouse IgG2a, κ isotype Ctrl	CTCCTACCTAAACTG
A0092	MPC-11	Mouse IgG2b, κ isotype Ctrl	ATATGTATCACGCGA
A0095	RTK4530	Rat IgG2b, κ Isotype Ctrl	GATTCTTGACGACCT
A0100	2H7	anti-human CD20	TTCTGGGTCCCTAGA
A0101	9E2	anti-human CD335 (NKp46)	ACAATTTGAACAGCG
A0102	BM16	anti-human CD294 (CRTH2)	TGTTTACGAGAGCCC
A0103	RA3-6B2	anti-mouse/human CD45R/B220	CCTACACCTCATAAT
A0123	9C4	anti-human CD326 (Ep-CAM)	TTCCGAGCAAGTATC
A0124	WM59	anti-human CD31	ACCTTTATGCCACGG
A0127	NC-08	anti-Human Podoplanin	GGTTACTCGTTGTGT
A0128	16A1	anti-human CD140a (PDGFR α)	ATGCGCCGAGAATTA
A0129	18A2	Hu CD140b (PDGFR β)	CAATGGTTCCTGCC
A0132	AY13	anti-human EGFR	GCTTAACATTGGCAC
A0134	P1H12	anti-human CD146	CCTTGGATAACATCA
A0136	MHM-88	anti-human IgM	TAGCGAGCCCGTATA
A0138	UCHT2	anti-human CD5	CATTAACGGGATGCC
A0140	G025H7	anti-human CD183 (CXCR3)	GCGATGGTAGATTAT
A0141	J418F1	anti-human CD195 (CCR5)	CCAAAGTAAGAGCCA
A0142	FUN-2	anti-human CD32	GCTTCCGAATTACCG
A0144	J252D4	anti-human CD185 (CXCR5)	AATTCAACCGTCGCC
A0145	Ber-ACT8	anti-human CD103 (Integrin α E)	GACCTCATTGTGAAT
A0148	G043H7	anti-human CD197 (CCR7)	AGTTCAGTCAACCGA
A0149	HP-3G10	anti-human CD161	GTACGCAGTCCTTCT
A0151	BNI3	anti-human CD152 (CTLA-4)	ATGGTTCACGTAATC
A0153	SA231A2	anti-human KLRG1 (MAFA)	CTTATTTCTGCCCT
A0154	O323	anti-human CD27	GCACTCCTGCATGTA
A0155	H4A3	anti-human CD107a (LAMP- 1)	CAGCCCACTGCAATA
A0156	DX2	anti-human CD95 (Fas)	CCAGCTCATTAGAGC
A0159	L243	anti-human HLA-DR	AATAGCGAGCAAGTA
A0160	L161	anti-human CD1c	GAGCTACTTCACTCG
A0161	ICRF44	anti-human CD11b	GACAAGTGATCTGCA

DNA_ID	Clone	Target	barcode
A0162	10.1	anti-human CD64	AAGTATGCCCTACGA
A0163	M80	anti-human CD141 (Thrombomodulin)	GGATAACCGCGCTTT
A0165	1D11	Hu CD314 (NKG2D)	CGTGTTTGTTCTCA
A0166	6/40c	anti-human CD66b	AGCTGTAAGTTTCGG
A0168	QA17A04	anti-human CD57 Recombinant	AACTCCCTATGGAGG
A0169	F38-2E2	anti-human CD366 (Tim-3)	TGTCCTACCCAACTT
A0170	MIH26	anti-human CD272 (BTLA)	GTTATTGGACTAAGG
A0171	C398.4A	anti-human/mouse/rat CD278 (ICOS)	CGCGCACCCATTAAA
A0172	9F.8A4	anti-human CD275 (B7-H2, B7-RP1, ICOSL)	GTTAGTGTTAGCTTG
A0175	NK92.39	anti-human CD96 (TACTILE)	TGGCCTATAAATGGT
A0176	A1	anti-human CD39	TTACCTGGTATCCGT
A0177	NOK-1	anti-human CD178 (Fas-L)	CCGGTCCTCTGTATT
A0179	K0124E1	anti-human CX3CR1	AGTATCGTCTCTGGG
A0181	Bu32	anti-human CD21	AACCTAGTAGTTCGG
A0185	TS2/4	anti-human CD11a	TATATCCTTGTGAGC
A0187	CB3-1	anti-human CD79b (Igβ)	ATTCTTCAACCGAAG
A0189	C1.7	anti-human CD244 (2B4)	TCGCTTGGATGGTAG
A0196	HIR2	anti-human CD235ab	GCTCCTTTACACGTA
A0205	15-2	anti-human CD206 (MMR)	TCAGAACGTCTAACT
A0207	8F9	anti-human CD370 (CLEC9A/DNGR1)	CTGCATTTAGTAAG
A0215	11C1	anti-human CD268 (BAFF-R)	CGAAGTCGATCCGTA
A0216	HIP1	anti-human CD42b	TCCTAGTACCGAAGT
A0217	HA58	anti-human CD54	CTGATAGACTTGAGT
A0218	AK4	anti-human CD62P (P- Selectin)	CCTTCCGTATCCCTT
A0219	GIR-208	anti-human CD119 (IFN-γ R α chain)	TGTGTATTCCCTTGT
A0224	IP26	anti-human TCR α/β	CGTAACGTAGAGCGA
A0233	MHN3-21	anti-human Notch 3	CTATTGGACGTATCT
A0236	RTK2071	Rat IgG1, κ Isotype Ctrl	ATCAGATGCCCTCAT
A0237	G0114F7	Rat IgG1, λ Isotype Ctrl	GGGAGCGATTCAACT
A0238	RTK2758	Rat IgG2a, κ Isotype Ctrl	AAGTCAGGTTTCGTTT
A0240	RTK4174	Rat IgG2c, κ Isotype Ctrl	TCCAGGCTAGTCATT
A0241	HTK888	Armenian Hamster IgG Isotype Ctrl	CCTGTCATTAAGACT
A0242	K036C2	anti-human CD192 (CCR2)	GAGTTCCCTTACCTG
A0244	CBR-IC2/2	anti-human CD102 (ICAM-2)	TGACCTTCCTCTCCT

