Title

Improving single-cell genomics scalability and data interpretability for applications in single-cell chemical transcriptomics

Author

McGinnis, Christopher

Publication Date

2021

Peer reviewed|Thesis/dissertation

Improving single-cell genomics scalability and data interpretability for applications in single-cell chemical transcriptomics

by
Christopher Mcginnis

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Biochemistry and Molecular Biology

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Michael J Keiser
4DF1BD06D670465...

Michael J Keiser
Chair

DocuSigned by:

Zev Gartner

DocuSigned by:
Max Krummel
630992D3070C445...

Max Krummel

Committee Members

Acknowledgments

Over the course of my undergraduate studies at Wesleyan University, my time at the Institute for Systems Biology, and my doctoral work at UCSF, I have been extremely fortunate to be surrounded by great mentors, collaborators, friends and family, without whom I would have never completed this dissertation.

First and foremost, I would like to thank my thesis advisor Dr. Zev Gartner for his guidance through many diverse research projects over the last four years. During my time in his lab, I have been impressed by Zev's ability to strike the balance between being available but not stifling, to be open to "big-picture" thinking while recognizing the need for tangible results, and to be professional and ambitious while leaving space for jokes and fun. The training and mentorship I have received from Zev are the best I've had over my short scientific career, and I am excited to continue building on the foundation he has helped me lay during my time at UCSF.

I also appreciate Zev for fostering a lab environment that is highly collaborative, positive, and full of great colleagues. The entire Gartner Lab has been pivotal for my success during graduate school though constructive dialogues had during lab and mini-meetings, exam preparations, and (usually frantic) periods of grant writing. However, I specifically want to thank Dr. Dave Patterson and Dr. Lyndsay Murrow for mentoring me when I first joined the lab – the DoubletFinder and MULTI-seq projects would not have been nearly as efficient or fun without you two. I would also like to thank Danny Conrad and Hikaru Miyazaki for being great lab-mates and collaborators on miscellaneous sample multiplexing projects over the years.

I am also grateful for the broader UCSF academic community. I specifically want to thank Dr. Eric Chow for his unparalleled knowledge of next-generation sequencing, being willing to share server space and answer my steady stream of my questions, and for being a generally awesome human. I also want to thank Dr. Rachel Zwick and Dr. Juliane Winkler for helping with

my numerous fellowship applications and being amazing collaborators who helped plan and execute the most epic MULTI-seq experiments I have been involved with to date. I am also grateful to Dr. Sulggi Lee and Dr. Nadia Roan for helping to organize the UCSF single-cell interest group meetings, which has served as a great opportunity to build my academic network and hear about great science.

There are also many others outside of UCSF who have been integral to my scientific journey so far. First, I am grateful to Dr. Bob Lane at Wesleyan University for his mentorship during my undergraduate studies and for facilitating my move to Dr. Lee Hood's lab at the Institute for Systems Biology after graduation – I don't think any of this would have happened without your early vote of confidence. At the Institute for Systems Biology, I appreciate Dr. Sui Huang and Dr. Kalli Trachana for being great mentors and for hosting the weekly 'Complexity Salons' which formed the basis for how I think about biology to this day. And at Caltech, I would like to thank Dr. Matt Thomson, Dr. Sisi Chen, Dr. Tami Khazei, Dr. Jeff Park, and others in the Thomson Lab for being fantastic mentors, collaborators, and friends as we pushed the limits of single-cell analysis.

Finally, I am beyond grateful for my friends and family who were incredible supportive during my graduate studies and were patient as I rambled about doublets, droplets, LMOs, and macrophages over the years. Specifically, I am grateful to my siblings(-in-law), Ryan, Ellen, Annie, and Maz, as well as my parents, Barb and Michael McGinnis, for their unwavering love and support. I am especially grateful for the career advice and commiseration from fellow-academics Ryan and Ellen, and for the late-night immunology questions from my nephew, Sully. I would also like to thank my East Bay friends including Isaac, Will, Miles Mac, Miles, Davey, Zia, Tennessee, Tobias, Noah, and Tory for the dinners, dancing, music, and games and for helping me make the West Coast feel like home. I am also deeply grateful to my partner Jimmy, who has helped me grow immensely as a person and has been an amazing source of love, laughter, and homemade sourdough over the last two years.

Contributions

Chapters 2-5 contain work previously published in peer reviewed journals:

McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. Cell Systems. 2019; 8(4): 329-37.e4.

McGinnis CS, Patterson DM, Winkler J, Conrad DN, Hein MY, Srivastavan V, Hu JL, Murrow LM, Weissman JS, Werb Z, Chow ED, Gartner ZJ. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nature Methods. 2019; 16: 619-26.

McGinnis CS, Siegel DA, Xie G, Hartoularos D, Stone M, Ye CJ, Gartner ZJ, Roan NR, Lee SA. No detectable alloreactive transcriptional responses under standard sample preparation conditions during donor-multiplexed single-cell RNA sequencing of peripheral blood mononuclear cells. BMC Biology. 2021; 19(1): 10.

Thibodeau A, Eroglu A, **McGinnis CS**, Lawlor N, Nehar-Belaid D, Kursawe R, Marches R, Conrad DN, Kuchel GA, Gartner ZJ, Banchereau J, Stitzel ML, Cicek AE, Ucar D. AMULET: a novel read count-based method for effective multiplet detection from single nucleus ATAC-seq data. Genome Biology. 2021; *in press*.

Improving single-cell genomics scalability and data interpretability for applications in single-cell chemical transcriptomics

Christopher Swart McGinnis

Abstract

Cellular biology has traditionally relied upon simple, low-dimensional single-cell measurements (e.g., microscopy and flow cytometry) which fail to adequately address cellular complexity, or high-dimensional aggregative measurements (e.g., bulk RNA-sequencing) which obscure cellular heterogeneity. Single-cell genomics technologies strike the ideal balance between measurement complexity and resolution, but have historically been hampered by two technological limitations: cell-cell doublets which confound data interpretation, and scalability limitations due to high reagent costs and complex parallel sample preparation workflows.

In this dissertation, I present solutions to both of these issues. First, I describe DoubletFinder, a machine learning approach for finding cell-cell doublets in scRNA-seq data by identifying real cells with heightened similarity to *in silico*-generated artificial doublets. Second, I describe MULTI-seq, a method enabling pooled single-cell genomics sample processing by labeling plasma membranes with sample-specific DNA barcodes prior to cellular isolation. After describing these technologies, I demonstrate three MULTI-seq applications. First, I explore the effects of sample pooling on single-cell RNA-sequencing (scRNA-seq) data-quality. Second, I extend MULTI-seq to single-cell epigenomics assays. And third, I leverage MULTI-seq to perform the largest-ever single-cell screen-by-sequencing experiment on PBMCs. Collectively, this dissertation documents molecular and computational tools for improving single-cell genomics scalability and data interpretability, and illustrates how these improvements expand the boundaries of feasibility for single-cell genomics experiments.

Table of Contents

List of Figures

Chapter 2:

Chapter 3:

Chapter 4:

Chapter 5:

Chapter 6:

List of Tables

Chapter 3:

Chapter 4:

Chapter 6:

List of Abbreviations

ATAC-seq: Assay for transposase-accessible Chromatin using sequencing

AUC: Area under the curve

BC: Bimodality coefficient

BC_{MVN}: Mean-variance normalized bimodality coefficient

BSA: Bovine serum albumin

CM: Classical monocyte

CS: Classification stability

CMO: Cholesterol-modified oligonucleotide

DC: Dendritic cell

DEG: Differentially-expressed gene

DMSO: Dimethyl sulfoxide

FACS: Fluorescence activated cell sorting

gDNA: Genomic DNA

GSEA: Gene set enrichment analysis

HDAC: Histone deacetylase

HEK: Human embryonic kidney cell

HMEC: Human mammary epithelial cell

JSD: Jensen-Shannon divergence

K_d: Dissociation constant

LMO: Lipid-modified oligonucleotide

LSI: Latent semantic indexing

MEF: Mouse embryonic fibroblasts

MPEC: Memory precursor cell

MULTI-seq: Single-cell RNA sequencing sample multiplexing using lipid tagged indices

MULTI-ATAC-seq: Single-cell assay for transposase-accessible chromatin using sequencing sample multiplexing using lipid tagged indices

MULTI-C&T: Single-cell cleavage under targets and tagmentation sample multiplexing using lipid tagged indices

NK: Natural killer cell

NSAID: Non-steroidal anti-inflammatory drug

pANN: Proportion of artificial nearest neighbors

PBMC: Peripheral blood mononuclear cell

PCA: Principal component analysis

PCR: Polymerase chain reaction

PDF: Probability density function

PDX: Patient-derived xenograft

PMA: Phorbol 12-myristate 13-acetate

ROC: Receiver operator curve

ROGUE: Ratio of Global Unshifted Entropy

RPC: Next-generation sequencing reads per cell

RT: Reverse transcription

scATAC-seq: Single-cell assay for transposase-accessible chromatin using sequencing

scC&T: Single-cell cleavage under targets and tagmentation

SCMK: Single-cell multiplexing kit (BD Biosciences)

scRNA-seq: Single-cell RNA sequencing

SIK: Salt-inducible kinase

SLEC: Short-lived effector cell

SNP: Single nucleotide polymorphism

TCM: Central memory T-cell

TCR: T-cell receptor

TEM: Effector memory T-cell

TKI: Tyrosine kinase inhibitor

Treg: Regulatory T-cell

t-SNE: T-distributed stochastic neighbor embedding

UMAP: Uniform Manifold Approximation and Projection

UMI: unique molecular identifier

Chapter 1: Introduction

1.1 The Promises and Pitfalls of Single-Cell Genomics

Cells are the irreducible components of biological systems, and cellular heterogeneity and communication define the line between order and chaos — between life and death. The simplest unicellular organisms leverage cell-cell communication to form diverse collectives which are robust to changing environments. In multicellular organisms, cellular communities develop into networks of tissues and organs comprised of functionally-distinct "cell types" which satisfy diverse physiological demands. A core tenet of the field of Cell Biology is to systematically characterize the structure, function, and behavior of these cell types; a lofty goal which has been recently accelerated by the field of single-cell genomics.

Traditionally, cells have been studied using simple, low-dimensional measurements. For example, qualitative cellular features such as shape, size, and color are readily identified using microscopy, as are the quantitative abundances of small sets of biomolecules using methods such as flow cytometry. However, low-dimensional measurements necessarily fail to encompass cellular complexity, as each individual human cell can contain tens of thousands of different genes, hundreds of thousands of mRNA transcripts, and many billion protein molecules. Next-generation sequencing technologies identify and quantify these biomolecules in an unbiased manner, and thereby provide the requisite dimensionality to understand what makes a cell assume a particular identity. However, until recently, such approaches have been limited to profiling only large pools of cells, which in turn obscures the cellular complexity these methods are employed to study.

In 2011, the first method for generating high-dimensional measurements of individual cells – single-cell RNA-sequencing (scRNA-seq) – was described [1]. Despite only analyzing 85 total cells, this achievement sparked an avalanche of technological advances which have vastly

increased the scalability and applicability of single-cell analysis. For example, droplet microfluidics [2-4] and combinatorial indexing technologies [5,6] have increased the numbers of cells that can be analyzed in a single experiment by multiple orders of magnitude. Moreover, methods for analyzing diverse levels of biological information including the genome [7], epigenome [8-14], and proteome [15-17] have been described, providing the blueprint for reducing cellular identity to its component parts. As a result, researchers have now generated cell type "atlases" of entire organs and organisms [18,19]; uncovered variability within ostensibly homogenous cell types which initiate humoral immune responses [20]; and pinpointed the cellular origins of human diseases such as cystic fibrosis [21].

While such descriptive analyses of biological systems remain of paramount importance, technical and analytical limitations have resulted in questions of mechanism and causality being historically left untouched by the single-cell genomics community. For instance, while it is undeniably important to categorize the component building blocks of the human body, understanding how these building blocks differ between individuals (both healthy and diseased), vary in space (e.g., regions of the gastrointestinal tract responsible for differential nutrient absorption), shift over time (e.g., during development and aging), and respond to perturbation (e.g., pharmaceuticals or environmental toxins) represents a scale of single-cell genomics experiments that surpasses the capacity of existing technologies. Moreover, the process of identifying these cellular building blocks is confounded by technical artifacts (e.g., cell-cell doublets) introduced by single-cell genomics technologies, themselves. Developing molecular and analytical tools which circumvent these limitations in single-cell genomics scalability and interpretability is the focus of this thesis.

1.2 Distinguishing Single Cells from "Single" Cells with DoubletFinder

The one shared requirement of all single-cell genomics workflows is a mechanism for isolating individual cells. The earliest single-cell genomics workflows used fluorescence-activated cell sorting (FACS) to isolate individual cells in microwells, to which molecular biology reagents were added for analyte capture, barcoding, and amplification. Since cells were sorted one-by-one, these approaches were limited to relatively low numbers of cells (e.g., 10^1-10^3) per experiment, precluding their applicability to many biological questions. Many subsequent cell isolation methods [2-6,22-24] raised the upper-limit on cell-throughput to 10^4-10^6 cells per experiment by capturing cells in a pooled format. For example, in droplet microfluidics-based scRNA-seq, single-cell isolation is achieved by Poisson loading cells into emulsion oil droplets. Because the number of droplets greatly exceeds of the number of loaded cells, the chance that any two cells are co-encapsulated in a single droplet is relatively low.

However, due to the inherently random nature of pooled single-cell isolation workflows, the occurrence of cell-cell co-encapsulation is unavoidable and generates technical artifacts known as doublets. Doublets confound single-cell genomics analysis by producing multi-cell measurements which masquerade as individual cells, making analytical tasks such as cell type identification challenging. For example, it can be nearly impossible to discern doublets from *bona fide* cell states co-expressing known cell type markers (e.g., developmental intermediates). In Chapter 2 of this thesis, I describe a machine learning-based solution to this problem called DoubletFinder [25]. DoubletFinder "finds" doublets in scRNA-seq data by identifying cells which co-localize with *in silico*-generated artificial doublets in gene expression space. As a result, DoubletFinder improves the interpretability of single-cell genomics datasets.

1.3 Transitioning from Descriptive to Mechanistic Single-Cell Genomics

High-throughput single-cell isolation methods have increased single-cell genomics cell-throughput by 4-5 orders of magnitude relative to FACS-based approaches. Despite this achievement, sample-throughput (i.e., the number of technical or biological replicates, time-points, perturbations, etc. assayed in a single experiment) is severely limited in standard single-cell genomics workflows. This limitation is rooted in details of next-generation library preparation workflows, wherein oligonucleotide indices recording each cell's sample-of-origin are introduced at the end of the procedure. As a result, traditional single-cell genomics experiments require individual samples to be processed in parallel, which introduces undesirable technical variability, limits scalability, and translates to prohibitively-high reagent costs for even modestly-large experimental designs. One potential solution to this problem is to tag cells with sample-specific DNA barcodes prior to single-cell isolation. Such a 'sample multiplexing' approach would enable any number of experimental samples to be processed in a single pool, thus circumventing existing limitations on single-cell genomics sample-throughput.

In Chapter 3 of this thesis, I describe a scRNA-seq sample multiplexing method called MULTI-seq [26], which uses lipid-modified oligonucleotides to introduce DNA barcodes to plasma membranes. MULTI-seq increases scRNA-seq sample throughput by simplifying sample preparation workflows and lowering costs, while additionally improving data-quality through doublet identification and batch effect minimization. In turn, MULTI-seq fundamentally expands the purview of single-cell genomics experiments beyond descriptive analyses of biological systems. For instance, when multiplexing samples is not prohibitively expensive, questions of mechanism and causality which require perturbations, diverse sample sources, and experimental replicates can be adequately interrogated. Moreover, technical concerns related to single-cell genomics sample processing (as discussed in Chapter 4) which have traditionally been de-

prioritized due to high experiment costs can be economically addressed. Building on this work, in Chapter 5 of this thesis I also describe an extension of MULTI-seq called MULTI-ATAC-seq, which is a method for multiplexing samples during single-cell epigenomics experiments.

1.4 Single-Cell 'Screen-by-Sequencing'

Integrating computational doublet detection and sample multiplexing technologies into single-cell genomics workflows makes it possible to generate readily-interpretable datasets at unprecedented scale. To date, these technologies have been applied successfully across many experimental parameters including time [26], space [27], individuals [28], and perturbations [29-32]. Amongst all of these parameters, the use of sample multiplexing approaches for incorporating scRNA-seq into high-throughput chemical perturbation screens has particularly gained momentum. For instance, in one pioneering study [29], large numbers of samples consisting of drug-perturbed cancer cell lines undergoing an epithelial-to-mesenchymal transition were analyzed using MULTI-seq. In contrast to traditional high-throughput screening workflows which are limited to simple read-outs such as cell growth or viability [33], this study demonstrates how single-cell genomics sample multiplexing enables screening against more nuanced cellular phenotypes. In another recent study, Zhao and colleagues used patient-derived glioblastoma slice cultures in a small scRNA-seq-coupled chemical screen, highlighting how the resolution of scRNA-seq facilitates the use of *in vitro* systems comprised of diverse interacting cell types which better recapitulate *in vivo* disease biology. This represents an improvement over existing bulk 'screen-by-sequencing' platforms [34-36] which obscure cellular heterogeneity and, as a consequence, are predominantly limited to simple *in vitro* cell line systems.

In Chapter 6 of this thesis, I build on this growing body of work by using MULTI-seq to perform the largest per-sample single-cell 'screen-by-sequencing' experiment to date.

Specifically, I describe the analysis of ~1,000,000 single-cell transcriptomes of resting or CD3/CD28-stimulated peripheral blood mononuclear cells responding to ~800 unique perturbations including immunomodulatory and FDA-approved drugs. These analyses reveal the primary modes of immunomodulation in T lymphocytes and myeloid cells, and highlight key insights that can be gained using single-cell screen-by-sequencing. For example, I discuss how scRNA-seq can capture the 'polypharmacology' of off-target drug activities on cellular phenotypes, enabling more accurate drug classifications beyond what is possible from known primary molecular targets. Moreover, I highlight a subset of macrophage-depleting drugs which achieve this phenotype by enhancing T-cell-mediated macrophage apoptosis to demonstrate the importance of performing high-throughput screens on complex *in vitro* systems. As a result, these analyses outline the potential for single-cell 'screen-by-sequencing' to revolutionize drug development.

1.5 References

1. Islam S, Kjällquist U, Moliner A, Zajac P, Fan J, Lönnerberg P, Linnarsson S. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Research. 2011; 21: 1160-7.
2. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, et al. Massively parallel digital transcriptional profiling of single cells. Nature Communications. 2017; 8: 14049.
3. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015; 161(5): 1202-14.
4. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015; 161(5): 1187-1201.
5. Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. Science. 2017; 357(6352): 661-7.
6. Rosenberg AB, Roco CM, Muscat RA, Kuchina A, Sample P, Yao Z, et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science. 2018; 360(6385): 176-82.
7. Leung ML, Davis A, Gao R, Casasent A, Wang Y, Sei E, et al. Single-cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer. Genome Research. 2017; 27(8): 1287-99.
8. Lareau CA, Duarte FM, Chew JG, Kartha VK, Burkett ZD, Kohlway AS, et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. Nature Biotechnology. 2019; 37(8): 916-24.

9. Satpathy AT, Granja JM, Yost KE, Qi Y, Meschi FM, McDermott GP, et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nature Biotechnology. 2019; 37(8): 925-36.
10. Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. Science. 2015; 348(6237): 910-4.
11. Wu SJ, Furlan SN, Mihalas AB, Kaya-Okud HS, Feroze AH, Emerson SN, et al. Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression. Nature Biotechnology. 2021. doi: 10.1038/s41587-021-00865-z.
12. Bartosovic M, Kabbe M, Castelo-Branco G. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. Nature Biotechnology. 2021. doi: 10.1038/s41587-021-00869-9.
13. Grosselin K, Durand A, Marsolier J, Poitou A, Marangoni E, Nemati F, et al. High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. Nature Genetics. 2019; 51(6): 1060-6.
14. Ramani V, Deng X, Qiu R, Gunderson KL, Steemers FJ, Disteche C, et al. Massively multiplex single-cell Hi-C. Nature Methods. 2017; 14(3): 263-6.
15. Bendall SC, Simond E, Qiu P, Amir ED, Krutzik PO, Finck R, et al. Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. Science. 2011; 332(6030): 687-96.
16. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P. Simultaneous epitope and transcriptome measurement in single cells. Nature Methods. 2017; 14(9): 865-8.
17. Peterson VM, Zhang KX, Kumar N, Wong J, Li L, Wilson DC, et al. Multiplexed quantification of proteins and transcripts in single cells. Nature Biotechnology. 2017; 35(10): 936-9.

18. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The Human Cell Atlas. Elife. 2017; 6:e27041.
19. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Schaum N, Karkanias J, Neff NF, May AP, Quake SR, Wyss-Coray T, et al. Nature. 2018; 562(7727): 367-72.
20. Shalek AK, Satija R, Shuga J, Trombetta JJ, Gennert D, Lu D, et al. Single cell RNA Seq reveals dynamic paracrine control of cellular variation. Nature. 2014; 510(7505): 363-9.
21. Plasschaert LW, Žilionis R, Choo-Wing R, Savova V, Kneh Jr, Roma G, et al. A single cell atlas of the tracheal epithelium reveals the CFTR-rich pulmonary ionocyte. Nature. 2018; 560(7718): 37-81.
22. Gierahn TM, Wadsworth MH 2nd, Hughes TK, Bryson BD, Butler A, Satija R, et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Nature Methods. 2017; 14(4): 395-8. 8.
23. Hughes TK, Wadsworth MH 2nd, Gierahn TM, Do T, Weiss D, Andrade PR, et al. Second-Strand Synthesis-Based Massively Parallel scRNA-Seq Reveals Cellular States and Molecular Features of Human Inflammatory Skin Pathologies. Immunity. 2020; 53(4): 878-94.
24. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science. 2014; 343(6172): 776-9.
25. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. Cell Systems. 2019; 8(4): 329-37.e4.
26. McGinnis CS, Patterson DM, Winkler J, Conrad DN, Hein MY, Srivastavan V, et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nature Methods. 2019; 16: 619-26.

27. Hu KH, Eichorst JP, McGinnis CS, Patterson DM, Chow ED, Kersten K, et al. ZipSeq: barcoding for real-time mapping of single cell transcriptomes. Nature Methods. 2020; 17(8): 833-43.
28. van der Wijst MGP, de Vries DH, Groot HE, Trynka G, Hon CC, Bonder MJ, et al. Single-cell eQTLGen Consortium: a personalized understanding of disease. Elife. 2020; 9: e52155.
29. Zhao W, Dovas A, Spinazzi EF, Levitin HM, Banu MA, Upadhyayula P, et al. Deconvolution of cell type-specific drug responses in human tumor tissue with single-cell RNA-seq. Genome Medicine. 2021; 13(1): 82.
30. Cook DP, Vanderhyden BC. Context specificity of the EMT transcriptional response. Nature Communications. 2020; 11(1): 2142.
31. Srivatsan SR, McFaline-Figueroa JL, Ramani V, Saunders L, Cao J, et al. Massively multiplex chemical transcriptomics at single-cell resolution. Science. 2020; 367(6473): 45-51.
32. McFarland JM, Paolella BR, Warren A, Geiger-Schuller K, Shibue T, Rothberg M, et al. Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. Nature Communications. 2020; 11(1): 4296.
33. Nair NU, Greninger P, Friedman A, Amzallag A, Cortez E, Sahu AD, et al. A landscape of synergistic drug combinations in non-small-cell lung cancer. bioRxiv. 2021. doi: 10.1101/2021.06.03.447011.
34. Lamb J, Crawford ED, Peck DD, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006; 313(5795): 1929-35.
35. Chen B, Ma L, Paik H, Sirota M, Wei W, Chua M, So S, Butte AJ. Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. Nature Communications. 2017; 8: 16022.

36. Ye C, Ho DJ, Neri M, Yang C, Kulkarni T, Randhawa R, et al. DRUG-seq for miniaturized high-throughput transcriptome profiling in drug discovery. Nature Communcations. 2018; 9(1): 4307.

Chapter 2: scRNA-seq doublet prediction using DoubletFinder

Elements of the following chapter are reprinted from the manuscript “DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors” by Christopher S. McGinnis, Lyndsay M. Murrow, and Zev J. Gartner, published in *Cell Systems* on April 24, 2019, Vol. 8, Issue 4, pages 329-337.

2.1 Abstract

Single-cell RNA sequencing (scRNA-seq) data are commonly affected by technical artifacts known as “doublets,” which limit cell throughput and lead to spurious biological conclusions. Here, we present a computational doublet detection tool—DoubletFinder—that identifies doublets using only gene expression data. DoubletFinder predicts doublets according to each real cell’s proximity in gene expression space to artificial doublets created by averaging the transcriptional profile of randomly chosen cell pairs. We first use scRNA-seq datasets where the identity of doublets is known to show that DoubletFinder identifies doublets formed from transcriptionally distinct cells. When these doublets are removed, the identification of differentially expressed genes is enhanced. Second, we provide a method for estimating DoubletFinder input parameters, allowing its application across scRNA-seq datasets with diverse distributions of cell types. Lastly, we present “best practices” for DoubletFinder applications and illustrate that DoubletFinder is insensitive to an experimentally validated kidney cell type with “hybrid” expression features.

2.2 Introduction

scRNA-seq has evolved into a powerful and scalable assay through the development of combinatorial cell indexing techniques [1] and cellular isolation strategies that utilize nanowells

[2] and droplet microfluidics [3-5]. In droplet microfluidics and nanowell-based scRNA-seq modalities, Poisson loading is used to co-encapsulate individual cells and mRNA capture beads in emulsion oil droplets where the cells are lysed, mRNA is captured on the bead, and transcripts are barcoded by reverse transcription. Since cells are randomly apportioned into droplets, the frequency at which droplets are filled with two cells—forming technical artifacts known as "doublets"—varies according to the input cell concentration with a frequency that follows Poisson statistics [6]. Doublets are known to confound scRNA-seq data analysis [7-8], and it is common practice to mitigate these effects by sequencing far fewer cells than is theoretically possible in order to minimize doublet formation rates. For this reason, doublet formation fundamentally limits scRNA-seq cell throughput.

Recently developed sample multiplexing approaches can overcome this limitation in some circumstances. For example, genomic [9-14] and cellular sample multiplexing techniques [15-18] directly detect most doublets in scRNA-seq data by identifying cells associated with orthogonal sample barcodes or single nucleotide polymorphisms (SNPs). By identifying and removing doublets, these techniques minimize technical artifacts while enabling users to "super-load" droplet microfluidics devices for increased scRNA-seq cell throughput. However, sample multiplexing techniques have limitations in the context of doublet detection. For instance, doublets formed from cells associated with identical sample indices or SNPs cannot be detected. Moreover, sample multiplexing cannot be applied retroactively to existing scRNA-seq datasets.

To address these limitations, we developed DoubletFinder: a computational doublet detection tool that relies solely on gene expression data. DoubletFinder begins by simulating artificial doublets and incorporating these "cells" into existing scRNA-seq datasets that have been processed using the popular "Seurat" analysis pipeline [19,20]. DoubletFinder then distinguishes real doublets from singlets by identifying real cells with high proportions of artificial neighbors in gene expression space. In this study, we describe the development and validation of

DoubletFinder in three parts. In the first part, we benchmark DoubletFinder against “ground-truth” scRNA-seq datasets where doublets are empirically defined by the sample multiplexing approaches Demuxlet [9] and Cell Hashing [15]. These comparisons reveal that DoubletFinder detects ground-truth false negatives and improves downstream differential gene expression analyses. Moreover, ground-truth comparisons illustrate that DoubletFinder predominantly detects doublets derived from transcriptionally distinct cells—referred to here as “heterotypic” doublets—and is less sensitive to “homotypic” doublets formed from transcriptionally similar cells. In the second part, we leverage scRNA-seq data simulations to demonstrate that DoubletFinder input parameters must be tailored to data with different numbers of cell types and magnitudes of transcriptional heterogeneity. These analyses facilitated the development of a parameter estimation strategy for datasets without ground-truth while also revealing that DoubletFinder is most accurately applied to scRNA-seq data with well-resolved clusters in gene expression space.

In the third part, we apply DoubletFinder to “real-world” data lacking ground-truth doublet labels. Specifically, we test DoubletFinder on an existing mouse kidney scRNA-seq dataset [21] containing an experimentally validated intermediate cell state that shares gene expression features with two other kidney cell types. We chose this dataset in order to explicitly test whether this strategy for artificial doublet generation (i.e., averaging of expression profiles) leads to DoubletFinder false positives in context with bona fide “hybrid” cell states. DoubletFinder correctly classifies this “hybrid” cell state as singlets, which suggests that DoubletFinder can be broadly applied to scRNA-seq data describing cell-state transitions. This case study also illustrates “best practices” for DoubletFinder application and emphasizes how results should be interpreted with methodological limitations in mind (e.g., undetectable doublets and poor performance on homogeneous data).

2.3 Results

2.3.1 DoubletFinder algorithm overview.

DoubletFinder predicts doublets in a fashion that can be split into five distinct steps (**Fig. 2-1**). First, DoubletFinder simulates artificial doublets from existing scRNA-seq data by averaging the gene expression profiles of random pairs of cells. Simulating doublets in this fashion preserves cell composition while recapitulating the intermixing of mRNAs from two cells that occurs during doublet formation. Second, DoubletFinder merges and pre-processes real and artificial data using the "Seurat" single-cell analysis pipeline [19,20]. Notably, pre-processing parameters are held constant between the original and merged real-artificial datasets. Third, DoubletFinder performs dimensionality reduction on the merged real-artificial data using principal-component analysis (PCA), producing a low-dimensional space that describes the similarity between real and artificial cells. Fourth, DoubletFinder detects the *k* nearest neighbors for every real cell in principal component (PC) space, and this information is used to compute each cell's proportion of artificial nearest neighbors (pANN). Finally, building on the assumption that real and artificial doublets co-localize in PC space, DoubletFinder predicts real doublets as cells with the top *n* pANN values, where *n* is set to the total number of expected doublets.

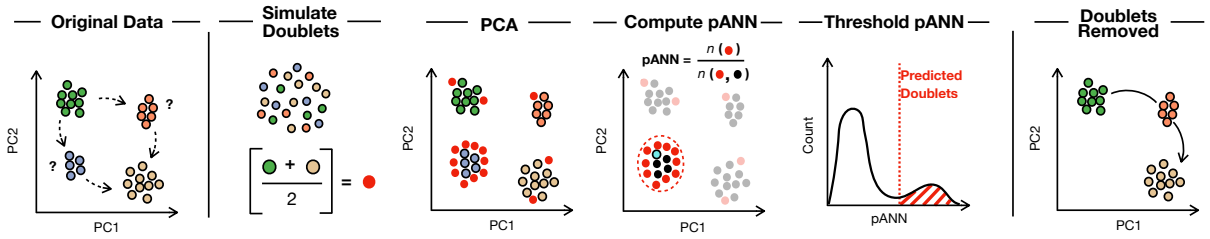

Figure 2-1: Schematic overview of DoubletFinder workflow.

Doublet detection is necessary to correctly interpret intermediate cell states (blue, orange) in scRNA-seq data, which could represent developmental intermediates or technical artifacts. Starting with scRNA-seq data pre-processed using Seurat, DoubletFinder integrates artificial doublets (red) into the existing data at a defined proportion (pN). DoubletFinder then defines each cell's neighborhood in gene expression space (pK, example neighborhood seed in bright blue). The proportion of artificial nearest neighbors (pANN) is then defined, and cells with the top pANN values are predicted as doublets. Doublet removal aids in scRNA-seq data interpretation – e.g., when discerning doublets from legitimate differentiation intermediates.

2.3.2 DoubletFinder parameter interrogation, performance benchmarking on Cell Hashing and Demuxlet PBMC datasets.

DoubletFinder requires three input parameters expressed as proportions of the merged real-artificial dataset: the number of expected real doublets, the number of artificial doublets (pN) and the neighborhood size (pK) used to compute the number of artificial nearest neighbors. For example, in a dataset with 15,000 real cells, a pN of 0.25 would represent the integration of 5,000 artificial doublets, and a pK of 0.01 would represent a pK of 200 cells. To explore how parameter variation influences DoubletFinder performance, we used existing datasets of peripheral blood mononuclear cells (PBMCs) generated using sample multiplexing techniques (Demuxlet and Cell Hashing). Demuxlet identifies cells belonging to each sample group according to sample-specific SNPs and identifies doublets as cell barcodes associated with mutually exclusive sets of SNPs [9]. Cell Hashing identifies doublets using a conceptually analogous strategy, except sample-specific SNPs are replaced by sample-specific DNA barcodes that are linked to cells by conjugation to antibodies targeting cell-surface proteins [15]. Notably, neither method can detect doublets formed from cells associated with the same SNPs or sample barcodes.

We selected these two datasets because, at the time, they were the only publicly-available datasets where within-species doublets were empirically determined. Moreover, since each dataset was sequenced at variable depths (Demuxlet = 2,438 unique molecular identifiers (UMIs), Cell Hashing = 676 UMIs), we could assess whether sequencing depth influenced DoubletFinder performance. We compared the predictive capacity of DoubletFinder outputs (i.e., a vector of every real cell's pANN) across a sweep of pN (0.05-0.3) and pK (5e4-0.3) values using receiver operating characteristic curve (ROC) analysis (**Fig. 2-2a**). Comparing the relative areas under the curve (AUCs) demonstrates that DoubletFinder performance is largely invariant of pN (**Fig. 2-2b**). Moreover, optimal parameter regimes are similar for each dataset, suggesting that DoubletFinder performance is insensitive to sequencing depth. These observations demonstrate that pK is the

main parameter that must be tuned when applying DoubletFinder to different scRNA-seq data. Therefore, we set pN to 0.25 for all DoubletFinder applications and optimized pK for each dataset.

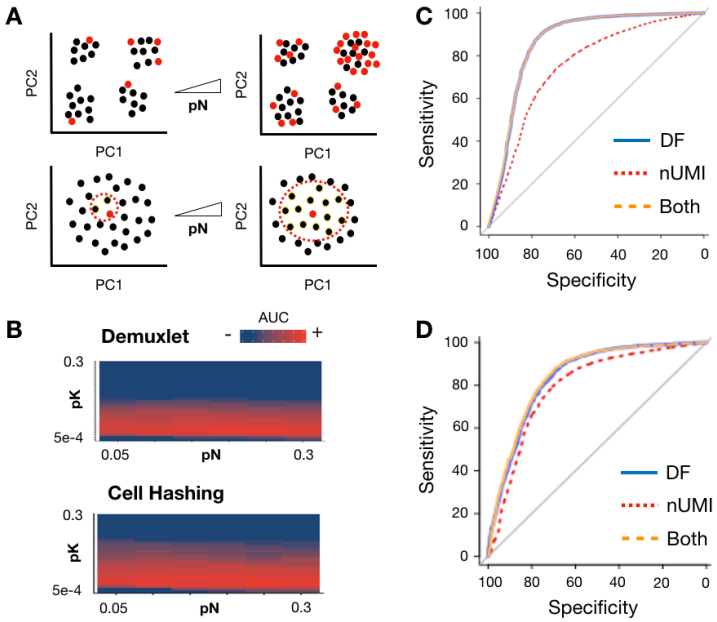

Figure 2-2: Benchmarking DoubletFinder parameters and predictive capacity relative to nUMIs using ROC analysis.

(A) Schematic describing pN-pK parameter sweep. Increasing pN corresponds with increasing numbers of artificial doublets (red) relative to singlets (black). Increasing pK corresponds with larger neighborhood sizes (red dotted circle, neighbors highlighted in yellow) used during pANN computation.
(B) pN-pK parameter sweep AUC heat map for Demuxlet and Cell Hashing data.
(C) ROC analysis of logistic regression models trained on Demuxlet data using DoubletFinder alone (blue), nUMIs alone (red), and both nUMIs and DoubletFinder (orange).
(C) ROC analysis of logistic regression models trained on Cell Hashing data using DoubletFinder alone (blue), nUMIs alone (red), and both nUMIs and DoubletFinder (orange).

Using pK values with the highest AUC from Demuxlet and Cell Hashing ROC analysis (pK = 0.01 for both datasets), we next benchmarked DoubletFinder against a commonly used feature for doublet identification in real-world scRNA-seq data—the number of UMIs [22,23]. UMIs are uniquely associated with individual mRNA transcripts via reverse transcription and enable PCR amplification bias correction as UMIs link each sequenced molecule back to its original mRNA. Since droplets associated with two cells often have more total mRNA molecules than droplets associated with single cells, doublets are commonly removed by setting an upper nUMI threshold. However, this approach has well-established limitations [9], as it does not consider technical variability in mRNA capture efficiency or biological variability in cellular RNA content.

To compare the relative predictive capacities of DoubletFinder and nUMIs for doublet detection, we first randomly split the Demuxlet and Cell Hashing datasets into evenly sized test and training sets. Next, we used ROC analysis to compare logistic regression models trained using DoubletFinder alone (i.e., pANN values for every cell), nUMI alone, or a linear combination of both features. DoubletFinder-based models outperformed nUMI-based models for predicting ground-truth doublets in both the Demuxlet (**Fig. 2-2c**) and Cell Hashing data (**Fig. 2-2d**). Moreover, models trained with both DoubletFinder and nUMIs performed nearly indistinguishably to DoubletFinder-alone models, demonstrating that the method captures all of the doublet-specific information inherent to nUMIs in this context.

Although DoubletFinder predicts doublets better than nUMIs, it remained unclear whether DoubletFinder results accurately recapitulated the ground-truth doublet labels provided by Demuxlet or Cell Hashing sample classifications. To make these comparisons, we needed to convert the DoubletFinder output (i.e., pANN values for every cell) into a list of singlet and doublet labels. To generate this list, we assigned doublet labels to cells in the Demuxlet and Cell Hashing datasets with the top *n* pANN values, where *n* was set to the total number of doublets expected from the empirical sample multiplexing results. For example, since 6,045 doublets were defined by Demuxlet SNP profiling of 8 individuals, and because doublets formed from cells with the same SNPs are classified as singlets by Demuxlet, we estimated that 12.5% of real doublets (864 cells) remained unclassified. To account for both the ground-truth false negatives and the true, Demuxlet-identified doublets, we assigned doublet labels to cells with the top 6,909 pANN values. A similar list was made for the Cell Hashing dataset.