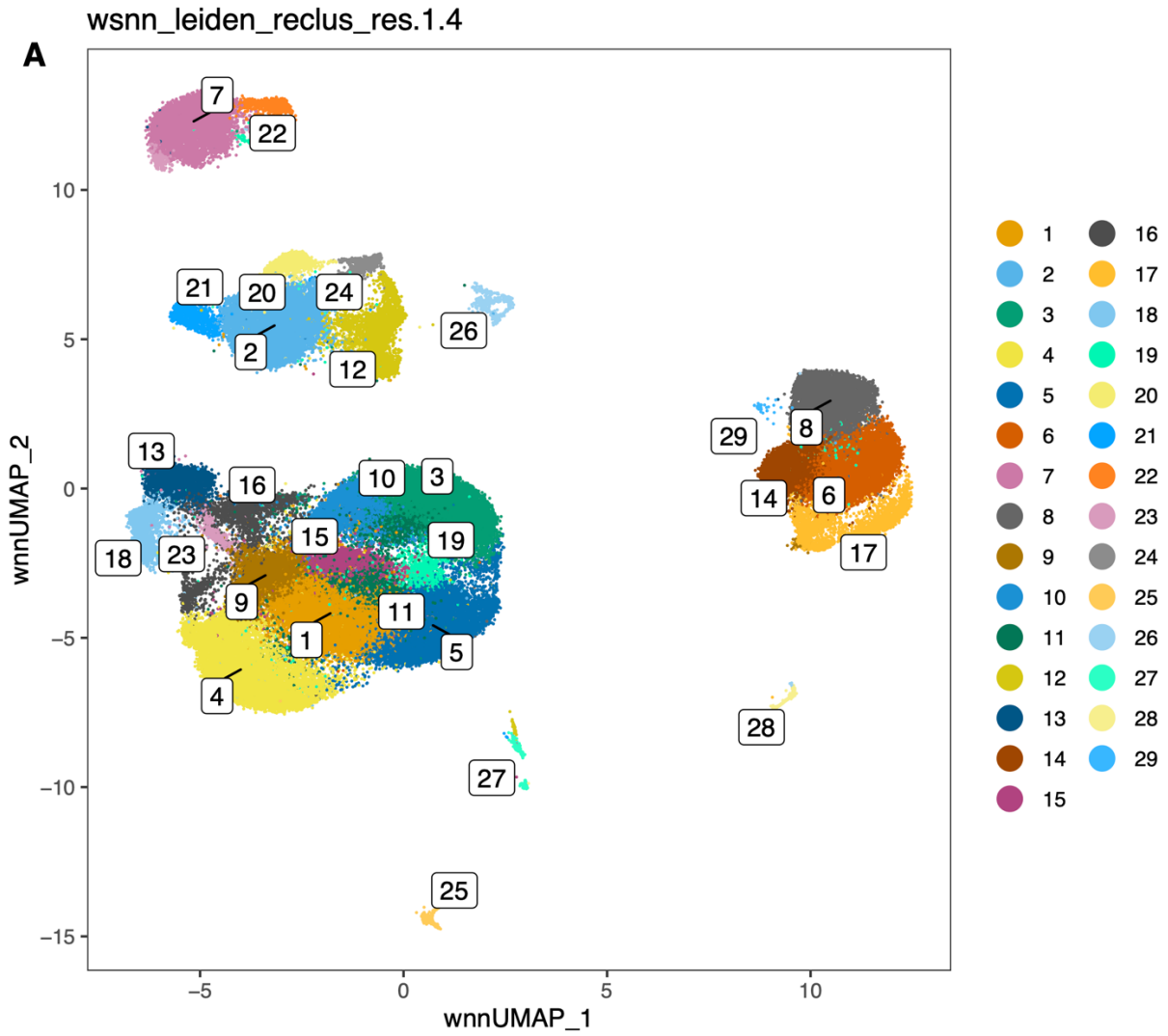
DNA_ID	Clone	Target	barcode
A0245	STA	anti-human CD106	TCACAGTTCCTTGGA
A0246	TU27	anti-human CD122 (IL-2R β)	TCATTTCCCTCCGATT
A0247	1A1	anti-human CD267 (TACI)	AGTGATGGAGCGAAC
A0352	AER-37 (CRA-1)	anti-human Fc ϵ R1 α	CTCGTTTCCGTATCG
A0353	HIP8	anti-human CD41	ACGTTGTGGCCTTGT
A0355	4B4-1	anti-human CD137 (4-1BB)	CAGTAAGTTCGGGAC
A0356	MIH24	anti-human CD254 (TRANCE, RANKL)	TCCGTGTTAGTTTGT
A0357	CD43-10G7	anti-human CD43	GATTAACCAGCTCAT
A0358	GHI/61	anti-human CD163	GCTTCTCCTTCCTTA
A0359	HB15e	anti-human CD83	CCACTCATTTCGGGT
A0361	p282 (H19)	anti-human CD59	AATTAGCCGTCGAGA
A0364	WM15	anti-human CD13	TTTCAACGCCCTTTC
A0366	12G5	anti-human CD184 (CXCR4)	TCAGGTCCTTTC AAC
A0367	TS1/8	anti-human CD2	TACGATTTGTCAGGG
A0369	TS2/16	anti-human CD29	GTATCCCTCAGTCA
A0371	P1E6-C5	anti-human CD49b	GCTTTCTTCAGTATG
A0372	VI-PL2	anti-human CD61	AGGTTGGAGTAGACT
A0373	5A6	anti-human CD81 (TAPA-1)	GTATCCTTCCTTGGC
A0374	MEM-108	anti-human CD98	GCACCAACAGCCATT
A0375	M1310G05	anti-human IgG Fc	CTGGAGCGATTAGAA
A0382	MEM-166	anti-human CD177	AGTATGGAGCCATAT
A0383	JS11	anti-human CD55	GCTCATTACCCATTA
A0384	IA6-2	anti-human IgD	CAGTCTCCGTAGAGT
A0385	TS1/18	anti-human CD18	TATTGGGACACTTCT
A0386	CD28.2	anti-human CD28	TGAGAACGACCCTAA
A0387	1D3	anti-human TSLPR (TSLP-R)	CAGTCCTCTCTGTCA
A0389	HIT2	anti-human CD38	TGTACCCGCTTGTGA
A0392	W6D3	anti-human CD15 (SSEA-1)	TCACCAGTACCTAGT
A0393	S-HCL-1	anti-human CD22	GGGTTGTTGTCTTTG
A0395	MIH43	anti-human B7-H4	TGTATGTCTGCCTTG
A0396	BA5b	anti-human CD26	GGTGGCTAGATAATG
A0397	5E8	anti-human CD193 (CCR3)	ACCAATCCTTTCGTC
A0398	9-4D2-1E4	anti-human CD115 (CSF-1R)	AATCACGGTCCTTGT
A0399	7C9C20	anti-human CD204	TAGCGAGCCAGATGT
A0401	H037G3	anti-human CD301 (CLEC10A)	ACCTAGAAATCAGCA
A0402	HI149	anti-human CD1a	GATCGTGTTGTGTTA
A0404	H5C6	anti-human CD63	GAGATGTCTGCAACT

DNA_ID	Clone	Target	barcode
A0406	12C2	anti-human CD304 (Neuropilin-1)	GGACTAAGTTTCGTT
A0407	5-271	anti-human CD36	TTCTTTGCCTTGCCA
A0409	17G10.2	anti-human CD85g (ILT7)	TGTCAGTTCCTATGA
A0418	4E3.16	anti-human CD243 (ABCB1)	TGACCCGACCTTTAG
A0419	3F3	anti-human CD72	CAGTCGTGGTAGATA
A0420	HP-MA4	anti-human CD158 (KIR2DL1/S1/S3/S5)	TATCAACCAACGCTT
A0423	590H11G1E3	anti-human MERTK	TCCTGCATGTACCCA
A0427	94b/FOLR2	anti-human Folate Receptor β (FR- β)	TGTGGCTAGTCAGTT
A0430	L1-OV198.5	anti-human CD171 (LICAM)	GATGGACGACAATTC
A0432	3F4	anti-CD230 (Prion)	CAGGTCCCTTATTTTC
A0433	8C11	anti-human CD325 (N-Cadherin)	CCTTCCCTTTCTCT
A0446	VIMD2	anti-human CD93	GCGCTACTTCCTTGA
A0569	5D3	anti-human CD338 (ABCG2)	TAAGACTTGCCGTC
A0574	HI264	anti-human CD235a (Glycophorin A)	AGAGTATGTATGGGA
A0575	TS2/7	anti-human CD49a	ACTGATGGACTCAGA
A0576	9F10	anti-human CD49d	CCATTCAACTTCCGG
A0577	AD2	anti-human CD73 (Ecto-5'-nucleotidase)	CAGTTCCTCAGTTCG
A0579	HI9a	anti-human CD9	GAGTCACCAATCTGC
A0580	AA1	anti-human mast cell tryptase	ACTGATAGACCCGCT
A0581	3C10	anti-human TCR V α 7.2	TACGAGCAGTATTCA
A0582	B6	anti-human TCR V δ 2	TCAGTCAGATGGTAT
A0583	B3	anti-human TCR V γ 9	AAGTGATGGTATCTG
A0586	TREM-26	anti-human CD354 (TREM-1)	TAGCCGTTTCCTTTG
A0588	33.1 (Ab33)	anti-human CD202b (Tie2/Tek)	CGATCCCTTACCTAT
A0590	NKTA255	anti-human CD305 (LAIR1)	ATTTCCATTCCCTGT
A0591	15C4	anti-human LOX-1	ACCCTTTACCGAATA
A0592	DX27	anti-human CD158b (KIR2DL2/L3, NKAT2)	GACCCGTAGTTTGAT
A0597	9E9A8	anti-human CD209 (DC-SIGN)	TCACTGGACACTTAA
A0598	S16017E	anti-human CD110	TGTTGTAAGATGCCA
A0599	DX9	anti-human CD158e1 (KIR3DL1, NKB1)	GGACGCTTTCCTTGA
A0801	P30-15	anti-human CD337 (NKp30)	AAAGTCACTCTGCCG
A0803	RIK-2	anti-human CD253 (Trail)	GCCATTCTGCCTAA

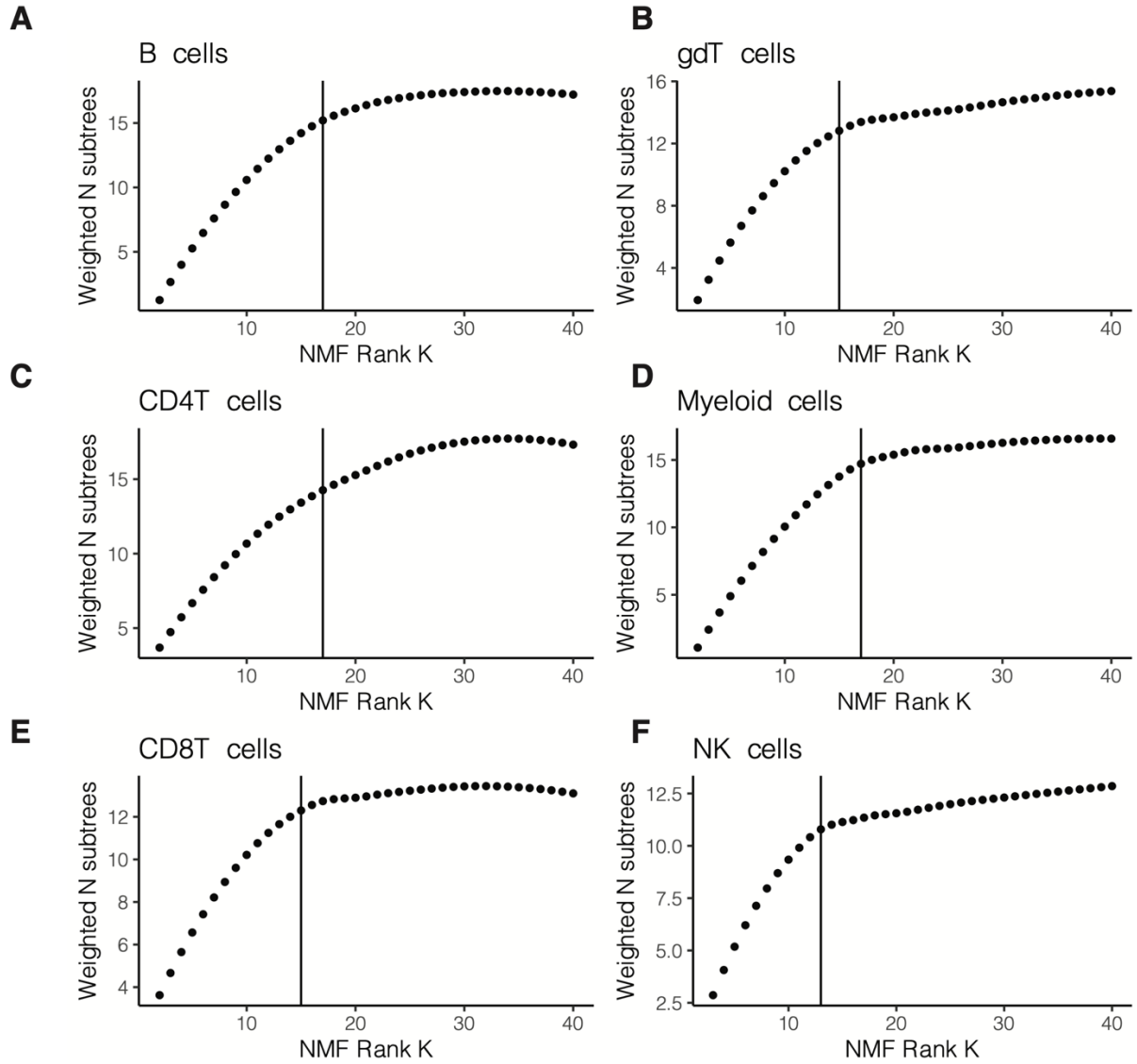
DNA_ID	Clone	Target	barcode
A0804	K041E5	anti-human CD186 (CXCR6)	GACAGTCGATGCAAC
A0805	TX25	anti-human CD226 (DNAM-1)	AGACCAACTCATTCA
A0814	HD30	anti-human CD205 (DEC-205)	CTATCGTTTGATGCA
A0817	W7C5	anti-human CD109	CACTTAACTCTGGGT
A0819	UV4	anti-human CD126 (IL-6R α)	TGATGGGAGCTTATC
A0820	2E1B02	anti-human GP130	CACGAGAATTTTCAGT
A0821	67D2	anti-human CD164	GAGGCACTTAACATA
A0822	NY2	anti-human CD142	CACTGCCGTCGATTA
A0826	H5/FcRL3	anti-human CD307c/FcRL3	GCCTAGTTTGAACGC
A0828	413D12	anti-human CD307d (FcRL4)	CGATTTGATCTGCCT
A0829	509f6	anti-human CD307e (FcRL5)	TCACGCAGTCCTCAA
A0843	L053E8	anti-human CD199 (CCR9)	ATTCCTCATTCTGA
A0844	MEM-55	anti-human CD45RB	AGATGGGACTCACCA
A0845	3B2/TA8	anti-human CD99	ACCCGTCCCTAAGAA
A0853	50C1	anti-human CLEC12A	CATTAGAGTCTGCCA
A0858	TRA-2-10	anti-human CD46	ACAGTACGACCTTCT
A0861	50-6	anti-human CD151 (PETA-3)	CTTACCTAGTCATTC
A0862	H44	anti-human CD218a (IL-18R α)	TTGTTGTATCCGATC
A0864	NT-7	anti-human CD352 (NTB-A)	AGTTTCCACTCAGGC
A0866	AYP1	anti-human CLEC1B (CLEC2)	TGCCAGTATCACGTA
A0867	DX22	anti-human CD94	CTTTCCGGTCCTACA
A0868	MHE-18	anti-human IgE	GGATGTACCGCGTAT
A0869	1D12	anti-human CD365 (Tim-1)	CTTCTGGGATTCTGG
A0870	A12 (7D4)	anti-human CD150 (SLAM)	GTCATTGTATGTCTG
A0871	KPL-1	anti-human CD162	ATATGTCAGAGCACC
A0872	CD84.1.21	anti-human CD84	CTCCCTAGTTCTTTT
A0894	MHK-49	anti-human Ig light chain κ	AGCTCAGCCAGTATG
A0895	M3/38	anti-mouse/human Mac-2 (Galectin-3)	GATGCAATTAGCCGG
A0896	GHI/75	anti-human CD85j (ILT2)	CCTTGTGAGGCTATG
A0897	EBVCS-5	anti-human CD23	TCTGTATAACCGTCT
A0898	MHL-38	anti-human Ig light chain λ	CAGCCAGTAAGTCAC
A0899	BB7.2	anti-human HLA-A2	GAACATTTCCGACAA
A0900	L263G8	anti-human CD198 (CCR8)	AGCCCGGATGTATTT
A0901	7B11	anti-human GARP (LRRC32)	AGGTATGGTAGAGTA
A0902	6-434	anti-human CD328 (Siglec-7)	CTTAGCATTTCACTG
A0908	H131	anti-human TCR V β 13.1	TTATGGACGTATGGT
A0912	CG4	anti-human GPR56	GCCTAGTTTCCGTTT
A0918	3D12	anti-human CD19	GAGTCGAGAAATCAT

DNA_ID	Clone	Target	barcode
A0919	6D4	anti-human MICA/MICB	CCCGCAGTATAACGA
A0920	ASL-24	anti-human CD82	TCCCCTTCCGCTTT
A0923	5D12	anti-human NKp80	TATAGTTCCTCTGTG
A0928	SA17RN1	anti-human TSPAN33 (BAAM)	GAGTTCGTTGTTCCA
A0931	1C1	anti-human CD131	CTGCATGAGACCAAA
A0932	31G4D8	anti-human Lymphotoxin β Receptor (LT- β R)	CCTCTATTCAGAGCA
A0933	74/3	anti-Annexin A1	CCCCTGGAGCAATT
A0934	HSL96	anti-human CD179a (VpreB)	TAGATGGGATTCCGG
A0935	LN2	anti-human CD74	CTGTAGCATTTCCT
A0936	RS38E	anti-human CD317 (BST2, Tetherin)	AAGAGCCGTTGTGAA
A0939	TW4-2F8	anti-human LAP (TGF- β 1)	ATCCTTCCGATTGTG
A0940	4H1	anti-human CD116	ATGGACAGTTCGTGT
A0941	M-B371	anti-human CD37	ACAGTCACTGGGCAA
A0944	BB27	anti-human CD101 (BB27)	CTACTTCCCTGTCAA
A0945	MAb11	anti-human TNF- α	CCTATGAACGTAACG
A0948	OV-5B8	anti-human CD321	GACAGTACCGACT
A1018	Tü39	anti-human HLA-DR, DP, DQ	AGCTACGAGCAGTAG
A0028	BY88	anti-human CD30	TCAGGGTGTGCTGTA
A0056	19F2	anti-human CD269 (BCMA)	CAGATGATCCACCAT
A0131	16G5	anti-human Cadherin 11	CGTTGCCATTAACCA
A0135	67A4	anti-human CD324 (E- Cadherin)	ATCCTTCTCCCTTTC
A0139	B1	anti-human TCR γ/δ	CTTCCGATTCATTCA
A0152	11C3C65	anti-human CD223 (LAG-3)	CATTTGTCTGCCGGT
A0158	Ber-ACT35 (ACT35)	anti-human CD134 (OX40)	AACCCACCGTTGTTA
A0174	TS2/9	anti-human CD58 (LFA-3)	GTTCCCTATGGACGAC
A0206	7-239	anti-human CD169 (Sialoadhesin, Siglec-1)	TACTCAGCGTGTTTG
A0208	S15046E	anti-human XCR1	AAGACGCATGTCAAC
A0213	MHN1-519	anti-human Notch 1	AATCTGTAGTGCGTT
A0248	HAE-1f	anti-human CD62E	CTCCCTGTGGCTTAA
A0351	BV10A4H2	anti-human CD135 (Flt-3/Flk- 2)	CAGTAGATGGAGCAT
A0360	108-17	anti-human CD357 (GITR)	ACCTTTCGACTCG
A0362	7D4-6	anti-human CD309 (VEGFR2)	TTCACGCAGTAAGAT
A0363	G077F6	anti-human CD124 (IL-4R α)	CCGTCCTGATAGATG
A0370	201A	anti-human CD303 (BDCA-2)	GAGATGTCCGAATTT