Running DoubletFinder on the same data and visualizing doublets on identical t-stochastic neighbor embedding (t-SNE) plots revealed that Demuxlet (**Fig. 2-3a**), Cell Hashing (**Fig. 2-3b**), and DoubletFinder doublet classifications were generally concordant, with DoubletFinder identifying few false positives relative to ground-truth (Demuxlet specificity = 0.91, Cell Hashing

= 0.91). However, DoubletFinder was insensitive to many ground-truth doublets exhibiting similar gene expression profiles to singlets (Demuxlet sensitivity = 0.73, **Fig. 2-3a**, orange inset, gold dots; Cell Hashing = 0.64). We hypothesized that these cells represented homotypic doublets—i.e., doublets formed from transcriptionally similar cells that cluster among their composite cell-type singlets in gene expression space. Since DoubletFinder requires putative doublets to cluster separately from singlets in PC space, we did not expect DoubletFinder to robustly detect homotypic doublets. Supporting this hypothesis, DoubletFinder sensitivity was increased when homotypic doublets were identified (see Methods for homotypic doublet identification strategy) and excluded from the Demuxlet (**Fig. 2-3c**; sensitivity = 0.93) and Cell Hashing datasets (**Fig. 2-3d**; sensitivity = 0.82), while specificity remained unchanged. Collectively, these results illustrate that DoubletFinder primarily detects heterotypic doublets—i.e., doublets formed from transcriptionally distinct cells.

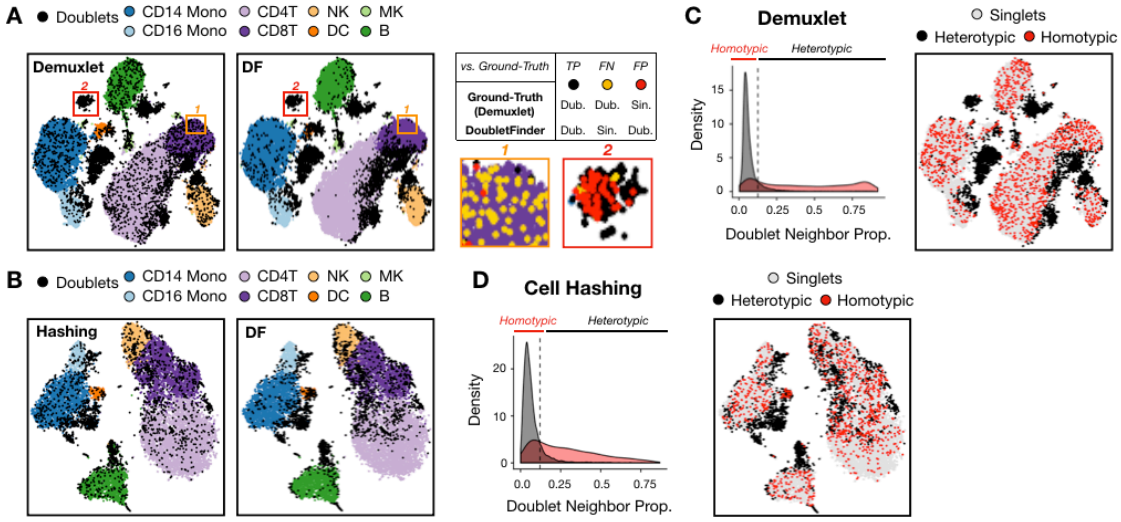

Figure 2-3: Benchmarking DoubletFinder predictions against Demuxlet and Cell Hashing classifications reveals concordance and DoubletFinder insensitivity to homotypic doublets.

(A) t-SNE visualizations of Demuxlet and DoubletFinder doublets (black) among PBMC cell types. Inset regions exemplify two types of discordance. False-negative DoubletFinder classifications (gold) localize among singlets in gene expression space, while putative false-positive DoubletFinder classifications (red) localize among heterotypic doublets. Mono, monocytes; NK, natural killer cells; MK, megakaryocytes; and DC, dendritic cells.

(B) t-SNE visualizations of Cell Hashing and DoubletFinder doublets (black) among PBMC cell types.

(C) Density plot (left) describing the proportion of ground-truth doublet neighbors in Demuxlet gene expression space among ground-truth singlets and doublets. Singlets (gray) have low doublet neighbor proportions, whereas doublets (red) have widely variable doublet neighbor proportions. Homotypic and heterotypic doublets were thresholded at the intersection of singlet and doublet densities (black dotted line). t-SNE visualization (right) demonstrates that homotypic doublets (red) localize among singlets (gray), unlike heterotypic doublets (black).

(D) Analyses presented in Fig 2-3c, but for Cell Hashing data.

DoubletFinder additionally identified a set of doublets left unclassified by Cell Hashing and Demuxlet (**Fig. 2-3a**, red inset, red dots). As described above, we had estimated that 12.5% of real doublets were formed from cells with the same SNPs or sample barcodes. Thus, these doublets would have remained unclassified by Demuxlet or Cell Hashing but should be detected efficiently by DoubletFinder. If these apparent DoubletFinder false-positive cells were in fact ground-truth false negatives, two predictions would follow. First, if putative ground-truth false negatives were real doublets, then these cells should exhibit enriched nUMIs relative to singlets. Second, ground-truth false negatives should express marker genes associated with multiple distinct cell states. In line with these predictions, putative false negatives had nUMI levels indistinguishable from true-positive doublets (Wilcoxon rank-sum test, $p = 0.4$) and were enriched relative to singlets ($p < 2e16$) in both datasets (**Fig. 2-4a**). Moreover, these ground-truth false negatives expressed marker genes associated with hematopoietic cell types that do not share a common progenitor in peripheral blood (**Fig. 2-4b**). Collectively, these results suggest that DoubletFinder recapitulates heterotypic doublet classifications made by Demuxlet and Cell Hashing and accurately predicts sample multiplexing false negatives formed from cells associated with identical SNPs or sample barcodes.

2.3.3 DoubletFinder improves differential gene expression analysis performance.

A common application of scRNA-seq is to discover genes that are differentially expressed among distinct cell types that are obscured in bulk transcriptomic assays [19-21]. Doublets hinder differential gene expression analyses because doublets often cluster separately in gene expression space while sharing transcriptional features with the cell types from which they are derived. To demonstrate this effect, we compared differential gene expression analysis results between Demuxlet and Cell Hashing datasets before and after removing doublets. Doublet removal results in pronounced increases in the total number of differentially expressed genes

(Demuxlet with and without doublets = 2,567 and 4,339; Cell Hashing = 2,185 and 5,598) across nearly every PBMC cell type (**Fig. 2-4c**). Importantly, many newly identified differentially expressed genes are PBMC cell-type marker genes supported in the literature (**Fig. 2-4d**) [19,20,24-28]. These results illustrate how doublet detection and removal improves scRNA-seq analysis workflows.

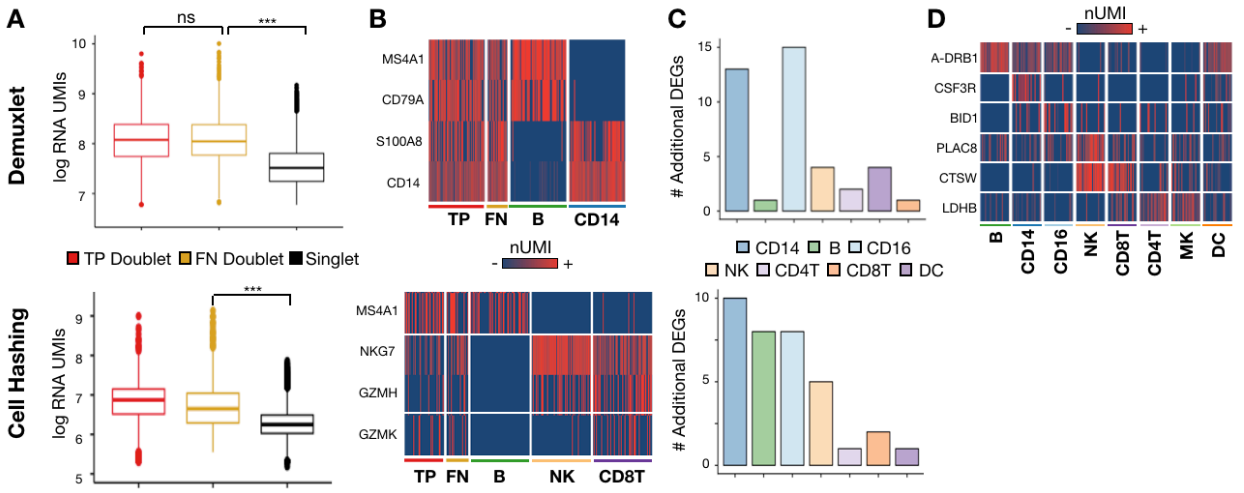

Figure 2-4: DoubletFinder identifies ground-truth false-negative doublet classifications, improves differential gene expression analysis performance.

(A) RNA UMI box plots for true positive doublets (red), putative false negative doublets (gold), and singlets (black) in Demuxlet (top) and Cell Hashing datasets (bottom). Data are represented as mean ± SEM. *** = statistically-significant, ns = not significant.
(B) Marker gene heat map for true positive Demuxlet doublets, false negative doublets, B cells, and CD14 monocytes (top) as well as for true positive Cell Hashing doublets, false negative doublets, B cells, NK cells, and CD8+ T-cells (bottom).
(C) Bar chart describing the number of additional differentially-expressed genes identified following doublet removal for Demuxlet (top) and Cell Hashing datasets (bottom).
(D) Heat map of differentially-expressed genes identified following doublet removal that have previously been described as marker genes in the literature.

2.3.4 Defining the relationship between DoubletFinder parameter selection and performance using scRNA-seq data simulations.

DoubletFinder performance is demonstrably sensitive to changes in the input parameter specifying pK used to compute each cell's pANN (**Figs. 2-2a,b**). To understand the relationship between scRNA-seq data structure and DoubletFinder performance, we used the "splatter" R package [29] to generate simulated scRNA-seq datasets with 3-8 distinct cell clusters that ranged from being intermixed to completely separated in gene expression space (**Fig. 2-5a**). Real

doublets were simulated by adding the gene expression profiles of randomly selected cells such that 10% of the final data was doublets. We visualized parameter performance by finding the mean AUC for each pK value across all pN since pK selection was previously shown to dominate DoubletFinder performance.

For most simulations, mean AUC distributions featured an inflection point representing the point at which pKs became too large to enable accurate doublet prediction (**Fig. 2-5b**). Mean AUC inflection point positions differed for simulations with variable numbers of cell states, suggesting that pK parameter selection is sensitive to the inherent diversity of scRNA-seq data (**Fig. 2-5b**, **top**). Moreover, among simulations with the same number of cell-state clusters but varying degrees of cluster separation, mean AUC inflection points were only observed for simulations with well-separated clusters (**Fig. 2-5b**, **bottom**).

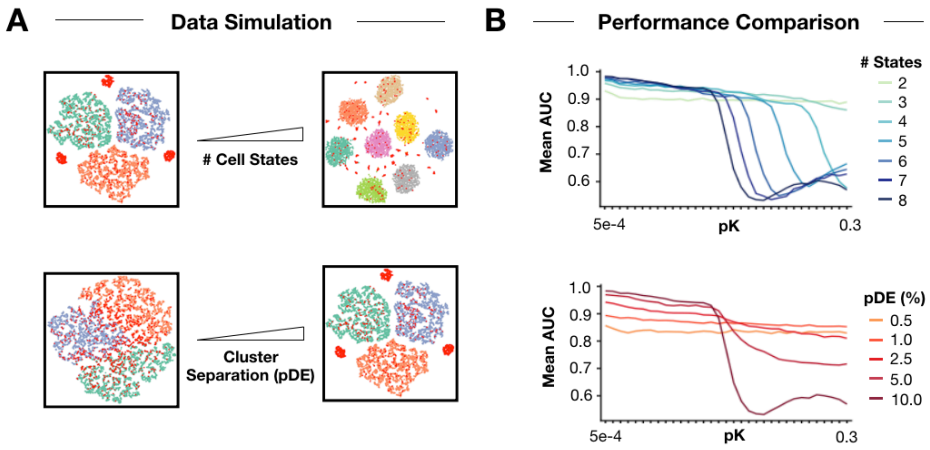

Figure 2-5: scRNA-seq data simulations highlight relationship between pK parameter selection, scRNA-seq data structure, and DoubletFinder performance.

(A) Schematic overview of data simulation strategy. scRNA-seq data including doublets (red) with different numbers of cell states (top) and extent of cluster separation in gene expression space (bottom) were simulated. pDE, probability of differential expression. (B) Simulated pN-pK parameter sweep results. Range of pK values coinciding with high mean AUC differ between simulated data with varying numbers of equally separated cell states (pDE, 10.0% for all simulations, top). DoubletFinder performance suffers on the whole when applied to simulated data with variable degrees of cluster separation (number of cell states = 8 for all simulations, bottom).

This observation suggests that DoubletFinder performance suffers as a whole when applied to data describing transcriptionally homogeneous cell states. This decrease in performance is common to other computational doublet detection strategies that utilize

transcriptomic information alone [30] and illustrates a key methodological limitation that should be carefully considered by all prospective users.

2.3.5 pK optimization using mean-variance-normalized bimodality coefficient (BC_{MVN}).

To identify optimal pK values for real-world scRNA-seq data when ground-truth doublet information is not known (precluding ROC analysis and AUC maximization for pK selection), we suggest that DoubletFinder users calculate the mean-variance-normalized bimodality coefficient (BC_{MVN}) [31] of pANN distributions produced during pN-pK parameter sweeps of their data (**Fig. 2-6a**). BC_{MVN} can be used to identify the pK that separates singlets and doublets effectively, without being sensitive to local density differences in gene expression space (**Fig. 2-6b,c**).

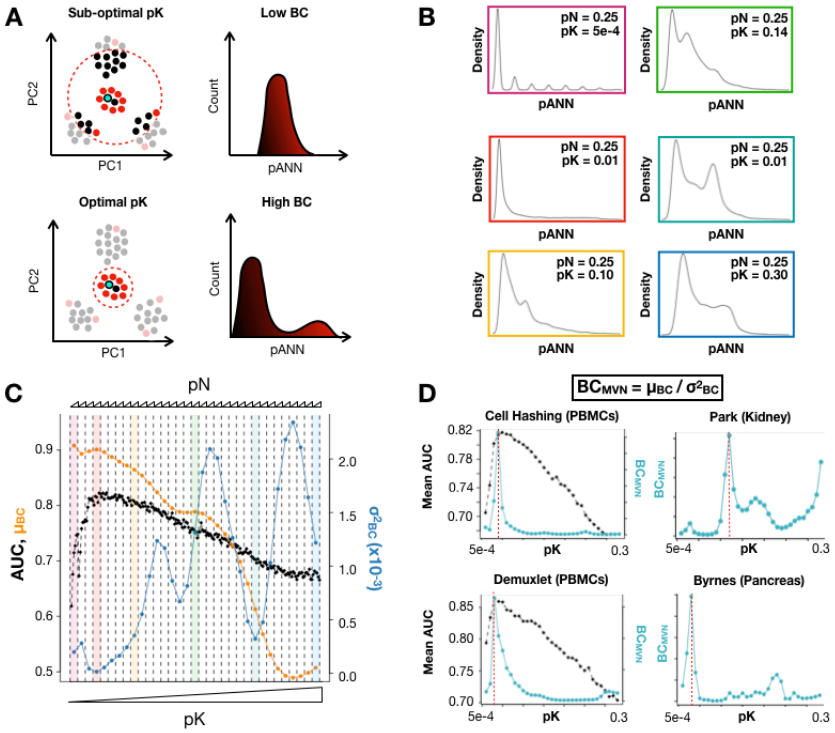

Figure 2-6: DoubletFinder pK optimization using BC_{MVN} maximization.
(A) Schematic overview of relationships between pK, BC, and pANN. As pK becomes too large, pANN for doublets and singlets becomes more similar, resulting in unimodal pANN distributions with low BC (top). Neighborhood sizes that reflect the structure of clusters in gene expression space correspond to distinct pANN regimes for real singlets and doublets, resulting in non-unimodal pANN distributions with high BC (bottom).
(B) Representative pANN distributions across the Cell Hashing pN-pK sweep. Borders correspond to pK bins in Fig. 2-6c.
(C) Maximizing BC mean ($\mu_.$, orange) while minimizing BC variance ($\sigma_.$, blue) enables identification of the optimal pK value for Cell Hashing data, as measured using AUC from ROC analysis (black). Highlighted pK bins correspond to pANN distribution borders in Fig. 2-6b. pK values are separated into pN bins by the black dashed lines.
(D) Comparison of $BC_{MVN.}$ (teal) and mean AUC distributions (black) enables identification of high-AUC pK values. BC_{MVN} distributions for mouse kidney and pancreas data inform pK parameter selection. Red dotted lines denote optimal pK values.

To demonstrate the utility of BC_{MVN} for DoubletFinder parameter selection, we benchmarked BC_{MVN} against the previously computed ROC results for Demuxlet and Cell Hashing data, as well as two scRNA-seq datasets generated without sample multiplexing [21,32]. Across all datasets tested, BC_{MVN} distributions featured a single maximum that for the Cell Hashing and Demuxlet datasets, coincided with the pK range maximizing AUC (**Fig. 2-6d**). Therefore, we propose that BC_{MVN} maximization selects a near-optimal DoubletFinder parameter across a range of scRNA-seq datasets.

2.3.6 DoubletFinder application to mouse kidney scRNA-seq data illustrates insensitivity to bona fide 'hybrid' cell states.

We next sought to demonstrate DoubletFinder's capabilities in a real-world context where ground-truth doublets are not known and BC_{MVN} maximization must be used to determine a reasonable pK. To this end, we applied DoubletFinder to a previously published scRNA-seq dataset describing the mouse kidney. In this study, the authors discover and experimentally validate the existence of a novel cell type—collecting duct transitional cells (CDTCs)—which expresses genes characteristic of two other kidney cell types: collecting duct principal cells (CDPCs) and collecting duct intercalated cells (CDIC) [21]. This dataset represented an intriguing "challenge-case" for DoubletFinder, as we reasoned that legitimate cell types with "hybrid" gene expression profiles may resemble artificial doublets, triggering DoubletFinder false positives.

Beginning with a pre-processed Seurat object (**Fig. 2-7a**), we first used BC_{MVN} maximization to identify a suitable pK value for these data (pK = 0.09, **Fig. 2-6d**). We then applied DoubletFinder to the full dataset and classified doublets as cells with the top *n* pANN values. Initially, *n* was set according to the Poisson doublet formation rate, as specified for the particular cell-loading density used in the study [21]. This resulted in 913 total doublet predictions, which were highly enriched for nUMIs (**Fig. 2-7b**) and included a region of heterotypic doublets

characterized by the co-expression of proximal tubule and distal convoluted tubule marker genes (**Fig. 2-7c**).

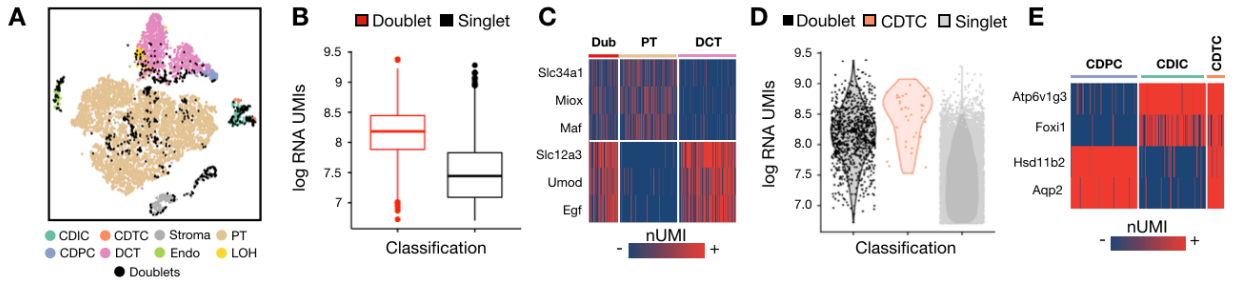

Figure 2-7: DoubletFinder is insensitive to bona fide 'hybrid' cell states in mouse kidney scRNA-seq data.

(A) t-SNE visualization of DoubletFinder doublet predictions (black) amongst mouse kidney cell types. DCT = distal convoluted tubule, PT = proximal tubule, Endo = endothelial, LOH = loop of Henle.
(B) RNA UMI box plots for doublets (red), and singlets (black). Data are represented as mean ± SEM.
(C) Marker gene heat maps for doublets, PT cells (beige), and DCT cells (pink).
(D) Mouse kidney CDTCs (peach) exhibit elevated nUMIs similar to doublets (black) and distinct from singlets (grey).
(E) Mouse kidney CDTCs (peach) co-express marker genes associated with CDPCs (blue) and CDICs (turquoise).

DoubletFinder correctly identified 64% of CDTCs as singlets, despite CDTCs having exceptionally high nUMIs (**Fig. 2-7d**) and co-expressing both CDPC and CDIC marker genes (**Fig. 2-7e**). However, these ini3tial results represented an overestimation of the true number of detectable doublets as DoubletFinder was applied without taking homotypic doublets into account. We therefore adjusted the expected doublet number to account for homotypic doublets. Specifically, we grouped cells according to literature-supported cell-type annotations and estimated the proportion of homotypic doublets as the sum of squared cell-type frequencies. This strategy assumes (1) that a cell type's frequency in the final dataset reflects that cell type's contribution to the doublet pool, and (2) that user-defined cell-type groups approximate the magnitude of transcriptional divergence necessary to make a detectable heterotypic doublet. In contexts where such annotations are inaccurate or unavailable, unsupervised clustering results should be used instead. This analysis resulted in a revised 473 total heterotypic doublet predictions and allowed us to identify 97% of CDTCs as singlets. This result suggests that DoubletFinder can be insensitive to legitimate cell states with intermediate expression profiles.

We suggest that the heterotypic doublet frequency and Poisson doublet formation rates be used as lower and upper bounds for estimating the number of detectable doublets, respectively. We urge DoubletFinder users to interrogate the results of each thresholding strategy against the known biology of the system under study.

2.4 Discussion

DoubletFinder is a computational doublet detection method that integrates artificial doublets into existing scRNA-seq data and identifies real doublets as cells enriched for artificial nearest neighbors in gene expression space. DoubletFinder is implemented in the R programming language and is written to interface with the popular Seurat scRNA-seq analysis package [19,20]. However, DoubletFinder is prospectively generalizable to scRNA-seq data analyzed using alternative pipelines as well. In this study, we benchmarked DoubletFinder against ground-truth scRNA-seq data where doublets are directly measured using sample multiplexing techniques such as Demuxlet [9] and Cell Hashing [15]. We leveraged these results to define “best practices” for how DoubletFinder should be applied to real-world scRNA-seq data without ground-truth doublet labels. We then successfully demonstrated these practices on mouse kidney data featuring an experimentally validated cell state that could trigger DoubletFinder false positives [21].

Ground-truth comparisons revealed a number of DoubletFinder strengths and limitations. For example, DoubletFinder outperforms nUMI thresholding in these data and accurately predicts heterotypic doublets with >90% sensitivity. In contrast, DoubletFinder is insensitive to homotypic doublets, as these cells do not diverge significantly from real singlets in gene expression space. DoubletFinder also identifies false negatives in the Demuxlet and Cell Hashing datasets that are formed from cells associated with identical sample barcodes. For this reason, we view

DoubletFinder and sample multiplexing as complementary doublet removal approaches, especially in experimental contexts with relatively low sample numbers. When used in concert, sample multiplexing and computational doublet detection techniques provide an effective solution to the issue of doublets in scRNA-seq data, enabling users to “super-load” droplet microfluidic devices and thereby further increase scRNA-seq cell throughput.

In contexts where sample multiplexing information is unavailable, DoubletFinder detects and removes the preponderance of heterotypic doublets while homotypic doublets remain. The presence of homotypic doublets is unlikely to negatively influence cell-type classification and differential gene expression analysis, as homotypic doublets cluster together with bona fide cell singlets. In fact, simply removing heterotypic doublets improved differential gene expression analysis results in every dataset tested in this study. However, certain scRNA-seq analyses may also benefit from the removal of homotypic doublets. For example, imputation uses the average gene expression profiles of transcriptionally similar cells to infer missing values caused by transcript dropout events [33,34]. It is possible that the structure of missing values in singlets and homotypic doublets is distinct, and thus, it remains unclear how the presence of homotypic doublets influences imputation performance.

Beyond exposing DoubletFinder strengths and limitations, ground-truth benchmarking revealed three methodological features that should be considered as users apply DoubletFinder to scRNA-seq data lacking ground-truth doublet labels. First, applying DoubletFinder to simulated scRNA-seq data with poorly resolved clusters demonstrates that DoubletFinder cannot be accurately applied to scRNA-seq data describing transcriptionally similar cells. DoubletFinder users should therefore carefully consider the diversity of their dataset prior to using the method. Second, DoubletFinder input parameters (e.g., pK) must be tuned to datasets with variable numbers of cell states. We predicted pK for ground-truth scRNA-seq data using ROC analysis, which is not possible for real-world data. Instead, we developed a ground-truth-agnostic

parameter selection strategy—termed BC_{MVN} maximization—which finds the pK value that optimally separates singlet and doublet pANN distributions. Third, since DoubletFinder is insensitive to homotypic doublets, thresholding DoubletFinder results according to the total number of doublets estimated via Poisson loading statistics will necessarily result in false positives. To account for this issue, we describe how the proportion of homotypic doublets can be estimated from cell-type frequencies described using existing annotations or unsupervised clustering. Using this strategy, one can threshold DoubletFinder results in a fashion that accounts for homotypic doublets and thereby limits false positives.

2.5 Materials and Methods

2.5.1 scRNA-seq data pre-processing using Seurat.

DoubletFinder was implemented in the R programming language in a fashion that purposefully interfaces with the Seurat analysis package. DoubletFinder takes as an input a Seurat object that has been pre-processed using the standard Seurat analysis pipeline. Briefly, raw RNA UMI counts are normalized (e.g., log2-transform), centered, and scaled before regression is used to remove undesired sources of variability (e.g., total nUMI). In the standard Seurat workflow, variably expressed genes are then defined via dispersion and mean expression thresholds. In this study, thresholds were chosen that identified ~2000 total genes, as described previously [19,20]. PCA is then performed using this set of variably expressed genes, and statistically-significant PCs are selected (e.g., via inflection point estimation on PC elbow plots). These are the minimum pre-processing requirements prior to running DoubletFinder, although further dimensionality reduction (e.g., t-SNE) and unsupervised clustering were also utilized in this study.

2.5.2 DoubletFinder algorithm overview.

The DoubletFinder workflow begins with a pre-processed Seurat object, prepared as described above. Artificial doublets are then generated from raw UMI count matrices by averaging the gene expression profiles of cell pairs selected via random sampling with replacement. Sufficient artificial doublets are then generated to comprise 25% of the resulting merged data (pN = 0.25). Next, real and artificial data are merged and pre-processed using the same normalization, scaling, and variable gene definition parameters employed during the original data analysis workflow. Notably, nUMI regression is not performed during merged real-artificial dataset pre-processing in order to preserve differences between singlets and doublets. Using the same number of statistically-significant PCs selected during original data pre-processing, PC cell embeddings are then converted into a Euclidean distance matrix using the 'rdist' function from the 'fields' R package [35]. Each cell's nearest neighbors are then defined from this distance matrix, and the proportion of artificial nearest neighbors (pANN) is computed for every real cell by dividing its number of artificial neighbors by the neighborhood size (pK). Final doublet classifications are then assigned to the cells with the *n* highest pANN, where *n* was set to the total number of expected doublets.

2.5.3 ROC analysis for optimizing pK selection.

For Cell Hashing and Demuxlet scRNA-seq data, optimal parameters were selected by maximizing the AUC from ROC analysis of pN-pK parameter sweeps. Specifically, Cell Hashing and Demuxlet datasets were first randomly sub-sampled to 10,000 cells in order to maximize computational efficiency during the parameter sweep. Second, artificial doublets were integrated at varying proportions (pN = 0.05-0.30), and merged real-artificial data was pre-processed as described above. Third, the proportion of artificial nearest neighbors was computed for varying neighborhood sizes (pK = 5e4-0.3) for each real cell. This produced a list of pANN vectors

corresponding to each pN-pK combination. Fourth, ground-truth doublet labels and pANN vectors were then evenly split into test and training sets via random sampling without replacement. Fifth, logistic regression models were fit on training cells using the 'glm' R function with the 'family' and 'link' arguments set to 'binomial' and 'logit', respectively. Logistic regression was used because this technique specifically models the binary nature of singlet/doublet classifications. Sixth, models were applied to test cells and the predictive capacity of each model was compared by computing AUC during ROC analysis, as implemented in the 'ROCR' [36] and 'pROC' [37] R packages. Notably, in cases where ground-truth classifications are not available for DoubletFinder parameter selection, users should employ BC_{MVN} maximization (see Section 2.5.8)

2.5.4 ROC analysis for comparing predictive capacity of DoubletFinder and nUMIs.

DoubletFinder parameters optimized for the Cell Hashing and Demuxlet datasets using ROC analysis were then used to benchmark the method against nUMI thresholding. Test and training sets were defined as described above, and logistic regression models were fit using DoubletFinder alone, nUMI alone, or a linear combination of both features. Trained models were then applied to test cells, and ROC analysis was used to compare each of the three models.

2.5.5 Estimating doublet numbers according to sample multiplexing results.

For Demuxlet and Cell Hashing data, cells with the *n* highest pANN values were classified as doublets, where *n* was defined as the number of ground-truth doublets adjusted according to the expected ground-truth false-negative rate. Notably, we utilized this strategy prior to discovering that DoubletFinder is insensitive to homotypic doublets. Thus, we suggest users interpret DoubletFinder results using the Poisson doublet formation rate with and without adjustment for homotypic doublet proportions (see Section 2.5.9).

2.5.6 Predicting homotypic doublets via pANN thresholding.

To illustrate that DoubletFinder is predominantly sensitive to heterotypic doublets, we sought a strategy to directly distinguish homotypic and heterotypic doublets within ground-truth doublet classifications. Specifically, since homotypic and heterotypic doublets respectively co-localize with singlets and doublets in gene expression space, we reasoned that the two doublet types could be discerned according to the proportion of nearest neighbors that were real doublets. We computed this proportion for each real cell using the pK value optimized by ROC analysis (pK = 0.01). We then visualized the density distributions of doublet neighborhood proportions for real doublets and singlets. We then used the intersection of these distributions as a threshold to split real doublets into homotypic and heterotypic subsets, following the assumption that homotypic doublets and real singlets would have similar doublet neighborhood proportions (**Figs. 2-3c, 2-3d**). Homotypic doublets identified using this strategy localize amongst singlets in gene expression space, as expected.

2.5.7 scRNA-Seq data simulation using 'splatter'.

scRNA-seq data was simulated using the 'splatter' R package [29] as in [30]. Datasets were simulated with 3-8 equally-proportioned cell states, and cluster separation in gene expression space was controlled using the 'de.prob' parameter (0.005-0.1) of the 'splatSimulate' R function. Simulated doublets were added to these data by adding the UMI counts for random pairs of cells such that 10% of the final data were doublets. Simulated datasets containing doublets were then pre-processed using 'Seurat' as described previously, with the number of statistically-significant PCs set to the total number of cell states. Following pre-processing, parameter sweeps, logistic regression modeling, and ROC analysis were performed on each simulated dataset, as described above. Since pK is the main parameter requiring adjustment in

different contexts, we visualized our results by finding the mean AUC across all pN values for each pK.

2.5.8 pK optimization with BC_{MVN} maximization.

When ground-truth doublets labels are unavailable for identifying optimal DoubletFinder pK values, pK must be selected using a ground-truth-agnostic strategy called mean-variance-normalized bimodality coefficient (BC_{MVN}) maximization. Bimodality coefficient (BC) measures deviations from unimodality in data distributions [31]. For DoubletFinder parameter fitting, we reasoned that parameter sets that produced non-unimodal pANN distributions would optimally separate singlets from doublets and, as a result, would perform the best. Thus, for the Demuxlet [9], Cell Hashing [15], mouse kidney [21], and mouse pancreas [32] scRNA-seq datasets, we tested every pANN distribution generated during pN-pK parameter sweeps to find those with elevated BC values. Specifically, we computed BC as is implemented in the 'bimodality_coefficient' function in the 'modes' R package [38], which is formalized as:

$$BC = \frac{\gamma^2 + 1}{\kappa + \frac{3(n-1)^2}{(n-2)(n-3)}}$$

Where γ is the pANN distribution skewness (i.e., peak width), k is the kurtosis (i.e., peak sharpness), and n is the sample size. We then measured the BC mean and variance for each pK across all pN values tested, as it was previously shown that DoubletFinder performance is not influenced by the number of generated artificial doublets.

We documented the results of this workflow when applied to the Cell Hashing data as a representative example (**Fig. 2-6**). When pK values are too high, singlets and doublets have

similar proportions of artificial nearest neighbors, and the resulting pANN distributions are associated with low BC and AUC. In contrast, when pK values are too low, DoubletFinder performance suffers because neighborhoods in gene expression space are dominated by local effects that result in multimodal pANN distributions (**Fig. 2-6b**, top left). Since these distributions are not unimodal, they are associated with high BC. However, local effects are sensitive to the number of artificial doublets integrated into the dataset (pN), resulting in elevated BC variance for the associated pK values (**Fig. 2-6c**, pink). Finally, ideal pK values generate long-tailed pANN distributions (**Fig. 2-6b**, mid left) that are characterized by high AUC and high BC with low variance (**Fig. 2-6c**, red). Since high BC values with low-variance predicted high AUC parameter sets in the Cell Hashing data, we leveraged these observations to devise a new metric for pK parameter selection – BCMVN – formalized as:

$$BC_{MVN} = \frac{\mu_{BC}}{\sigma^2_{BC}}$$

Where μ_{BC} and σ^2_{BC} are the BC mean and variance, respectively, for each pK across pN values. BC_{MVN} distributions feature a single, visually-discernible maximum for the four datasets tested in this study (**Fig. 2-6d**). For ground-truth datasets, this maximum corresponds with the ideal pK value identified via ROC analysis.

2.5.9 Adjusting estimated doublet numbers to account for homotypic doublets.

Since DoubletFinder is insensitive to homotypic doublets, thresholding results based on the Poisson doublet formation rate will necessarily result in false-positives. Thus, DoubletFinder results can be interpreted after adjusting the number of expected doublets to account for the estimated proportion of homotypic doublets in the data. To this end, homotypic doublet

proportions were modeled as the sum of squared cell state frequencies. For example, consider a scRNA-seq dataset with five unique cell states present at the following proportions:

$$p_{ci} = \{0.40, 0.25, 0.15, 0.1, 0.1\}$$

Where p_{ci} is the proportion of cell state *i*. The proportion of homotypic doublets present in this data, p_{homo} is then estimated as:

$$p_{homo} = \sum(p_{c1}^2 + p_{c2}^2 + p_{c3}^2 + p_{c4}^2 + p_{c5}^2) = 0.265$$

The final number of detectable (i.e., heterotypic) doublets is then defined by adjusting the total number of doublets (i.e., as determined by the Poisson doublet formation rate) by the homotypic doublet proportion:

$$DDR = (1 - p_{homo}) * TDR$$

Where DDR and TDR are the detectable and total doublet rates, respectively. This strategy follows the assumption that, during droplet microfluidics-based cell capture, the probability that an emulsion oil droplet is filled with a cell from state *i* matches the proportion of cell state *i* in the final scRNA-seq dataset. This assumption does not consider differential doublet formation propensities between cell types (e.g., due to adhesive properties, cell size, etc.). Notably, the accuracy of this strategy depends on whether cell state annotations accurately group cells with transcriptional profiles that are sufficiently similar to preclude formation of a heterotypic doublet. This magnitude of transcriptional similarity is difficult to define and is likely dataset-

dependent. However, unsupervised clustering results and/or literature-supported cell state annotations from existing scRNA-seq data are the best approximations, and represent a lower bound for the total number of doublets.

For the mouse kidney data with which this strategy was implemented, we assigned doublet labels to cells with the top *n* pANN values, where *n* was set to the Poisson doublet formation rate with and without homotypic doublet adjustment. This strategy results in two sets of doublet predictions associated with varying stringencies. With the unadjusted Poisson threshold, DoubletFinder users can be confident that all heterotypic doublets were removed, albeit along with a subset of real singlets. In contrast, the homotypic-adjusted threshold preserves the most existing data while potentially leaving real doublets remaining.

2.5.10 Quantification and statistical analyses.

Statistically-significant differences between nUMIs in Demuxlet and Cell Hashing singlets and doublets were defined using the Wilcoxon rank sum test implemented with the 'pairwisewilcox.test' R function. Multiple comparison correction was performed using the Benjamini-Hochberg procedure. In this context, *n* represents the total nUMIs associated with one cell (Demuxlet n = 33,328; Cell Hashing n = 15,178). Sensitivity and specificity were computed for ground-truth scRNA-seq data before and after homotypic doublet definition, as described above. Sensitivity and specific calculations were performed using the 'caret' R package [39]. Differential gene expression analysis comparisons between scRNA-seq datasets before and after doublet removal was performed with the 'FindMarkers' function in 'Seurat'. Statistical significance was tested using the likelihood-ratio test for single-cell gene expression [40], and marker genes were defined as statistically-significant genes with 3-fold expression enrichment. For the Cell Hashing and Demuxlet datasets, doublet removal included all doublets classified either by

sample-multiplexing or DoubletFinder. For mouse kidney scRNA-seq data, only DoubletFinder-defined doublets were removed.

2.5.11 Data and software availability.

Cell Hashing (GSE108313), Demuxlet (GSE96583), mouse kidney (GSE107585), and mouse pancreas (GSE101099) UMI count matrices were downloaded from the Gene Expression Omnibus. DoubletFinder is implemented as a fast, easy-to-use R package that interfaces with Seurat version 2.0 and higher. DoubletFinder can be downloaded from GitHub (https://github.com/chris-mcginnis-ucsf/DoubletFinder) and is available as an executable Compute Capsule on Code Ocean (DOI: https://doi.org/10.24433/CO.4902498.v1).

2.6 Perspective

Since DoubletFinder and Scrublet were co-published by *Cell Systems* in 2019, computational doublet detection has been readily adopted by the single-cell genomics community, and is now considered a standard component of scRNA-seq quality-control workflows [41-44]. Moreover, considerable effort and resources has been dedicated to developing alternative doublet detection algorithms. To date, 12 distinct doublet detection workflows have been described, including DoubletFinder, Scrublet [30], doubletCells [45], DoubletDetection [46], scds [47], Solo [48], DoubletDecon [49], BIRD [50], Chord [51], MLtiplet [52], GMM-Demux [53], and DoubletCollection [54]. Xi and Li eloquently summarize the differences between many of these methods in their pioneering benchmarking effort [54], wherein the authors apply each method to 16 real scRNA-seq datasets and compute a variety of performance metrics related to prediction accuracy, improvements in downstream analyses (e.g., differential gene expression testing, unsupervised clustering, cell trajectory inference, etc.), and computational efficiency.