DNA_ID	Clone	Target	barcode
A0400	BV9	anti-human CD144 (VE-Cadherin)	TCCACTCATTCTGTA
A0403	10E2	anti-human CD207 (Langerin)	CATTCTTCACGGGAT
A0405	HTA125	anti-human CD284 (TLR4)	GCTTAGCTGTATCCG
A0447	OX-104	anti-human CD200 (OX2)	CACGTAGACCTTTGC
A0593	NP4D6	anti-human CD203c (E-NPP3)	TAACCGTACCTGCAT
A0815	6588-5	anti-human CCR10	ATCTGTATGTCACAG
A0816	ME20.4	anti-human CD271 (NGFR)	AACCGCGCTTCAGAT
A0831	DL-101	anti-human CD138 (Syndecan-1)	GTATAGACCAAAGCC
A0863	1D6	anti-human CD257 (BAFF, BLYS)	CAGAGCACCCATTAA
A0180	ML5	anti-human CD24	AGATTCCTTCGTGTT
A0394	CY1G4	anti-human CD71	CCGTGTTCCCTCATTA
A0143	G034E3	anti-human CD196 (CCR6)	GATCCCTTTGTCACT
A0027	113-16	anti-human CD70	CGCGAACATAAGAAG
A0070	GoH3	anti-human/mouse CD49f	TTCCGAGGATGATCT
A0428	9F4	anti-human TIM-4	CGTCATATAGTATGG
A0408	15-414	anti-human CD172a (SIRP α)	CGTGTTTAACTTGAG
A0830	162.1	anti-human CD319 (CRACC)	AGTATGCCATGTCTT
A0600	UP-R1	anti-human CD158f (KIR2DL5)	AAAGTGATGCCACTG
A0594	S16016B	anti-human CD133	GTAAGACGCCTATGC
A0390	A019D5	anti-human CD127 (IL-7R α)	GTGTGTTGTCCTATG
A0572	1D9-M12	anti-human C5L2	ACAATTTGTCTGCGA
A0147	DREG-56	anti-human CD62L	GTCCCTGCAACTTGA
A0164	51.1	anti-human CD1d	TCGAGTCGCTTATCA
A0146	FN50	anti-human CD69	GTCTCTTGGCTTAAA
A0167	E11	anti-human CD35	ACTTCCGTGCGATCTT
A0214	FIB504	anti-human/mouse integrin β 7	TCCTTGATGTACCG
A0865	9D9F9	anti-human VEGFR-3 (FLT-4)	TGATCCGAAGTCGTG
A0072	RPA-T4	anti-human CD4	TGTTCCCGCTCAACT
A0133	24D2	anti-human CD340 (erbB2/HER-2)	CTGTAGCCGCCTATT



Supplemental Figure 4.36 Original multi-modal wnnUMAP, using Leiden clustering at a resolution of 1.4.



Supplemental Figure 4.37 Elbow plots for rank selection for NMF ran on each major cell type, with rank k indicated by vertical line.

CHAPTER 5 CONCLUSION

Tissues, as complex living systems, exhibit hierarchical organization and orchestrated state changes to execute their biological functions. These transitions are driven by various factors, including local cell-to-cell signaling and microenvironmental cues such as hormonal fluctuations and mechanotransduction. Single-cell genomics emerged as a powerful tool for uncovering cellular heterogeneity and molecular underpinnings across various biological processes. Nevertheless, analyzing single-cell genomics data continues to pose computational challenges due to its high dimensionality and noise. This dissertation addressed these challenges by introducing DECIPHER, a machine learning framework for network inference applied to single-cell RNA sequencing data, specifically targeting cellular coordination and interaction networks from single-cell genomics data.

Summary of Advancements

During this PhD, I originally set out to tackle the looming question in this current era of ‘big data’ biology where the availability of high-dimensional data is scaling in accord with Moore’s law: given the sheer amount of data, how do scientists go from big data to biological insight? To that end, in this dissertation I have described and demonstrated the utility of the DECIPHER algorithm as a tool to address this challenge in single-cell RNA sequencing data.

Chapter 2 introduced the algorithm and its R implementation, `deciphR`, enabling the reconstruction of simulated cell state networks from high-dimensional molecular profiles. Expanding upon this, Chapter 3 applied DECIPHER to unveil cell-cell interaction networks within human breast tissue, shedding light on the dynamic interplay between different cell types in response to hormonal fluctuations and validating a subset of these interactions

nominated from transcriptomic data through IHC imaging. Chapter 4 extended DECIPHER's application to investigate coordinated immune dysregulation in a rare autoimmune disease, uncovering functional immune imbalances and potential therapeutic avenues.

In particular, the successful application of the DECIPHER framework to a rare disease with poorly understood etiology—and subsequent validation of key DECIPHER-discovered signatures in a separate cohort with a different computational method—exemplifies the technical advancements enabled by this machine learning approach. By deploying NMF in a way that both integrates data and extracts key biological signals, i.e. combining the advantages of iNMF and consensus NMF, we were able to maximize the biological information learned from a limited patient cohort and resolve heterogeneity that could not be parsed by traditional differential expression approaches.

Beyond the technical advantages of this method, I have prioritized interpretability of the algorithm and downstream results in order to strengthen its utility for biologists. For example, an initial version of the method relied on identifying correlations between PCs because first, the dimensionality reduction of PCA provides a unique solution and second, the PCs are ranked according to the amount of variability explained by that component and thus provide a starting point for prioritizing biological interpretation. However, PCs contain both 'positive' and 'negative' components, i.e. genes that are both highly and lowly expressed which together explain a fraction of the observed variability. Thus, interpreting the meaning of correlations between PCs becomes more difficult, as one must parse the contributions of both up- and down-regulated genes to the observed co-variation across biological contexts. As another example of how I have prioritized usability of this method,

I developed the method to interface with Seurat, a popular single-cell analysis pipeline. Finally, in the *deciphR* package implementation, we have included visualization functions for both the network outputs and intermediate metrics that are tuned during rank optimization. With these design choices, we hope that the DECIPHER algorithm will be a useful tool to not just computational and systems biologists, but also biologists across disciplines who are generating single-cell genomics data.

Limitations and Assumptions

However, the downstream results and technical advantages of DECIPHER should be interpreted in the context of the method's limitations and assumptions. First, as has been previously mentioned, NMF is distinct from principal component analysis in that there is no single solution for the number of patterns or components into which the data is segmented. As such, it is necessary to optimize the parameter 'rank K' such that the NMF results capture the relevant biology at an appropriate granularity. By developing the DEW metric, we intended to address this limitation of NMF as none of the previously reported metrics for rank selection such as KL divergence or reconstruction error produced a clear 'elbow' or saturation point when used in k-sweeps. However, further investigation, including broader use of the algorithm, would help identify cases and general 'data structures' that the DEW metric does not cleanly resolve.

Additionally, the NMF algorithm is computationally intensive, particularly because our consensus approach requires replicate iterations of NMF with unique random seeds. Although some of the time required for dimensionality reduction has been reduced by adopting the online iNMF approach⁷⁸ developed by Gao et al. and parallelization of the replicates, running the K-sweep remains a major process bottleneck in the workflow. We

hope that by making the method open source before manuscript acceptance, the deciphR package can be further refined through the input of the broader computational biology community.

Finally, the network approach used in the second half of the DECIPHER workflow makes several fundamental biological assumptions in order to apply the chosen mathematical framework. First, by filtering nodes based on modularity, i.e. programs that are poorly connected to the rest of the network, we assume that biologically relevant programs co-vary with multiple other activity programs. Thus, we assume the low dimensionality of gene expression on multiple levels: first on the level of individual genes during NMF and second on the level of coordinated activity programs during network inference. As previously mentioned in Chapter 2, if one is interested in querying single-cell data for rare cell types or isolated biological processes, this method is likely not appropriate.