DoubletFinder performed the best amongst all doublet detection methods in this context, although it was one of the worst methods in terms of computational efficiency (an observation recapitulated elsewhere [55]). It is difficult to pinpoint exactly why DoubletFinder is the among the best (and slowest) methods. However, I speculate that this is because after artificial doublet generation, DoubletFinder incorporates end-to-end scRNA-seq data normalization and parameter optimization, which slows down the overall workflow while ensuring that the artificial doublets more closely recapitulate real scRNA-seq doublets.

I anticipate that efforts invested in scRNA-seq doublet detection algorithm development and optimization will soon peter out as these methods approach their sensitivity limit and consensus-based methods [51,54], which incorporate doublet predictions from large numbers of algorithms, become the gold-standard. However, as other single-cell genomics assay modalities such as proteomics [56-58] and epigenomics [59-61] continue to gain popularity, existing computational doublet detection algorithms will undoubtedly require further reformulation.

For example, a doublet detection method conceptually-inspired by Scrublet and DoubletFinder was implemented in the recently-described single-cell epigenomics analysis software package, ArchR [62]. While ArchR-based doublet detection worked well on a cell-line mixing single-cell ATAC-seq dataset and PBMC Multiome (RNA+ATAC) dataset generated for the study, an alternative approach called ATAC-DoubletDetector [63] has exhibited superior performance in broader contexts (see Chapter 5 of this thesis). The key conceptual shift for ATAC-DoubletDetector relative to ArchR is that instead of identifying real doublets via similarity to artificial doublets, ATAC-DoubletDetector leverages the trinary nature of single-cell ATAC-seq data (i.e., 0, 1, or 2 counts for closed, open in one allele, and open in two alleles, respectively) to predict doublets as cells with >2 next-generation sequencing reads aligning to significant numbers of genomic loci. This approach yields higher global doublet prediction accuracy and the ability to identify homotypic doublets which are otherwise missed by neighborhood detection algorithms. I

speculate that this theme – i.e., developing doublet detection algorithms which reflect the underlying structure of the datasets to which they are applied – will be repeatedly observed as new single-cell genomics assay modalities become more mainstream.

2.7 References

1. Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. Science. 2017; 357(6352): 661-7.
2. Gierahn TM, Wadsworth MH 2nd, Hughes TK, Bryson BD, Butler A, Satija R, et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Nature Methods. 2017; 14(4): 395-8.
3. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015; 161(5): 1202-14.
4. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015; 161(5): 1187-1201.
5. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, et al. Massively parallel digital transcriptional profiling of single cells. Nature Communications. 2017; 8: 14049.
6. Bloom JD. Estimating the frequency of multiplets in single-cell RNA sequencing from cell-mixing experiments. PeerJ. 2018; 6: e5578.
7. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nature Reviews Genetics. 2016; 16(3): 133-45.
8. Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, Teichmann SA. Classification of low quality cells from single-cell RNA-seq data. Genome Biology. 2016; 17: 29.
9. Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nature Biotechnology. 2017; 36(1): 89-94.

10. Heaton H, Talman AM, Knights A, Imaz M, Gaffney DJ, Durbin R, et al. Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. Nature Methods. 2020; 17(6): 615-20.
11. Huang Y, McCarthy DJ, Stegle O Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. Genome Biology. 2019; 20(1): 273.
12. Xu J, Falconer C, Nguyen Q, Crawford J, McKinnon BD, Mortlock S, et al. Genotype-free demultiplexing of pooled single-cell RNA-seq. Genome Biology. 2019; 20(1): 290.
13. Guo C, Kong W, Kamimoto K, Rivera-Gonzalez GC, Yang X, Kirita Y, Morris SA. CellTag Indexing: genetic barcode-based sample multiplexing for single-cell genomics. Genome Biology. 2019; 20(1): 90.
14. Shin D, Lee W, Lee JH, Bang D. Multiplexed single-cell RNA-seq via transient barcoding for simultaneous expression profiling of various drug perturbations. Science Advances. 2019; 5(5): eaav2249.
15. Stoeckius M, Zheng S, Houck-Loomis B, Hao S, Yeung BZ, Mauck WM 3rd, et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biology. 2018; 19(1): 224.
16. Gehring J, Park JH, Chen S, Thomson M, Pachter L. Highly multiplexed single-cell RNA-seq by DNA oligonucleotide tagging of cellular proteins. Nature Biotechnology. 2019; 38(1): 35-8.
17. McGinnis CS, Patterson DM, Winkler J, Conrad DN, Hein MY, Srivastavan V, et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nature Methods. 2019 June 17; 16(7): 619-26.
18. Gaublomme JT, Li B, McCabe C, Knecht A, Yang Y, Drokhlyanksy E, et al. Nuclei multiplexing with barcoded antibodies for single-nucleus genomics. Nature Communications. 2019 Jul 2; 10(1): 2907.

19. Satija R, Farrell JA, Gennert D, Schier AF, and Regev A (2015). Spatial reconstruction of single-cell gene expression data. Nature Biotechnology. 2015 Apr 13; 33(5): 495-502.
20. Butler A, Hoffman P, Smibert P, Papalexi E, and Satija R (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology. 2018; 36(5): 411-20.
21. Park J, Shrestha R, Qiu C, Kondo A, Huang S, Werth M, et al. Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. Science. 2018; 360(6390): 758-63.
22. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nature Methods. 2014; 11(2): 163-6.
23. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. Molecular Cell. 2017; 65(4): 631-43.e4.
24. Clark HL, Banks R, Jones L, Hornick TR, Higgins PA, Burant CJ, Canaday DH. Characterization of MHC-II antigen presentation by B cells and monocytes from older individuals. Clinical Immunology. 2012; 144(2): 172-7.
25. Ancuta P, Liu KY, Misra V, Wacleche VS, Gosselin A, Zhou X, Gabuzda D. Transcriptional profiling reveals developmental relationship and distinct biological functions of CD16+ and CD16- monocyte subsets. BMC Genomics. 2009; 10: 403.
26. Zhao C, Tan Y, Wong W, Sem X, Zhang H, Han H, et al. The CD14+/lowCD16+ monocyte subset is more susceptible to spontaneous and oxidant-induced apoptosis than the CD14+CD16 subset. Cell Death & Disease. 2010; 1(11): e95.
27. Jeevan-Raj B, Gehrig J, Charmoy M, Chennupati V, Grandclément C, Angelino P, et al. The transcription factor Tcf1 contributes to normal NK cell development and function by limiting the expression of granzymes. Cell Reports. 2017; 20(3): 613-26.

28. Stoeckle C, Gouttefangeas C, Hammer M, Weber E, Melms A, Tolosa E. Cathepsin W expressed exclusively in CD8+ T cells and NK cells, is secreted during target cell killing but is not essential for cytotoxicity in human CTLs. Experimental Hematology. 2009; 37(2): 266-75.
29. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. Genome Biology. 2017; 18(1): 174.
30. Wolock SL, Lopez R, Klein AM. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Systems. 2019; 8(4):281-91.e9.
31. Pfister R, Schwarz KA, Janczyk M, Dale R, Freeman JB. Good things peak in pairs: a note on the bimodality coefficient. Frontiers in Psychology. 2013; 4: 700.
32. Byrnes LE, Wong DM, Subramaniam M, Meyer NP, Gilchrist CL, Knox SM, et al. Lineage dynamics of murine pancreatic development at single-cell resolution. Nature Communications. 2018; 9(1): 3922.
33. Huang M, Wang J, Torre E, Dueck H, Shaffer S, Bonasio R, et al. SAVER: gene expression recovery for single-cell RNA sequencing. Nature Methods. 2018; 15(7): 539-42.
34. van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, et al. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell. 2018; 174(3): 716-29.e27.
35. Nychka D, Furrer R, Paige J, Sain S. Fields: tools for spatial data. R package, version 9.6. 2017. https://cran.r-project.org/web/packages/fields/index.html.
36. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005; 21: 3940-1.
37. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011; 12: 77.

38. Deevi S. Modes: Find the Modes and Assess the Modality of Complex and Mixture Distributions, Especially with Big Datasets. R package, version 0.7.0. 2016. https://rdrr.io/cran/modes/
39. Kuhn M. Building predictive models in R using the caret package. Journal of Statistical Software. 2008. 28; 1-26.
40. McDavid A, Finak G, Chattopadyay PK, Dominguez M, Lamoreaux L, Ma SS, et al. Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. Bioinformatics. 2013. 29; 461-7.
41. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Molecular Systems Biology. 2019; 15(6): e8746.
42. Andrews TS, Kiselev VY, McCarthy D, Hemberg M. Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. Nature Protocols. 2021; 16(1): 1-9.
43. Clarke ZA, Andrews TS, Atif J, Pouyabahar D, Innes BT, MacParland SA, Bader GD. Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. Nature Protocols. 2021; 16(6): 2749-64.
44. Hie B, Peters J, Nyquist SK, Shalek AK, Berger B, Bryson BD. Computational Methods for Single-Cell RNA Sequencing. Annual Review of Biomedial Data Science. 2020; 3(1): 339-64.
45. Lun ATL, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Res. 2016; 5: 2122.
46. Gayoso A, Shor J. DoubletDetection (Zenodo). 2018. doi.org/ 10.5281/zenodo.2678042.
47. Bais AS, Kostka D. scds: computational annotation of doublets in single-cell RNA sequencing data. Bioinformatics. 2020; 36(4): 1150-8.
48. Bernstein NJ, Fong NL, Lam I, Roy MA, Hendrickson DG, Kelley DR. Solo: Doublet Identification in Single-Cell RNA-Seq via Semi-Supervised Deep Learning. Cell Systems. 2020; 11(1): 95-101.e5.

49. DePasquale EAK, Schnell DJ, Van Camp P, Valiente-Alandí I, Blaxall BC, Grimes HL, et al. DoubletDecon: Deconvoluting Doublets from Single-Cell RNA-Sequencing Data. Cell Reports. 2019; 29(6): 1718-27.e8.
50. Wainer-Katsir K, Linial M. BIRD: identifying cell doublets via biallelic expression from single cells. Bioinformatics. 2020; 36: i251-7.
51. Xiong K, Zhou H, Yin J, Kristianses K, Yang H, Li G. Chord: Identifying Doublets in Single-Cell RNA Sequencing Data by an Ensemble Machine Learning Algorithm. bioRxiv. 2021. doi: 10.1101/2021.05.07.442884.
52. Sun B, Bugarin-Estrada E, Overend LE, Walker CE, Tucci FA, Bashford-Rogers RJM. Double-jeopardy: scRNA-seq doublet/multiplet detection using multi-omic profiling. Cell Reports Methods. 2021; 1(1): 100008.
53. Xin H, Lian Q, Jiang Y, Luo J, Wang X, Erb C, Xu Z, et al. GMM-Demux: sample demultiplexing, multiplet detection, experiment planning, and novel cell-type verification in single cell sequencing. Genome Biology. 2020; 21(1): 188.
54. Xi NM, Li JJ. Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data. Cell Systems. 2021; 12(2): 176-94.e6.
55. Germain P, Sonreal A, Robinson MD. pipeComp, a general framework for the evaluation of computational pipelines, reveals performant single cell RNA-seq preprocessing tools. Genome Biology. 2020; 21(1): 227.
56. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P. Simultaneous epitope and transcriptome measurement in single cells. Nature Methods. 2017; 14(9): 865-8.
57. Peterson VM, Zhang KX, Kumar N, Wong J, Li L, Wilson DC, et al. Multiplexed quantification of proteins and transcripts in single cells. Nature Biotechnology. 2017; 35(10): 936-9.

58. Mimitou EP, Lareau CA, Chen KY, Zorzetto-Fernandes AL, Hao Y, Takeshima Y, et al. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. Nature Biotechnology. 2021. doi: 10.1038/s41587-021-00927-2.
59. Lareau CA, Duarte FM, Chew JG, Kartha VK, Burkett ZD, Kohlway AS, et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. Nature Biotechnology. 2019; 37(8): 916-24.
60. Satpathy AT, Granja JM, Yost KE, Qi Y, Meschi FM, McDermott GP, et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nature Biotechnology. 2019; 37(8): 925-36.
61. Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. Science. 2015; 348(6237): 910-4.
62. Granja JM, Corces MR, Pierce SE, Bagdatli ST, Choudhry H, Chang HY, Greenleaf WJ. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nature Genetics. 2021; 53(3): 403-11.
63. Thibodeau A, Eroglu A, McGinnis CS, Lawlor N, Nehar-Belaid D, Kursawe R, et al. AMULET: a novel read count-based method for effective multiplet detection from single nucleus ATAC-seq data. Genome Biology. 2021; *in press*.

Chapter 3: MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices

Elements of the following chapter are reprinted from the manuscript “MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices” by Christopher S. McGinnis, David M. Patterson, Juliane Winkler, Daniel N. Conrad, Marco Y. Hein, Vasudha Srivastava, Jennifer L. Hu, Lyndsay M. Murrow, Jonathan S. Weissman, Zena Werb, Eric D. Chow, and Zev J. Gartner, published in *Nature Methods* on June 17, 2019, Vol. 16, Issue 7, pages 619-626.

3.1 Abstract

Sample multiplexing facilitates scRNA-seq by reducing costs and identifying artifacts such as cell doublets. However, universal and scalable sample barcoding strategies have not been described. We there3fore developed MULTI-seq: multiplexing using lipid-tagged indices for single-cell and single-nucleus RNA sequencing. MULTI-seq reagents can barcode any cell type or nucleus from any species with an accessible plasma membrane. The method involves minimal sample processing, thereby preserving cell viability and endogenous gene expression patterns. When cells are classified into sample groups using MULTI-seq barcode abundances, data quality is improved through doublet identification and recovery of cells with low RNA content that would otherwise be discarded by standard quality-control workflows. We use MULTI-seq to track the dynamics of T-cell activation, perform a 96-plex perturbation experiment with primary human mammary epithelial cells and multiplex cryopreserved tumors and metastatic sites isolated from a patient-derived xenograft mouse model of triple-negative breast cancer.

3.2 Introduction

Single-cell and single-nucleus RNA sequencing (scRNA-seq, snRNA-seq) have emerged as powerful technologies for interrogating the heterogeneous transcriptional profiles of multicellular systems. Early scRNA-seq workflows were limited to analyzing tens to hundreds of single-cell transcriptomes at a time [1,2]. With the advent of single-cell sequencing technologies based on microwell [3], split-pool barcoding [4], and droplet-microfluidics [5-8] the parallel transcriptional analysis of 10^3-10^5 cells or nuclei is now routine. This increase in cell-throughput has catalyzed efforts to characterize the composition of organs [9] and entire organisms [4,10].

These technologies will increasingly be used to reveal the mechanisms by which cell populations interact to promote development, homeostasis, and disease. This shift from descriptive to mechanistic analyses requires integrating spatiotemporal information, diverse perturbations, and experimental replicates in order to draw strong conclusions [11,12]. While existing methods can assay many thousands of cells, sample-specific barcodes (e.g., Illumina library indices) are incorporated at the very end of standard library preparation workflows, which limits scRNA-seq sample-throughput due to reagent costs and the physical constraints of droplet microfluidics devices. Sample multiplexing approaches address this limitation by labeling cells with sample-specific barcodes prior to pooling and single-cell isolation. Several multiplexing methods have been described that distinguish samples using pre-existing genetic diversity [13-16], or introduce sample barcodes using either genetic [17-22] or non-genetic [23-25] mechanisms. However, each of these methods has liabilities, including issues with scalability, universality, and the potential to introduce secondary perturbations to experiments.

We identified lipid- and cholesterol-modified oligonucleotides (LMOs, CMOs) as reagents that circumvent many of the limitations of other sample multiplexing techniques. We previously described LMO and CMO scaffolds that rapidly and stably incorporate into the plasma membrane

of live cells by step-wise assembly [28]. Here, we adapt LMOs and CMOs into MULTI-seq – scRNA-seq and snRNA-seq sample multiplexing using lipid-tagged indices. MULTI-seq localizes sample barcodes to live cells and nuclei regardless of species or genetic background. MULTI-seq is non-perturbative, rapid, and involves minimal sample processing. Here, MULTI-seq simplicity and modularity enabled the analysis of a T-cell activation time-course, 96 human mammary epithelial cell (HMEC) culture conditions, and cryopreserved primary cells isolated from patient-derived xenograft (PDX) mouse models at varying stages of metastatic progression.

3.3 Results

3.3.1 MULTI-seq overview, comparison of LMO and CMO performance on intact cells and isolated nuclei using flow cytometry

MULTI-seq localizes DNA barcodes to plasma membranes by hybridization to an 'anchor' LMO. The 'anchor' LMO associates with membranes through a hydrophobic 5' lignoceric acid amide. Subsequent hybridization to a 'co-anchor' LMO incorporating a 3' palmitic acid amide increases the hydrophobicity of the complex and prolongs membrane retention (**Fig. 3-1**).

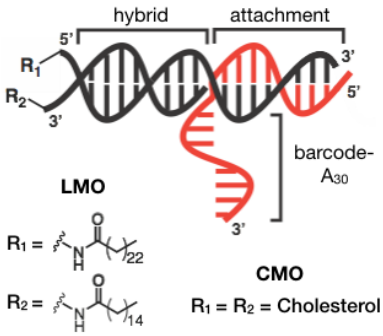

Figure 3-1: MULTI-seq Design. Diagram of the anchor/co-anchor scaffolds (black) with hybridized sample barcode oligonucleotide (red). LMOs and CMOs are distinguished by their unique lipophilic moieties (e.g., lignoceric acid, palmitic acid, or cholesterol).

MULTI-seq sample barcodes include a 3' poly-A capture sequence, an 8 base-pair sample barcode, and a 5' PCR handle necessary for library preparation and anchor hybridization. Cells

or nuclei carry membrane-associated MULTI-seq barcodes into emulsion droplets where the 3' poly-A domain mimics endogenous transcripts during hybridization to mRNA capture beads. Endogenous transcripts and MULTI-seq barcodes are then linked to a common cell- or nucleus-specific barcode during reverse transcription, which enables sample demultiplexing. MULTI-seq barcode and endogenous expression libraries are separated by size selection prior to next-generation sequencing library construction, enabling pooled sequencing at user-defined proportions. The same strategy can be applied to commercially-available CMOs.

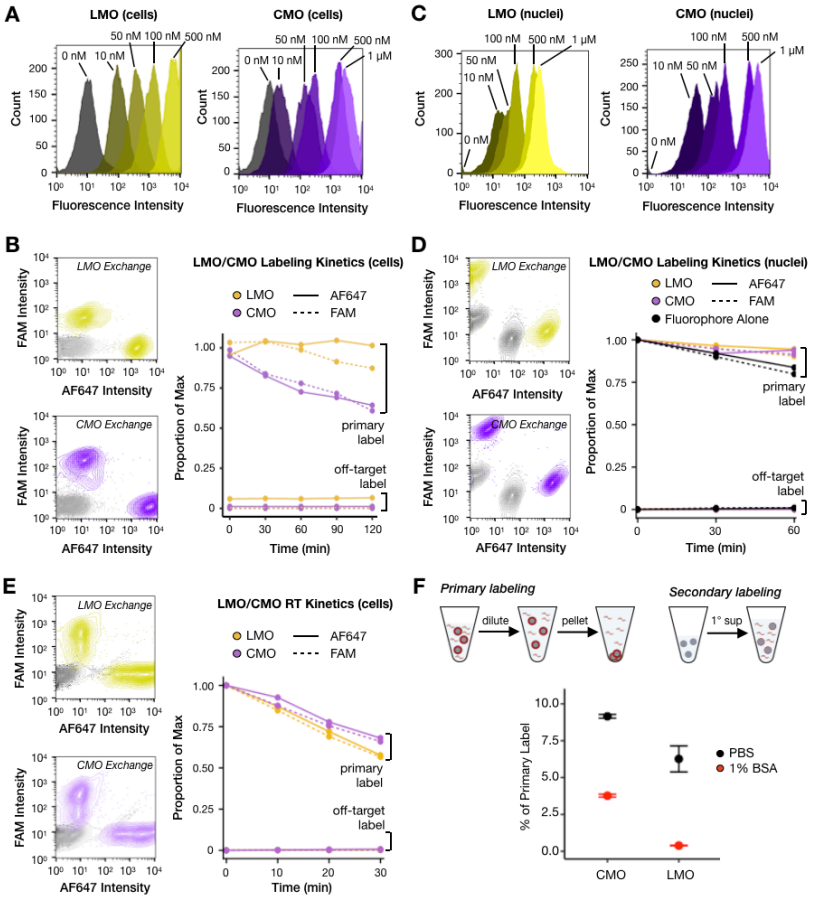

Figure 3-2: Flow cytometry demonstrates robust LMO and CMO labeling efficiency on living cells and nuclei, label stability over time, and LMO quenching with BSA.
(A) Live-cell LMO (gold) and CMO (purple) labeling efficiency varies predictably across anchor and co-anchor concentrations. n = 10,000 events/sample.
(B) Qualitative trends (contour plots, left) and quantitative analysis (scatter plot, right) of time-course experiment with live cells on ice. LMO (gold) and CMO (purple) labeled cells maintain fluorescence signal over unlabeled control cells (grey) over time, with LMOs exhibiting more membrane stability.
(C) As in (B) but with nuclei on ice. LMOs and CMOs exhibit similar nuclear membrane stability.
(D) As in (B) except with live cells at room temperature (RT). The LMO advantage in label stability shown at 4 °C is lost at RT as both CMO (purple) and LMO (gold) labels decrease at similar rates.
(E) LMO- or CMO-labeled cells were diluted with PBS (black) or 1% BSA in PBS (red). Supernatant was then transferred to unlabeled cells to extent of BSA quenching off-target membrane labeling.

We used flow cytometry to evaluate whether LMOs and CMOs predictably label and minimally exchange between live cells at typical sample preparation temperatures of 4 °C (**Figs. 3-2a, 3-2b**). Identical experiments were also performed using freshly-isolated nuclei (**Figs. 3-2c, 3-2d**). These data revealed that LMOs exhibit longer membrane residency times than CMOs on live-cell membranes at 4 °C, whereas LMOs and CMOs exchange comparably between live cells at room temperature (**Fig. 3-2e**), suggesting cells should be maintained on ice to achieve optimal sample multiplexing results. For nuclei, both oligonucleotide conjugates showed minimal exchange between nuclear membranes (**Fig. 3-2d**), however, bovine serum albumin (BSA) in nuclei isolation buffer specifically quenched LMOs, reducing labeling efficiency (**Fig. 3-2b**). While problematic during nuclei labeling, we reasoned that LMO quenching could be strategically employed to reduce off-target barcoding and potentially minimize washes prior to sample pooling. Indeed, we found that diluting LMO-labeling reactions with 1% BSA in PBS resulted in 18-fold lower off-target labeling following pooling compared to dilution with PBS (**Fig. 3-2f**).

3.3.2 MULTI-seq enables live-cell scRNA-seq sample demultiplexing

We tested the capacity of MULTI-seq to demultiplex scRNA-seq samples by performing a proof-of-concept experiment using HEK293 cells (HEKs) and primary human mammary epithelial cells (HMECs) cultured in the presence or absence of TGF-β (**Fig. 3-3**). Cells were trypsinized,

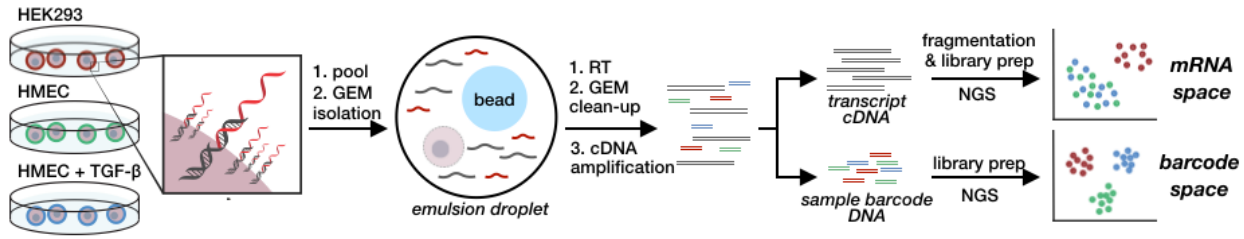

Figure 3-3: Proof-of-concept MULTI-seq experimental design. Three samples (HEKs and HMECs with and without TGF-β stimulation) were barcoded with either LMOs or CMOs and sequenced alongside unlabeled controls. MULTI-seq barcodes are captured alongside mRNA during reverse transcription (RT) within gel bead-in-emulsions (GEMs) and size-separated from endogenous cDNA following library amplification. MULTI-seq and cDNA libraries are then sequenced as a pool using next-generation sequencing (NGS), producing coupled gene expression and MULTI-seq barcode count matrices.

barcoded with LMOs or CMOs, and pooled prior to droplet microfluidic emulsion with the 10X Genomics Chromium system. In parallel, we prepared unbarcoded replicates to test whether MULTI-seq influenced gene expression or mRNA capture efficiency.

Following data pre-processing, we analyzed a final scRNA-seq dataset containing 14,377 total cells. We identified clusters in gene expression space according to known markers for HEKs as well as the two cellular components of HMECs, myoepithelial (MEPs) and luminal epithelial cells (LEPs, **Fig. 3-4a**). Projecting MULTI-seq barcode classifications onto gene expression space for LMO- and CMO-labeled cells (**Fig. 3-4b**) illustrates that both membrane scaffolds successfully demultiplexed each sample. Moreover, HMECs predicted to have been cultured with TGF-β exhibited enriched expression of the TGF-β-induced gen, TGFBI (data not shown), further

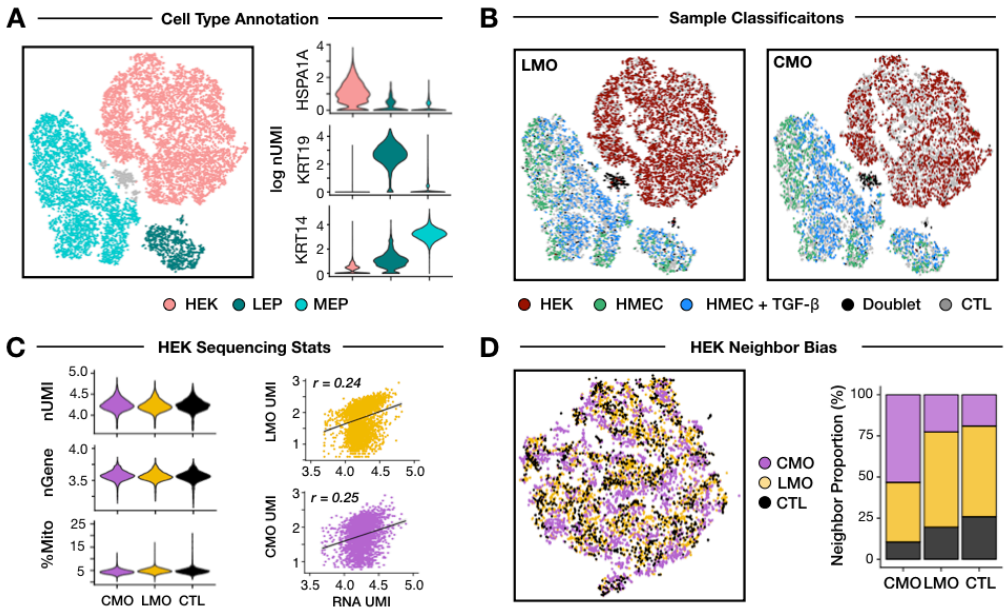

Figure 3-4: MULTI-seq using LMOs and CMOs successfully multiplexes scRNA-seq samples, CMOs induce subtle transcriptional response to labeling.
(A) Cell state annotations for aggregated LMO, CMO, and unlabeled control scRNA-seq data mapped onto t-SNE gene expression space (left). Violin plots (right) depict marker genes used to define HEKs (pink), MEPs (cyan), and LEPs (dark teal).
(B) MULTI-seq classifications for LMO- and CMO-labeled cells mapped onto gene expression space. Sample classifications match their expected cell type annotations.
(C) Violin plots (left) describing the number of detected UMIs and genes and the percentage of mitochondrial gene expression for LMO-labeled (gold), CMO-labeled (purple), and unlabeled control HEKs (black). Scatter plots (right) illustrating that MULTI-seq and RNA UMIs are not negatively correlated. These two observations suggest that MULTI-seq barcodes do not significantly interfere with mRNA capture.
(D) Gene expression space (left) for LMO-labeled (gold), CMO-labeled (purple), or unlabeled HEKs (black) reveals sub-structure specific to CMO-labeling. Sub-structure manifests as CMO-labeled cells having more CMO-labeled neighbors amongst each cell's 100 nearest neighbors in gene expression space (right).

demonstrating the accuracy of MULTI-seq classification results. Importantly, RNA and MULTI-seq barcode UMI counts in HEKs were not negatively correlated (**Fig. 3-4c**) and LMO-labeled HEKs did not exhibit distinct gene expression space localization patterns relative to unlabeled controls (**Fig. 3-4d**). Collectively, these results suggest that MULTI-seq does not impair mRNA capture, and that LMO-labeling (unlike CMO-labeling) does not induce a significant transcriptional response under the test conditions (Methods).

3.3.3 Demultiplexing single nucleus RNA-seq (snRNA-seq) and time-course experiments

snRNA-seq is widely used for the analyses of solid tissues that are difficult to dissociate [29]. We explored whether MULTI-seq could demultiplex snRNA-seq samples by purifying nuclei from HEKs and mouse embryonic fibroblasts (MEFs) and labeling each pool of nuclei with LMOs or CMOs prior to snRNA-seq. In parallel, we multiplexed Jurkat cells treated with ionomycin and phorbol 12-myristate 13-acetate (PMA) at eight time points (0-24 hours) to track T-cell activation dynamics (**Fig. 3-5**).

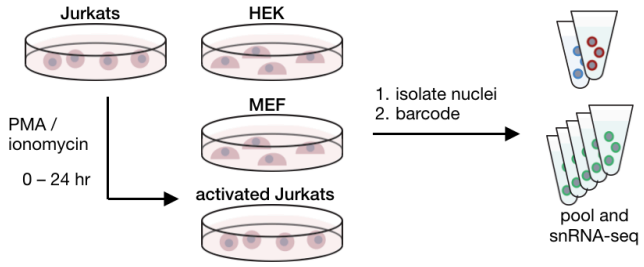

Figure 3-5: Schematic overview of a proof-of-concept snRNA-seq experiment using MULTI-seq. Nuclei were isolated from HEKs (red), MEFs (blue), and Jurkats stimulated with ionomycin and PMA for 8 distinct time points (green) prior to LMO barcoding and sequencing. CMO-labeled and unlabeled HEK and MEF nuclei were sequenced in parallel.

MULTI-seq sample classifications matched their intended cell type clusters with a ~0.5% misclassification rate (**Fig. 3-6a**). Notably, MULTI-seq classifications were species-specific and predicted ~85% of mouse-human doublets, which approximates the theoretical doublet detection limit of ~92%. Matching live-cell results, MULTI-seq barcoding did not impair mRNA capture (**Fig.**

3-6b). In contrast to live-cell results, both CMO- and LMO-labeled nuclei were transcriptionally indistinguishable from unbarcoded controls (**Fig. 3-6c**). Moreover, CMO-labeled nuclei had higher average signal-to-noise ratio (SNR) and total number of barcode UMIs relative to LMO-labeled nuclei (CMO SNR = 91.5, LMO SNR = 2.3), consistent with previous flow cytometry results.

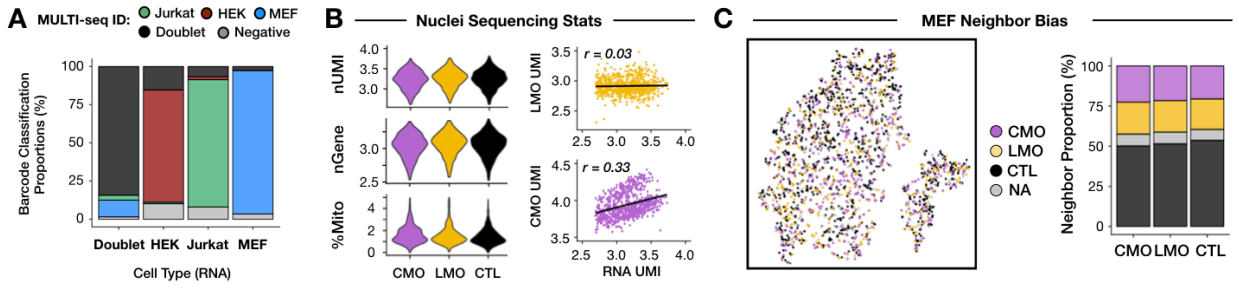

Figure 3-6: MULTI-seq using LMOs and CMOs successfully multiplexes snRNA-seq samples.
(A) MULTI-seq sample classification proportions for each cell type identified by clustering in gene expression space.
(B) Same analysis as described in Fig. 3-4c on snRNA-seq data.
(C) Same analysis as described in Fig. 3-4d on MEFs present in multiplexed snRNA-seq data.

Upon demultiplexing individual time points along the trajectory of T-cell activation (**Fig. 3-7a**) we observed multiple literature-supported transcriptional dynamics (**Fig. 3-7b**). For example, genes undergoing early down-regulation (e.g., TSHR) [30] and transient (e.g., DUSP2) [31], sustained (e.g., CD69) [32], and late (e.g., GRZA) [33] up-regulation were readily identified.

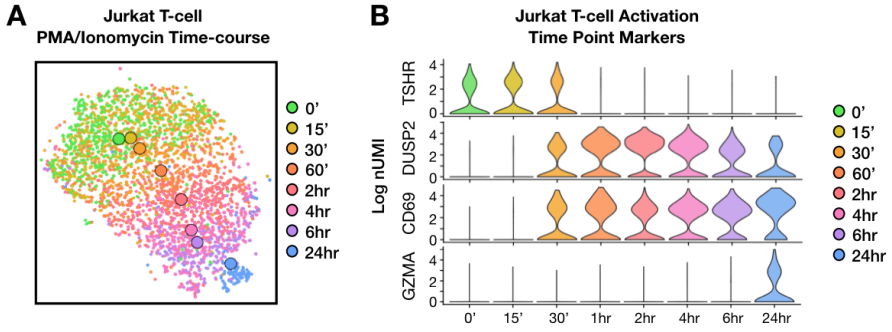

Figure 3-7: MULTI-seq enables snRNA-seq time-course analysis of Jurkat T-cell activation with PMA and ionomycin.
(A) MULTI-seq sample classifications illuminate temporal gene expression patterns in Jurkat cells following activation with ionomycin and PMA for varying amounts of time. Time-point centroids in gene expression space are denoted with larger circles.
(B) Violin plots of gene expression marking different stages of T-cell cell activation.

3.3.4 MULTI-seq sample classification and doublet identification algorithm

We next sought to demonstrate MULTI-seq scalability by multiplexing 96 unique HMEC samples spanning a range of microenvironmental conditions. We exposed duplicate cultures consisting of MEPs, LEPs, and both cell types grown in M87A media [34] without EGF to 15 physiologically-relevant signaling molecules [35] or signaling molecule combinations (**Fig. 3-8**). We barcoded each sample before pooling and loaded cells across three 10X microfluidics lanes, resulting in a 32-fold reduction in reagent use relative to standard practices.

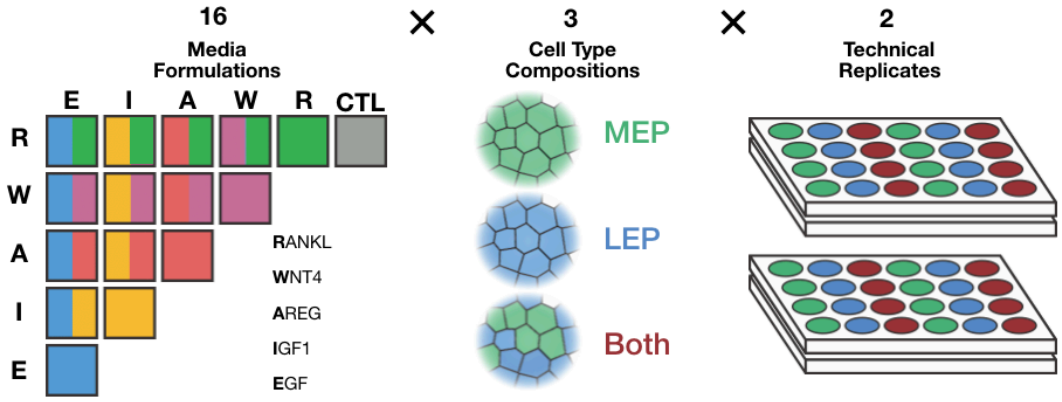

Figure 3-8: Schematic overview of 96-plex HMEC MULTI-seq experimental design. 96 distinct HMEC cultures consisting of LEPs alone (blue), MEPs alone (green), or both cell types together (dark red) were grown in media supplemented with 15 distinct signaling molecules or signaling molecule combinations and one control.

To classify HMECs into sample groups, we implemented a sample classification workflow inspired by previous strategies [17,18,23]. Briefly, following pre-processing of the raw MULTI-seq data (**Fig 3-9**, Step 1), MULTI-seq barcode count matrices are normalized and a probability density function (PDF) is computed for each sample barcode.. Next, local maxima for each PDF are identified before 'negative' and 'positive' maxima are assigned to the most frequent and highest local maxima, respectively. Next, with positive and negative bounds for each MULTI-seq barcode count distribution identified, an inter-maxima threshold quantile sweep is performed wherein the number of singlets (i.e., cells surpassing a single barcode threshold), doublets (i.e., cells surpassing >1 threshold), and negatives (i.e., cells surpassing 0 thresholds) is determined

for each quantile threshold value. Next, the inter-maxima quantile threshold resulting in the highest number of singlet classifications is applied to all barcode PDFs to codify sample classifications. Negative cells are then removed and this procedure is repeated until all cells are classified as singlets or doublets (**Fig. 3-9**, Step 2). Finally, subsets of negative cells are then reclassified using a semi-supervised classification approach, where singlets defined during the initial workflow are used to initialize cluster centers during k-means clustering of negative cells (**Fig. 3-9**, Step 3). See the Materials & Methods section for more information.

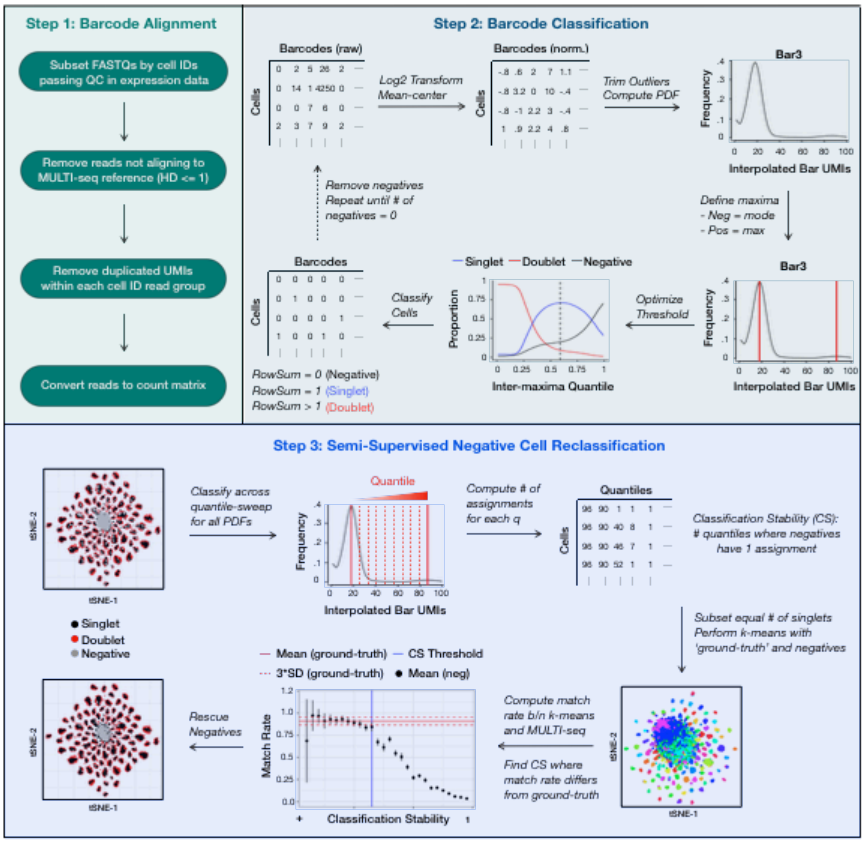

Figure 3-9: MULTI-seq barcode pre-processing and sample classification workflows. Results from the 96-plex HMEC experiment are used as representative examples for the barcode classification workflow. Results from the 96-plex technical replicate HMEC experiment as used as representative examples for the semi-supervised negative cell reclassification workflow. PDF = probability density function.

Application of this sample classification workflow to our 96-plex HMEC MULTI-seq dataset resulted in the identification of 76 sample groups consisting of 26,439 total cells. Each group was

exclusively enriched for a single barcode (**Figs. 3-10a, 3-10b**) an average of ~199-fold above the most abundant off-target barcode (**Fig. 3-10c**). Unlike sample multiplexing data with relatively few samples, MULTI-seq-defined doublets localized to the peripheries of singlet clusters in barcode space for this experiment (**Fig. 3-10a**, right). To understand why 96 samples were not robustly identified in this experiment, we visualized the total number of classified singlets mapping to each well of the original 96-well plate (**Fig. 3-10d**). This analysis revealed that a significant proportion of missing barcodes came from a single column of the 96-well plate, which suggests that samples were lost due to pipette handling errors.

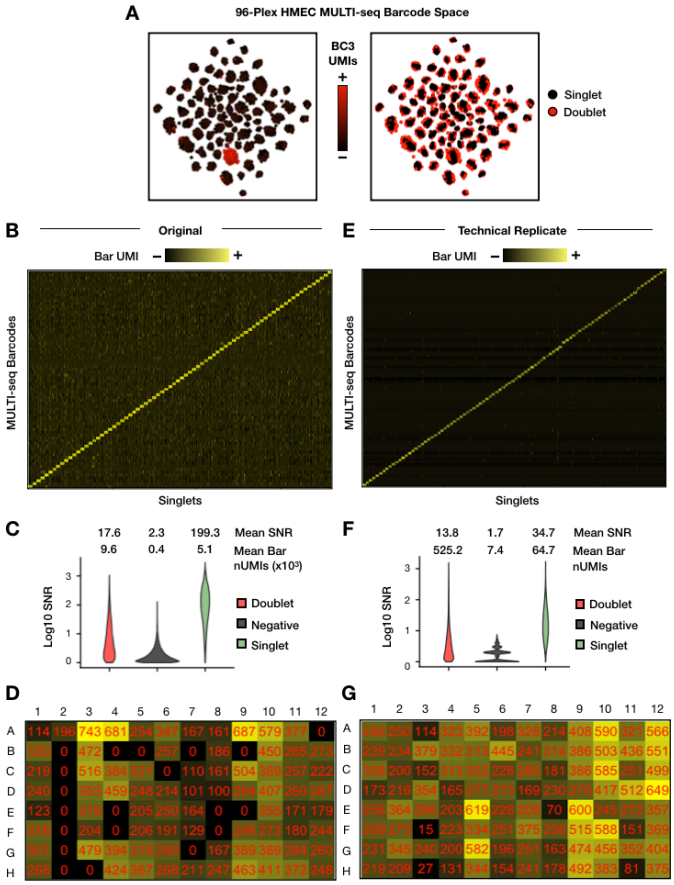

Figure 3-10: 96-plex HMEC MULTI-seq sample classification results.
(A) Barcode UMI abundances (left) and doublet classifications (right) mapped onto barcode space. MULTI-seq barcode #3 is used as a representative example. Doublets localize to the peripheries of sample groups in large-scale sample multiplexing experiments.
(B) Normalized barcode UMI heat map demonstrating that sample groups are predominantly associated with single MULTI-seq barcodes.
(C) Violin plots describing the mean barcode UMIs and SNR for negative cells, doublets, and singlets.
(D) 96-well plate schematic overlaid with a heat map showing the number of cells assigned to each sample barcode group. Twenty samples — predominantly those arising from column 2 — were not represented in the original large-scale HMEC experiment due to technical error during sample preparation.
(E-G) Same analyses as described in 3-10b, 3-10c, and 3-10d except for 96-plex HMEC technical replicate experiment.

Indeed, a technical replicate of this experimental design yielded successful classifications for all 96 MULTI-seq barcodes (**Figs. 3-10e, 3-10f, 3-10g**).

To assess demultiplexing accuracy, we grouped MULTI-seq classifications according to cell type composition (e.g., MEPs alone, LEPs alone, or both) and visualized these groups in gene expression space. Unsupervised clustering and marker analysis of the resulting transcriptome data distinguished LEPs from MEPs along with a subset of ambiguous cells expressing markers for both cell types (**Fig. 3-11a**). MULTI-seq classifications matched their expected cell type clusters (**Fig. 3-11b**), while cells co-expressing MEP and LEP markers were predominantly defined as doublets. MULTI-seq identified doublets that were overlooked when predicting doublets using marker genes (**Fig. 3-11b,** arrow).

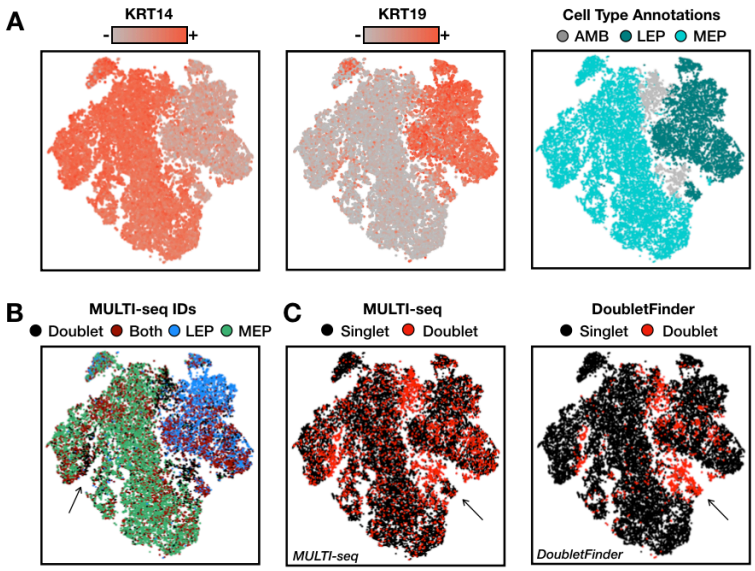

Figure 3-11: Benchmarking 96-plex HMEC MULTI-seq sample classification and doublet identification results against marker-based cell type annotations and computational doublet prediction algorithms.
(A) Distributions of marker gene expression used to identify MEPs (KRT14, left) and LEPs (KRT19, middle) in gene expression space. Cell type annotations for MEPs (cyan) and LEPs (dark teal) in gene expression space (right). Ambiguous cells positive for multiple marker genes are displayed in grey.
(B) MULTI-seq classifications grouped by culture composition: LEP-alone (blue), MEP-alone (green), and both cell types together (dark red). Discordant region where annotated MEPs are classified as MULTI-seq doublets (black) is indicated with arrow.
(C) MULTI-seq doublet classifications (left) and computational predictions produced by DoubletFinder (right). Discordant region where DoubletFinder-defined doublets are classified as MULTI-seq singlets is indicated with arrows.

Additionally, MULTI-seq doublet classifications generally agreed with computational predictions (**Fig. 3-11c**, Sensitivity = 0.283 Specificity = 0.965), with the exception of 'homotypic'

doublets, to which computational doublet detection techniques are insensitive [36,37]. Moreover, DoubletFinder erroneously classified proliferative LEPs as doublets (**Fig. 3-11c**, arrow), illustrating how computational doublet inference performance suffers when applied to datasets with low cell type numbers [36,37].

3.3.5 MULTI-seq identifies transcriptional responses to co-culture conditions and signaling molecules in HMECs

Sample demultiplexing, doublet removal, and quality-control filtering resulted in a final scRNA-seq dataset including 21,753 total cells, revealing two transcriptional responses linked to culture composition. First, we observed that LEPs co-cultured with MEPs exhibited enriched proliferation relative to LEPs cultured alone (**Fig. 3-12a**). In contrast, MEPs were equally

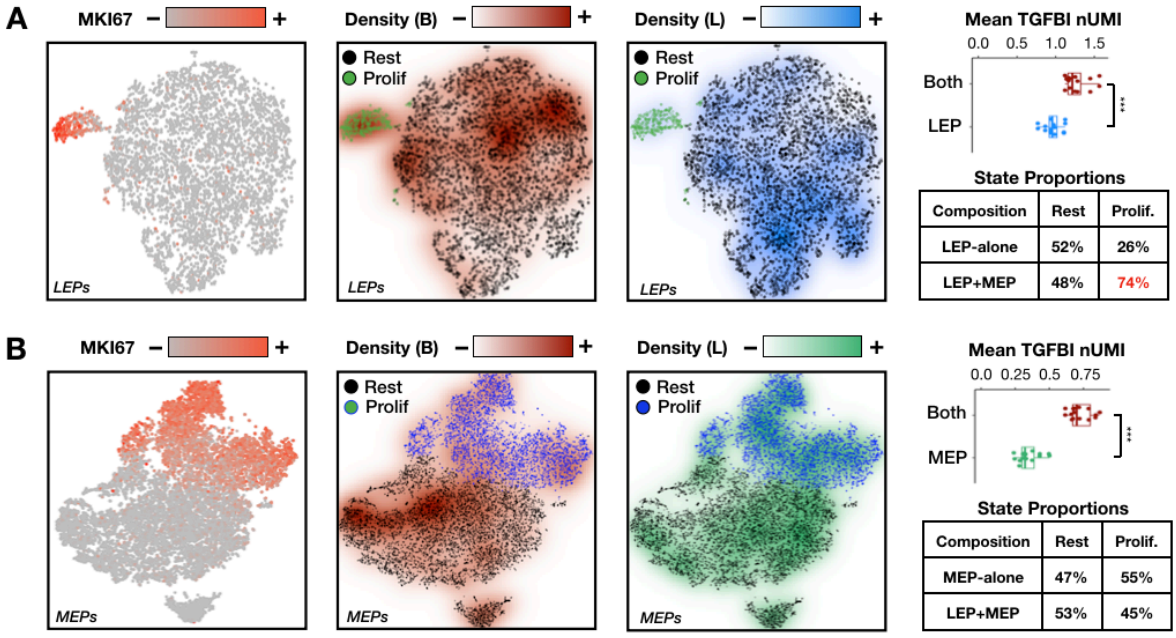

Figure 3-12: LEP-MEP co-culture induces TGF-β paracrine signaling, enrichment in proliferative LEPs.
(A) Sample classification densities for co-cultured LEPs (dark red, middle left) and LEPs cultured alone (blue, middle right) projected onto gene expression space containing resting (black) and proliferative (green) LEPs identified using MKI67 expression (left). LEP-MEP co-culture conditions are associated with enhanced TGFBI expression (right; *** = Wilcoxon rank sum test (two-sided), $p = 3.1 \times 10^{-6}$) and proliferation (table).
(B) Sample classification densities for co-cultured MEPs (dark red, middle left) and MEPs cultured alone (green, middle right) projected onto gene expression space containing resting (black) and proliferative (blue) MEPs identified using MKI67 expression (left). LEP-MEP co-culture conditions are associated with enhanced TGFBI expression (right; *** = Wilcoxon rank sum test (two-sided), $p = 1.5 \times 10^{-6}$) but not proliferation (table).

proliferative when cultured alone or with LEPs (**Fig. 3-12b**). Second, we observed that non-proliferative co-cultured MEPs and LEPs are enriched for TGFBI expression relative to MEPs and LEPs cultured alone (**Fig. 3-12c**).

We next used hierarchical clustering to assess how LEPs or MEPs responded to signaling molecule exposure. HMECs exposed to the EGFR ligands AREG and EGF exhibited gene expression profiles that were significantly different from control cells. AREG- and EGF-stimulated LEPs expressed increased levels of EGFR signaling genes (e.g., DUSP4) [38] and genes up-regulated in HER2+ breast cancers (e.g., PHLDA1) [39] relative to control LEPs (**Fig. 3-13a**. AREG- and EGF-stimulated MEPs also express high levels of known EGFR-regulated genes (e.g., ANGPTL4; **Fig. 3-13b**) [40].

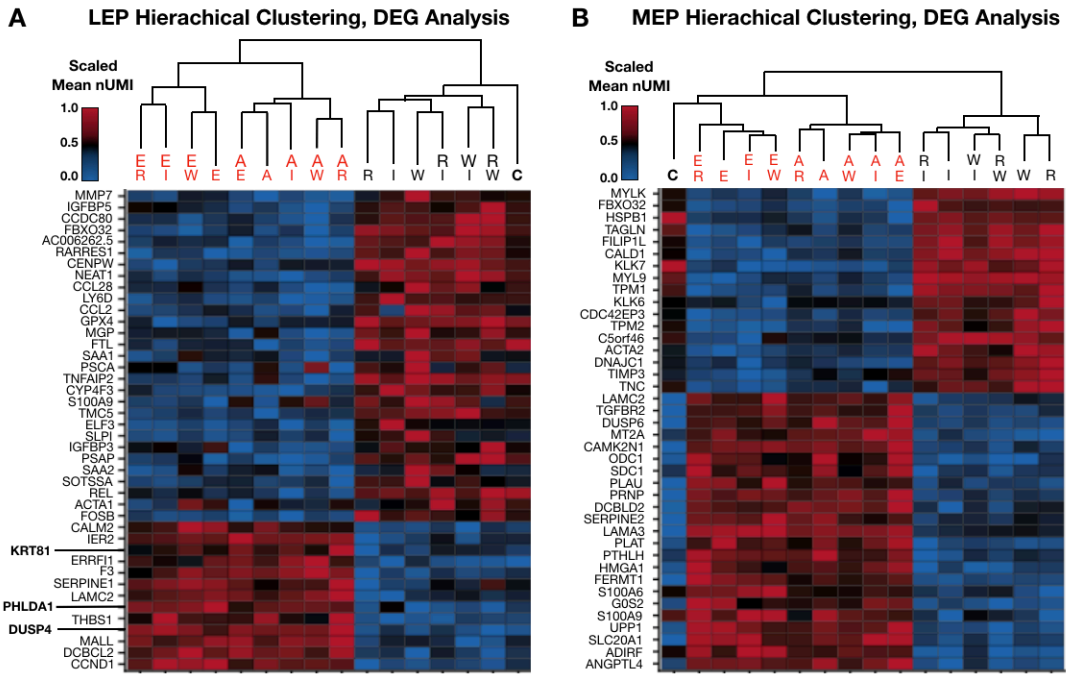

Figure 3-13: Detection of EGFR signaling responses in MEPs and LEPs.
(A) Hierarchical clustering and differential gene expression (DEG) analysis results summarized as annotated heatmap for resting LEPs grouped by treatment. Emphasized genes are known EGFR signaling targets. RNA UMI abundances are scaled from 0-1 for each gene. Values correspond to the average expression within each signaling molecule treatment group. Dendrogram labels: E = EGF, W = WNT4, A = AREG, I = IGF-1, R = RANKL, C = Control.
(B) Same analysis as described in Fig. 3-13a except for resting MEPs.

3.3.6 MULTI-seq identifies low-RNA cells in cryopreserved, primary PDX samples

Using scRNA-seq to analyze archival primary tissue samples is often difficult because these samples can have low cell viability that is compounded during cryopreservation, thawing, enzymatic digestion, and scRNA-seq sample preparation. We investigated whether the rapid and non-perturbative nature of MULTI-seq barcoding would enable cryopreserved tissue multiplexing using samples dissected from a PDX mouse model of metastatic triple-negative breast cancer [41]. In this model system, the diameter of primary tumors was used as a proxy for metastatic

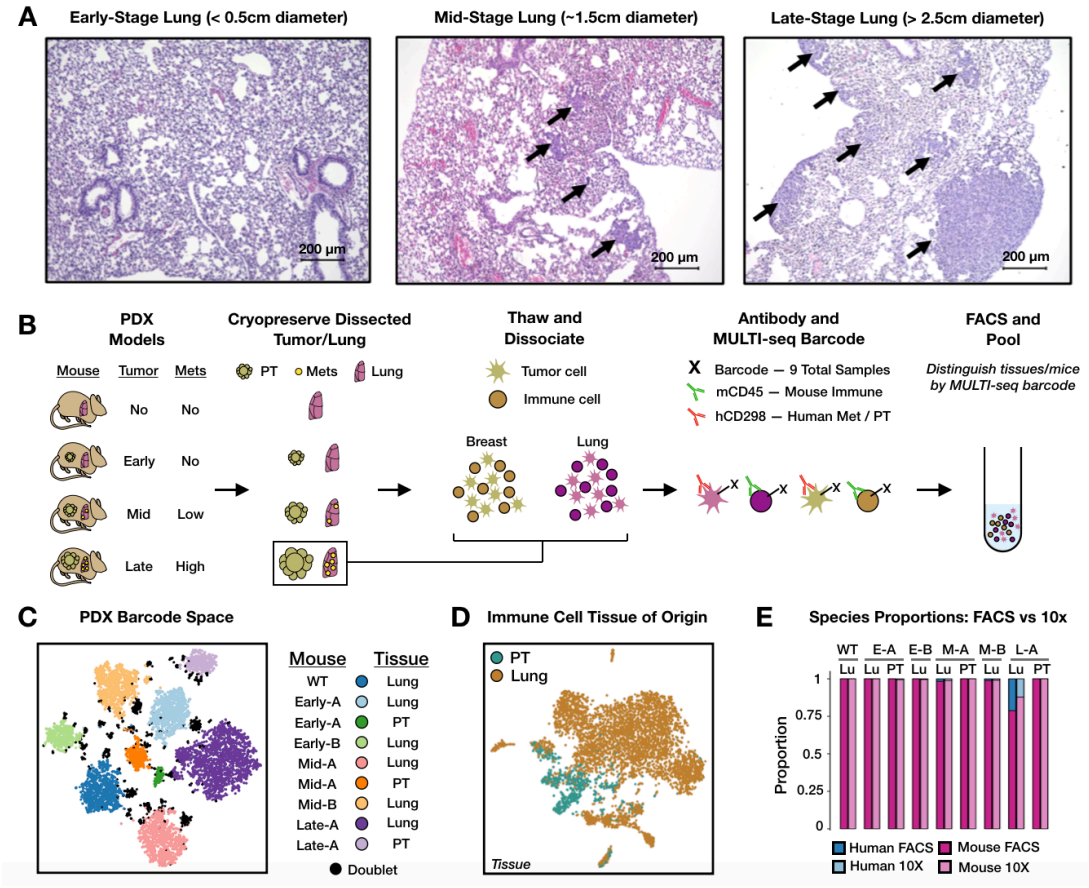

Figure 3-14. MULTI-seq application to cryopreserved organs isolated from PDX models of triple negative breast cancer.
(A) Representative histology of lung tissue illustrates metastatic progression in early, mid, and late-stage PDX mice. Metastases denoted with black arrows.
(B) Schematic overview of MULTI-seq PDX experiment.
(C) MULTI-seq sample classifications (WT, early, mid, late tumor progression) mapped onto barcode space. Replicate tissues are denoted as 'A' or 'B'.
(D) Mouse immune cells in gene expression space colored according to tissue of origin. Lung immune cells (brown) cluster separately from primary tumor immune cells (teal)
(E) Bar plots describing the proportion of mouse (pink) and human (blue) cells detected during FACS enrichment and detected in the final 10x dataset.

progression in the lung (**Fig. 3-14a**). We barcoded 9 distinct samples representing primary tumors and lungs from early- and mid-stage PDX mice (in duplicate), one late-stage PDX mouse, and a single lung from an immunodeficient mouse without tumors (**Fig. 3-14b**). We then pooled FACS-enriched populations of barcoded hCD298+ human metastases with mCD45+ mouse immune cells prior to "super-loading" a single 10X Genomics microfluidics lane.

Quality-control filtering, sample classification, and doublet removal resulted in a final scRNA-seq dataset of 9,110 mouse and human singlets spanning all 9 samples (**Fig. 3-14c**). Classification accuracy was supported by tissue-specific gene expression patterns (**Fig. 3-14d**) and comparisons to FACS enrichment results (**Fig. 3-14e**). Additionally, MULTI-seq classifications identified high-quality single-cell transcriptomes that would have been discarded using standard quality-control workflows (e.g., CellRanger RNA UMI inflection point threshold = 1350, **Fig. 3-15**). When comparing cells with 100-1350 RNA UMIs, classified cells included immune cell types that are difficult to detect using single-cell and bulk transcriptomics (e.g., neutrophils) [42]. Strikingly, 90.8% of sequenced neutrophils would have been discarded by

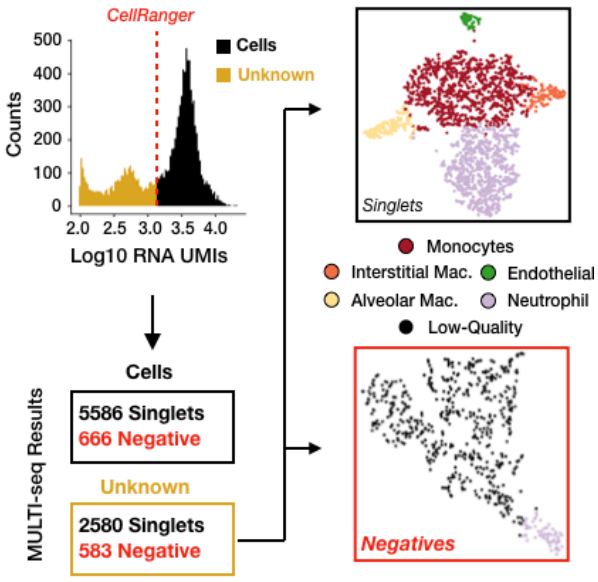

Figure 3-15. MULTI-seq classifications facilitate low-RNA and low-quality cell deconvolution. CellRanger discards cells barcodes with low RNA UMI counts (red dotted line). Gene expression profiles for classified low-RNA cells reflect established immune cell types (top right). Unclassified low-RNA cells resemble low-quality single-cell transcriptomes (bottom right).

CellRanger. In contrast, unclassified low-RNA cells had poor-quality gene expression profiles predominantly corresponding to broken cells (Methods).

3.3.7 Characterizing the lung immune response to metastatic progression

We next sought to describe how lung immune cells respond to metastatic progression. Beginning with a dataset comprised of 5,690 mCD45+ cells, we identified gene expression profiles associated with neutrophils, monocytes and macrophages (alveolar, interstitial, and (non)-classical monocytes), dendritic cells (mature, immature, Ccr7+, and plasmacytoid DCs), and endothelial cells [9,44] (**Fig. 3-16a**). The use of immunodeficient PDX mice resulted in a lack of lymphocytes (e.g., T, B, and NK cells).

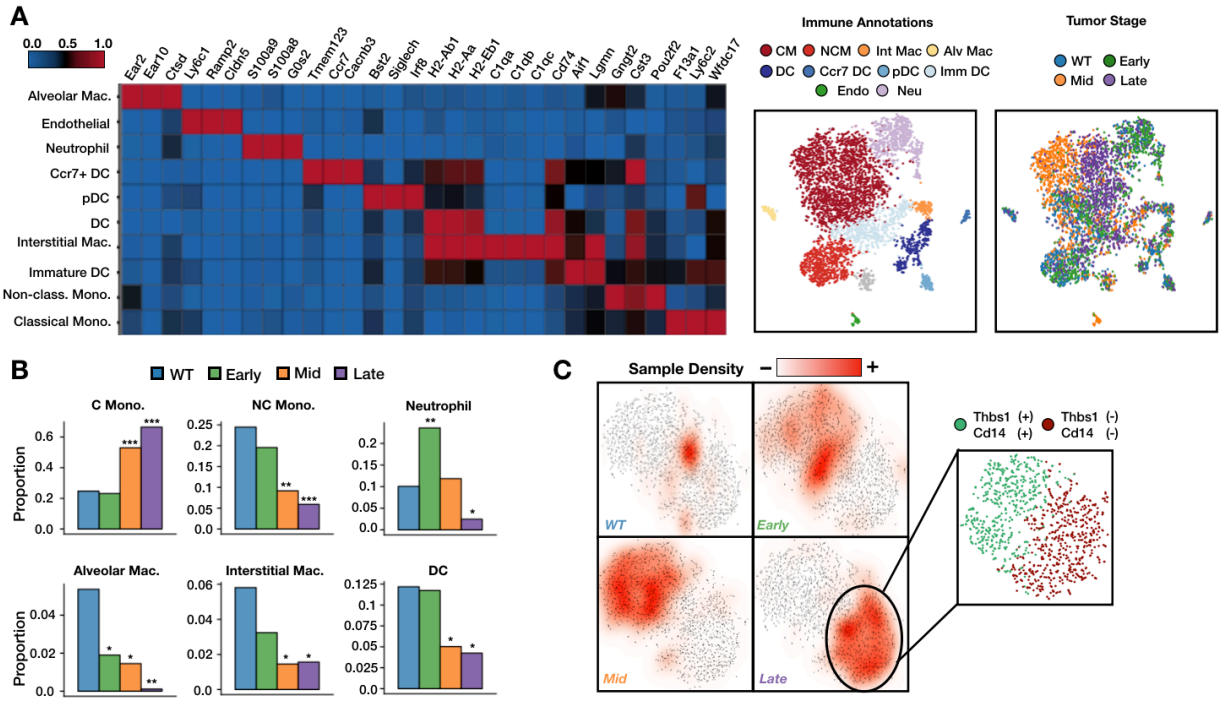

Figure 3-16. PDX sample multiplexing reveals immune cell proportional shifts and classical monocyte heterogeneity in the progressively metastatic lung.
(A) Marker gene heat map (left) describing markers utilized for defining immune cell type annotations. RNA UMI abundances are scaled from 0-1 for each gene. Mouse immune gene expression space embeddings colored by cell type annotations (left) and tumor stage (right) are shown to the right of the heat map. CM = Classical Monocyte, NCM = Non-Classical Monocyte, Mac = Macrophage, Int = Interstitial, Alv = Alveolar, DC = Dendritic Cell, pDC = Plasmacytoid DC, Imm DC = Immature DC, Endo = Endothelial cell, Neu = Neutrophil.
(B) Statistically-significant shifts in lung immune cell type proportions for each tumor stage relative to WT. Two-proportion z-test with Bonferroni multiple comparisons adjustment, * = $0.05 > p > 10^{-10}$; ** = $10^{-10} > p > 10^{-20}$; *** = $p < 10^{-20}$. n = 44 tumor-stage/cell type groups. Statistically-insignificant proportional shifts omitted.
(C) Subsetted classical monocyte gene expression space overlaid with sample classification densities corresponding to tumor stage. Inset illustrates heterogeneity within late-stage classical monocytes characterized by differential expression of Thbs1 and Cd14.

We observed literature-supported changes in immune cell proportions and transcriptional state at each tumor stage. For instance, neutrophils were enriched in early-stage PDX mice while alveolar macrophages were depleted over the course of metastasis (**Fig. 3-16b**) [45,46]. Moreover, stage-specific transcriptional heterogeneity among classical monocytes (CMs) reflects previous descriptions of lung CM state transitions in PDX breast cancer models [47-50]. Specifically, hierarchical clustering and DEG analysis of CMs revealed that CMs from late-stage PDX mice fell into two distinct transcriptional states discernible by Cd14 expression (**Fig. 3-16c**) matching previous observations [48]. Genes that are differentially-expressed between CM subsets include genes known to influence metastatic progression (e.g., Thbs1, S100a8/9, and Wfdc21) [47,49,50]. To discern whether the results were primarily attributable to inter-mouse variability, we used Earth Mover's Distance (EMD) [51] to quantify the magnitude of transcriptional dissimilarity between lung CMs from each mouse and tumor stage. These results illustrate that CMs from early- and mid-stage mouse replicates (scaled EMD = 0.16) were more similar than CMs from distinct tumor stages (scaled EMD = 0.69).

3.4 Discussion

MULTI-seq is an ideal sample multiplexing approach because it is scalable, universal, and improves scRNA-seq data quality. MULTI-seq is scalable because it uses inexpensive reagents, involves minimal sample handling, and is rapid and modular in design. MULTI-seq modularity enables any number of samples to be multiplexed with a single pair of 'anchor' and 'co-anchor' LMOs. Moreover, since LMOs are quenchable with BSA and can be incorporated during proteolytic dissociation, we anticipate that further method optimization will facilitate wash-free sample preparation workflows. When integrated with automated liquid handling, these features

position MULTI-seq as a powerful technology enabling 'screen-by-sequencing' applications (e.g., L10004 [52], DRUG-seq [53]) in multicellular systems (e.g., organoids, PBMCs, etc.).

In this study, we leveraged MULTI-seq scalability to perform a 96-plex HMEC perturbation assay, revealing noteworthy principles for future scRNA-seq sample multiplexing experiments. Specifically, we observed that responses to signaling molecules were less pronounced than responses linked to cellular composition. For instance, co-cultured MEPs and LEPs engage in TGF-β signaling that is absent in the associated monocultures. In contrast, MEPs and LEPs only exhibited pronounced transcriptional responses to the EGFR ligands AREG and EGF in these data, despite the established roles of all tested signaling molecules in mammary morphogenesis. We speculate that rich media formulations used to expand cells, such as the M87A media (-EGF) used here, likely buffer cells against microenvironmental perturbations. Thus, careful consideration of cell-type composition and media formulation will be essential to accurately interpret future scRNA-seq experiments.

Beyond its scalability, MULTI-seq improves scRNA-seq data quality in two distinct ways. First, MULTI-seq identifies doublets as cells associated with multiple sample indices. The ability to detect doublets allows for droplet-microfluidics devices to be "super-loaded", resulting in ~5-fold improvement in cellular throughput [13,23]. Moreover, unlike computational doublet prediction methods [36,37], MULTI-seq detects homotypic doublets and performs well on scRNA-seq data with minimal cell-type complexity. However, since computational doublet detection methods detect doublets formed from cells with shared sample barcodes, doublet detection should ideally involve a synergy of computational and molecular approaches.

Second, MULTI-seq improves scRNA-seq data quality by 'rescuing' cells that would otherwise be discarded by quality-control workflows utilizing RNA UMI thresholds. Such workflows are systematically biased against cell types with low RNA content [43]. MULTI-seq classifications provide an orthogonal metric to RNA UMIs for distinguishing low-RNA from low-quality cells. We

leveraged this feature (described initially by Stoeckius et al [23]) to improve the quality of the PDX dataset, where MULTI-seq classifications 'rescued' > 90% of the sequenced neutrophils while avoiding misclassification of broken cells.

Finally, MULTI-seq is universally applicable to any sample including cells or nuclei with an accessible plasma membrane. As a result, we used the same set of MULTI-seq reagents to multiplex 15 distinct cell types or nuclei from both mice and humans. Notably, CMOs outperformed LMOs in nuclei isolation buffers containing BSA because BSA sequesters LMOs. Additionally, we anticipate that MULTI-seq is compatible with sample preservation strategies such as flash-freezing and fixation.

We leveraged all three of these features – scalability, universality, and data quality improvement – to multiplex cryopreserved primary tumors and lungs dissected from PDX mouse models at varying stages of metastatic progression. PDX sample multiplexing requires barcoding cells from (i) multiple species that may (ii) down-regulate surface epitopes commonly targeted by antibody-based multiplexing techniques (e.g., MHC-1 [54]), and (iii) have intrinsically-low viability requiring minimal sample handling. MULTI-seq successfully demultiplexed every sample, revealing novel and literature-supported immune cell responses to metastatic progression in the lung. For example, while metastasis-associated shifts in neutrophil, alveolar macrophage, and CM proportions were previously observed, we described significant shifts in interstitial macrophages, dendritic cells, and non-classical monocytes that, to our knowledge, are novel and require further experimental validation.

Moreover, we identified CM subsets that were discernible by Cd14 expression and genes with diverse effects on metastatic progression. Perplexingly, Cd14-high CMs expressing the pro-metastatic gene Thbs1 [49] and CD14-low CMs expressing the anti-metastatic genes S100a8/9 and Wfdc21 [50] coexisted in metastasized lungs. Since we isolated immune cells from the whole lung in this study, we could not discern whether CD14-high and CD14-low states were spatially

correlated with metastatic sites. However, MULTI-seq could be employed to spatially barcode distinct regions of a single metastatic lung, enabling direct interrogation of CM spatial heterogeneity.

In summary, MULTI-seq broadly enables users to incorporate additional layers of information into scRNA-seq experiments. In the future, we anticipate that more diverse types of information will be targeted including spatial coordinates, time-points, species-of-origin, and subcellular structures (e.g., nuclei from multinucleated cells). We also anticipate that increasing LMO membrane residency time using alternative oligonucleotide conjugate designs may enable MULTI-seq applications for non-genetic lineage tracing and/or cellular competition assays.

3.5 Materials and Methods

3.5.1 Design of LMOs, CMOs, and sample barcode oligonucleotides

Anchor and co-anchor LMO and CMO designs were adapted from Weber et al [28]. Briefly, the anchor LMO has a 5' lignoceric acid (LA) modification with two oligonucleotide domains. The 5' end is complimentary to the co-anchor LMO, which bears a 3' palmitic acid (PA), and the 3' end is complimentary to the PCR handle of the sample barcode oligonucleotide. The sample barcode was designed to have three components (as in Stoeckius et al [55]): (1) a 5' PCR handle for barcode amplification and library preparation, (2) an 8 base-pair barcode with Hamming distance >3 relative to all other utilized barcodes, and (3) a 30 poly-A tail necessary for hybridization to the oligo-dT region of mRNA capture bead oligonucleotides. Identically designed anchor and co-anchor CMOs are conjugated to cholesterol at the 3' or 5' ends via a triethylene glycol (TEG) linker and are commercially available from Integrated DNA Technologies.

Anchor: {LA/Chol-TEG}–5′GTAACGATCCAGCTGTCACTTGGAATTCTCGGGTGCCAAGG-3′
Co-anchor: 5′-AGTGACAGCTGGATCGTTAC-3′-{PA/TEG-Chol}

Barcode oligo: 5′-CCTTGGCACCCGAGAATTCCANNNNNNNNA$_{30}$-3′

3.5.2 Anchor and co-anchor LMO synthesis

Oligonucleotides were synthesized on an Applied Biosystems Expedite 8909 DNA synthesizer, as previously described [28]. Specifically, Hexadecanoic (palmitic) acid, tetracosanoid (lignoceric) acid, N,N- diisopropylethylamine (DIPEA), N,N-diisopropylcarbodiimide (DIC), N,N-dimethylformamide (DMF), methylamine, ammonium hydroxide, and piperidine were obtained from Sigma-Aldrich. HPLC grade acetonitrile (CH3CN), triethylamine (NEt3), acetic acid, and anhydrous dichloromethane (CH2Cl2) were obtained from Fisher Scientific. 6-(4-Monomethoxytritylamino)hexyl-(2-cyanoethyl)-(N,N-diisopropyl)-phosphoramidite (5'-Amino-Modifier C6 Phopshoramidite), standard phosphoramidites, and DNA synthesis reagents were obtained from Glen Research. Controlled pore glass (CPG) supports (2-Dimethoxytrityloxymethyl-6-fluorenylmethoxycarbonylamino-hexane- 1-succinoyl)-long chain alkylamino-CPG (3'-Amino-Modifier C7 CPG 1000), 5'-Dimethoxytrityl-N-dimethylformamidine-2'-deoxyGuanosine, 3'-succinoyl-long chain alkylamino-CPG (dmf-dG-CPG 1000), and 5'-Dimethoxytrityl-N-Acetyl-2'-deoxyCytidine, 3'-succinoyl-long chain alkylamino-CPG (Ac-dC- CPG 1000) synthesis columns were obtained from Glen Research. All materials were used as received from manufacturer.

For the anchor LMO, after synthesis of the DNA sequence, the 5' end was modified with an amine using 5'-Amino-Modifier C6 Phosphoramidite (100 mM) and a custom 15-minute coupling protocol. After synthesis of 5′ amino-modified DNA, the MMT protecting group was removed manually on the synthesizer by priming alternately with deblock and dry CH3CN at least three times until yellow color disappears. CPG beads were dried by priming several times with dry Helium gas. For the 3′ FMOC-protected amino-modified CPG, prior to oligonucleotide synthesis, the FMOC group was removed by suspending the CPG in a solution of 20% piperidine

in dimethylformamide for 10 minutes at room temperature. The beads were then washed three times each with DMF and CH2Cl2. This procedure was repeated twice more to ensure complete deprotection of the FMOC protecting group prior to coupling to the fatty acid. Residual solvent was removed with reduced pressure on a Savant SPD121P SpeedVac System (ThermoFisher).