Second, the network construction from all samples in the dataset assumes that the underlying structure of the cellular coordination network is consistent across biological contexts and only the magnitude of correlation, i.e. proximity of individual nodes in graph representations, differs across the sample cohort. There are several clear cases where this assumption would not apply such as lesional versus healthy skin tissue within an individual or tumors, which are characterized by high mutational burden and markedly distinct microenvironments. Despite these limitations, I believe the DECIPHER method still contributes to the advancement of the field because it presents an interpretable and generally implementable machine learning tool to biologists interested in nominating hypotheses from high dimensional data in a principled manner.

Future Directions

Looking ahead, an exciting frontier lies at the integration of multi-modal data and application of transfer learning techniques within the DECIPHER framework. Multi-modal data fusion, encompassing transcriptomics, proteomics, and epigenomics, holds promise for a more holistic understanding of cellular dynamics and interactions, providing insights into tissue-specific responses and disease mechanisms. Transfer learning, meanwhile, enables the transfer of knowledge across datasets and tissues, facilitating the translation of findings from model systems to clinical contexts and is already deployed by the iNMF algorithm adapted for DECIPHER. By leveraging transfer learning, we can bridge the gap between experimental models and human biology, accelerating the development of clinically relevant insights and personalized treatment strategies.

Importantly, future research endeavors should prioritize experimental validation and perturbation studies to elucidate the functional implications of identified cell-cell interaction networks. Incorporating spatial information into single-cell genomics analysis will further enhance our understanding of tissue microenvironments and cellular interactions within spatial contexts.

Translating research findings into clinical applications remains paramount. Utilizing DECIPHER-derived insights for biomarker discovery, patient stratification, and therapeutic target identification could markedly improve clinical decision-making and treatment outcomes for complex diseases. Interdisciplinary collaborations between computational biologists, experimentalists, and clinicians will be crucial for ensuring the seamless integration of computational insights into clinical practice, ultimately improving patient care and outcomes.

REFERENCES

1. Wilkinson, D. J. Stochastic modelling for quantitative description of heterogeneous biological systems. *Nat. Rev. Genet.* 10, 122–133 (2009).
2. Kitano, H. Computational systems biology. *Nature* 420, 206–210 (2002).
3. Clarke, R. B., Howell, A., Potten, C. S. & Anderson, E. Dissociation between steroid receptor expression and cell proliferation in the human breast. *Cancer Res.* 57, 4987–4991 (1997).
4. Collaborative Group on Hormonal Factors in Breast Cancer. *Lancet Oncol.* 13, 1141–1151 (2012).
5. Chowdhury, D. Immune Network: An Example of Complex Adaptive Systems. in *Artificial Immune Systems and Their Applications* (ed. Dasgupta, D.) 89–104 (Springer, 1999).
6. Lee, A. S. & Bar-Or, R. L. Immunology Viewed as the Study of an Autonomous Decentralized System. in *Artificial Immune Systems and Their Applications* (ed. Dasgupta, D.) 65–88 (Springer Berlin Heidelberg, 1999). doi:10.1007/978-3-642-59901-9_4.
7. Shalek, A. K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* 510, 363–369 (2014).
8. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* 36, 89–94 (2018).
9. Elkou, K. B. & Stone, V. V. Type I interferon and systemic lupus erythematosus. *J. Interferon Cytokine Res.* 31, 803–812 (2011).

10. Taplin, C. E. & Barker, J. M. Autoantibodies in type 1 diabetes. *Autoimmunity* 41, 11–18 (2008).
11. Ferreira, R. C. *et al.* A type I interferon transcriptional signature precedes autoimmunity in children genetically at risk for type 1 diabetes. *Diabetes* 63, 2538–2550 (2014).
12. Apaolaza, P. S. *et al.* Islet expression of type I interferon response sensors is associated with immune infiltration and viral infection in type 1 diabetes. *Sci Adv* 7, (2021).
13. Savic, S., Caseley, E. A. & McDermott, M. F. Moving towards a systems-based classification of innate immune-mediated diseases. *Nat. Rev. Rheumatol.* 16, 222–237 (2020).
14. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 161, 1202–1214 (2015).
15. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* 10, 1096–1098 (2013).
16. Wilhelm, B. T. *et al.* Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 453, 1239–1243 (2008).
17. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–476 (2008).
18. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628 (2008).

19. McGinnis, C. S. *et al.* MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods* 30, 1 (2019).
20. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* 14, 865–868 (2017).
21. Bandura, D. R. *et al.* Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal. Chem.* 81, 6813–6822 (2009).
22. Cusanovich, D. A. *et al.* Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* 348, 910–914 (2015).
23. Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* 11, 817–820 (2014).
24. Combes, A. J. *et al.* Global absence and targeting of protective immune states in severe COVID-19. *Nature* 591, 124–130 (2021).
25. Li, H. *et al.* Dysfunctional CD8 T Cells Form a Proliferative, Dynamically Regulated Compartment within Human Melanoma. *Cell* 176, 775-789.e18 (2019).
26. Nehar-Belaid, D. *et al.* Mapping systemic lupus erythematosus heterogeneity at the single-cell level. *Nat. Immunol.* 21, 1094–1106 (2020).
27. Flynn, E., Almonte-Loya, A. & Fragiadakis, G. K. Single-Cell Multiomics. *Annu Rev Biomed Data Sci* 6, 313–337 (2023).

28. Eling, N., Morgan, M. D. & Marioni, J. C. Challenges in measuring and understanding biological noise. *Nat. Rev. Genet.* 20, 536–548 (2019).
29. Jerby-Arnon, L. & Regev, A. DIALOGUE maps multicellular programs in tissue from single-cell or spatial transcriptomics data. *Nat. Biotechnol.* (2022) doi:10.1038/s41587-022-01288-0.
30. Cui, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* (2024) doi:10.1038/s41592-024-02201-0.
31. Wang, J. *et al.* Data denoising with transfer learning in single-cell transcriptomics. *Nat. Methods* 16, 875–878 (2019).
32. Segel, L. A. The Immune System as a Prototype of Autonomous Decentralized Systems. in *IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation* vol. 1 375–385 (1997).
33. Dasgupta, D. Advances in artificial immune systems. *IEEE Comput. Intell. Mag.* 1, 40–49 (2006).
34. Michelan, R. & Von Zuben, F. J. Decentralized control system for autonomous navigation based on an evolved artificial immune network. in *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No.02TH8600)* vol. 2 1021–1026 vol.2 (IEEE, 2002).
35. Field, D. J. What is the goal of sensory coding? *Neural Comput.* 6, 559–601 (1994).

36. Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607–609 (1996).
37. Sejnowski, T. J., Koch, C. & Churchland, P. S. Computational neuroscience. *Science* 241, 1299–1306 (1988).
38. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* 20, 296–15 (2019).
39. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36, 411–420 (2018).
40. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* 177, 1888-1902.e21 (2019).
41. Zheng, A. & Casari, A. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. (“O’Reilly Media, Inc.,” 2018).
42. Jimenez, L. O. & Landgrebe, D. A. Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data. *IEEE Trans. Syst. Man Cybern. C Appl. Rev.* 28, 39–54 (1998).
43. Thippa Reddy, G. *et al.* Analysis of Dimensionality Reduction Techniques on Big Data. *IEEE Access* 8, 54776–54788 (2020).

44. Altman, N. & Krzywinski, M. The curse(s) of dimensionality. *Nat. Methods* 15, 399–400 (2018).
45. Pe'er, D., Regev, A., Elidan, G. & Friedman, N. Inferring subnetworks from perturbed expression profiles. *Bioinformatics* 17 Suppl 1, S215-24 (2001).
46. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788–791 (1999).
47. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2, 559–572 (1901).
48. Turk, M. & Pentland, A. Eigenfaces for recognition. *J. Cogn. Neurosci.* 3, 71–86 (1991).
49. Logothetis, N. K. & Sheinberg, D. L. Visual object recognition. *Annu. Rev. Neurosci.* 19, 577–621 (1996).
50. Wachsmuth, E., Oram, M. W. & Perrett, D. I. Recognition of objects and their component parts: responses of single units in the temporal cortex of the macaque. *Cereb. Cortex* 4, 509–522 (1994).
51. Lin, T.-Y. *et al.* Feature Pyramid Networks for Object Detection. *arXiv [cs.CV]* (2016).
52. Shokrollahi, M. & Krishnan, S. Non-negative matrix factorization and sparse representation for sleep signal classification. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2013, 4318–4321 (2013).

53. Hoyer, P. O. Non-negative matrix factorization with sparseness constraints. *arXiv [cs.LG]* 1457–1469 (2004).
54. Grechkin, M., Logsdon, B. A., Gentles, A. J. & Lee, S.-I. Identifying Network Perturbation in Cancer. *PLoS Comput. Biol.* 12, e1004888 (2016).
55. Heimberg, G., Bhatnagar, R., El-Samad, H. & Thomson, M. Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. *Cell Syst* 2, 239–250 (2016).
56. Segal, E., Battle, A. J. & Koller, D. Decomposing gene expression into cellular processes. in *Pacific Symposium on Biocomputing* vol. 8 89–100 (2003).
57. Segal, E., Taskar, B., Gasch, A., Friedman, N. & Koller, D. Rich probabilistic models for gene expression. *Bioinformatics* 17 Suppl 1, S243-52 (2001).
58. Segal, E. *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* 34, 166–176 (2003).
59. Bayes, T. & Price, N. LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London* 53, 370–418 (1763).

60. Pearl, J. *Probabilistic Reasoning in Intelligent Systems*. (Morgan Kaufmann, 1988).
61. *Probabilistic and Causal Inference: The Works of Judea Pearl*. vol. 36 (Association for Computing Machinery, 2022).
62. Pearl, J. Fusion, Propagation, and Structuring in Belief Networks. *Artif. Intell.* 29, 241–288 (1986).
63. Chen, M. *et al.* Inference for Network Structure and Dynamics from Time Series Data via Graph Neural Network. *arXiv:2001.06576 [cs, stat]* (2020).
64. Zhang, Z. *et al.* A general deep learning framework for network reconstruction and dynamics learning. *Applied Network Science* 4, 1–17 (2019).
65. Murrow, L. M. *et al.* Mapping hormone-regulated cell-cell interaction networks in the human breast at single-cell resolution. *Cell Syst* (2022) doi:10.1016/j.cels.2022.06.005.
66. Yuan, Y. & Bar-Joseph, Z. GCNG: graph convolutional networks for inferring gene interaction from spatial transcriptomics data. *Genome Biol.* 21, 300 (2020).
67. Dudoit, S., Shaffer, J. P. & Boldrick, J. C. Multiple Hypothesis Testing in Microarray Experiments. *Statistical Science* 18, 71–103 (2003).
68. Westfall, P., Kropf, S. & Finos, L. Weighted FWE-controlling methods in high-dimensional situations. *Institute of Mathematical Statistics* 47, 143–154 (2004).

69. Keselman, H. J., Cribbie, R. & Holland, B. Controlling the rate of Type I error over a large set of statistical tests. *Br. J. Math. Stat. Psychol.* 55, 27–39 (2002).
70. Davis, B. R. & Hardy, R. J. Upper bounds for type I and type II error rates in conditional power calculations. *Communications in Statistics - Theory and Methods* 19, 3571–3584 (1990).
71. W., T. J. The Problem of Multiple Comparisons. *Multiple Comparisons* (1953).
72. Berry, D. A. & Hochberg, Y. Bayesian perspectives on multiple comparisons. *J. Stat. Plan. Inference* 82, 215–227 (1999).
73. Halperin, M., Lan, K. K. G. & Hamdy, M. I. Some implications of an alternative definition of the multiple comparison problem. *Biometrika* 75, 773–778 (1988).
74. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* 57, 289–300 (1995).
75. Horn, M. & Dunnett, C. Power and sample size comparisons of stepwise FWE and FDR controlling test procedures in the normal many-one case. *Vol.* 48–64 (2004) doi:10.1214/LNMS/1196285625.
76. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14863–14868 (1998).

77. Kotliar, D. *et al.* Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *Elife* 8, (2019).
78. Gao, C. *et al.* Iterative single-cell multi-omic integration using online learning. *Nat. Biotechnol.* 39, 1000–1007 (2021).
79. Mairal, J., Bach, F., Ponce, J. & Sapiro, G. Online Learning for Matrix Factorization and Sparse Coding. *arXiv [stat.ML]* (2009).
80. Desper, R. & Gascuel, O. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J. Comput. Biol.* 9, 687–705 (2002).
81. Prospero, M. C. F. *et al.* A novel methodology for large-scale phylogeny partition. *Nat. Commun.* 2, 321 (2011).
82. Satopaa, V., Albrecht, J., Irwin, D. & Raghavan, B. Finding a “Kneedle” in a Haystack: Detecting Knee Points in System Behavior. in *2011 31st International Conference on Distributed Computing Systems Workshops* 166–171 (2011). doi:10.1109/ICDCSW.2011.20.
83. Cao, Y. *et al.* scDC: single cell differential composition analysis. *BMC Bioinformatics* 20, 721–12 (2019).
84. Traag, V. A., Dooren, P. & Nesterov, Y. Narrow scope for resolution-limit-free community detection. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 84, 016114 (2011).
85. Gysi, D. M., Voigt, A., Fragoso, T. de M., Almaas, E. & Nowick, K. wTO: an R package for computing weighted topological overlap and a

- consensus network with integrated visualization tool. *BMC Bioinformatics* 19, 392–16 (2018).
86. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* 9, 5233 (2019).
 87. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* 1, 417–425 (2015).
 88. Harris, M. A. *et al.* The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32, D258-61 (2004).
 89. Russo, J., Rivera, R. & Russo, I. H. Influence of age and parity on the development of the human breast. *Breast Cancer Res. Treat.* 23, 211–218 (1992).
 90. Nakshatri, H., Anjanappa, M. & Bhat-Nakshatri, P. Ethnicity-Dependent and -Independent Heterogeneity in Healthy Normal Breast Hierarchy Impacts Tumor Characterization. *Sci. Rep.* 5, 13526–14 (2015).
 91. Rosenbluth, J. M. *et al.* Organoid cultures from normal and cancer-prone human breast tissues preserve complex epithelial lineages. *Nat. Commun.* 11, 1711–14 (2020).
 92. Dunphy, K. A. *et al.* Inter-Individual Variation in Response to Estrogen in Human Breast Explants. *J. Mammary Gland Biol. Neoplasia* 25, 51–68 (2020).
 93. Muenst, S. *et al.* Pregnancy at early age is associated with a reduction of progesterone-responsive cells and epithelial Wnt signaling in human breast tissue. *Oncotarget* 8, 22353–22360 (2017).

94. Tanos, T. *et al.* Progesterone/RANKL is a major regulatory axis in the human breast. *Sci. Transl. Med.* 5, 182 55-182 55 (2013).
95. Anderson, T. J., Ferguson, D. J. & Raab, G. M. Cell turnover in the “resting” human breast: influence of parity, contraceptive pill, age and laterality. *Br. J. Cancer* 46, 376–382 (1982).
96. Jindal, S. *et al.* Postpartum breast involution reveals regression of secretory lobules mediated by tissue-remodeling. *Breast Cancer Res.* 16, 1–14 (2014).
97. Söderqvist, G. *et al.* Proliferation of breast epithelial cells in healthy women during the menstrual cycle. *Am. J. Obstet. Gynecol.* 176, 123–128 (1997).
98. Ramakrishnan, R., Khan, S. A. & Badve, S. Morphological changes in breast tissue with menstrual cycle. *Mod. Pathol.* 15, 1348–1356 (2002).
99. Vogel, P. M., Georgiade, N. G., Fetter, B. F., Vogel, F. S. & McCarty, K. S. The correlation of histologic changes in the human breast with the menstrual cycle. *Am. J. Pathol.* 104, 23–34 (1981).
100. Lyons, T. R. *et al.* Postpartum mammary gland involution drives progression of ductal carcinoma in situ through collagen and COX-2. *Nat. Med.* 17, 1109–1115 (2011).
101. O’Brien, J. *et al.* Alternatively activated macrophages and collagen remodeling characterize the postpartum involuting mammary gland across species. *Am. J. Pathol.* 176, 1241–1255 (2010).

102. Heaton, H. *et al.* Souporecell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat. Methods* 17, 615–620 (2020).
103. Bhat-Nakshatri, P., Gao, H., Sheng, L., Storniolo, A. M. V. & Nakshatri, H. A single-cell atlas of the healthy breast tissues reveals clinically relevant clusters of breast epithelial cells. *Cell Reports Medicine* 2, 100219 (2021).
104. Nguyen, Q. H. *et al.* Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. *Nat. Commun.* 9, 2028 (2018).
105. Battersby, S., Robertson, B. J., Anderson, T. J., King, R. J. & McPherson, K. Influence of menstrual cycle, parity and oral contraceptive use on steroid hormone receptors in normal breast. *Br. J. Cancer* 65, 601–607 (1992).
106. Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* 177, 1873-1887.e17 (2019).
107. O’Flanagan, C. H. *et al.* Dissociation of solid tumor tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses. *Genome Biol.* 20, 210–13 (2019).
108. Pardo, I. *et al.* Komen for the Cure Tissue Bank at the IU Simon Cancer Center. *Breast Cancer Res.* 16, 26 (2014).