Fatty acid conjugation was performed on solid support by coupling the carboxylic acid moiety of the fatty acid to the 3' or 5' free amine—lignoceric acid and palmitic acid for the anchor and co-anchor, respectively. The solid support was transferred to a microcentrifuge tube and resuspended in a solution of anhydrous dichloromethane containing 200 mM fatty acid, 400 mM DIPEA, and 200 mM DIC. The microcentrifuge tubes were sealed with parafilm, crowned with a cap lock, and shaken overnight at room temperature. The beads were then washed 3X with CH2Cl2, 3X with DMF, and 2X CH2Cl2. Oligonucleotides were then deprotected and cleaved from solid support by suspending the resin in a 1:1 mixture of ammonium hydroxide and 40% methylamine (AMA) for 15 minutes at 65 °C with a cap lock followed by evaporation of AMA with a Savant SPD121P SpeedVac System. Cleaved oligonucleotides were dissolved in 0.7 mL of 0.1 M triethylammonium acetate (TEAA) and filtered through 0.2 μM Ultrafree-MC Centrifugal Filter Units (Millipore) to remove any residual CPG support prior to HPLC purification.

Fatty acid-modified oligonucleotides were purified from unmodified oligonucleotides by reversed-phase high-performance liquid chromatography (HPLC) using an Agilent 1200 Series HPLC System outfitted with a C8 column (Hypersil Gold, Thermo Scientific) and equipped with a diode array detector (DAD) monitoring at 230 and 260 nm. For HPLC purification, Buffer A was 0.1 M TEAA at pH 7 and buffer B was CH3CN. running a gradient between 8 and 95% CH3CN over 30 minutes. Pure fractions were collected manually and lyophilized. The resulting powder was then resuspended in distilled water and lyophilized again two more times to remove residual TEAA salts prior to use. Purified fatty acid-modified oligonucleotides were resuspended in distilled

water and concentrations were determined by measuring their absorbance at 260 nm on a Thermo-Fischer NanoDrop 2000 series.

3.5.3 Cell culture

For the proof-of-concept scRNA-seq and snRNA-seq experiments, HEK293 cells, HMECs, Jurkat cells, and MEF cells were maintained at 37 °C with 5% CO2. HEK293 and MEF cells were cultured in Dulbecco's Modified Eagle's Medium, High Glucose (DMEM H-21) containing 4.5 g/L glucose, 0.584 g/L L-glutamine, 3.7 g/L NaHCO3, supplemented with 10% fetal bovine serum (FBS) and penicillin/streptomycin (100 U/mL and 100 µg/mL, respectively). HMECs were cultured in M87A media [34] with or without 24 hours of stimulation with 5 ng/mL human recombinant TGF-β (Peprotech). Jurkat cells were cultured in RPMI-1640 with 25 mM HEPES and 2.0 g/L NaHCO3 supplemented with 10% FBS and penicillin/streptomycin (100 U/mL and 100 µg/mL, respectively).

For the 96-sample HMEC experiments, fourth passage HMECs were lifted using 0.05% trypsin-EDTA for 5 minutes. The cell suspension was passed through a 45 µm cell strainer to remove any clumps. The cells were washed with M87A media once and resuspended at 10^7 cells/mL. The cells were incubated with 1:50 APC/Cy-7 anti-human/mouse CD49f (Biolegend, #313628) and 1:200 FITC anti-human CD326 (EpCAM) (Biolegend, #324204) antibodies for 30 minutes on ice. The cells were washed once with PBS and resuspended in PBS with 2% BSA with DAPI at 2-4 million cells/mL. Cells were sorted on BD FACSAria III. DAPI+ cells were discarded. LEPs were gated as $EpCAM^{hi}/CD49f^{lo}$ and MEPs were gated as $EpCAM^{lo}/CD49f^{hi}$ (**Fig. 3-17**) [56].

Notably, this gating strategy results in trace numbers of MEPs and LEPs sorted incorrectly. HMEC sub-populations were sorted into 24-well plates such that wells contained LEPs only,

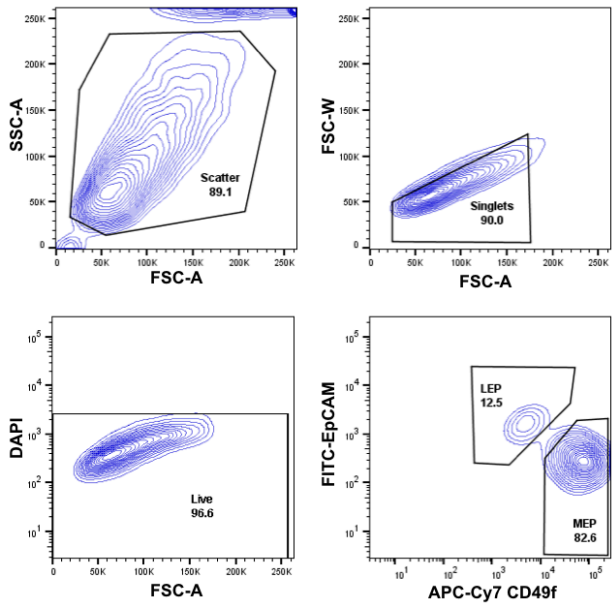

Figure 3-17. FACS purification of LEP and MEP cells from bulk HMECs. Bulk HMECs were labeled with FITC anti-EpCAM and APC-Cy7 anti-CD49f to identify and isolate LEPs and MEPs. LEPs are identified as EpCAM high and CD49f low, while MEPs are CD49f high and EpCAM low. Gating strategy causes minor cell type impurities in final sorted population [56].

MEPs only, or a 2:1 ratio of LEPs to MEPs. Sorted cell populations were cultured for 48 hours in M87A media before culturing for 72 hours in M87A media (-EGF) supplemented with different signaling molecules or signaling molecule combinations. Specifically, M87A media (-EGF) was supplemented with 100 ng/mL RANKL, 100 ng/mL WNT4, 100 ng/mL IGF-1, 113 ng/mL AREG, and/or 5 ng/mL EGF (all from Peprotech) alone or in all possible pairwise combinations. For the 96-sample HMEC technical replicate experiment, in vitro cultures were prepared as described above, except all sorted wells contained both LEPs and MEPs. Cultures were then grown in complete M87A media for 72 hours prior to isolation.

3.5.4 Analytical Flow Cytometry

The BD FACSCalibur instrument was used to perform analytical flow cytometry experiments measuring live-cell and nuclear membrane labeling efficiency, LMO and CMO membrane residency kinetics on ice and at room temperature, and efficacy of BSA quenching

(**Fig. 3-2**). HEK293 cells and nuclei were utilized for all experiments. Samples were prepared using the same workflows as proof-of-concept scRNA-seq and snRNA-seq experiments with one exception. In place of barcode oligonucleotides, anchor LMOs or CMOs were pre-hybridized to equimolar concentrations of FAM- or AF647-conjugated oligonucleotides matching the barcode oligonucleotide 5' PCR handle excluding the barcode and poly-A regions.

For titration experiments, $5x10^5$ cells or nuclei were suspended in 180 µL cold PBS followed by addition of 20 µL 10X anchor LMO or CMO pre-mixed with equimolar complimentary oligonucleotide conjugated to AF647 (final concentrations of 10 nM, 50 nM, 100 nM, 500 nM, or 1000 nM). Cells were incubated on ice for 5 minutes followed by addition of 20 µL of 10X stock corresponding co-anchor. The experiment was repeated three times, mean fluorescence intensity was calculated for each condition, and linear regression was performed. For exchange experiments, HEK293 cells were labeled with 200 nM LMOs or CMOs bearing FAM- or AF647-conjugated oligonucleotides. FAM- and AF647-labeled cells were then mixed and kept on ice for 2 hours in PBS with 1% BSA (2% for nuclei), during which cell aliquots were analyzed every 30 minutes. For room temperature experiments, cells were incubated for 30 minutes at room temperature and analyzed every 10 minutes. Label stability was computed as proportional differences between FAM or AF647 intensity relative to time zero. Off-target labeling was computed as FAM abundance on AF647-labeled cells (or vice versa). Fluorophore only controls were included in nuclei flow cytometry experiments because fluorophore-conjugated oligonucleotides demonstrate non-specific labeling.

For BSA quenching experiments, HEK293 cells were labeled with 200 nM LMOs or CMOs in 100 µL total volume PBS as described above. Prior to washing, each sample was diluted with ice cold PBS or PBS containing 1% BSA followed by centrifugation (160 rcf, 4 °C, 4 min). The 150 µL supernatant was removed from each primary labeling mixture and used to resuspend unlabeled HEK293 cells (secondary labeling). All primary and secondary labeled cells were

washed 3X with ice cold PBS containing 1% BSA and analyzed by flow cytometry. Each secondary labeled sample was plotted as a proportion of the primary labeled sample. All analytical flow cytometry data analyses were performed in FlowJo and R.

3.5.5 scRNA-seq sample preparation

For the proof-of-concept experiment, cells were first treated with trypsin for 5 minutes at 37 °C in 0.05% trypsin-EDTA before quenching with appropriate cell culture media. Single-cell suspensions were then pelleted for 4 minutes at 160 rcf and washed once with PBS before suspension in 90 µL of a 200 nM solution containing equimolar amounts of anchor LMO and sample barcode oligonucleotides in PBS. Anchor LMO-barcode labeling was performed for 5 minutes on ice before 10 µL of 2 µM co-anchor LMO in PBS (for a final concentration of 200 nM) was added to each cell pool. Following gentle mixing, the labeling reaction was continued on ice for another 5 minutes before cells were washed twice with PBS, resuspended in PBS with 0.04% BSA, filtered, and pooled. The same workflow was also performed with CMOs. LMO-, CMO-, and unlabeled control cells were then loaded into three distinct 10X microfluidics lanes.

For the original 96-plex HMEC experiment, LMO labeling was performed during trypsinization in order to minimize wash steps and thereby limit cell loss and preserve cell viability. HMECs cultured in 24-well plates were labeled for 5 minutes at 37 °C and 5% CO2 in 190 µL of a 200 nM solution containing equimolar amounts of anchor LMO and sample barcode oligonucleotides in 0.05% trypsin-EDTA. 10 µL of 4 µM co-anchor LMO in 0.05% trypsin-EDTA was then added to each well (for a final concentration of 200 nM) and labeling/trypsinization was continued for another 5 minutes at 37 °C and 5% CO2 before quenching with appropriate cell culture media. A similar labeling protocol was used for the technical replicate experiment, except LMOs were incorporated once the cells were in single-cell suspension. Cells were then transferred to a 96-well plate for washing with 0.04% BSA in PBS. Finally, cells were pooled into

a single aliquot, filtered through a 0.45 µm cell strainer, and counted before loading 10X microfluidics lanes.

For the PDX experiment, primary tumors and lungs were cryopreserved after dissection from triple-negative breast cancer PDX models generated in NOD-SCID gamma (NSG) mice as described previously [57]. The UCSF Institutional Animal Care and Use Committee (IACUC) reviewed and approved all animal experiments. On the day of the experiment, cryopreserved tissues were thawed and dissociated in digestion media containing 50 µg/mL Liberase TL (Sigma-Aldrich) and 2×10^4 U/mL DNase I (Sigma-Aldrich) in DMEM/F12 (Gibco) using standard GentleMacs protocols. Dissociated cells were then filtered through a 70 µm cell strainer to obtain a single-cell suspension prior to washing with PBS. Cells were then stained for 15 minutes on ice with 1:500 Zombie NIR (BioLegend, #423105) viability dye in PBS. Cells were then washed with 2% FBS in PBS prior to blocking for 5 minutes on ice with 100 µL 1:200 Fc-block (Tonbo, #70–0161-U500) in 2% FBS in PBS. After blocking, cells were stained for 45 minutes on ice with 100 µL of an antibody cocktail containing anti-mouse TER119 (FITC, ThermoFisher, #11-5921-82), anti-mouse CD31 (FITC, ThermoFisher, #11-0311-85), anti-mouse CD45 (BV450, Tonbo, #75–0451-U100), anti-mouse MHC-I (APC, eBioscience, #17-5999-82) and anti-human CD298 (PE, BioLegend, #341704). Cells were then washed with PBS prior to MULTI-seq labeling for 5 minutes on ice with 100 µL of 2.5 µM anchor LMO-barcode in PBS. 20 µL of 15 µM co- anchor LMO in PBS was added to each cell pool (for a final concentration of 2.5 µM) and labeling was continued for another 5 minutes.

Notably, we used a 10-fold greater LMO concentration for this experiment to account for increases in the total number of cells and lipophilic molecules remaining after dissociation. Following LMO labeling, cells were diluted with 100 µL of 2% FBS in PBS to ‘quench’ LMOs and washed once in 2% FBS in PBS. Finally, mCD45+ mouse immune cells and hCD298+ human metastases from dissociated primary tumors and lungs were pooled after FACS enrichment, as

described previously [57] (**Fig. 3-18**). Cell pools were then sequenced in a single 10X microfluidics lane.

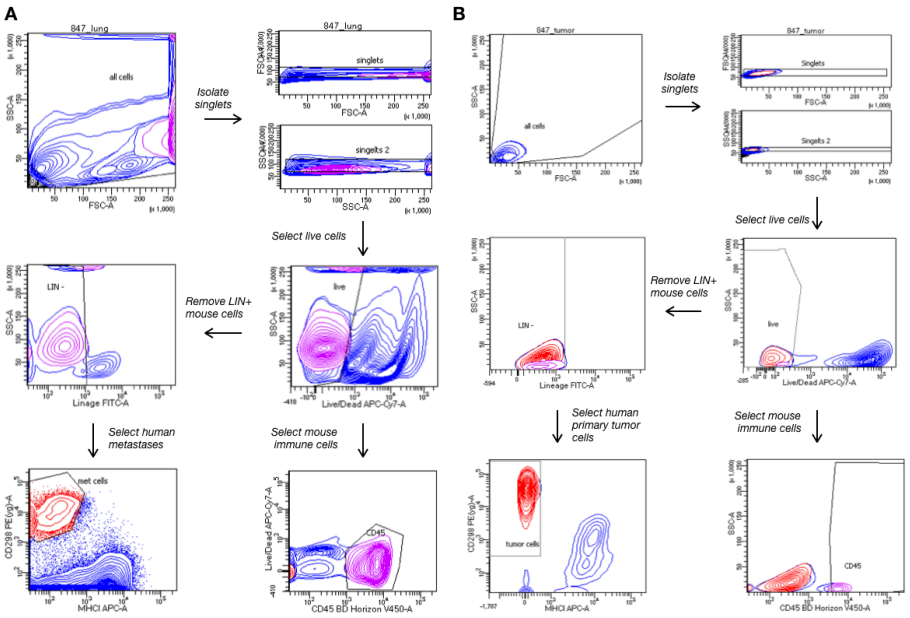

Figure S-18. FACS gating strategy for PDX lung and primary tumor samples
(A) Human metastases and mouse immune cells were separated from PDX mouse lungs using hCD298 and mCD45 following gating for live singlets. Mouse 847 (Sample L-A) is presented here as a representative example.
(B) Human primary tumor cells and mouse tumor-associated immune cells were separated using hCD298 and mCD45 following gating for live, singlets. Sample A is presented here as a representative example for all other primary tumor samples.

3.5.6 snRNA-seq sample preparation

For the Jurkat cell activation time-course, 2×10^5 Jurkat cells were added to 8 wells of a 12-well plate and treated with 10 ng/µL phorbol 12-myristate 13-acetate (PMA, Sigma-Aldrich #P8139) and 1.3 µM ionomycin (Sigma-Aldrich #I0634) at 15 min, 30 min, 1 hr, 2 hr, 4 hr, 6 hr, or 24 hr prior to barcoding with LMOs. A single well of Jurkat cells were left untreated. HEK293 and MEF cells were cultured as described above. Nuclei were isolated from cells using a protocol adapted from 10X Genomics. Briefly, suspensions of HEK293, MEF, or treated Jurkat cells were washed once with PBS, pelleted at 160 rcf (HEK293, MEFS) or 300 rcf (Jurkat) for 4 min at 4 °C and suspended in chilled lysis buffer (0.5% Nonidet P40 Substitute, 10 mM Tris-HCl, 10 mM NaCl, and 3 mM MgCl2 in milliQ water) to a density of 2.5×10^6 cells/mL. Lysis proceeded for 5 minutes

on ice, after which the lysate was pelleted (500 rcf, 4 °C, 4 minutes) and washed three times in chilled resuspension buffer (2% BSA in PBS). Nuclei were then diluted to a concentration of ~10^6 nuclei/mL prior to LMO or CMO labeling. HEK293 and MEF cells were each divided into two samples and labeled with LMOs or CMOs (500 nM in resuspension buffer) using the same procedure as described for live cells (presence of BSA during labeling is the lone alteration as it is required to prevent nuclei clumping). Each Jurkat sample was labeled with LMOs, alone. Each sample was washed 3X in 1mL resuspension buffer (500 rcf, 4 °C, 4 min). The four LMO- and CMO-labeled HEK293 and MEF samples were pooled in equal portions and, separately, Jurkat samples were pooled in equal proportions. These final two samples were combined in a 1:1 ratio and sequenced on a single 10X microfluidics lane.

3.5.7 scRNA-seq and snRNA-seq library preparation

Sequencing libraries were prepared using a custom protocol based on the 10X Genomics Single Cell V2 and CITE-seq [55] workflows. Briefly, the 10X workflow was followed up until cDNA amplification, where 1 µL of 2.5 µM MULTI-Seq additive primer was added to the cDNA amplification master mix:

MULTI-seq Additive Primer: 5′-CCTTGGCACCCGAGAATTCC-3′

This primer increases barcode sequencing yield by enabling the amplification of barcodes that successfully primed reverse transcription on mRNA capture beads but were not extended via template switching (**Fig. 3-19**). Notably, the MULTI-seq additive primer was erroneously excluded during the proof-of-concept snRNA-seq library preparation, and nuclei were still able to be robustly classified. Following amplification, barcode and endogenous cDNA fractions were separated using a 0.6X SPRI size selection. The endogenous cDNA fraction was then processed

according to the 10X workflow until next-generation sequencing (NGS) with the following formats: Proof-of-concept scRNA-seq (2x HiSeq 4000, 100%), Proof-of-concept snRNA-seq (NovaSeq, 20%), HMEC (NovaSeq S4, 100%), HMEC technical replicate (NovaSeq S4, 5%), PDX (NovaSeq S4, 70%)

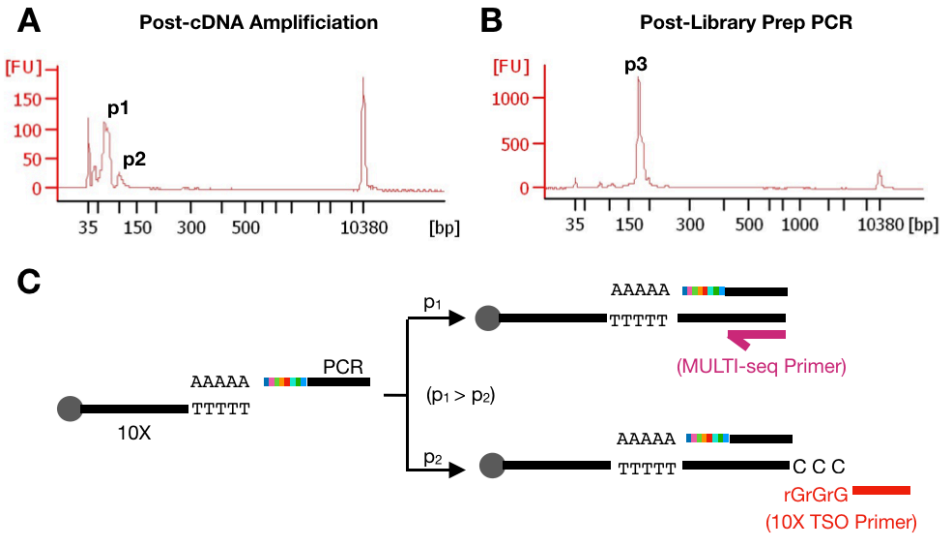

Figure 3-19. Bioanalyzer traces of representative MULTI-seq barcode library.
(A) Bioanalyzer traces following cDNA amplification and MULTI-seq barcode enrichment using 3.2X SPRI with 1.8X 100% isopropanol exhibits two distinct peaks. Bioanalyzer traces are representative of all datasets presented in this study (n = 4). The first peak (p1) is an average of 65-70 base-pairs in length and likely corresponds to barcodes amplified via the MULTI-seq additive primer. The second peak (p2) is an average of 100 base-pairs in length and likely corresponds to barcodes that successfully underwent MMLV-RTase template switching and were subsequently amplified by the standard 10X Genomics Single Cell V2 primer.
(B) Bioanalyzer analysis following library preparation PCR exhibits one distinct peak (p3) with an average length of 173 base-pairs, matching expectations. Bioanalyzer traces are representative of all datasets presented in this study (n = 4).
(C) Schematic illustrating the two species of reverse-transcribed MULTI-seq barcodes with and without template switching. Processive reverse-transcription without template switching (p1) is more likely than reverse-transcription with template switching (p2), resulting in relative enrichment of the 65-70 base-pairs product following cDNA amplification.

To prepare the barcode fraction for NGS, contaminating oligonucleotides remaining from cDNA amplification were first removed using an established small RNA enrichment protocol (Beckman Coulter). Specifically, we increased the final SPRI ratio in the barcode fraction to 3.2X reaction volumes and added 1.8X reaction volumes of 100% isopropanol (Sigma-Aldrich). Beads were then washed twice with 400 µL of 80% ethanol and allowed to air dry for 2-3 minutes before elution with 50 µL of Buffer EB (Qiagen, USA). Eluted barcode cDNA was then quantified using QuBit before library preparation PCR (95 °C, 5’; 98 °C, 15”; 60 °C, 30”; 72 °C, 30”; 8 cycles; 72

°C, 1’; 4 °C hold). Each reaction volume was a total of 50 µL containing 26.25 µL 2X KAPA HiFi HotStart master mix (Roche), 2.5 µL of 10 µM TruSeq RPIX primer (Illumina), 2.5 µL of 10 µM TruSeq Universal Adaptor primer (Illumina), 3.5 ng barcode cDNA, and nuclease-free water.

TruSeq RPIX:

5′-CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA-3′

TruSeq P5 Adaptor:

5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′

To prepare the barcode fraction for NGS, contaminating oligonucleotides remaining from cDNA amplification were first removed using an established small RNA enrichment protocol (Beckman Coulter). Specifically, we increased the final SPRI ratio in the barcode fraction to 3.2X reaction volumes and added 1.8X reaction volumes of 100% isopropanol (Sigma-Aldrich). Beads were then washed twice with 400 µL of 80% ethanol and allowed to air dry for 2–3 minutes before elution with 50 µL of Buffer EB (Qiagen, USA). Eluted barcode cDNA was then quantified using QuBit before library preparation PCR (95 °C, 5’; 98 °C, 15”; 60 °C, 30”; 72 °C, 30”; 8 cycles; 72 °C, 1’; 4 °C hold). Each reaction volume was a total of 50 µL containing 26.25 µL 2X KAPA HiFi HotStart master mix (Roche), 2.5 µL of 10 µM TruSeq RPIX primer (Illumina), 2.5 µL of 10 µM TruSeq Universal Adaptor primer (Illumina), 3.5 ng barcode cDNA, and nuclease-free water.

Following library preparation PCR, remaining sequencing primers and contaminating oligonucleotides were removed via a 1.6X SPRI clean-up. Barcode libraries were sequenced using the following NGS formats: Proof-of-concept scRNA-seq LMO (HiSeq 4000, 33.3%), Proof-of-concept scRNA-seq CMO (HiSeq 4000, 33.3%), HMEC (HiSeq 4000, 100%), HMEC technical replicate (HiSeq 4000 50%), PDX (NovaSeq S4, 2.5%), Proof-of-concept snRNA-seq LMO (NovaSeq S4, 1.25%), Proof-of-concept snRNA-seq CMO (NovaSeq S4, 1.25%). Notably, sequencing reads predominantly aligned to the barcode reference sequences (average of 97.0%

across all datasets), and resulted in high SNRs with low rates of duplicated UMIs (average of 16.07% across all datasets), suggesting that barcode libraries were not sequenced to saturation.

3.5.8 Expression library pre-processing

Expression library FASTQs were pre-processed using CellRanger (10X Genomics) and aligned to the hg19 (proof-of-concept scRNA-seq, HMEC), concatenated mm10-hg19 (PDX), or concatenated mm10-hg19 pre-mRNA (proof-of-concept snRNA-seq) reference transcriptomes. When multiple 10X lanes were sequenced in an experiment, CellRanger aggregate was used to perform read-depth normalization.

3.5.9 Cell/Nuclei calling

For the proof-of-concept scRNA-seq, snRNA-seq and HMEC technical replicate experiments, cell-associated barcodes were defined using CellRanger. For the original 96-plex HMEC experiment, cells were defined as cell barcodes (1) associated with >= 600 total RNA UMIs that (2) were successfully classified during MULTI-seq sample classification workflow. We manually selected 600 RNA UMIs as a threshold in order to exclude low-quality cell barcodes. For the PDX experiment, we defined cells as barcodes (1) associated with >= 100 total RNA UMIs that (2) were successfully classified during the MULTI-seq sample classification workflow.

3.5.10 Expression library analysis

Following pre-processing and cell/nuclei calling, RNA UMI count matrices were prepared for analysis using the 'Seurat' R package, as described previously [58,59]. Briefly, genes expressed in fewer than 3 cells were discarded before the percentage of reads mapping to mitochondrial genes (%Mito) was computed for each cell. Outlier cells with elevated %Mito were visually defined and discarded. Data was then log2-transformed, centered, and scaled before

variance due to %Mito and the total number of RNA UMIs were regressed out. Highly variable genes were then defined for each dataset by selecting mean expression and dispersion thresholds resulting in ~2000 total genes. These variable genes were then used during PCA, and statistically-significant PCs were defined by PC elbow plot inflection point estimation. Significant PCs were then utilized for unsupervised Louvian clustering and dimensionality reduction with t-SNE [60]. Following pre-processing, differential gene expression analysis was performed using the 'FindMarkers' command in 'Seurat', with 'test.use' set to 'bimod' [61].

3.5.11 Proof-of-concept scRNA-seq and snRNA-seq analyses

Testing the effects of MULTI-seq barcoding on scRNA-seq and snRNA-seq data: MULTI-seq could negatively influence scRNA-seq and snRNA-seq data in two main ways: by (1) competing with endogenous mRNAs for capture bead hybridization regions, or (2) inducing a transcriptional response to LMO or CMO labeling. To test these possibilities, we first parsed our proof-of-concept scRNA-seq and snRNA-seq datasets to include only HEK293 (HEK) cells and MEF nuclei, respectively. Focusing on individual cell types ensures that any observed performance differences are primarily due to technical and not biological reasons.

All HEK cells and MEF nuclei subsets were indistinguishable with regards to the total number of detected RNA UMIs and genes (**Figs. 3-4c, 3-6b**). Moreover, barcode and RNA UMIs were not negatively correlated (**Figs. 3-4c, 3-6b**). These observations suggest that MULTI-seq barcodes do not detrimentally compete with endogenous transcripts during mRNA capture. Additionally, LMO-, CMO-, and unlabeled HEK cells and MEF nuclei exhibited similar proportions of reads aligning to mitochondrial genes (**Figs. 3-4c, 3-6b**); therefore, LMO and CMO labeling are unlikely to induce an apoptotic cellular response. To explore whether MULTI-seq labeling perturbs endogenous gene expression in other ways, we compared the proportion of each cell/nuclei's 100 nearest neighbors in principal component (PC) space that were derived from LMO-, CMO-, or

unlabeled subsets. Neighborhoods were defined by computing the Euclidean distance matrix for statistically-significant PCs with the 'rdist' R function.

For HEK cells, neighborhood analysis revealed that CMO-labeled cells preferentially co-localized in gene expression space, while LMO-labeled and unlabeled neighborhoods were nearly indistinguishable (**Fig. 3-4d**). We then performed differential gene expression analysis between HEKs from each sample group, which demonstrated that 3 and 8 genes were 1.5-fold enriched in LMO- or CMO-labeled HEKs relative to unlabeled controls, respectively (**Table 3-1**). Intriguingly, even after only < 1 hour on ice, CMO-labeled HEKs exhibited differential expression of AP2B1, which has established roles in cholesterol and sphingolipid transport. When considered along with flow cytometry analyses demonstrating that CMOs exhibit reduced live-cell membrane residency compared to LMOs (**Fig. 3-2b**), these results collectively illustrate that LMOs are the preferred reagent for scRNA-seq sample multiplexing.

Table 3-1. List of genes with >1.5-fold expression difference between LMO/CMO-labeled and unlabeled HEKs from proof-of-concept scRNA-seq experiment.

Gene	FC
LMO vs Unlabeled	
MIF	2.8
KRTCAP2	1.5
TOMM5	1.5
LMO vs CMO	
SNORA76	1.7
NMT1	1.5
CMO vs Unlabeled	
MIF	1.9
S100A2	1.6
MT2A	1.5
AP2B1	-1.5
TOP2A	-1.5
MALAT1	-1.5
SNORA76	-1.7
NMT1	-1.8

In contrast to HEK cells, MEF nuclei from each labeling condition had uniform neighborhood proportions (**Fig. 3-4c**). Additionally, we did not detect any genes that were

differentially expressed > 0.7-fold between LMO-, CMO-, and unlabeled nuclei. These results demonstrate that the transcriptional response to CMO labeling observed in HEK cells was absent in nuclei. Moreover, we observed a ~10-fold increase in barcode nUMIs for CMO-labeled MEF nuclei relative to LMO-labeled nuclei (**Fig. 3-4b**). This observation was in-line with our previous flow cytometry titration experiments (**Fig. 3-2c**). We believe that this difference in sample barcode capture efficiency was due to the presence of BSA in nuclei resuspension buffer, which is necessary to prevent aggregation nuclei purification. BSA has a lipid-binding pocket which likely sequesters LMOs, leading to reduced sample barcode association with the nuclear membrane. When considered along with the commercial-availability of CMOs, these results collectively illustrate that CMOs are the preferred reagent for snRNA-seq sample multiplexing.

3.5.12 96-plex HMEC scRNA-seq analyses

Doublet analysis, comparison to computational doublet prediction methods: To fit DoubletFinder [36] parameters to our 96-plex HMEC scRNA-seq data, we began by performing a parameter sweep using the 'paramSweep' function in the 'DoubletFinder' R package. Ideal parameters were then defined using the 'summarizeSweep' function, which uses receiver operating curve analysis to compute the predictive capacity of each parameter set relative to ground-truth doublet labels. We used MULTI-seq doublet classifications as ground-truth in this application. With ideal parameters defined (e.g., pN = 0.25, pK = 0.03), we then thresholded DoubletFinder results by adjusting the total number of MULTI-seq-defined doublets to account for homotypic doublet formation. Homotypic doublets are doublets that are formed from transcriptionally-similar cells and are known to be undetectable using computational doublet detection methods that rely solely on gene expression features [36,37]. To account for homotypic doublets, we multiplied the total doublet number (3413) by the sum of squared cell type frequencies (0.51), resulting in 1738 total doublet predictions.

Exploring transcriptional responses to cell type composition: To explore the transcriptional responses to cell type composition, we began by pre-processing data subsets containing only MEPs or LEPs. Separating cell types revealed distinct resting and proliferative MEP and LEP subsets discernible by enriched MKI67 expression (**Fig. 3-12**). To assess whether co-culture influenced proliferation, we specified subsets of cells where each cell type composition (e.g., mono- or co-cultures) were equally abundant. We then determined whether mono- and co-cultured cells were evenly represented in the resting and proliferative states. Down-sampling in this fashion revealed that the co-culture-induced increase in proliferation was specific to LEPs while controlling for differences in the total numbers of cells from each comparison group.

Focusing on the resting cells, we next pre-processed data subsets containing equal numbers of resting MEPs or LEPs from each sample condition. We then computed the average TGFBI expression amongst MEPs and LEPs grouped by signaling molecule exposure and observed that co-cultured LEPs and MEPs were associated with elevated TGFBI expression independent of perturbation (**Fig. 3-12**). In our proof-of-concept scRNA-seq experiment, we observed that TGFBI expression is increased specifically in HMECs responding to TGF-β (data not shown). Thus, these results suggest that co-culturing induces paracrine-mediated TGF-β signaling in both LEPs and MEPs.

Exploring transcriptional responses to signaling molecule perturbation: Using the same dataset employed in the TGFBI analysis described above, we next sought to characterize transcriptional responses to signaling molecules. To this end, we grouped cells according to signaling molecule exposure and performed hierarchical clustering on the average gene expression profile for each group using the 'BuildClusterTree' function in 'Seurat'. Hierarchical clustering revealed two distinct clades corresponding to cells stimulated with (1) the EGFR ligands AREG and EGF and (2) RANKL, IGF1, or WNT4. Notably, AREG/EGF stimulation dominated the

effect of RANKL/IGF1/WNT4, as cells grown in media supplemented with both AREG/EGF and RANKL/IGF1/WNT4 remained members the EGFR ligand clade.

Differential gene expression analysis between these two clades revealed that AREG/EGF stimulated cells expressed elevated levels of a number of EGFR signaling target genes (**Fig. 3-13**), as expected. Differentially-expressed genes amongst RANKL/IGF1/WNT4 stimulated cells could not be as readily connected to their corresponding signaling pathways. This observation suggests that the rich media used to culture HMECs buffered the cells against RANKL/IGF1/WNT4 induction. This notion is further supported by the fact that cells stimulated with EGFR ligands – which were purposefully depleted from the M87A media used in this experiment – represented the most pronounced transcriptional signature amongst signaling molecule conditions.

Signal-to-noise ratio (SNR) computation: SNR for singlets, doublets, and negative cells was calculated as the quotient of the two most abundant raw barcode UMI abundances for each cell (**Figs. 3-10c, 3-10f**). Since cells are discarded as doublets when surpassing two or more barcode-specific thresholds during our sample classification workflow, we reasoned that the relative abundances of the top two barcodes was a sufficient SNR definition. In singlets from the original 96-plex HMEC experiment, on-target barcodes are an average of 199-fold higher than the most abundant off-target barcode. Doublets have much lower SNR but higher total barcode nUMIs. This observation matches expectations, as doublet formation results in the pooling of MULTI-seq barcodes from two cells. Negative cells exhibit very low SNR and total nUMIs, indicating that negative cells were not sufficiently labeled with LMOs to enable sample classification.

3.5.13 PDX scRNA-seq analyses

MULTI-seq sample classifications distinguish low-RNA from low-quality cells: Following expression library pre-processing, raw RNA UMI count matrices must be parsed to define cell barcodes associated with intact cells versus ambient mRNA and cell debris. This challenge is commonly addressed by identifying the inflection point of log-log RNA UMI by RNA UMI rank distributions, which follows the assumption that droplets containing intact cells should feature elevated nUMIs. This strategy is inherently biased against cells with intrinsically low RNA content, and may be confounded by distributions with multiple inflection points (e.g., datasets with many cell types) [43].

The RNA UMI distribution for mouse immune cells sequenced during our PDX experiment exemplifies this issue. Specifically, we observed a mode corresponding to cell barcodes with ~500 total RNA UMIs that was discarded by the standard CellRanger UMI threshold (**Fig. 3-15**, top left). To assess whether this region represented intact low-RNA cells, we performed the MULTI-seq sample classification workflow on all cell barcodes with at least 100 RNA UMIs. We selected this threshold because droplets with < 100 RNA UMIs can be confidently assumed to be empty [43]. Intriguingly, sample classifications produced 2,580 singlets and 583 negatives amongst cells with RNA UMIs between 100 and the CellRanger threshold (1350 RNA UMIs).

To test whether sample classification results could be used to distinguish low-RNA cells from ambient mRNA and cellular debris, we first pre-processed putative low-RNA singlets using 'Seurat' and used unsupervised clustering and differential gene expression analyses to reveal discrete clusters in gene expression space characterized by established marker genes for neutrophils, monocytes, alveolar macrophages and endothelial cells (**Fig. 3-16a**). In contrast, equivalent analyses of unclassified cell barcodes with 100-1350 RNA UMIs revealed clusters corresponding to broken cells and a small number of neutrophils. We annotated broken cells into two subsets – one with enriched mitochondrial gene expression and another with elevated levels

of lncRNAs (e.g., Xist) and ribosomal RNAs. We speculate that the latter represents nuclei released from cells due to shear stress (see [23] for analogous analyses).

Immune cell transcriptional and proportional shifts associated with metastatic progression: To assess whether lung immune cell type proportions shifted during metastatic progression in our PDX mice, we first defined a subset of cells where each tumor stage (e.g., WT, early, mid, and late) was equally represented. Down-sampling in this fashion controls for technical differences in the number of sequenced cells. We then computed the proportion of each cell type present in lung immune cells from each tumor stage. Statistically-significant proportional shifts relative to WT proportions were then defined using two-proportion z-tests ('prop.test' function in R) with Bonferroni multiple comparison correction ('p.adjust' function in R).

We additionally focused on changes in classical monocyte (CM) gene expression patterns during metastatic progression. We began by pre-processing a dataset including only CMs using 'Seurat'. Unsupervised clustering of these data revealed sub-structure demarcating each tumor stage (**Fig. 3-16c**). Early-stage CMs were distinct from WT CMs despite the lack of detectable metastases, which suggests that this data could provide insight into CM transcriptional behavior during metastatic colonization. Early-stage CMs were also transcriptionally-similar to a subset of mid-stage CMs. However, mid- and late-stage CMs manifested as two distinct sub-states featuring heterogeneous expression of many genes previously linked to metastatic/aggressive behavior (**Table 3-2**).

Computing inter-sample variability using Earth Mover's Distance: Earth Mover's Distance (EMD) measures the distance in gene expression space that is required to map two distinct high-dimensional manifolds onto one another. To this end, EMD is an emerging tool to quantify differences amongst sets of cells in scRNA-seq data. We used EMD, as implemented in the 'calculate_emd' function from the 'EMDomics' R package [51], to quantify the variability between lung immune cells from biological replicate mice and mice from distinct tumor stages.

Table 3-2. List of genes with >1.5-fold expression difference between classical monocytes at distinct stages of metastatic progression.