109. Rajaram, R. D. *et al.* Progesterone and Wnt4 control mammary stem cells via myoepithelial crosstalk. *EMBO J.* 34, 641–652 (2015).
110. Aupperlee, M. D., Leipprandt, J. R., Bennett, J. M., Schwartz, R. C. & Haslam, S. Z. Amphiregulin mediates progesterone-induced mammary ductal development during puberty. *Breast Cancer Res.* 15, 44–15 (2013).
111. Hyder, S. M., Nawaz, Z., Chiappetta, C. & Stancel, G. M. Identification of functional estrogen response elements in the gene coding for the potent angiogenic factor vascular endothelial growth factor. *Cancer Res.* 60, 3183–3190 (2000).
112. LaMarca, H. L. & Rosen, J. M. Estrogen regulation of mammary gland development and breast cancer: amphiregulin takes center stage. *Breast Cancer Res.* 9, 304–3 (2007).
113. Ribieras, S., Tomasetto, C. & Rio, M. C. The pS2/TFF1 trefoil factor, from basic research to clinical applications. *Biochim. Biophys. Acta* 61–77 (1998) doi:10.1016/s0304-419x(98)00016-x.
114. Dabrosin, C. Variability of vascular endothelial growth factor in normal human breast tissue in vivo during the menstrual cycle. *J. Clin. Endocrinol. Metab.* 88, 2695–2698 (2003).
115. Ferguson, J. E., Schor, A. M., Howell, A. & Ferguson, M. W. Changes in the extracellular matrix of the normal human breast during the menstrual cycle. *Cell Tissue Res.* 268, 167–177 (1992).

116. Hallberg, G., Andersson, E., Naessén, T. & Ordeberg, G. E. The expression of syndecan-1, syndecan-4 and decorin in healthy human breast tissue during the menstrual cycle. *Reprod. Biol. Endocrinol.* 8, 35 (2010).
117. Schedin, P., O'Brien, J., Rudolph, M., Stein, T. & Borges, V. Microenvironment of the Involuting Mammary Gland Mediates Mammary Cancer Progression. *J. Mammary Gland Biol. Neoplasia* 12, 71–82 (2007).
118. Joshi, P. A. *et al.* RANK Signaling Amplifies WNT-Responsive Mammary Progenitors through R-SPONDIN1. *STEMCR* 5, 31–44 (2015).
119. Stein, T., Salomonis, N., Nuyten, D. S. A., Vijver, M. J. & Gusterson, B. A. A mouse mammary gland involution mRNA signature identifies biological pathways potentially associated with breast cancer metastasis. *J. Mammary Gland Biol. Neoplasia* 14, 99–116 (2009).
120. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Systems* 8, 329-337.e4 (2019).
121. Barkas, N. *et al.* Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nat. Methods* 16, 695–698 (2019).
122. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *bioRxiv* 060012 (2021) doi:10.1101/060012.

123. Miller, F. W. *et al.* Genome-wide association study of dermatomyositis reveals genetic overlap with other autoimmune disorders. *Arthritis Rheum.* 65, 3239–3247 (2013).
124. Ravelli, A. *et al.* Long-term outcome and prognostic factors of juvenile dermatomyositis: a multinational, multicenter study of 490 patients. *Arthritis Care Res.* 62, 63–72 (2010).
125. Mathiesen, P. *et al.* Long-term outcome in patients with juvenile dermatomyositis: a cross-sectional follow-up study. *Scand. J. Rheumatol.* 41, 50–58 (2012).
126. Tsaltskan, V. *et al.* Long-term outcomes in Juvenile Myositis patients. *Semin. Arthritis Rheum.* 50, 149–155 (2020).
127. Ruperto, N. *et al.* The Paediatric Rheumatology International Trials Organisation provisional criteria for the evaluation of response to therapy in juvenile dermatomyositis. *Arthritis Care Res.* 62, 1533–1541 (2010).
128. Okiyama, N. *et al.* Immune response to dermatomyositis-specific autoantigen, transcriptional intermediary factor 1 γ can result in experimental myositis. *Ann. Rheum. Dis.* 80, 1201–1208 (2021).
129. Seto, N. *et al.* Neutrophil dysregulation is pathogenic in idiopathic inflammatory myopathies. *JCI Insight* 5, (2020).
130. Rider, L. G. *et al.* The myositis autoantibody phenotypes of the juvenile idiopathic inflammatory myopathies. *Medicine* 92, 223–243 (2013).

131. Piper, C. J. M. *et al.* CD19⁺CD24^{hi}CD38^{hi} B Cells Are Expanded in Juvenile Dermatomyositis and Exhibit a Pro-Inflammatory Phenotype After Activation Through Toll-Like Receptor 7 and Interferon- α . *Front. Immunol.* 9, 1372 (2018).
132. Throm, A. A. *et al.* Dysregulated NK cell PLC γ 2 signaling and activity in juvenile dermatomyositis. *JCI Insight* 3, (2018).
133. Neely, J. *et al.* Multi-Modal Single-Cell Sequencing Identifies Cellular Immunophenotypes Associated With Juvenile Dermatomyositis Disease Activity. *Front. Immunol.* 13, (2022).
134. Gofshteyn, J. S. *et al.* Association of Juvenile Dermatomyositis Disease Activity With the Expansion of Blood Memory B and T Cell Subsets Lacking Follicular Markers. *Arthritis Rheumatol* 75, 1246–1261 (2023).
135. Morita, R. *et al.* Human blood CXCR5⁽⁺⁾CD4⁽⁺⁾ T cells are counterparts of T follicular cells and contain specific subsets that differentially support antibody secretion. *Immunity* 34, 108–121 (2011).
136. López de Padilla, C. M. *et al.* Plasmacytoid dendritic cells in inflamed muscle of patients with juvenile dermatomyositis. *Arthritis Rheum.* 56, 1658–1668 (2007).
137. Nistala, K. *et al.* Myeloid related protein induces muscle derived inflammatory mediators in juvenile dermatomyositis. *Arthritis Res. Ther.* 15, R131 (2013).
138. Turnier, J. L. *et al.* Imaging Mass Cytometry Reveals Predominant Innate Immune Signature and Endothelial-Immune Cell Interaction in Juvenile

- Myositis Compared to Lupus Skin. *Arthritis Rheumatol* 74, 2024–2031 (2022).
139. Patel, J., Maddukuri, S., Li, Y., Bax, C. & Werth, V. P. Highly Multiplexed Mass Cytometry Identifies the Immunophenotype in the Skin of Dermatomyositis. *J. Invest. Dermatol.* 141, 2151–2160 (2021).
 140. Hilliard, K. A. *et al.* Expansion of a novel population of NK cells with low ribosome expression in juvenile dermatomyositis. *Front. Immunol.* 13, 1007022 (2022).
 141. Chen, X., Lian, D. & Zeng, H. Single-cell profiling of peripheral blood and muscle cells reveals inflammatory features of juvenile dermatomyositis. *Front Cell Dev Biol* 11, 1166017 (2023).
 142. Clavarino, G. *et al.* Novel Strategy for Phenotypic Characterization of Human B Lymphocytes from Precursors to Effector Cells by Flow Cytometry. *PLoS One* 11, e0162209 (2016).
 143. Shimizu, Y., Meunier, L. & Hendershot, L. M. pERp1 is significantly up-regulated during plasma cell differentiation and contributes to the oxidative folding of immunoglobulin. *Proc. Natl. Acad. Sci. U. S. A.* 106, 17013–17018 (2009).
 144. Richert-Spuhler, L. E. *et al.* CD101 genetic variants modify regulatory and conventional T cell phenotypes and functions. *Cell Rep Med* 2, 100322 (2021).
 145. Rao, D. A. *et al.* Pathologically expanded peripheral T helper cell subset drives B cells in rheumatoid arthritis. *Nature* 542, 110–114 (2017).