Gene	FC
WT	
Ear2	2.3
Rsrp1	2.1
Plaur	1.7
Rgcc	2.1
Klf4	1.7
Jund	1.6
Wsb1	1.6
Pdlim1	1.6
Tagln2	1.6
Pglyrp1	2.0
Fn1	1.5
Ezr	1.5
Tsc22d3	1.5
Cks2	1.8
Hspa1a	1.6
Early/Mid-1	
Isg15	1.8
Thbs1	1.7
Mid-2	
Tppp3	2.3
Adgre5	2.1
Tagln2	1.7
Crip1	2.1
Metrnl	1.7
Emp3	1.6
Cd74	1.6
Cd300a	1.6
Btg2	1.6
Pou2f2	2.0
Sgk1	1.5
Il1b	1.5
Gngt2	1.5
Hist1h1c	1.8

Gene	FC
Late-1	
Fos	2.6
Dusp1	2.3
Jun	2.6
Atf3	2.2
Ier3	2.0
Ccl3	2.6
Tsc22d3	1.7
S100a8	3.7
Ccl2	2.0
Saa3	2.5
Fosb	1.7
Socs3	1.8
Wfdc21	2.4
Klf6	1.5
S100a9	2.2
Egr1	1.5
Hspa1a	1.8
Lcn2	1.6
Lrg1	1.6
Ccl4	1.9
Late-2	
Rgcc	2.8
Rsrp1	2.8
Nfkbia	2.4
Mmp19	2.4
Cxcr4	2.3
Arg2	2.3
Lmna	2.2
Saa3	2.0

Specifically, we first down-sampled our existing data to include equal numbers of CMs from each tumor stage and mouse. Down-sampling in this fashion is necessary to control for differences in EMD results solely due to the total number of cells. We then extracted the PC space embeddings for this cell subset, and performed EMD on cells grouped by (1) tumor stage and (2) mouse ID. Notably, we only extracted embeddings for statistically-significant PCs (e.g., 10 for the CM-only dataset). We then scaled all of the EMD values from 0 to 1 and found the mean EMD between tumor stages and biological replicates (e.g., mice 1/4 and 2/5). CMs from biological replicates had a lower mean scaled EMD than CMs from each tumor stage (0.16 vs. 0.69),

demonstrating that the observed CM heterogeneity between different tumor stages is not solely attributable to variability between individual mice.

3.5.14 MULTI-seq library pre-processing

Raw barcode library FASTQs were converted to barcode UMI count matrices using custom scripts leveraging the 'ShortRead' [61] and 'stringdist' [62] R packages. Briefly, raw FASTQs were first parsed to discard reads where the first 16 bases of R1 did not perfectly match any of the cell barcodes associated a pre-defined list of cell barcodes. Second, reads where the first 8 bases of R2 did not align with < 1 mismatch to any reference barcode were discarded. Third, reads were binned by cell barcodes and duplicated UMIs were identified as reads where bases 17-26 of R2 exactly matched. Finally, reference barcode alignment results were then parsed to remove duplicated UMIs before being converted into a final barcode UMI count matrix.

3.5.15 MULTI-seq sample classification algorithm

MULTI-seq barcode UMI count matrices were used to classify cells into sample groups via a workflow inspired by previous scRNA-seq multiplexing approaches [17,18,23]. First, raw barcode reads were log2-transformed and mean-centered. The presence of each barcode was then visually inspected by performing t-SNE on the normalized barcode count matrix, as implemented in the 'Rtsne' R package with 'initial_dims' set to the total number of barcodes. Missing barcodes (observed only for the 96-plex HMEC experiment) were discerned as those lacking any enrichment in barcode space and were removed.

Next, the top and bottom 0.1% of values for each barcode were excluded and the probability density function (PDF) for each barcode was defined by applying the 'approxfun' R function to Gaussian kernel density estimations produced using the 'bkde' function from the 'KernSmooth' R package [63]. We then sought to classify cells according to the assumption that

groups of cells that are positive and negative for each barcode should manifest as local PDF maxima [17,18]. To this end, we computed all local maxima for each PDF and defined negative and positive maxima as the most frequent and highest local maxima, respectively. Notably, this strategy assumes that truly-barcoded cells will have the highest abundance for any given barcode, and that no individual sample group will have more members than the sum of all other groups.

With these positive and negative approximations in hand, we next sought to define barcode- specific UMI thresholds. To find the best inter-maxima quantile for threshold definition (e.g., an inter-maxima quantile of 0.5 corresponds to the mid-point), we iterated across 0.02-quantile increments and chose the value that maximized the number of singlet classifications. Sample classifications were then made using these barcode-specific UMI thresholds by discerning which thresholds each cell surpasses, with doublets being defined as cells surpassing > 1 threshold [23]. Negative cells (i.e., cells surpassing 0 thresholds) were then removed, and this procedure was repeated until all cells were classified as singlets or doublets.

Following this initial classification workflow, two varieties of negative classifications remain: true and false negatives. True negatives manifest in barcode space as high-density regions lacking enrichment for any particular barcode. True negatives result from cells with poor barcode labeling. In contrast, false negatives result from algorithmic misclassification. Since a single inter-maxima quantile threshold is applied to all barcodes during sample classification, we believe false negatives arise because this thresholding strategy may be sub-optimal for a subset of barcode distributions. Thus, although false negatives have poor absolute signal in comparison to high-confidence singlets, we reasoned that false negatives could be ‘rescued’ using a semi-supervised approach (akin to [23]) by computing the relative strength of each barcode signal on a cell-by-cell basis.

To distinguish which negative cells are the best candidates for reclassification before reclassifying negatives into their appropriate barcode groups, we used the following strategy:

1. Repeat the original sample classification workflow, recording the total number of 94 thresholds that each negative cell surpasses at each quantile.
2. Compute each cell's classification stability (CS) – defined as the number of quantiles across which a cell surpasses a single threshold.
3. Subset equal numbers of 'ground-truth' cells from the original classification results.
4. Perform semi-supervised k-means clustering on merged data including 'ground-truth' and negative cells. Clustering is semi-supervised because one member of each 'ground-truth' sample group is used to initialize cluster centers.
5. Compute the rate at which 'ground-truth' and negative cell classifications match the k-means results.
6. Iteratively repeat steps 4 and 5 using a different 'ground-truth' cell to initialize cluster centers during each iteration. Repeat until all 'ground-truth' cells have been used.
7. Compare k-means matching rates between 'ground-truth' and negative cells binned according to CS values. Negative cells with CS values resulting in matching rates that approximate 'ground-truth' matching rates are reclassified.

Negative cell reclassification rescues 10%-20% of negative cells across the different datasets presented in this study (data not shown). While not insignificant, we believe that further optimization will improve performance.

3.6 Perspective

Since MULTI-seq was published in 2019, efforts in the single-cell genomics sample multiplexing space has accelerated considerably. A slew of commercial sample multiplexing

products are now available from companies such as 10x Genomics, Beckton Dickinson, BioLegend, Singleron Biotechnologies, and Millipore-Sigma. These products will undoubtedly increase the adoption of sample multiplexing technologies by the broader single-cell genomics community, as will emerging benchmarking studies which aim to quantitatively clarify the strengths and weaknesses of each approach [64]. Moreover, technological development in the sample-multiplexing space remains active, as is evinced by methods such as Zip-Seq [65], which localize barcodes to user-defined spatial locations using photo-uncageable LMOs and antibody-conjugated oligonucleotides; and CASB [66], which tags cells or nuclei with streptavidin-conjugated concanavalin A beads bound to biotinylated sample barcode oligonucleotides.

Outside of these public demonstrations, at the time of writing this thesis, I have personally sent MULTI-seq reagents and protocols to over 350 academic labs in 21 territories around the world where researchers have applied MULTI-seq to 26 distinct model and non-model organisms. Moreover, I have directly collaborated on a wide diversity of MULTI-seq projects within the UCSF community and beyond. I have leveraged MULTI-seq to do spatial transcriptomic analyses of the developing and adult mouse gastrointestinal tract, perform high-throughput chemical transcriptomics on PBMCs (discussed in Chapter 6) and induced pluripotent stem cell-derived neurons, study primary samples comprised of mammary gland tissue, bone marrow biopsies, primary and metastasized tumors, and other tissue sources, and have adapted MULTI-seq for paired single-cell transcriptomic/epigenomic assays (discussed in Chapter 5) – all of which will be described in future publications. These collaborative experiences have clarified to me the diverse biological questions to which sample multiplexing technologies can be applied, while additionally highlighting a key problem in sample-multiplexed single-cell genomics data analysis that should be prioritized in the coming years.

While many statistical methods for inter-sample comparisons have recently been described including PopAlign [67], MELD [68], PhEMD [69], and others (as reviewed in [70]), a

foundational issue that supersedes this important and growing family of methods relates to the computational task of sample classification. There are currently 7 different sample classification algorithms which use a variety of strategies to assign cells to their sample-of-origin. For example, HTODemux [23] and FBA [71] use unsupervised clustering to identify cells marked by each barcode in sample barcode space, and use these clustering results to compute barcode-specific thresholds for downstream singlet, doublet, and negative identification (analogous to the semi-supervised negative cell reclassification workflow described here). hashedDrops [43] classifies cells according to each cell's most abundant barcode, and identifies doublets via manual or mixture model-informed thresholding of the log2 fold-change distribution describing the difference between each cell's two most abundant sample barcodes. demuxEM [25], GMM-Demux [72], and Bimodal Flexible Fitting [73] also use gaussian mixture models, but instead apply this technique to normalized barcode counts (instead of log2 fold-change distributions, as in hashedDrops) to identify cells that are 'positive' and 'negative' for each barcode prior to singlet, doublet, and negative identification (conceptually analogous to the local maxima detection and quantile thresholding approach described here).

Despite this diverse landscape of sample classification algorithms, however, only minimal effort (with the notable exception of [72]) has been dedicated to systemically comparing the performance of each method on single-cell genomics datasets of varying structure and data quality. While some methods (e.g., HTODemux) are believed to over-estimate doublets (as described here and in [72]), it is unclear if this a general property of the method or a consequence of insufficient benchmarking and optimization. I believe that thorough 'stress-testing' of each algorithm across single-cell genomics datasets of varying structure, sequencing depth, ambient background, and overall sample quality will be central to ensuring that sample-multiplexed datasets are utilized to their fullest potential by the single-cell genomics community. Moreover, I believe that definitive guidelines on how to 'rescue' poor-quality sample multiplexing data using

different quality-control metrics (e.g., SNR thresholding, barcode-specific quality checks, normalization method selection to maximize bimodality, removal of unbarcoded cells and debris-filled droplets, ambient background subtraction, etc.) will further enable the wide adoption of sample multiplexing.

3.7 References

1. Ramsköld D, Luo S, Wang YC, Li R, Deng Q, Faridani OR, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nature Biotechnology. 2012; 30(8): 777-82.
2. Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Reports. 2012; 2(3): 666-73.
3. Gierahn TM, Wadsworth MH 2nd, Hughes TK, Bryson BD, Butler A, Satija R, et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Nature Methods. 2017; 14(4): 395-8.
4. Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. Science. 2017; 357(6352): 661-7.
5. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015; 161(5): 1202-14.
6. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015; 161(5): 1187-1201.
7. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, et al. Massively parallel digital transcriptional profiling of single cells. Nature Communications. 2017; 8: 14049.
8. Habib N, Avraham-Davidi I, Basu A, Burks T, Shekhar K, Hofree M, et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. Nature Methods. 2017; 14(10): 955-8.
9. Tabula Muris Consortium. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature. 2018; 562(7727): 367-72.
10. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The Human Cell Atlas. Elife. 2017; 6: e27041.

11. Wagner DE, Weinreb C, Collins ZM, Briggs JA, Megason SG, Klein AM. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. Science. 2018; 360(6392): 981-7.
12. Ordovas-Montanes J, Dwyer DF, Nyquist SK, Buchheit KM, Vukovic M, Deb C, et al. Allergic inflammatory memory in human respiratory epithelial progenitor cells. Nature. 2018; 560(7720): 649-54.
13. Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nature Biotechnology. 2017; 36(1): 89-94.
14. Heaton H, Talman AM, Knights A, Imaz M, Gaffney DJ, Durbin R, et al. Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. Nature Methods. 2020; 17(6): 615-20.
15. Huang Y, McCarthy DJ, Stegle O Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. Genome Biology. 2019; 20(1): 273.
16. Xu J, Falconer C, Nguyen Q, Crawford J, McKinnon BD, Mortlock S, et al. Genotype-free demultiplexing of pooled single-cell RNA-seq. Genome Biology. 2019; 20(1): 290.
17. Dixit A, Parnas O, Li B, Chen K, Fulco CP, Jerby-Arnon L, et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell. 2016; 167(7): 1853-66.e17.
18. Adamson B, Norman TM, Jost M, Cho MY, Nuñez JK, Chen Y, et al. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. Cell. 2016; 167(7): 1867-82.e21.
19. Jaitin DA, Weiner A, Yofe I, Lara-Astiaso D, Keren-Shaul H, David E, et al. Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. Cell. 2016; 167(7): 1883-96.e15.

20. Aarts M, Georgilis A, Beniazza M, Beolochi P, Banito A, Carroll T, et al. Coupling shRNA screens with single-cell RNA-seq identifies a dual role for mTOR in reprogramming-induced senescence. Genes & Development. 2017; 31(20): 2085-98.
21. Guo C, Kong W, Kamimoto K, Rivera-Gonzalez GC, Yang X, Kirita Y, Morris SA. CellTag Indexing: genetic barcode-based sample multiplexing for single-cell genomics. Genome Biology. 2019; 20(1): 90.
22. Shin D, Lee W, Lee JH, Bang D. Multiplexed single-cell RNA-seq via transient barcoding for simultaneous expression profiling of various drug perturbations. Science Advances. 2019; 5(5): eaav2249.
23. Stoeckius M, Zheng S, Houck-Loomis B, Hao S, Yeung BZ, Mauck WM 3rd, et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biology. 2018; 19(1): 224.
24. Gehring J, Park JH, Chen S, Thomson M, Pachter L. Highly multiplexed single-cell RNA-seq by DNA oligonucleotide tagging of cellular proteins. Nature Biotechnology. 2019; 38(1): 35-8.
25. Gaublomme JT, Li B, McCabe C, Knecht A, Yang Y, Drokhlyanksy E, et al. Nuclei multiplexing with barcoded antibodies for single-nucleus genomics. Nature Communications. 2019; 10(1): 2907.
26. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nature Biotechnology. 2015; 33(5): 495-502.
27. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology. 2018; 36(5): 411-20.
28. Weber RJ, Liang SI, Selden NS, Desai TA, Gartner ZJ. Efficient targeting of fatty-acid modified oligonucleotides to live cell membranes through stepwise assembly. Biomacromolecules. 2014; 15(12): 4621-6.

29. Wu H, Kirita Y, Donnelly EL, Humphreys BD. Advantages of Single-Nucleus over Single-Cell RNA Sequencing of Adult Kidney: Rare Cell Types and Novel Cell States Revealed in Fibrosis. Journal of the American Society of Nephrology. 2019; 30(1): 23-32.
30. Coutelier JP, Kehrl JH, Bellur SS, Kohn LD, Notkins AL, Prabhakar BS. Binding and functional effects of thyroid stimulating hormone on human immune cells. Journal of Clinical Immunology. 1990; 10(4): 204-10.
31. Jeffrey KL, Brummer T, Rolph MS, Liu SM, Callejas NA, Grumont RJ, et al. Positive regulation of immune cell function and inflammatory responses by phosphatase PAC-1. Nature Immunology. 2006; 7(3): 274-83.
32. Ziegler SF, Ramsdell F, Alderson MR. The activation antigen CD69. Stem Cells. 1994; 12(5): 456-65.
33. Lieberman J, Fan Z. Nuclear war: the granzyme A-bomb. Current Opinions in Immunology. 2003; 15(5): 553-9.
34. Garbe JC, Bhattacharya S, Merchant B, Basset E, Swisshelm K, Feiler HS, et al. Molecular distinctions between stasis and telomere attrition senescence barriers shown by long-term culture of normal human mammary epithelial cells. Cancer Research. 2009; 69(19): 7557-68.
35. Brisken C. Progesterone signalling in breast cancer: a neglected hormone coming into the limelight. Nature Reviews Cancer. 2013; 13(6): 385-96.
36. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. Cell Systems. 2019; 8(4): 329-37.e4.
37. Wolock SL, Lopez R, Klein AM. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. Cell Systems. 2019; 8(4): 281-91.e9.
38. Chitale D, Gong Y, Taylor BS, Broderick S, Brennan C, Somwar R, et al. An integrated genomic analysis of lung cancer reveals loss of DUSP4 in EGFR-mutant tumors. Oncogene. 2009; 28(31): 2773-83.

39. Fearon AE, Carter EP, Clayton NS, Wilkes EH, Baker A, Kapitonova E, et al. PHLDA1 Mediates Drug Resistance in Receptor Tyrosine Kinase-Driven Cancer. Cell Reports. 2018; 22(9): 2469-81.
40. Savage P, Blanchet-Cohen A, Revil T, Badescu D, Saleh SMI, Wang Y, et al. A Targetable EGFR-Dependent Tumor-Initiating Program in Breast Cancer. Cell Reports. 2017; 21(5): 1140-9.
41. DeRose YS, Wang G, Lin YC, Bernard PS, Buys SS, Ebbert MTW, et al. Tumor grafts derived from women with breast cancer authentically reflect tumor pathology, growth, metastasis and disease outcomes. Nature Medicine. 2011; 17(11): 1514-20.
42. Jiang K, Sun X, Chen Y, Shen Y, Jarvis JN. RNA sequencing from human neutrophils reveals distinct transcriptional differences associated with chronic inflammatory states. BMC Medical Genomics. 2015; 8:55.
43. Lun ATL, Riesenfeld S, Andrews T, Dao TP, Gomes T, Marioni JC. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome Biology. 2019; 20: 63.
44. Reyfman PA, Walter JM, Joshi N, Anekalla KR, McQuattie-Pimentel AC, Chiu S, et al. Single-Cell Transcriptomic Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis. American Journal of Respiratory and Critical Care Medicine. 2019; 199(12): 1517-36.
45. Jablonska J, Lang S, Sionov RV, Granot Z. The regulation of pre-metastatic niche formation by neutrophils. Oncotarget. 2017; 8(67): 112132-44.
46. Sharma SK, Chintala NK, Vadrevu SK, Patel J, Karbowniczek M, Markiewski MM. Pulmonary alveolar macrophages contribute to the premetastatic niche by suppressing antitumor T cell responses in the lungs. Journal of Immunology. 2015; 194: 5529-38.

47. Condamine T, Ramachandran I, Youn J, Gabrilovich DI. Regulation of Tumor Metastasis by Myeloid-derived Suppressor Cells. Annual Review of Medicine. 2015; 66: 97-110.
48. Kitamura T, Doughty-Shenton D, Cassetta L, Fragkogianni S, Brownlie D, Kato Y, Carragher N, Pollard JW. Monocytes Differentiate to Immune Suppressive Precursors of Metastasis-Associated Macrophages in Mouse Models of Metastatic Breast Cancer. Frontiers in Immunology. 2018; 8: 2004.
49. Catena R, Bhattacharya N, El Rayes T, Wang S, Choi H, Gao D, et al. Bone marrow-derived Gr1+ cells can generate a metastasis-resistant microenvironment via induced secretion of thrombospondin-1. Cancer Discovery. 2013; 3: 578-89.
50. Ouzounova M, Lee E, Piranlioglu R, El Andaloussi A, Kolhe R, Demirci MF, et al. Monocytic and granulocytic myeloid derived suppressor cells differentially regulate spatiotemporal tumour plasticity during metastatic cascade. Nature Communications. 2017; 8: 14979.
51. Nabavi S, SChmolze D, Maitituoheti M, Malladi S, Beck AH. EMDomics: a robust and powerful method for the identification of genes differentially expressed between heterogeneous classes. Bioinformatics. 2016; 32(4): 533-41.
52. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. 2017; 171(6): 143-52.e17.
53. Ye C, Ho DJ, Neri M, Yang C, Kulkarni T, Randhawa R, et al. DRUG-seq for miniaturized high- throughput transcriptome profiling in drug discovery. Nature Communications. 2018; 9(1): 4307.
54. Romero JM, Jiménez P, Cabrera T, Cóza JMr, Pedrinaci S, Tallada M, et al. Coordinated downregulation of the antigen presentation machinery and HLA class I/beta2-microglobulin complex is responsible for HLA-ABC loss in bladder cancer. International Journal of Cancer. 2005; 113(4): 605-10.

55. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P. Simultaneous epitope and transcriptome measurement in single cells. Nature Methods. 2017; 14(9): 865-8.
56. Lim E, Vaillant F, Wu D, Forrest NC, Pal B, Hart AH, et al. Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. Nature Medicine. 2009; 15(8): 907-13.
57. Lawson DA, Bhakta NR, Kessenbrock K, Prummel KD, Yu Y, Takai K, et al. Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. Nature. 2015; 526(7571): 131-5.
58. Satija R, Ferrel JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nature Biotechnology. 2015; 33(5): 495-502.
59. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology. 2018; 36(5): 411-20.
60. van der Maaten LJP. Accelerating t-SNE using Tree-Based Algorithms. Journal of Machine Learning Research. 2014; 15: 3221-45.
61. Morgan M, Anders S, Lawrence M, Aboyoun P, Pagès H, Gentleman R. ShortRead: a Bioconductor package for input, quality assessment and exploration of high-throughput sequence data. Bioinformatics. 2009; 25: 2607-8.
62. van der Loo M. The stringdist package for approximate string matching. The R Journal. 2014; 6: 111-22.
63. Wand MP, Jones MC. Kernel Smoothing Monographs on Statistics and Applied Probability. Chapman & Hall, 1995.

64. Mylka V, Aerts J, Matetovici I, Poovathingal S, Vandamme N, Seurinck R, et al. Comparative analysis of antibody- and lipid-based multiplexing methods for single-cell RNA-seq. bioRxiv. 2020. doi: 10.1101/2020.11.16.384222.

65. Hu KH, Eichorst JP, McGinnis CS, Patterson DM, Chow ED, Kersten K, et al. ZipSeq: barcoding for real-time mapping of single cell transcriptomes. Nature Methods. 2020; 17(8): 833-43.

66. Fang L, Li G, Sun Z, Zhu Q, Cui H, Li Y, et al. CASB: a concanavalin A-based sample barcoding strategy for single-cell sequencing. Molecular Systems Biology. 2021; 17: e10060.

67. Chen S, Rivaud P, Park JH, Tsou T, Charles E, Haliburton JR, et al. Dissecting heterogeneous cell populations across drug and disease conditions with PopAlign. PNAS. 2020; 117(46): 28784-94.

68. Burkhardt DB, Stanley JS 3rd, Tong A, Perdigoto AL, Gigante SA, Herold KC, et al. Quantifying the effect of experimental perturbations at single-cell resolution. Nature Biotechnology. 2021; 39(5): 619-29.

69. Chen WS, Zivanovic N, van Dijk D, Wolf G, Bodenmiller B, Krishnaswamy S. Uncovering axes of variation among single-cell cancer specimens. Nature Methods. 2020; 17(3): 302-10.

70. Ji Y, Lotfollahi M, Wolf FA, Theis FJ. Machine learning for perturbational single-cell omics. Cell Systems. 2021; 12(6): 522-37.

71. Duan J, Hon G. FBA: feature barcoding analysis for single cell RNA-Seq. Bioinformatics. 2021. doi: 10.1093/bioinformatics/btab375.

72. Xin H, Lian W, Jaing Y, Luo J, Wang X, Erb C, et al. GMM-Demux: sample demultiplexing, multiplet detection, experiment planning, and novel cell-type verification in single cell sequencing. Genome Biology. 2020; 21: 188.

73. Bimber B. cellhashR. 2021. Github Repository. https://github.com/BimberLab/cellhashR.

Chapter 4: Introduction

Elements of the following chapter are reprinted from the manuscript “No detectable alloreactive transcriptional responses under standard sample preparation conditions during donor-multiplexed single-cell RNA sequencing of peripheral blood mononuclear cells” by Christopher S. McGinnis, David A. Siegel, Guorui Xie, George Hartoularos, Mars Stone, Chun J. Ye, Zev J. Gartner, Nadia R. Roan, and Sulggi A. Lee, published in *BMC Biology* on January 20, 2021, Volume 19, Issue 1, Article Number 10.

4.1 Abstract

Single-cell RNA sequencing (scRNA-seq) provides high-dimensional measurements of transcript counts in individual cells. However, high assay costs and artifacts associated with analyzing samples across multiple sequencing runs limit the study of large numbers of samples. Sample multiplexing technologies such as MULTI-seq and antibody hashing using single-cell multiplexing kit (SCMK) reagents (BD Biosciences) use sample-specific sequence tags to enable individual samples to be sequenced in a pooled format, markedly lowering per-sample processing and sequencing costs while minimizing technical artifacts. Critically, however, pooling samples could introduce new artifacts, partially negating the benefits of sample multiplexing. In particular, no study to date has evaluated whether pooling peripheral blood mononuclear cells (PBMCs) from unrelated donors under standard scRNA-seq sample preparation conditions (e.g., 30 minute co-incubation at 4 °C) results in significant changes in gene expression resulting from alloreactivity (i.e., response to non-self). The ability to demonstrate minimal to no alloreactivity is crucial to avoid confounded data analyses, particularly for cross-sectional studies evaluating changes in immunologic gene signatures. Here, we applied the 10x Genomics scRNA-seq platform to MULTI-seq and/or SCMK-labeled PBMCs from a single donor with and without pooling

with PBMCs from unrelated donors for 30 minutes at 4 °C. We did not detect any alloreactivity signal between mixed and unmixed PBMCs across a variety of metrics, including alloreactivity marker gene expression in CD4+ T-cells, cell type proportion shifts, and global gene expression profile comparisons using Gene Set Enrichment Analysis and Jensen-Shannon Divergence. These results were additionally mirrored in publicly-available scRNA-seq data generated using a similar experimental design. Moreover, we identified confounding gene expression signatures linked to PBMC preparation method (e.g., Trima apheresis), as well as SCMK sample classification biases against activated CD4+ T-cells which were recapitulated in two other SCMK-incorporating scRNA-seq datasets. Collectively, these observations establish important benchmarks for future cross-sectional immunological scRNA-seq experiments.

4.2 Introduction

Recent advances in single-cell RNA sequencing (scRNA-seq) technologies have dramatically increased assay throughput from ~10^2 to 10^4-10^6 cells per experiment [1]. However, many applications of scRNA-seq workflows (e.g., 10x Genomics) require individual samples to be processed in parallel, which translates to prohibitively-high assay costs for population-scale studies requiring large numbers of samples. Several scRNA-seq sample multiplexing techniques have been developed which enable users to circumvent this limitation by processing samples in a pooled format [2-12]. By avoiding the usual requirement for processing distinct samples individually, these technologies increase scRNA-seq cell and sample throughput while minimizing technical confounders (e.g., doublets and batch effects). Two main types of sample multiplexing approaches have been described: (i) *in silico* genotyping using natural [7-10] or artificial [11,12] genomic variants and (ii) tagging cell membranes with sample-specific DNA barcodes using lipid-modified oligonucleotides (LMOs; e.g., MULTI-seq [2]), DNA-conjugated antibodies [3-5] (e.g.,

BD single-cell multiplexing kit (SCMK) [5]), or methyltetrazine-modified DNA 'ClickTags' [6]). Despite the increasing popularity of sample multiplexing, benchmarking studies aiming to measure transcriptional changes induced by mixing cell suspensions during scRNA-seq sample preparation have not been described. Determining the extent to which these changes might occur is critical, as they would confound cross-sectional data interpretation.

Mixing-specific transcriptional responses could occur when peripheral blood mononuclear cells (PBMCs) from unrelated donors are pooled during scRNA-seq sample preparation. Co-culturing human leukocyte antigen (HLA) mismatched PBMCs causes a rapid and potent allogeneic response wherein T lymphocytes are stimulated through T-cell receptor binding to 'non-self' major and minor histocompatibility complex proteins [13-16]. For example, CD154+ alloreactive CD4+ T-cells were detected within 2 hours after HLA-mismatched lymphocyte mixing [13], while bulk transcriptomics identified a ~5-fold increase within 24 hours of alloreactivity-associated gene expression relative to HLA-matched lymphocytes [14]. Although pooled samples are maintained on ice for short durations during scRNA-seq sample preparation, it is unclear whether the allogeneic response may occur at low temperatures or whether transient periods of warming (e.g., during droplet emulsion at room temperature) are sufficient to drive alloreactivity. Considering that scRNA-seq is sensitive to transcriptional responses in rare cell sub-populations which are obscured by bulk assays, directly assessing whether alloreactivity will confound downstream scRNA-seq analyses is a critical benchmark for large immunological studies [17].

Here, we performed scRNA-seq using the 10x Genomics platform on PBMC samples isolated from eight unrelated healthy donors pooled under conditions where cells from a single donor were processed in isolation or after donor pooling. Donor identities for each cell were assigned using SCMK and MULTI-seq data, as well as the *in silico* genotyping pipeline, souporcell [8]. We observed cell-type biases amongst SCMK classification results which were not due to sub-optimal antibody labeling conditions or the presence of MULTI-seq LMOs. We additionally

did not observe robust, mixing-associated changes in PBMC cell type frequencies, global transcriptional profiles, or alloreactivity-associated gene expression in any PBMC cell type. Finally, we validated the observed lack of alloreactivity in a publicly-available scRNA-seq dataset where PBMCs from two unrelated donors were sequenced in isolation and after pooling [18]. As a result, we conclude that pooling PBMCs from unrelated donors under standard 10x Genomics-based scRNA-seq sample preparation conditions (e.g., 30-minute co-incubation at 4 °C) does not result in any detectable alloreactivity at the RNA level.

4.3 Results

4.3.1 Study Design

To assess whether mixing PBMCs from unrelated donors causes alloreactivity during scRNA-seq, we performed a cross-sectional study of PBMCs isolated from 8 unrelated healthy donors (**Fig. 4-1**). To record the donor-of-origin for each cell, PBMC samples were tagged with donor-specific MULTI-seq [2] and/or SCMK antibody-DNA [5] barcodes. PBMCs were mixed for 30 minutes at 4 °C prior to emulsion across four droplet microfluidics lanes (10x Genomics) at room temperature. The 30-minute pooled incubation was chosen to mimic the typical processing

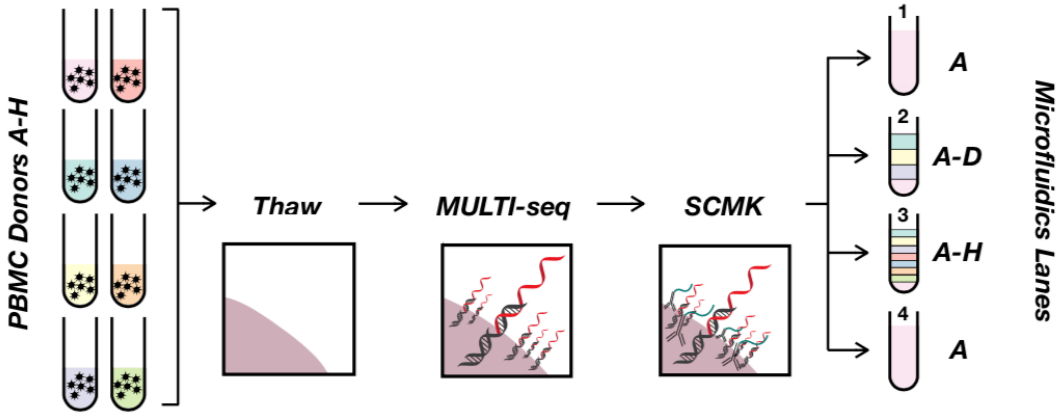

Figure 4-1. Schematic overview of experimental design. PBMCs from 8 HLA-mismatched donors were barcoded with MULTI-seq LMOs (black double-helix with red DNA barcode) and SCMK antibodies (black antibody conjugated to teal DNA barcode). Cells were then strategically pooled to directly assess whether mixing HLA-mismatched PBMCs during scRNA-seq causes alloreactivity.

time required for preparing samples for multiplexed scRNA-seq analysis. Following scRNA-seq data pre-processing, quality-control, cell type annotation and sample demultiplexing (Computational Methods), we compared the expression profile of unmixed donor A PBMCs (microfluidic lane #1) to donor A PBMCs mixed with donors B-D (microfluidic lane #2), donors B-H (microfluidic lane #3), and an unmixed donor A PBMC technical replicate prepared without antibody-DNA labeling (microfluidic lane #4). We hypothesized that if co-incubation of PBMCs from unrelated donors for 30 minutes at 4 °C causes detectable alloreactivity, then mixed and unmixed donor A PBMCs would exhibit more variable gene expression profiles than what is observed due to technical variation

4.3.2 MULTI-seq classifies PBMCs more accurately than SCMK

We first assessed the performance of MULTI-seq and SCMK by comparing the results of three distinct demultiplexing workflows on donor A-H PBMCs from microfluidic lane #3: (i) deMULTIplex, (ii) demuxEM, and (iii) souporcell. deMULTIplex [2] and demuxEM [4] are algorithms that function on sample barcode count matrices, while souporcell is an in silico genotyping pipeline that functions on gene expression data [8]. MULTI-seq and SCMK classifications were largely consistent with souporcell (**Fig. 4-2**) – e.g., amongst cells classified

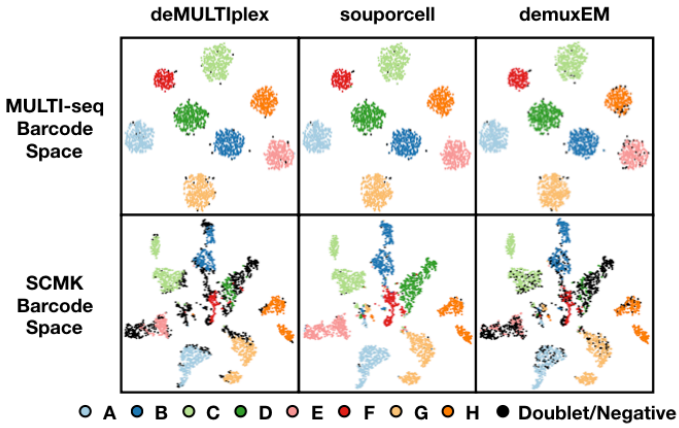

Figure 4-2. MULTI-seq and SCMK classifications largely match in silico genotyping, with lower SCMK classification efficiency. Sample classification results from three demultiplexing pipelines (e.g., deMULTIplex, souporcell, and demuxEM) projected onto MULTI-seq (top) and SCMK (bottom) sample barcode space for microfluidic lane #3.

as donors A-H using souporcell, 99.9% and 99.0% of donor classifications were consistent for MULTI-seq and SCMK, respectively. However, while 1.5% of cells remained unclassified following MULTI-seq demultiplexing, 36.2% of cells remained unclassified after SCMK demultiplexing. This decrease in classification efficiency was also observed when compared to the demuxEM results.

To assess whether cells that remained unclassified following SCMK demultiplexing were randomly distributed throughout the scRNA-seq data, we computed the frequency of unclassified cells for each annotated PBMC cell type (**Fig. 4-3a)**. This analysis revealed that T lymphocytes and NK cells were especially likely to remain unclassified in SCMK data (**Fig. 4-3b**). Moreover, annotation of CD4+ T-cell subsets (**Fig. 4-3c**) revealed that activated CD4+ T-cells were particularly prominent among the unclassified CD4+ T-cells (**Fig. 4-3d**).

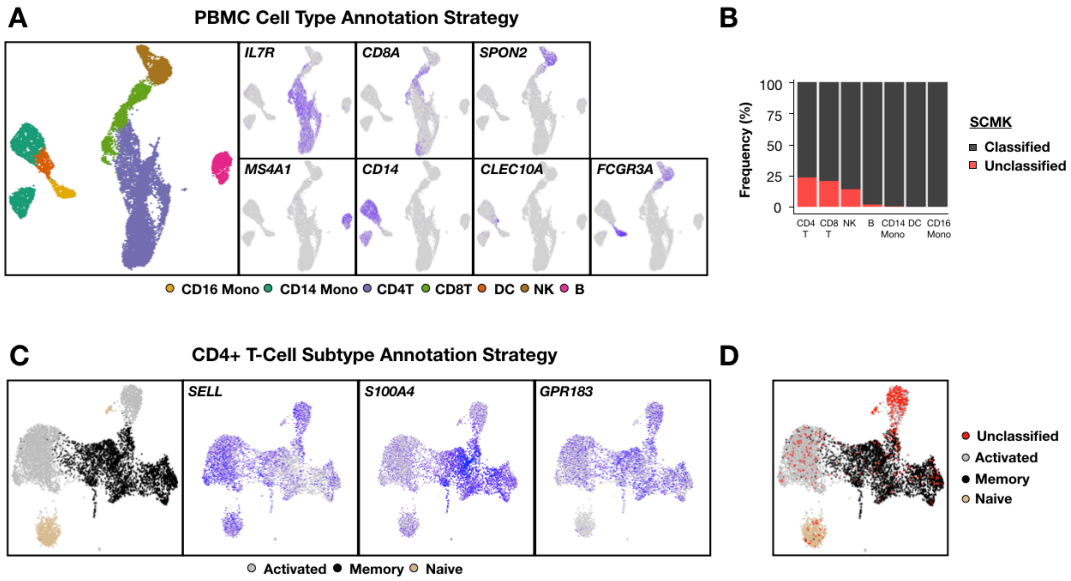

Figure 4-3. SCMK classifications are biased against activated CD4+ T lymphocytes.
(A) UMAP gene expression space for PBMCs colored by cell type annotations (left) or associated marker genes: CD4+ T-cells (IL7R), CD8+ T-cells (CD8A), NK cells (SPON2), B cells (MS4A1), CD14+ classical monocytes (CD14), dendritic cells (CLEC10A), and CD16+ patrolling monocytes (FCGR3A).
(B) Classification frequencies across all PBMC cell types following SCMK sample demultiplexing.
(C) UMAP gene expression space for CD4+ T-cells colored by subtype annotations (left) or associated marker genes: activated (SELL-high, S100A4-low, GPR183-high), memory (SELL-low, S100A4-high, GPR183-high), and naïve (SELL-low, S100A4-high, GPR183-high).
(D) UMAP embedding for CD4+ T-cells colored by annotation and SCMK classification status.

It is conceivable that the presence of LMOs and/or sub-optimal SCMK labeling buffer conditions caused the observed classification biases in PBMCs. To address whether LMOs

interfere with SCMK labeling, we generated scRNA-seq data where cells from 7 PBMC donors were pooled after labeling with SCMK reagents but not LMOs. As was observed previously, SCMK classifications were similarly biased against T lymphocytes and NK cells (**Fig. 4-4a**), with activated CD4+ T-cells being particularly difficult to classify (**Fig. 4-4b**).