146. Rao, D. A. The rise of peripheral T helper cells in autoimmune disease. *Nature reviews. Rheumatology* vol. 15 453–454 (2019).
147. Lerkvaleekul, B. *et al.* Siglec-1 expression on monocytes is associated with the interferon signature in juvenile dermatomyositis and can predict treatment response. *Rheumatology* 61, 2144–2155 (2022).
148. Kannan, K. *et al.* Lysosome-associated membrane proteins h-LAMP1 (CD107a) and h-LAMP2 (CD107b) are activation-dependent cell surface glycoproteins in human peripheral blood mononuclear cells which mediate cell adhesion to vascular endothelium. *Cell. Immunol.* 171, 10–19 (1996).
149. Kim, H. Updates on interferon in juvenile dermatomyositis: pathogenesis and therapy. *Curr. Opin. Rheumatol.* 33, 371–377 (2021).
150. Neely, J. *et al.* Gene Expression Meta-Analysis Reveals Concordance in Gene Activation, Pathway, and Cell-Type Enrichment in Dermatomyositis Target Tissues. *ACR Open Rheumatol* 1, 657–666 (2019).
151. 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).
152. Mootha, V. K. *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* 34, 267–273 (2003).

153. Kurata, I. *et al.* Potential involvement of OX40 in the regulation of autoantibody sialylation in arthritis. *Ann. Rheum. Dis.* 78, 1488–1496 (2019).
154. Adam, L. *et al.* Follicular T Helper Cell Signatures in Primary Biliary Cholangitis and Primary Sclerosing Cholangitis. *Hepatol Commun* 2, 1051–1063 (2018).
155. Sun, J. *et al.* Aberrant GITR expression on different T cell subsets and the regulation by glucocorticoid in systemic lupus erythematosus. *Int. J. Rheum. Dis.* 19, 199–204 (2016).
156. Tian, J., Zhang, B., Rui, K. & Wang, S. The Role of GITR/GITRL Interaction in Autoimmune Diseases. *Front. Immunol.* 11, 588682 (2020).
157. Fu, N., Xie, F., Sun, Z. & Wang, Q. The OX40/OX40L Axis Regulates T Follicular Helper Cell Differentiation: Implications for Autoimmune Diseases. *Front. Immunol.* 12, 670637 (2021).
158. Li, Q. *et al.* Two major genes associated with autoimmune arthritis, *Ncf1* and *Fcgr2b*, additively protect mice by strengthening T cell tolerance. *Cell. Mol. Life Sci.* 79, 482 (2022).
159. Good, C. R. *et al.* An NK-like CAR T cell transition in CAR T cell dysfunction. *Cell* 184, 6081-6100.e26 (2021).
160. Luo, Y. *et al.* Single-cell transcriptomic analysis reveals disparate effector differentiation pathways in human Treg compartment. *Nat. Commun.* 12, 3913 (2021).

161. Sather, B. D. *et al.* Altering the distribution of Foxp3(+) regulatory T cells results in tissue-specific inflammatory disease. *J. Exp. Med.* 204, 1335–1347 (2007).
162. Pizzolato, G. *et al.* Single-cell RNA sequencing unveils the shared and the distinct cytotoxic hallmarks of human TCRV δ 1 and TCRV δ 2 $\gamma\delta$ T lymphocytes. *Proc. Natl. Acad. Sci. U. S. A.* 116, 11906–11915 (2019).
163. Lazarevic, V., Glimcher, L. H. & Lord, G. M. T-bet: a bridge between innate and adaptive immunity. *Nat. Rev. Immunol.* 13, 777–789 (2013).
164. Bendersky, A. *et al.* Cellular interactions of synovial fluid $\gamma\delta$ T cells in juvenile idiopathic arthritis. *J. Immunol.* 188, 4349–4359 (2012).
165. Chen, B. *et al.* A Cuproptosis Activation Scoring model predicts neoplasm-immunity interactions and personalized treatments in glioma. *Comput. Biol. Med.* 148, 105924 (2022).
166. Baechler, E. C. *et al.* An interferon signature in the peripheral blood of dermatomyositis patients is associated with disease activity. *Mol. Med.* 13, 59–68 (2007).
167. Wong, D. *et al.* Interferon and biologic signatures in dermatomyositis skin: specificity and heterogeneity across diseases. *PLoS One* 7, e29161 (2012).
168. Greenberg, S. A. *et al.* Interferon-alpha/beta-mediated innate immune mechanisms in dermatomyositis. *Ann. Neurol.* 57, 664–678 (2005).

169. 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).
170. Preuße, C. *et al.* Skeletal muscle provides the immunological micro-milieu for specific plasma cells in anti-synthetase syndrome-associated myositis. *Acta Neuropathol.* 144, 353–372 (2022).
171. Dzangué-Tchoupou, G., Allenbach, Y., Preuße, C., Stenzel, W. & Benveniste, O. Mass cytometry reveals an impairment of B cell homeostasis in anti-synthetase syndrome. *J. Neuroimmunol.* 332, 212–215 (2019).
172. Vercoulen, Y. *et al.* Increased presence of FOXP3+ regulatory T cells in inflamed muscle of patients with active juvenile dermatomyositis compared to peripheral blood. *PLoS One* 9, e105353 (2014).
173. Koch, M. A. *et al.* The transcription factor T-bet controls regulatory T cell homeostasis and function during type 1 inflammation. *Nat. Immunol.* 10, 595–602 (2009).
174. Chaudhry, A. *et al.* CD4+ regulatory T cells control TH17 responses in a Stat3-dependent manner. *Science* 326, 986–991 (2009).
175. Zheng, Y. *et al.* Regulatory T-cell suppressor program co-opts transcription factor IRF4 to control T(H)2 responses. *Nature* 458, 351–356 (2009).

176. Terry, L. V. & Oo, Y. H. The Next Frontier of Regulatory T Cells: Promising Immunotherapy for Autoimmune Diseases and Organ Transplantations. *Front. Immunol.* 11, 565518 (2020).
177. Beheshti, S. A., Shamsasenjan, K., Ahmadi, M. & Abbasi, B. CAR Treg: A new approach in the treatment of autoimmune diseases. *Int. Immunopharmacol.* 102, 108409 (2022).
178. Biron, C. A., Sonnenfeld, G. & Welsh, R. M. Interferon induces natural killer cell blastogenesis in vivo. *J. Leukoc. Biol.* 35, 31–37 (1984).
179. Nguyen, K. B. *et al.* Coordinated and distinct roles for IFN- α beta, IL-12, and IL-15 regulation of NK cell responses to viral infection. *J. Immunol.* 169, 4279–4287 (2002).
180. Martinez, J., Huang, X. & Yang, Y. Direct action of type I IFN on NK cells is required for their activation in response to vaccinia viral infection in vivo. *J. Immunol.* 180, 1592–1597 (2008).
181. Moneta, G. M. *et al.* Muscle Expression of Type I and Type II Interferons Is Increased in Juvenile Dermatomyositis and Related to Clinical and Histologic Features. *Arthritis Rheumatol* 71, 1011–1021 (2019).
182. Lundberg, I. E. *et al.* 2017 European League Against Rheumatism/American College of Rheumatology classification criteria for adult and juvenile idiopathic inflammatory myopathies and their major subgroups. *Ann. Rheum. Dis.* 76, 1955–1964 (2017).
183. Lazarevic, D. *et al.* The PRINTO criteria for clinically inactive disease in juvenile dermatomyositis. *Ann. Rheum. Dis.* 72, 686–693 (2013).

184. Ahmed, S., Chen, K. L. & Werth, V. P. The validity and utility of the Cutaneous Disease Area and Severity Index (CDASI) as a clinical outcome instrument in dermatomyositis: A comprehensive review. *Semin. Arthritis Rheum.* 50, 458–462 (2020).
185. Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* 9, (2020).
186. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* 16, 1289–1296 (2019).
187. Mulè, M. P., Martins, A. J. & Tsang, J. S. Normalizing and denoising protein expression data from droplet-based single cell profiling. *Nat. Commun.* 13, 2099 (2022).
188. Merah-Mourah, F., Cohen, S. O., Charron, D., Mooney, N. & Haziot, A. Identification of Novel Human Monocyte Subsets and Evidence for Phenotypic Groups Defined by Interindividual Variations of Expression of Adhesion Molecules. *Sci. Rep.* 10, 4397 (2020).
189. 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).
190. Bunis, D. G., Andrews, J., Fragiadakis, G. K., Burt, T. D. & Sirota, M. dittoSeq: universal user-friendly single-cell and bulk RNA sequencing visualization toolkit. *Bioinformatics* 36, 5535–5536 (2021).

191. Sakai, R., Winand, R., Verbeiren, T., Moere, A. V. & Aerts, J. dendsort: modular leaf ordering methods for dendrogram representations in R. *F1000Res.* 3, 177 (2014).
192. ggplot2. <https://ggplot2.tidyverse.org/>.
193. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287 (2012).
194. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* 14, 1083–1086 (2017).

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:

Gabrielle Kabadam

118B0813BA70431...

Author Signature

5/28/2024

Date