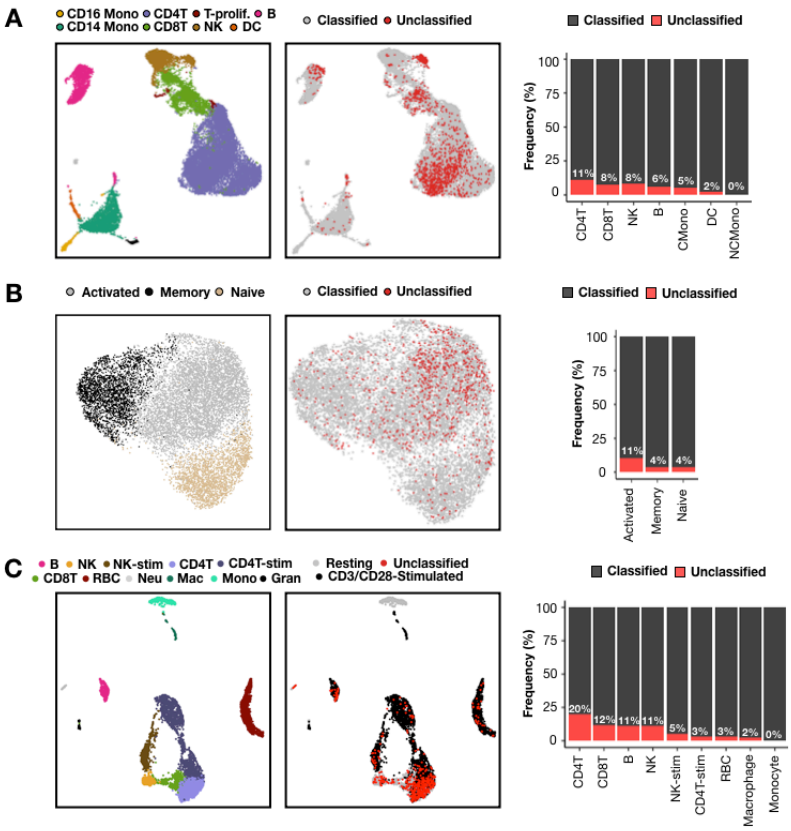

Figure 4-4. Validating SCMK classification biases in independent scRNA-seq datasets generated without LMOs.
(A) UMAP gene expression space for PBMCs from 7-donor multiplexed scRNA-seq experiment colored by cell type annotations (left) or SCMK classification status (middle) with per-cell-type classification frequencies summarized as bar plots (right).
(B) Same analysis as in Fig. 4-4a except for CD4+ T-cells from the 7-donor multiplexed scRNA-seq experiment. Proliferating T-cells (T-prolif) identified using MKI67 expression (data not shown).
(C) Same analysis as in Fig. 4-4a except for PBMCs from the CD3/CD28-stimulation scRNA-seq experiment. Red blood cells (RBC), neutrophils (Neu), macrophages (Mac) and granulocytes (Gran) identified using HBB, LTF, GBP1, and GATA2, respectively.

To address whether SCMK classification biases in PBMCs is due to sub-optimal antibody labeling conditions, we determined the extent of classification bias in a publicly-available scRNA-seq dataset provided by the SCMK reagent supplier where PBMCs were cultured in vitro for 24 hours in the presence or absence of anti-CD3/anti-CD28 antibodies [19]. In these data, SCMK classifications were biased against T lymphocytes and NK cells, despite optimal SCMK labeling

conditions (**Fig. 4-4c**). Collectively, these results illustrate that SCMK reagents produce biased classifications when applied to PBMCs. For these reasons, MULTI-seq donor classifications were used for all subsequent gene expression analyses.

4.3.3 Trima apheresis introduces biologically-relevant confounders into PBMC scRNA-seq data

The PBMCs that were used in this study came from whole blood that was processed using Ficoll-Paque density gradient centrifugation. Notably, these samples either underwent (donors D-H) or did not undergo (donors A-C) apheresis using Trima filtration, a method to enhance leukocyte yield during sample preparation [20,21]. Initial inspection of MULTI-seq donor classifications revealed that PBMCs predominantly clustered according to processing method – e.g., Trima vs. Ficoll (**Fig. 4-5a**). Upon sub-clustering CD14+ classical monocytes and NK cells, we observed that Trima and Ficoll classical monocytes expressed variable levels of the histone component gene HIST1H1C, as well as two genes involved in monocyte differentiation, MNDA and CEBPB (**Fig. 4-5b**, left) [22]. Moreover, we observed that Trima and Ficoll NK cells differentially expressed the immune cytokine IFNG, cytolytic genes GZMA and PRF1, and the stress marker JUN (**Fig. 4-5b**, right) [23].

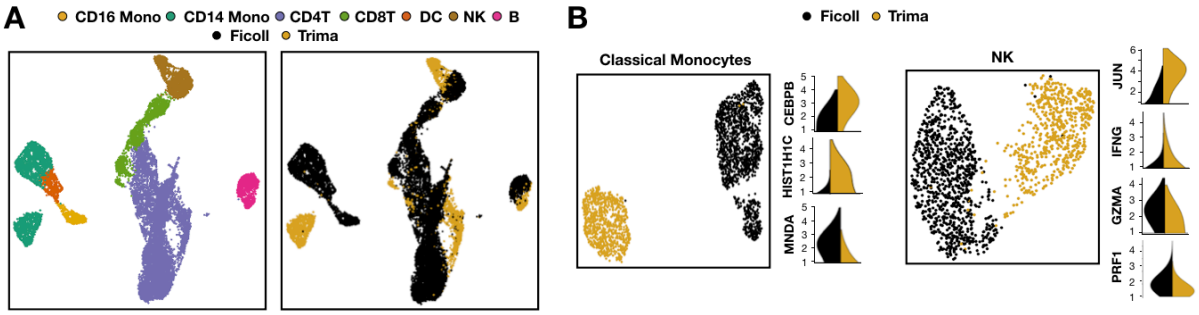

Figure 4-5. Trima-associated gene expression signatures
(A) UMAP gene expression space for PBMCs colored by cell type (left) or method of isolation (e.g., Ficoll or Trima; right).
(B) UMAP gene expression space for classical monocytes (left) and NK cells (right) with violin plots illustrating Trima-specific marker gene expression depicted as log1p-normalized counts.

These results suggest that apheresis using Trima filters induces confounding changes in gene expression patterns associated with differentiation state, cytolytic activity, and stress across multiple PBMC cell types. These signatures are consistent with prior observations [24] and should be accounted for in future analyses. Thus, to avoid these confounding effects when comparing donor- and mixing-specific expression profiles, we restricted our subsequent analyses to PBMC samples processed without Trima filtration.

4.3.4 Mixing PBMCs from unrelated healthy donors during scRNA-seq sample preparation does not cause a detectable allogeneic transcriptional response

To assess whether mixing PBMCs from unrelated donors induces alloreactivity during multiplexed scRNA-seq, we compared the expression profiles of mixed and unmixed donor A PBMCs. Mapping the densities of mixed and unmixed donor A sample classifications onto PBMC gene expression space (**Fig. 4-6**, top left) did not reveal any qualitative shifts in global gene expression profiles (**Fig. 4-6**, bottom). Notably, such shifts in classification densities were observed when including PBMCs from donors B and C (**Fig. 4-6**, top right), suggesting that natural

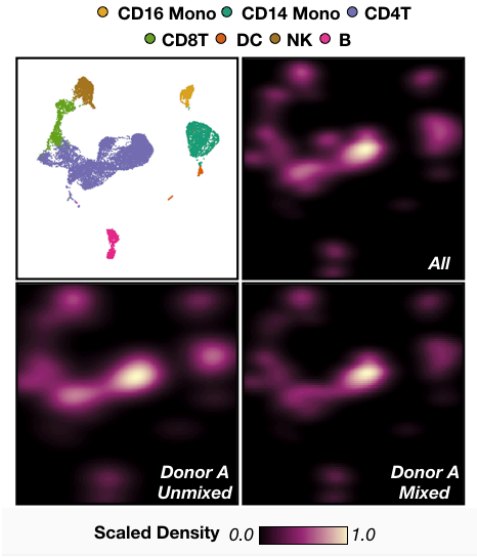

Figure 4-6. Qualitative assessment of allogeneic transcriptional response. UMAP gene expression space for Ficoll-isolated PBMCs colored by cell type (top left) or as sample classification densities for unmixed donor A (bottom left), mixed Donor A (bottom right), and Donors A-C (top right).

inter-donor variation is more pronounced than intra-donor variation due to PBMC mixing. Indeed, PBMC cell-type frequencies were similarly-variable between donors, while no statistically-significant shifts in cell-type frequencies were linked to mixing status (**Fig. 4-7**).

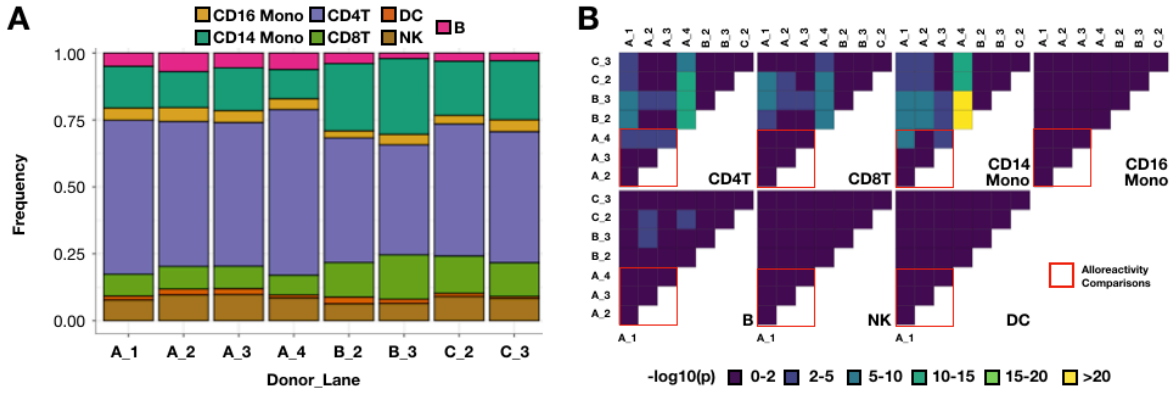

Figure 4-7. PBMC cell type proportions are not influenced by mixing status
(A) PBMC cell-type frequencies for each donor (e.g., A, B, C) and 10x lane (e.g., 1, 2, 3, 4).
(B) Pairwise proportion test of PBMC cell type proportion differences between each donor and microfluidic lane, visualized as -log10 p-value heatmaps for each cell type. Red boxes denote key comparisons for assessing putative effects of alloreactivity on cell type proportions.

Next, we focused on CD4+ T-cells because of their known involvement in alloreactivity [13-16]. Mixed and unmixed donor A CD4+ T-cells clustered together in CD4+ T-cell gene expression space, as was observed in the full dataset (**Fig. 4-8a**). Moreover, mixed and unmixed donor A CD4+ T-cells expressed genes known to be involved in an allogeneic response [13-16] at similar levels (**Fig. 4-8b**). Finally, no statistically-significant shifts in CD4+ T-cell subtype frequencies were linked to mixing status (**Figs. 4-8c, 4-8d**).

Expanding our analysis to other PBMC cell types, we next applied two unbiased approaches to measure any putative allogeneic response signature. First, we used the dissimilarity metric Jensen-Shannon Divergence (JSD) [25] to compute sample-level differences for each PBMC cell type. To control for differences in cell type proportions, we randomly subsetted equal numbers of each cell type from each experimental group during PBMC sub-clustering and repeated this workflow 100 times. Across the 100 iterations, we then computed the average JSD

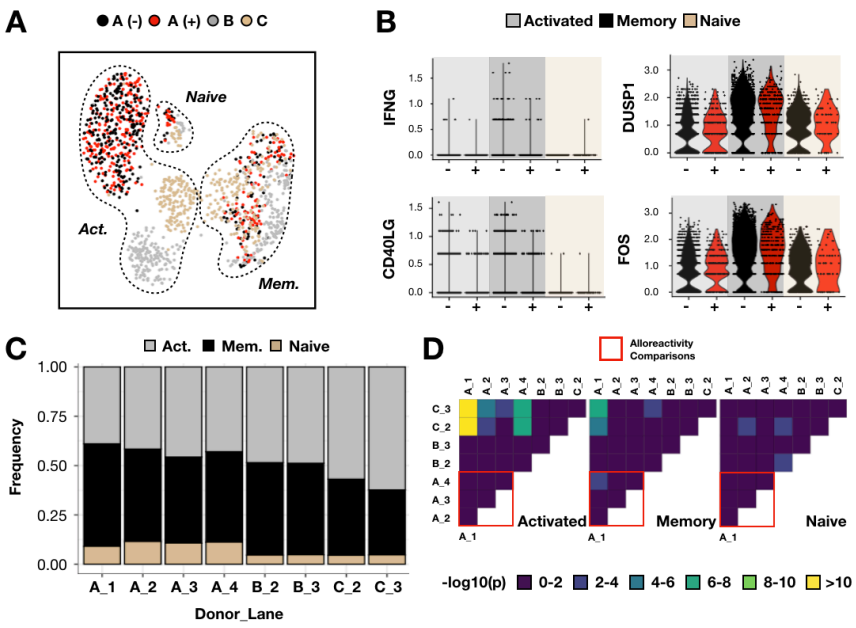

Figure 4-8. Evidence of allogeneic response not detected in CD4+ T-cell gene expression state or subtype proportions.
(A) UMAP gene expression space for Ficoll-isolated CD4+ T-cells colored by donor ID (e.g., A, B, C) and mixing status e.g., - = unmixed, + = mixed).
(B) Expression of genes known to be up-regulated (e.g., IFNG and CD40LG) or down-regulated (e.g., DUSP1 and FOS) by CD4+ T lymphocytes during an allogenic response across unmixed (black) and mixed (red) donor A CD4+ T-cell subsets.
(C) CD4+ T-cell subtype frequencies for each donor and 10x lane.
(D) Pairwise proportion test of CD4+ T-cell subtype proportion differences between each donor and microfluidic lane, visualized as -log10 p-value heatmaps for each cell type. Red boxes denote key comparisons for assessing putative effects of alloreactivity on CD4+ T-cell subtype proportions.

scores between donors, unmixed and mixed donor A cells, and technical replicates (Methods). For all cell types, inter-donor JSD scores were greater than those linked to mixing status and technical replicate, while mixing status JSD scores were greater than technical replicate JSD scores for CD8+ T-cells, CD14+ monocytes, and NK cells (**Fig. 4-9**).

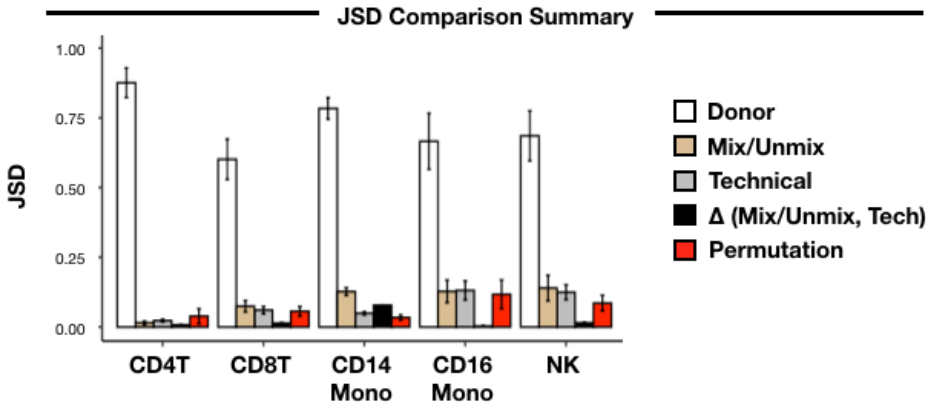

Figure 4-9. Iterative inter-sample JSD comparison analysis quantifies lack of allogeneic response to donor mixing.
Bar plots denote average JSD between PBMC donors (white), mixed/unmixed donor A cells (beige), technical replicates (grey), and donor A cells following label permutation (red). Difference in JSD scores between mixed/unmixed and technical replicates depicted in black. Error bars denote +/- 1 standard deviation. n = 100 iterations.

To determine the likelihood of observing elevated mixing status JSD scores relative to technical replicates by chance, we repeated this workflow after permuting donor A classifications. Specifically, we reasoned that if permuted JSD scores were greater than the difference between observed mixing status and technical replicate JSD scores, then the observed differences are not significant. To this end, JSD scores after donor A label permutation were larger than the experimental JSD score differential in all cell types except CD14+ classical monocytes (**Fig. 4-9**).

The second unbiased approach we utilized to look for allogeneic response signatures was Gene Set Enrichment Analysis (GSEA) [26,27]. Specifically, we applied GSEA to donor A cells from each PBMC cell type to determine whether pathways involved in immune activation and/or alloreactivity were enriched in mixed relative to unmixed cells. This analysis revealed that amongst unmixed donor A cells, activated CD4+ T-cells were enriched for humoral immune response genes and dendritic cells were enriched for epigenetic regulation and cell killing (**Table 4-1**). Notably, these detected gene sets in unmixed cells are not consistent with an allogeneic response, and no enriched gene sets were identified amongst any mixed donor A cell types.

Table 4-1. Enriched GO gene sets in unmixed PBMCs, 8-donor PBMC scRNA-seq data.

Cell Type	Gene Set	P-value	FDR q
CD4T-Act.	Humoral Immune Response	0	0.045
DC	Histone H3K4 Methylation	0	0.009
DC	Positive Regulation of Cell Killing	0	0.04
DC	Peptidyl Lysine Methylation	0	0.049
DC	Regulation of Cell Killing	0	0.045

It is conceivable that the presence of LMOs and antibody-DNA barcodes could delay or block any allogeneic response between PBMCs from unrelated donors. To explore this possibility, we repeated our analytical workflow on a publicly-available scRNA-seq dataset where PBMCs from two unrelated healthy donors were sequenced in isolation and after pooling and incubation on ice for 30 minutes [18]. Mirroring our previous observations, mixing was not robustly associated with any statistically-significant shifts in PBMC cell type proportions (**Figs. 4-10a, 4-10b**) or CD4+

T-cell subtypes (**Figs. 4-10c, 4-10d**). Moreover, cells clustered primarily by donor and not mixing status (**Fig. 4-10e**), and inter-donor JSD scores were greater than mixing status JSD scores for all cell types (**Fig. 4-10f**). Although this experimental design did not allow comparisons between JSD scores linked to mixing status and technical replicates, permuted and mixing status JSD scores were on-par for most cell types (including CD14+ monocytes; **Fig. 4-10f**).

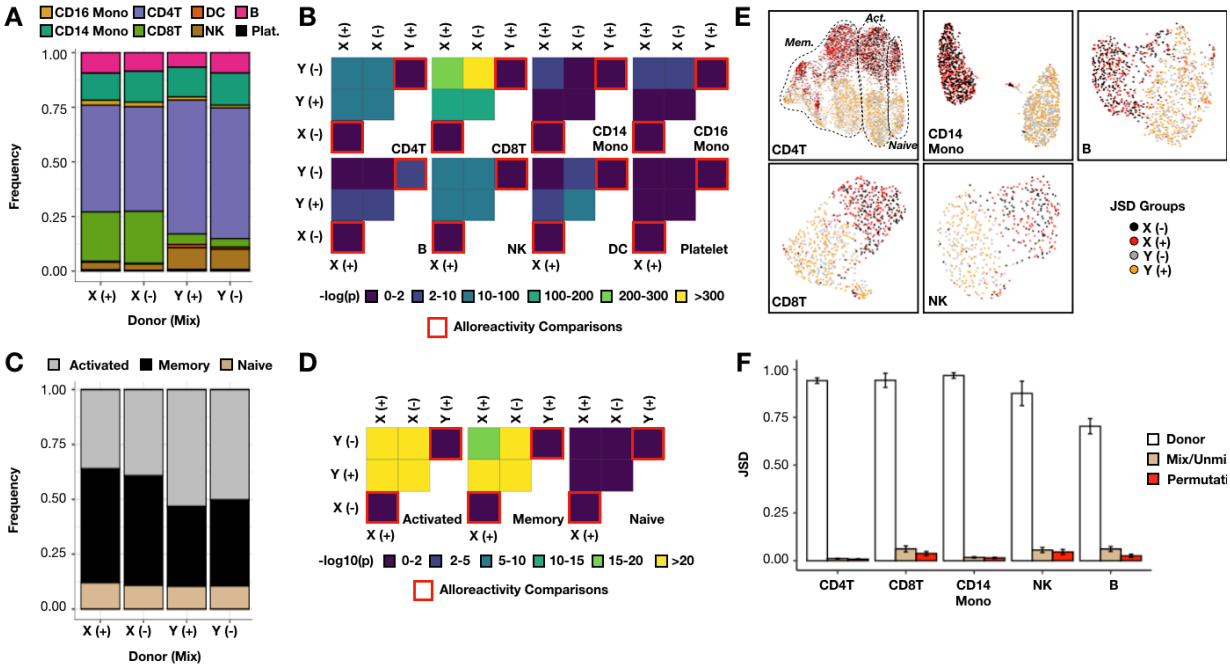

Figure 4-10. No detectable differences in PBMC cell type proportions, CD4+ T-cell proportions, or gene expression state linked to alloreactivity in Zheng et al scRNA-seq data.
(A) PBMC cell-type frequencies grouped by donor (e.g., X, Y) and mixing status (e.g., - = unmixed, + = mixed).
(B) Pairwise proportion test of PBMC cell type proportion differences between mixed and unmixed donors, visualized as -log10 p-value heatmaps for each cell type. Red boxes denote key comparisons for assessing putative effects of alloreactivity on PBMC cell type proportions. Significant differences in cell type proportions are mostly linked to donor, while the singular difference between mixed and unmixed donor Y B-cells is not recapitulated in donor X B-cells.
(C) CD4+ T-cell subtype frequencies grouped as in Fig. 4-10a.
(D) Pairwise proportion test of CD4+ T-cell subtypes proportion differences between mixed and unmixed donors, visualized as in Fig. 4-10b. Significant differences in cell type proportions are entirely linked to donor.
(E) Representative CD4+ T-cell, CD8+ T-cell, CD14+ monocyte, NK cell, and B-cell UMAP gene expression embeddings following iterative subsetting to select equal numbers from each JSD comparison group. Cells are colored according to donor ID (e.g., X, Y) and mixing status (e.g., - = unmixed, + = mixed).
(F) JSD analysis summary. Bar plots denote average JSD between PBMC donors (white), mixed/unmixed cells (beige), and cells following label permutation (red). Error bars denote +/- 1 standard deviation. n = 100 iterations.

Finally, while GSEA identified a number of enriched gene sets amongst mixed PBMCs in these data (e.g., protein trafficking, translation, non-sense mediated decay, viral gene expression, and amino acid metabolism; **Table 4-2**), these gene sets were unrelated to alloreactivity and were

shared across most PBMC cell types, suggesting they were caused by batch effects between the mixed and unmixed scRNA-seq libraries.

Table 4-2. Enriched GO gene sets in mixed PBMCs, Zheng et al PBMC scRNA-seq data.

Gene Set	P	FDR
B Cells		
Establishment of protein localization to ER	0	0.001
Protein targeting to membrane	0	0.001
Protein localization to ER	0	0.001
Cytoplasmic translation	0	0.001
Cotranslational protein targeting to membrane	0	0.001
Nuclear transcribed mRNA catabolic process nonsense mediated decay	0	0.001
Translational initiation	0	0.002
Viral gene expression	0	0.002
Nuclear transcribed mRNA catabolic process	0	0.002
Establishment of protein localization to membrane	0	0.003
Protein targeting	0	0.004
Protein localization to membrane	0	0.007
RNA catabolic proces	0	0.021
Ribosomal small subunit biogenesis	0.001	0.021
Establishment of protein localization to organelle	0	0.024
Organic cyclic compound catabolic process	0	0.025
Cellular macromolecule catabolic process	0	0.045
CD16 Monocytes		
Establishment of protein localization to ER	0	0.001
Cotranslational protein targeting to membrane	0	0.001
Protein localization to ER	0	0.001
Viral gene expression	0	0.001
Nuclear transcribed mRNA catabolic process nonsense mediated decay	0	0.004
Protein targeting to membrane	0	0.003
Nuclear transcribed mRNA catabolic process	0	0.009
Translation initiaion	0	0.041
Establishment of protein localization to membrane	0	0.048
Activated CD4+ T-cells		
Cytoplasmic translation	0	0.001
Translation initiaion	0	0.001
Establishment of protein localization to ER	0	0.002
Protein targeting	0	0.001
Cotranslational protein targeting to membrane	0	0.001
Protein localization to membrane	0	0.001
Protein localization to ER	0	0.001
Nuclear transcribed mRNA catabolic process nonsense mediated decay	0	0.001
Protein localization to organelle	0	0.001
Protein targeting to membrane	0	0.001
Nuclear transcribed mRNA catabolic process	0	0.001
Establishment of protein localization to organelle	0	0.001
RNA catabolic process	0	0.001
Establishment of protein localization to membrane	0	0.001
Intracellular protein transport	0	0.001
Viral gene expression	0	0.001
Organic cyclic comoound catabolic process	0	0.001
Cellular macromolecule catabolic process	0	0.003
Macromolecule catabolic process	0	0.003
Peptide biosynthetic process	0	0.004
Amide biosynthetic process	0	0.004
Intracellular transport	0	0.007
Cellular amide biosynthetic process	0	0.009
Interspecies interaction between organisms	0	0.012
Cellular macromolecule localization	0	0.013
CD8+ T-cells		
Protein localization to ER	0	0
Protein targeting to membrane	0	0
Establishment of protein localization to ER	0	0
Cotranslational protein targeting to membrane	0	0

Gene Set	P	FDR
Memory CD4+ T-cells		
Protein localization to ER	0	0.003
Establishment of protein localization to ER	0	0.002
Nuclear transcribed mRNA catabolic process nonsense mediated decay	0	0.001
Cotranslational protein targeting to membrane	0	0.001
Protein targeting	0	0.001
Nuclear transcribed mRNA catabolic process	0	0.001
Establishment of protein localization to membrane	0	0.001
Viral gene expression	0	0.001
Protein targeting to membrane	0	0.001
Cytoplasmic translation	0	0.001
Translational initiation	0	0.003
Ribosomal small subunit biogenesis	0	0.003
Establishment of protein localization to organelle	0	0.003
Protein localization to membrane	0	0.003
Organic cyclic compound catabolic process	0	0.008
RNA catabolic proces	0	0.008
Protein localization to organelle	0	0.018
Peptide biosynthetic process	0	0.031
Intracellular protein transport	0	0.031
Amide biosynthetic process	0	0.046
Naive CD4+ T-cells		
Establishment of protein localization to membrane	0	0
Cytoplasmic translation	0	0
Cotranslational protein targeting to membrane	0	0
Nuclear transcribed mRNA catabolic process nonsense mediated decay	0	0
Establishment of protein localization to ER	0	0
Nuclear transcribed mRNA catabolic process	0	0
Viral gene expression	0	0
Protein localization to ER	0	0
Protein targeting to membrane	0	0
Translation initiaion	0	0
Protein localization to membrane	0	0
Protein targeting	0	0
Establishment of protein localization to organelle	0	0
RNA catabolic proces	0	0
Organic cyclic compound catabolic process	0	0
Protein localization to organelle	0	0
Peptide biosynthetic process	0	0.001
Intracellular protein transport	0	0.002
Amide biosynthetic process	0	0.002
Cellular macromolecule catabolic process	0	0.003
Interspecies interaction between organisms	0	0.003
Cellular amide metabolic process	0	0.004
Macromolecule catabolic process	0	0.004
Intracellular transport	0	0.004
Cellular macromolecule localization	0	0.011
Organonitrogen compound biosynthetic proces	0	0.019
Ribosomal small subunit biogenesis	0.009	0.031
NK Cells		
Cotranslational protein targeting to membrane	0	0.012
Nuclear transcribed mRNA catabolic process nonsense mediated decay	0	0.008
Protein localization to ER	0	0.008
Establishment of protein localization to ER	0	0.008
Viral gene expression	0	0.01
Protein targeting to membrane	0	0.016
Translational initiation	0	0.025
Nuclear transcribed mRNA catabolic process	0	0.025
DCs		
Cotranslational protein targeting to membrane	0	0.005
Establishment of protein localization to ER	0	0.014
Protein localization to ER	0	0.02
Nuclear transcribed mRNA catabolic process nonsense mediated decay	0	0.017

Collectively, these targeted and unbiased quantitative comparisons across all PBMC cell types in two, independently-generated scRNA-seq datasets demonstrate that mixing PBMCs from unrelated donors under standard multiplexed scRNA-seq sample preparation conditions (e.g., 30 minute co-incubation at 4 °C) does not result in a detectable allogeneic transcriptional response.

4.4 Discussion

Sample multiplexing approaches for scRNA-seq are being increasingly utilized by the single-cell genomics field to reduce assay costs while improving data breadth and quality. However, the impact of pooling PBMCs from unrelated donors during scRNA-seq sample preparation on gene expression patterns has not yet been adequately quantified. Here, we used the 10x Genomics scRNA-seq platform to directly compare the gene expression profiles of PBMCs prepared for sequencing alone or after mixing with PBMCs from unrelated donors for 30 minutes at 4 °C. We found no evidence of global changes in gene expression profiles in any PBMC cell type (quantified using JSD and GSEA), PBMC cell type proportions, or alloreactivity marker gene expression in CD4+ T-cells linked to PBMC mixing status. Although PBMCs actively participating in an allogeneic response were not included in this study, these observations were mirrored in an independently-generated, publicly-available PBMC scRNA-seq dataset [18], demonstrating that mixing unrelated PBMCs during sample-multiplexed scRNA-seq sample preparation does not result in a detectable allogeneic response. Notably, it is possible that cellular responses to pooling could be detected by assays measuring levels of biological information with faster regulatory kinetics (e.g., cell surface protein assays [28,29]) or under different scRNA-seq experimental conditions (longer periods of co-incubation at higher temperatures, mixing cells from distinct species, etc.). To this end, the experimental design employed in this study can be used

to benchmark the prevalence of sample mixing-specific confounders in future single-cell genomics experiments.

In addition to the alloreactivity analysis, we found that Trima apheresis can introduce confounding variables into scRNA-seq data, which suggests that this PBMC preparation method should be avoided in future experiments. Moreover, we found that SCMK demultiplexing results were biased against activated CD4+ T-cells and other lymphoid cell types. This observation was mirrored in two scRNA-seq datasets generated (i) in the absence of LMOs and (ii) with optimized SCMK antibody-DNA labeling conditions. These findings are in contrast to the original Cell Hashing report [3], where PBMCs were systematically demultiplexed following incubation with a panel of DNA-conjugated antibodies selected for their uniform targeting of all known PBMC cell populations. Notably, the exact antigens targeted by the commercial SCMK reagents used in this study are proprietary and unknown, but our findings suggest that the “universal” antigens targeted by these antibody-DNA conjugates may be differentially-expressed by distinct cell types in ways that interfere with sample classification. It remains to be determined whether “universal” Cell Hashing reagents from BioLegend, which target human beta-2-microglobulin and CD298, suffer from similar performance issues. Thus, users should exercise caution before using SCMK reagents, for example by testing the uniformity of antibody binding conducting flow-cytometry experiments with fluorophore-conjugated DNA probes that hybridize to SCMK oligonucleotide domains. In any case, validation of surface antigen expression across all cells in a given experimental system and/or careful data quality-control is necessary to avoid systematically-biased interpretations.

Collectively, this study proposes three critical benchmarks for future sample-multiplexed scRNA-seq analyses of PBMCs. First, we demonstrate that alloreactivity can be disregarded as a potential confounder when analyzing scRNA-seq data from PBMCs of unrelated donors pooled under standard multiplexed scRNA-seq sample preparation conditions. These conclusions may

not, however, be generalizable to all single-cell genomics assays or sample preparation workflows. Second, we demonstrate that Trima apheresis of PBMCs introduces artifactual gene expression signatures which can confound downstream scRNA-seq data analyses. Third, we demonstrate that SCMK reagents are biased against certain PBMC cell types, which illustrates the importance of validating antibody-based sample multiplexing technology performance.

4.5 Materials and Methods

4.5.1 scRNA-seq sample preparation, 8-donor MULTI-seq/SCMK PBMC experiment

PBMCs were provided by the Vitalant Research Institute. PBMCs were thawed at 37°C and washed one time with warm media (RPMI (Corning, Cat#10-040-CV), supplemented with 10% FBS (VWR, Cat#97068-085) and Benzonase (1:1000, Sigma-Aldrich, Cat#E1014)) and one time with 2% FBS in PBS (Ca^{++} and Mg^{+} free, Corning, Cat#21-031-CV) before counting cells (Nexcelom K2). Live cells were then enriched using a dead-cell removal kit (STEM Cell, Cat#17899). Live cells were then washed with PBS and labeled with LMOs, as described previously [2]. LMOs were then quenched while washing cells with 1% BSA in cold PBS. Cells were then incubated with 5ul human Fc Block with 95ul 2% FBS in PBS at 4°C for 15 minutes before staining with SCMK and AbSeq antibodies (BD Biosciences) at 4°C for 60 minutes. Notably, AbSeq data was not analyzed in this study, and a subset of donor A PBMCs were not labeled with antibody-DNA conjugates (sequenced in microfluidics lane 4). Cells were then washed twice by using 0.04% BSA (Non-acetylate, Sigma-Aldrich; B6917)) in cold media before incubation for 30 minutes at 4°C either alone (e.g., donor A) or in a pooled format (e.g., donors A-D or A-H). Cell viabilities for each donor prior to pooling ranged from 89%-97%. Finally, cells were isolated via droplet emulsion across four 10x Genomics microfluidic lanes (V2) to yield 5,000 cells.

4.5.2 scRNA-seq sample preparation, 7-donor SCMK PBMC experiment

Healthy donor PBMCs were used from the ImmVar project [30], isolated from whole blood and frozen as described therein. Vials from 7 patients, each with 1 million, were thawed at 37°C and washed once with warm media before staining with SCMK antibodies. Briefly, cells were stained for 20 minutes at room temperature before being washed 3 times in 2mL BD stain buffer. Cells were then counted, pooled, resuspended in 0.04% BSA in PBS, and isolated via droplet emulsion across a single 10x Genomics microfluidic lane (V2) to yield 50,000 cells.

4.5.3 Next-generation sequencing and library preparation

cDNA expression, MULTI-seq, and SCMK libraries were prepared as described previously [2] or according to supplier recommendations. Notably, following size-selection of MULTI-seq and SCMK oligos after cDNA amplification, two separate sample-index PCRs were performed for the MULTI-seq and SCMK oligos using separate i7 indices. For the 8-donor and 7-donor PBMC experiments, cDNA expression and SCMK libraries were pooled and sequenced on a single NovaSeq 6000 lane (one lane per experiment). MULTI-seq libraries were sequenced separately using the MiSeq (V3).

4.5.4 scRNA-seq data pre-processing

Eight next-generation sequencing libraries from four separate experiments were analyzed in this study. Data pre-processing details for each library are summarized in **Table 4-3**. Notably, because the MULTI-seq and SCMK barcode sequences are 8 and 40 nucleotides in length, respectively, the Hamming Distance alignment threshold applied to SCMK data was increased to 5 (default = 1) to account for the increased probability of random sequencing errors.

Table 4-3. Data pre-processing details for all presented datasets and modalities.

Experiment	Library	Details
8-Donor PBMC	scRNA-seq	Cell Ranger (v3.0.0), hg19 reference, read-dpeth normaliation. *In silico* genotypicing using souprocell [8].
8-Donor PBMC	MULTI-seq	deMULTIplex (v1.0.2), Hamming Distance = 1.
8-Donor PBMC	SCMK	deMULTIplex (v1.0.2), Hamming Distance = 5.
7-Donor PBMC	scRNA-seq	Cell Ranger (v3.0.0), custom hg19 reference containing SCMK barcoes. *In silico* genotypicing using Demuxlet [7] (genotype error offset = 0.1, alpha = 0.5, mapping quality = 255).
7-Donor PBMC	SCMK	Cell Ranger (v3.0.0), custom hg19 reference containing SCMK barcoes. R2 FASTQs trimmed using Trimmomatic[31] (single-end mode. HEADCROP = 25, CROP = 45).
Zheng et al PBMC	scRNA-seq	Cell Ranger (v3.0.0), hg19 reference, read-dpeth normaliation.
2-Condition PBMC (BD)	scRNA-seq	Downloaded from provider [19].
2-Condition PBMC (BD)	SCMK	Downloaded from provider [19].

4.5.5 scRNA-seq data quality-control

The same quality-control workflows were applied to the 8-donor and Zheng et al PBMC datasets using Seurat [32,33]. First, cells with fewer than 250 RNA UMIs and genes with fewer than 3 UMIs across all cells were discarded. These parsed datasets were then normalized using 'SCTransform' prior to unsupervised clustering and dimensionality reduction using PCA and UMAP. Low-quality cells selected via membership in clusters associated with low total RNA UMIs and/or high proportions of mitochondrial gene expression were then removed.

Next, we split the cleaned datasets by microfluidic lane-of-origin and identified heterotypic doublets using DoubletFinder [34]. Notably, DoubletFinder was run on each lane independently to ensure that representative artificial doublets were constructed for each lane (e.g., multi-donor doublets were not generated for the unmixed data subsets). Moreover, we did not use MULTI-seq, SCMK, or souporcell classification results for doublet detection because each approach would produce different results for each lane (e.g., no doublets would be detected for single-donor datasets). DoubletFinder resulted in the removal of 1,287 and 1,832 heterotypic doublets in the 8-donor PBMC and Zheng et al PBMC datasets, respectively. DoubletFinder parameters were optimized using the 'paramSweep_v3', 'summarizeSweep', and 'find.pK' functions in the 'DoubletFinder' R package, as described previously [34]. DoubletFinder pK parameters employed

are as follows (pN = 0.25 for all datasets): 8-Donor PBMC (pK = 0.01), Zheng et al., Lane 1 (0.07), Zheng et al., Lane 2 (0.09), and Zheng et al., Lane 3 (0.08).

Notably, a simplified quality-control workflow was applied to the 7-donor and 2-condition PBMC datasets to assess the influence of (i) LMO labeling and (ii) SCMK antibody-DNA labeling conditions on SCMK demultiplexing performance. More stringent quality-control steps were not employed because these datasets were not being used to assess alloreactivity gene expression signatures. Briefly, raw gene expression matrices were parsed as described above before the data was log2-transformed, centered, and scaled. Following unsupervised clustering, the top 2,000 variable genes (selection.method = 'vst') were then used for dimensionality reduction using PCA and UMAP. Finally, low-quality cells were removed as described above. Summary statistics for each dataset following quality-control are as follows: 8-Donor PBMC (5,042 mean UMIs; 1,265 mean genes; 15,340 total cells), Zheng et al PBMC (1,883; 681; 25,140), 7-Donor PBMC (2,222; 695; 25,140), 2-Condition PBMC (2,836; 996; 5,419).

4.5.6 PBMC cell type annotation

We annotated cell types within each PBMC dataset using literature-supported cell type marker genes [32,33,35]. Marker genes employed are as follows: CD4+ T lymphocytes (IL7R), CD8+ T lymphocytes (CD8A), NK cells (SPON2), B lymphocytes (MS4A1), classical monocytes (CD14), non-classical monocytes (FCGR3A), dendritic cells (CLEC10A), platelets (PF4), proliferative cells (MKI67), plasma cells (MZB1), plasmacytoid dendritic cells (LILRA4), granulocytes (GATA2), neutrophils (LTF), erythrocytes (HBB), macrophages (GBP1), CD3/CD28-stimulated NK cells (GNLY), and CD3/CD28-stimulated T-cells (ENO1). Marker genes employed for CD4+ T-cell subtype annotation are as follows: activated (SELL-high, S100A4-low, GPR183-high), memory (SELL-low, S100A4-high, GPR183-high), and naïve CD4+ T-cells (SELL-high, S100A4-low, GPR183-low).

4.5.7 MULTI-seq, SCMK, and souporcell classification

For the 8-donor PBMC dataset, cells were classified into donor groups using three different workflows. First, MULTI-seq and SCMK barcode count matrices were fed into the 'classifyCells' and 'findThresh' functions in the deMULTIplex R package [2]. Second, MULTI-seq and SCMK barcode count matrices and the raw .h5 file (from Cell Ranger) were fed into demuxEM (p=8), an alternative sample classification pipeline [4]. Third, position-sorted BAM files (from Cell Ranger) were fed into the *in silico* genotyping pipeline, souporcell (k=8) [8]. For the 7-donor PBMC dataset, SCMK barcode count matrices were only analyzed using deMUTIplex, as deMULTIplex, DemuxEM, and souporcell results were observed to be consistent. For the Zheng et al PBMC dataset, donor identifies were inferred using souporcell (k=2), as MULTI-seq/SCMK barcode count matrices were unavailable. For the 2-condition PBMC dataset, classifications were provided from the supplier.

4.5.8 PBMC cell type proportion analysis

To determine whether mixing PBMCs from unrelated donors results in changes in PBMC cell type proportions in the 8-donor and Zheng et al PBMC datasets, we first computed the frequency of each cell type grouped according to donor and microfluidic lane. Statistically-significant proportional differences between groups were then identified on a per-cell-type basis using the 'pairwise.prop.test' function in the stats R package using default arguments. Evidence of alloreactivity-associated shifts in cell type proportions were assessed by comparing p-values for donor A cell type proportions. Statistically-significant shifts were never identified between donor A cells from microfluidic lane 1 (A1) and A2/A3 cells, although shifts were detected between A1/A2/A3 and A4, perhaps due to technical variability. This workflow was additionally repeated for CD4+ T-cell subsets yielding similar results.

4.5.9 Jensen Shannon Divergence (JSD) analysis

To perform global comparisons of gene expression profiles between mixed and unmixed PBMCs in the 8-donor and Zheng et al PBMC datasets, we used JSD in the following workflow. First, each PBMC cell type was randomly down-sampled to include equal numbers of cells from each donor and microfluidic lane. Down-sampling in this fashion ensures that any observed differences are due to gene expression state and not cell type proportions. Next, UMAP embeddings were computed for each cell type, and UMAP coordinates for each donor/lane group were used to compute group-wise 2-dimensional kernel density estimations with the 'kde2d' function in the 'MASS' R package [36]. Next, kernel density estimations were fed into the 'JSD' function in the 'philentropy' R package [37] to generate a JSD matrix representing the global dissimilarity between each donor/lane group. Finally, JSD scores for each cell type were scaled from 0-1, and this process was repeated 100 times. Notably, CD4+ T-cells were down-sampled to include equal numbers of each CD4+ T-cell subtype from each donor/lane group, and cell types with <50 cells in any donor/lane group were excluded (e.g., 8-donor: B cells and dendritic cells; Zheng et al: CD16+ monocytes, dendritic cells, and platelets).

Global differences in gene expression were then summarized as the average and standard deviation of JSD scores across the 100 iterations. Specifically, we quantified the difference between donors (donor A cells from microfluidics lane 1 (A1) vs B2/B3/C2/C3), between mixed and unmixed donor A cells (A1/A4 vs A2/A4), and between technical replicates (A1 vs A4). We then quantified the magnitude of variability due to algorithm performance by repeating this entire workflow after permuting donor A classifications 100 times. Finally, we contextualized the significance of differences in JSD scores associated with mixing status and technical noise via comparison to the average and standard deviation of permuted JSD scores.

4.5.10 Gene set enrichment analysis (GSEA)

To perform global comparisons of gene expression profiles between mixed and unmixed PBMCs in the 8-donor and Zheng et al PBMC datasets, we used GSEA in the following workflow. First, we used the 'FindMarkers' differential gene expression analysis function (test.use = 'MAST'; logfc.threshold = 0) in Seurat to compute p-values for every expressed gene amongst each mixed and unmixed donor A PBMC cell type. Signed p-values were then fed into GSEA using 'pre-ranked' mode, and enriched gene sets were identified as those with nominal p-values and false discovery rate q-values below 0.05.

4.6 References

1. Ding J, Adiconis X, Simmons SK, Kowalczyk MS, Hession CC, Marjanovic ND, et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. Nature Biotechnology. 2020; 38: 737-46.
2. McGinnis CS, Patterson DM, Winkler J, Conrad DN, Hein MY, Srivastavan V, et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nature Methods. 2019; 16: 619-26.
3. Stoeckius M, Zheng S, Houck-Loomis B, Hao S, Yeung BZ, Mauck WM 3rd, et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biology. 2018; 19(1): 224.
4. Gaublomme JT, Li B, McCabe C, Knecht A, Yang Y, Drokhlyanksy E, et al. Nuclei multiplexing with barcoded antibodies for single-nucleus genomics. Nature Communications. 2019; 10(1): 2907.
5. Mair F, Erickson JR, Voillet V, Simoni Y, Bi T, Tyznik AJ, et al. A Targeted Multi-omic Analysis Approach Measures Protein Expression and Low-Abundance Transcripts on the Single-Cell Level. Cell Repots. 2020; 31: 107499.
6. Gehring J, Park JH, Chen S, Thomson M, Pachter L. Highly multiplexed single-cell RNA-seq by DNA oligonucleotide tagging of cellular proteins. Nature Biotechnology. 2019; 38(1): 35-8.
7. Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nature Biotechnology. 2017; 36(1): 89-94.
8. Heaton H, Talman AM, Knights A, Imaz M, Gaffney DJ, Durbin R, et al. Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. Nature Methods. 2020; 17(6): 615-20.

9. Huang Y, McCarthy DJ, Stegle O Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. Genome Biology. 2019; 20(1): 273.
10. Xu J, Falconer C, Nguyen Q, Crawford J, McKinnon BD, Mortlock S, et al. Genotype-free demultiplexing of pooled single-cell RNA-seq. Genome Biology. 2019; 20(1): 290.
11. Guo C, Kong W, Kamimoto K, Rivera-Gonzalez GC, Yang X, Kirita Y, Morris SA. CellTag Indexing: genetic barcode-based sample multiplexing for single-cell genomics. Genome Biology. 2019; 20(1): 90.
12. Shin D, Lee W, Lee JH, Bang D. Multiplexed single-cell RNA-seq via transient barcoding for simultaneous expression profiling of various drug perturbations. Science Advances. 2019; 5(5): eaav2249.
13. Litjens NH, van de Wetering J, van Besouw NM, Betjes MG The human alloreactive CD4+ T-cell repertoire is biased to a Th17 response and the frequency is inversely related to the number of HLA class II mismatches. Blood. 2009; 114: 3947-55.
14. Nicolaidou V, Stylianou C, Koumas L, Vassiliou GS, Bodman-Smith KB, Costeas P. Gene expression changes in HLA mismatched mixed lymphocyte cultures reveal genes associated with allorecognition. Tissue Antigens. 2015; 85: 267-77.
15. DeWolf S, Shen Y, Sykes M. A New Window into the Human Alloresponse. Transplantation. 2016; 100: 1639-49.
16. Lakkis FG, Lechler RI. Origin and biology of the allogeneic response. Cold Spring Harbor Perspectives in Medicine. 2013; 3: a014993.
17. van der Wijst MGP, de Vries DH, Groot HE, Trynka G, Hon CC, Bonder MJ, et al. Single-cell eQTLGen Consortium: a personalized understanding of disease. Elife. 2020; 9: e52155.
18. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, et al. Massively parallel digital transcriptional profiling of single cells. Nature Communications. 2017; 8: 14049.

19. *WTA + Sample Multiplexing + AbSeq* publicly-available dataset downloaded from supplier (scomix.bd.com/hc/en-us/articles/360034192672-Rhapsody-WTA-Demo-Datasets).
20. Bueno JL, García F, Castro E, Barea L, González R. A randomized crossover trial comparing three plateletpheresis machines. Transfusion. 2005; 45: 1373-81.
21. Hubbard B, Fulmer B. Increasing Donor Satisfaction Through the Use of Single-Needle Procedures. Transfusion. 2009; 49: 250A.
22. Ramirez RN, El-Ali NC, Mager MA, Wyman D, Conesa A, Mortazavi A. Dynamic Gene Regulatory Networks of Human Myeloid Differentiation. Cell Systems. 2017; 4: 416-29.
23. van der Brink, SC, Sage F, Vértesy A, Spanjaard B, Peterson-Maduro J, Baron CD, et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. Nature Methods. 2017; 14: 935-6.
24. Knörck A, Marx S, Friedmann KS, Zöphel S, Lieblang L, Hässig C, et al. Quantity, quality, and functionality of peripheral blood cells derived from residual blood of different apheresis kits. Transfusion. 2018; 58: 1516-26.
25. Lin J. Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory. 1991; 37: 145-51.
26. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS. 2005; 102: 15545-50.
27. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics. 2003; 34: 267-73.
28. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, et al. Simultaneous epitope and transcriptome measurement in single cells. Nature Methods. 2017; 14: 865-8.

29. Peterson VM, Zhang KX, Kumar N, Wong J, Li L, Wilson DC, et al. Multiplexed quantification of proteins and transcripts in single cells. Nature Biotechnology. 2017; 35: 936-9.
30. De Jager PL, Hacohen N, Mathis D, Regev A, Stranger BE, Benoist C. ImmVar Project: Insights and Design ConsideratiKons for Future Studies of "Healthy" Immune Variation. Seminars in Immunology. 2015; 27: 51-7.
31. Bolger AM, Lohse M, Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30 : 2114-20.
32. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology. 2018; 36: 411-20.
33. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biology. 2019; 23: 296.
34. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. Cell Systems. 2019; 8: 329-37.e4.
35. Schmiedel BJ, Singh D, Madrigal A, Valdovino-Gonzalez AG, White BM, Zapardiel-Gonzalo J, et al. Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. Cell. 2018; 175: 1701-15.
36. Venables WN, Ripley BD. Modern Applied Statistics with S, Fourth edition. 2002.
37. Dorst H. Philentropy: Information Theory and Distance Quantification with R. Journal of Open Source Software. 2018; 3: 765.

Chapter 5: MULTI-ATAC-seq: sample multiplexing for single-cell epigenomics using lipid-tagged indices

5.1 Abstract

Single-cell genomics sample multiplexing technologies dramatically increase the number of samples that can be analyzed in a single experiment and thereby expand the types of biological questions that can addressed at single-cell resolution. As has been described for single-cell transcriptomics, sample multiplexing technologies additionally improve data quality through doublet detection and batch effect minimization. Single-cell epigenomics assays such as single-nucleus ATAC-seq (snATAC-seq) have also been described which leverage existing cellular isolation and library preparation workflows developed for single-cell transcriptomics. However, single-cell epigenomics sample multiplexing techniques remain in relative infancy. Here, we adapt our previously-documented singe-cell transcriptomics sample multiplexing technique, MULTI-seq, for snATAC-seq applications. Specifically, we benchmark this technology – MULTI-ATAC-seq – on pooled peripheral blood mononuclear cells (PBMCs) from three unrelated healthy donors. These analyses demonstrate the accuracy of MULTI-ATAC-seq classifications and outline preferred practices for doublet identification in snATAC-seq data. We then scale MULTI-ATAC-seq to perform a snATAC-seq-coupled drug screen analyzing titrating doses of two epigenetic modifying agents (the pan histone deacetylase inhibitor, SAHA, and the EZH2 methyltransferase inhibitor, GSK126) on resting and CD3/CD28-stimulated PBMCs. Collectively, this work demonstrates the utility of MULTI-ATAC-seq for snATAC-seq sample multiplexing.

5.2 Introduction

Single-cell genomics provides high-dimensional measurements of diverse levels of biological information in individual cells. Building on early single-cell RNA sequencing (scRNA-seq) efforts which were limited to tens or hundreds of cells in a single experiment [1,2], advances in droplet microfluidics [3-6], microwell [7,8], and combinatorial indexing [9,10] technologies routinized the simultaneous analysis of 10^3-10^6 cells. Leveraging this tremendous gain in cell-throughput, sample multiplexing technologies such as Cell and Nuclei Hashing [11,12], MULTI-seq [13], and ClickTags [14] have produced marked improvements in scRNA-seq sample throughput without requiring pre-existing genomic variability (as in [15-18]) or manipulation (as in [19,20]). Beyond improving scRNA-seq data quality through batch effect minimization, doublet identification [21,22], and improved data quality control [11,13]; sample multiplexing increases the sample-throughout of single-cell genomics assays by 1-2 orders of magnitude and, thus, enables entirely new types and scales of experiments. For example, sample multiplexing has been used to perform scRNA-seq-coupled chemical screens [23-25], time-course and perturbation analyses of organismal development and differentiation [26-29], and spatial transcriptomics [30,31].

More recently, technologies enabling high-throughput single-cell analyses of the epigenome have been described, including methods measuring global patterns of chromatin accessibility such as single-nucleus assay for transposase-accessible chromatin with sequencing (snATAC-seq) [32-34], methods measuring genomic loci associated with specific transcription factors and chemical histone modifications such as single-nucleus Cleavage Under Targets and Tagmentation (snCUT&Tag) [34,35] and single-cell ChIP-seq [36], as well as methods which enable paired measurements of single-cell transcriptomes and epigenomes [38-41]. Single-cell epigenomics methods have been successfully employed to answer fundamental questions in the field of cancer immunology [42], expand perturbation-free single-cell lineage tracing capabilities

[43], and infer the gene regulatory networks underlying cellular identity [44]. However, snATAC-seq sample multiplexing technologies are only beginning to be described [45], resulting in foundational limitations in the types of questions these assays can be used to answer.

We previously showed that lipid-modified oligonucleotides (LMOs) rapidly and stably incorporate into the nuclear membrane in a fashion that can be leveraged for scRNA-seq sample multiplexing using MULTI-seq [13]. Here, we expand on this platform to develop MULTI-ATAC-seq – snATAC-seq sample **m**ultiplexing **u**sing **l**ipid-**t**agged **i**ndices. We begin by demonstrating the ability of MULTI-ATAC-seq to accurately demultiplex PBMCs from unrelated healthy donors. These analyses additionally highlight the accuracy of MULTI-ATAC-seq doublet identification, while providing insight into algorithm selection for computational snATAC-seq doublet detection. Next, we leverage MULTI-ATAC-seq to analyze chromatin accessibility changes in PBMCs stimulated *in vitro* with titrating doses of epigenetic modifier compounds in both resting and inflammatory conditions. These analyses revealed immunomodulatory responses to the histone deacetylase inhibitor, SAHA, as well as nuanced shifts in T-cell differentiation induced by the EZH2 methyltransferase inhibitor, GSK126.

5.3 Results

5.3.1 MULTI-ATAC-seq overview.

MULTI-ATAC-seq localizes DNA barcodes to nuclear membranes using hybridization to LMO scaffolds which spontaneously and stably embed into the plasma membrane of cells or nuclei (**Fig. 5-1a**). LMO scaffolds are comprised of complementary ‘anchor’ and ‘co-anchor’ LMOs that are conjugated to a lignoceric acid and palmitic acid moiety, respectively. MULTI-ATAC-seq sample barcodes include a 5’ anchor LMO hybridization site, a TruSeq R2 PCR handle used for barcode amplification and library preparation, an 8 base-pair sample barcode, an 8 base-pair

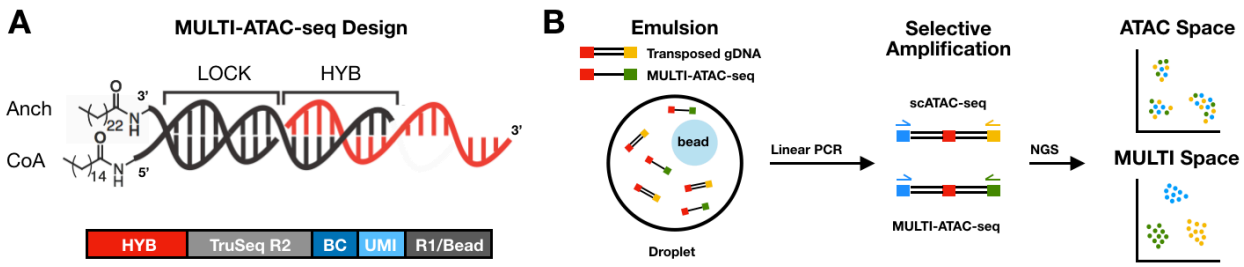

Figure 5-1: MULTI-ATAC-seq design and library preparation workflow.

(A) Diagram of the anchor/co-anchor LMO scaffold (black) with hybridized sample barcode (red). LMO scaffolds assemble via hybridization of anchor and co-anchor LMOs (LOCK). Sample barcode oligonucleotides hybridize to the anchor LMO (HYB), and include a TruSeq R2 PCR handle, a unique molecular identifier (UMI), an 8-bp sample barcode, and a R1/bead-capture domain.

(B) MULTI-ATAC-seq library preparation overview. Following capture of transposed gDNA and MULTI-ATAC-seq sample barcodes by the shared R1 domain (red), both oligonucleotide fractions are amplified during in-droplet linear PCR. Prior to next generation sequencing (NGS), the scATAC-seq and MULTI-ATAC-seq libraries are prepared via selective amplification off of the distinct R2 domains (orange and green), enabling pooled sequencing at user-defined proportions. The final result are couple chromatin accessibility and MULTI-ATAC-seq barcode count matrices.

unique molecular identifier (UMI) [46], and a 3' R1 PCR handle which also serves as the capture domain for the 10x Genomics snATAC-seq bead. MULTI-ATAC-seq barcodes differ from the original MULTI-seq sample barcode design due to the inclusion of distinct PCR handles and a UMI domain, which is not included on snATAC-seq bead oligonucleotides. During snATAC-seq, nuclei bearing MULTI-ATAC-seq barcodes are isolated in emulsion oil droplets, where transposed genomic DNA (gDNA) and sample barcodes are captured and associated with a common nucleus barcode during linear PCR which is used for downstream sample-of-origin classification. Nuclei-barcoded MULTI-ATAC-seq oligos and transposed gDNA are then separated via selective amplification during library preparation and can be sequenced in a pooled format at user-defined proportions (Methods; **Fig. 5-1b**).

5.3.2 MULTI-ATAC-seq prototyping using flow cytometry.

We previously observed that cholesterol-modified oligonucleotides (CMOs) were preferred to LMOs for labeling nuclear membranes prepared for scRNA-seq [13]. To determine whether LMOs or CMOs were preferred for labeling nuclei purified using snATAC-seq buffers, we isolated and transposed nuclei according to standard snATAC-seq protocols (Methods) before

labeling with LMOs or CMOs hybridized to fluorophore-conjugated MULTI-ATAC-seq sample barcodes. We then used flow cytometry to measure membrane labeling kinetics over 1 hour on ice. This analysis showed that LMOs are retained better than CMOs on transposed nuclear membranes (**Fig. 5-2a**).

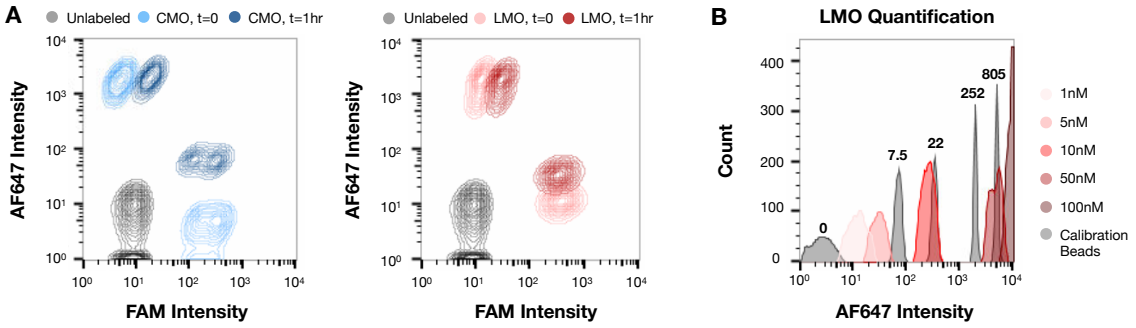

Figure 5-2: Flow cytometry using fluorophore-conjugated MULTI-ATAC-seq oligonucleotide probes demonstrates robust and quantitative LMO and CMO nuclear membrane tagging following transposition.

(A) Time-course analysis of CMO (blue) and LMO (red) abundances following mixture of AF647- and FAM-labeled transposed nuclei on ice for 1hr. n = 10,000 events/time-point.

(B) LMO titration (red gradient) and calibration using Bang's fluorescence calibration beads (grey) facilitates absolute quantification of LMO molecules localized to the nuclear membrane following transposition across varying LMO concentrations. Absolute number of molecules (x10^3) present on calibration beads are denoted with bold text.

Next, we prepared nuclei in the same fashion before labeling with titrating doses of LMOs and computed the absolute number of molecules localized to the nuclear membrane using fluorescence calibration beads. This analysis enabled the identification of an optimal LMO labeling concentration resulting in similar numbers of MULTI-ATAC-seq barcode and transposed gDNA molecules (e.g., 10^4-10^5 molecules; **Fig. 5-2b**), which we anticipated would avoid competition for bead capture sites and library preparation primers between MULTI-ATAC-seq barcodes and transposed gDNA.

5.3.3 Proof-of-concept MULTI-ATAC-seq experimental design.

After determining optimal labeling conditions for transposed nuclei, we next performed a proof-of-concept MULTI-ATAC-seq experiment on the 10x Genomics snATAC-seq platform which aimed to answer three key questions (**Fig. 5-3**). First, we sought to demonstrate whether MULTI-

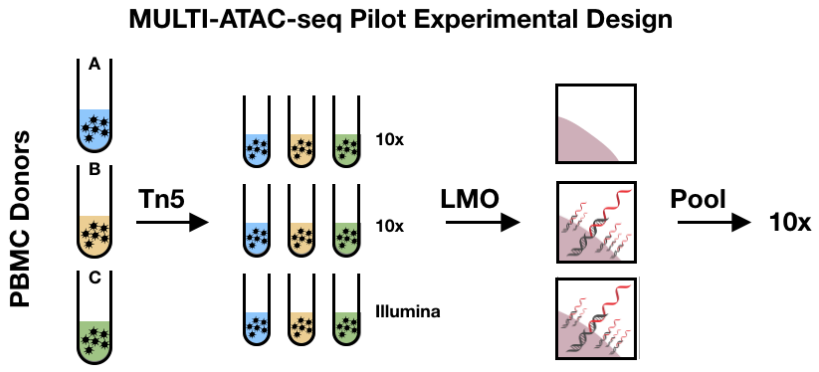

Figure 5-3: Proof-of-concept MULTI-ATAC-seq experimental design.
Nuclei were purified from PBMCs from three healthy donors, and split into three sets. Two sets were transposed using 10x reagents, one set was transposed using Illumina reagents. Illumina-transposed and one 10x-transposed set of nuclei were barcoded with LMOs, while the other 10x-transposed set was left unlabeled. Nuclei from each donor were then pooled according to sample preparation workflow prior to 10x Genomics snATAC-seq analysis.

ATAC-seq could accurately demultiplex snATAC-seq samples relative to an orthogonal set of ground-truth sample classifications. To this end, we processed PBMC nuclei isolated from three healthy donors which were computationally demultiplexed according to donor-specific genomic variants inferred using the *in silico* genotyping tool, Vireo [17]. Second, we sought to determine whether transposition reagents from other commercial vendors (e.g., Illumina) could be used during 10x Genomics-based snATAC-seq, as 10x Genomics provides limited reagent stocks with their snATAC-seq kits which are rapidly depleted when performing large-scale sample multiplexed experiments. To this end, we purified nuclei from all three PBMC donors and split them into three sets – two sets were transposed with 10x Genomics reagents (10x-Tn5) and the third set was transposed with Illumina reagents (Illumina-Tn5). Third, to determine whether MULTI-ATAC-seq barcodes interfere with gDNA capture and/or library preparation, we barcoded Illumina-Tn5 and one set of 10x-Tn5 nuclei to enable downstream comparisons to unbarcoded 10x-Tn5 controls. Finally, PBMC donors from each experimental condition (e.g., 10x-Tn5, 10x-Tn5+LMO, and Illumina-Tn5+LMO) were pooled and sequenced across three separate 10x Genomics snATAC-seq microfluidic lanes.

5.3.4 MULTI-ATAC-seq demultiplexes PBMC donors during snATAC-seq.

Following pre-processing and quality control (Methods), we first analyzed the 10x-Tn5+LMO snATAC-seq data to determine whether MULTI-ATAC-seq can successfully demultiplex PBMC donors. Amongst the 2,081 total nuclei passing quality-control, literature-supported marker analysis revealed diverse immune cell types including naïve and cytotoxic T-cells, natural killer (NK) cells, B-cells, monocytes, dendritic cells (DCs), and doublets predicted by the co-detection of multiple cell-type-specific markers (**Figs. 5-4a, 5-4b**). Mapping MULTI-ATAC-seq sample classifications (Methods) onto ATAC space illustrated that nuclei from each donor were spread across all PBMC cell type clusters while exhibiting donor-specific sub-structure, matching expectations (**Fig. 5-4c**). Moreover, MULTI-ATAC-seq doublet classifications

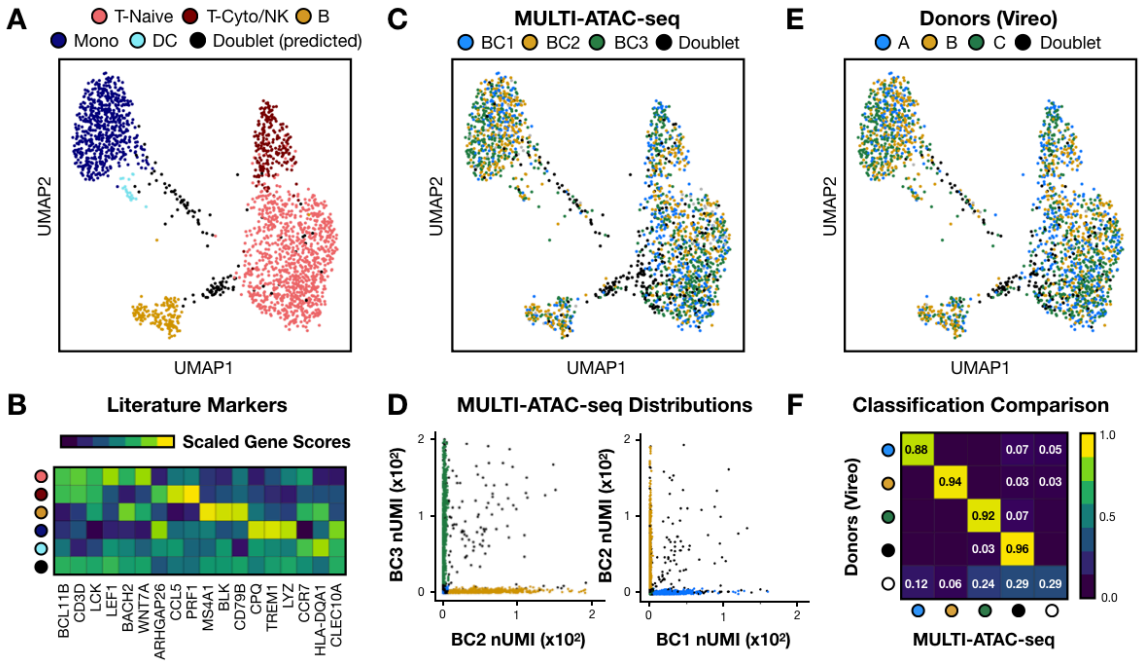

Figure 5-4: MULTI-ATAC-seq accurately demultiplexes PBMC donors during snATAC-seq.

(A,C,E) ATAC space for 10x-Tn5+LMO nuclei colored by (A) cell type annotation, (C) MULTI-ATAC-seq sample classification, and (E) Vireo *in silico* genotyping classifications. n = 2,081 nuclei.

(B) Literature-supported marker gene heatmap. Three literature-supported marker genes for each detected PBMC cell type are arranged in the following column order: T, naïve T, cytotoxic T/NK, B, monocyte, and DC. Averaged scaled gene scores are shown for each annotated cell type (coral = naïve T, dark red = cytotoxic T/NK, gold = B, dark blue = monocyte, cyan = DC, black = doublet).

(D) Scatter plots of MULTI-ATAC-seq barcode (BC) nUMI distributions colored by MULTI-ATAC-seq sample classifications (as in Fig. 5-4c). Nuclei with greater than 200 BC UMIs are excluded for visualization purposes.

(F) Heatmap summary of concordance between MULTI-ATAC-seq and Vireo classifications. Association proportions less than 1% are not annotated. Blank circles correspond to cells left unclassified by MULTI-ATAC-seq or Vireo.

included heterotypic doublets (formed between transcriptionally-distinct cell types) predicted from marker analysis, as well as homotypic doublets (formed between transcriptionally-similar cell types) that are difficult to predict using marker analysis [21,22].

To assess the accuracy of these MULTI-ATAC-seq classifications, we first visualized the raw MULTI-ATAC-seq barcode UMI distributions and observed (i) barcode orthogonality between classification groups and (ii) co-detection of multiple MULTI-ATAC-seq barcodes amongst classified doublets (**Fig. 5-4d**). These patterns are reminiscent of the "barnyard" plots generated during the original scRNA-seq mouse-human species mixing experiments [3-5], and support the accuracy of MULTI-ATAC-seq classifications. We next compared MULTI-ATAC-seq classification results to *in silico* genotyping predictions generated using Vireo. Qualitative comparison of Vireo and MULTI-ATAC-seq donor classifications in ATAC space suggests concordance between the two methods (**Fig. 5-4e**). Quantification of these results revealed that 99.8% (1752/1753) of singlets classified by both methods were assigned to the same donor (**Fig. 5-4f**). Moreover, MULTI-ATAC-seq classified 48 Vireo-unclassified nuclei, while classifying 106/1908 Vireo singlets as doublets.

5.3.5 Benchmarking MULTI-ATAC-seq doublet classifications.

To further explore the discrepancy between MULTI-ATAC-seq and Vireo doublets, we benchmarked these results against two computational snATAC-seq doublet prediction tools: ArchR [47] and AMULET [48]. Notably, ArchR and AMULET use distinct approaches to identify snATAC-seq doublets. Specifically, ArchR identifies doublets as real nuclei co-localizing with *in silico*-generated artificial doublets in ATAC space, as has been described for analogous scRNA-seq doublet detection methods [21,22]. In contrast, AMULET identifies doublets as nuclei with elevated levels of bi-allelic read counts. Due to the relative novelty of these approaches, the preferred method for snATAC-seq doublet detection remained unclear.

To explore the differences between MULTI-ATAC-seq, Vireo, AMULET, and ArchR-defined doublets, we first qualitatively compared the distributions of doublets in ATAC space (**Fig. 5-5a**). This analysis, along with subsequent quantification (**Fig. 5-5b**), revealed high concordance between MULTI-ATAC-seq, Vireo, and AMULET predictions (**Fig. 5-5b**, bracket *). Notably, while AMULET predictions aligned closely with Vireo and MULTI-ATAC-seq, many putative doublets were left undetected (**Fig. 5-5b**, bracket **). In contrast, default ArchR predictions resulted in DCs

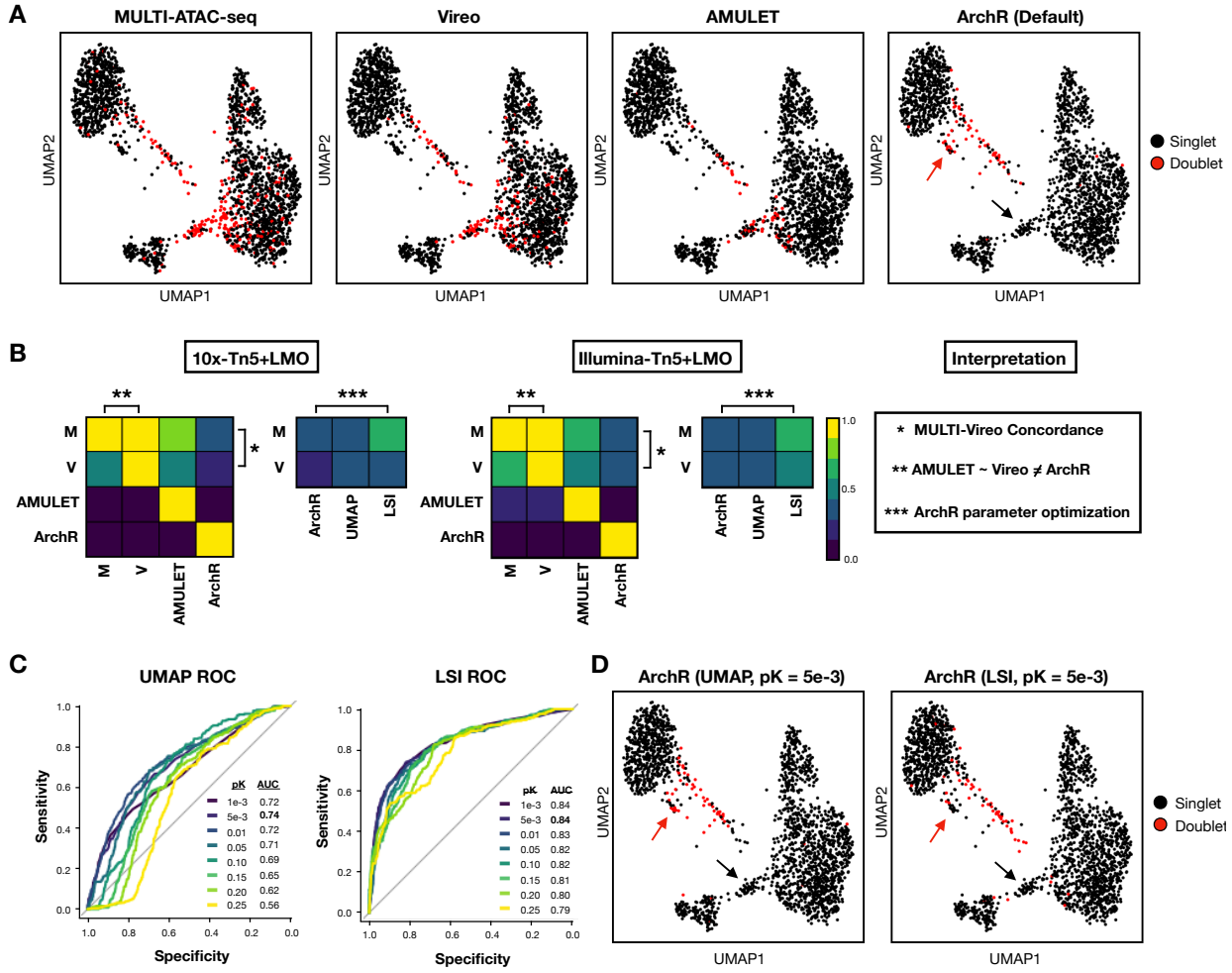

Figure 5-5: MULTI-ATAC-seq doublet detection benchmarking against Vireo, AMULET, and ArchR.

(A) ATAC space for 10x-Tn5+LMO nuclei colored by doublet predictions for MULTI-ATAC-seq, Vireo, AMULET, and ArchR (using default parameters).

(B) Heatmap summary of doublet prediction concordance between MULTI-ATAC-seq (M), Vireo (V), AMULET, default Archr, and pK-optimized ArchR run on UMAP and LSI embeddings for 10x-Tn5+LMO (left) and Illumina-Tn5+LMO datasets (middle). Heatmap asterisks correspond to key interpretations from this analysis, as summarized in the accompanying legend (right).

(C) ROC curves of ArchR doublet enrichment scores generated across multiple neighborhood sizes (pK) with associated AUC values for the 10x-Tn5+LMO dataset. Highest AUC value highlighted in bold.

(D) ATAC space for 10x-Tn5+LMO nuclei colored by doublet predictions following ArchR pK optimization.

being called as doublets (**Fig. 5-5a**, right, red arrow) while entirely missing the T-cell/B-cell doublet cluster (**Fig. 5-5a**, right, black arrow).

We previously showed that the neighborhood size used to compute similarity to artificial doublets is the key parameter for community-detection-based doublet detection algorithms [21]. Thus, to explore whether ArchR predictions could be improved through parameter optimization, we submitted ArchR outputs generated across a range of neighborhood sizes (measured as the proportion of the full dataset, pK) to receiver operator curve (ROC) analysis and identified optimal pK values for doublet identification in both UMAP and latent-semantic indexing (LSI) space (**Fig. 5-5c**). Optimizing neighborhood size only modestly improved ArchR doublet detection results across all datasets tested (**Fig. 5-5b**, bracket ***) and did not lead to the correct classification of DCs and T-cell/B-cell doublets (**Fig. 5-5d**), suggesting fundamental limitations of community-detection-based doublet detection for snATAC-seq data.

We next sought to untangle whether discordant MULTI-ATAC-seq and Vireo doublet classifications represented MULTI-ATAC-seq misclassifications. To this end, we computed the number of detected snATAC-seq fragments and total MULTI-ATAC-seq nUMIs amongst concordant and discordant classifications between the two methods, as true doublets would exhibit elevated levels for both metrics. This analysis revealed that MULTI-ATAC-seq-specific doublets exhibited levels of snATAC-seq fragments and MULTI-ATAC-seq nUMIs that resembled shared (i.e., high-confidence) doublets, suggesting that they were Vireo false-negatives (**Fig. 5-6a**, gold box). In contrast, Vireo-specific doublets had lower levels fragments and nUMIs than high-confidence singlets, suggesting that they were false-positives due to low per-nucleus sequencing depth (**Fig. 5-6a**, green box). Finally, MULTI-ATAC-seq-specific doublets localized preferentially to singlet rather than heterotypic doublet clusters in ATAC space (**Fig. 5-6b**), which suggests that Vireo suffers from insensitivity to homotypic doublets. These results were additionally observed in the Illumina-Tn5+LMO dataset (**Fig. 5-6c**; **Fig. 5-6d**).

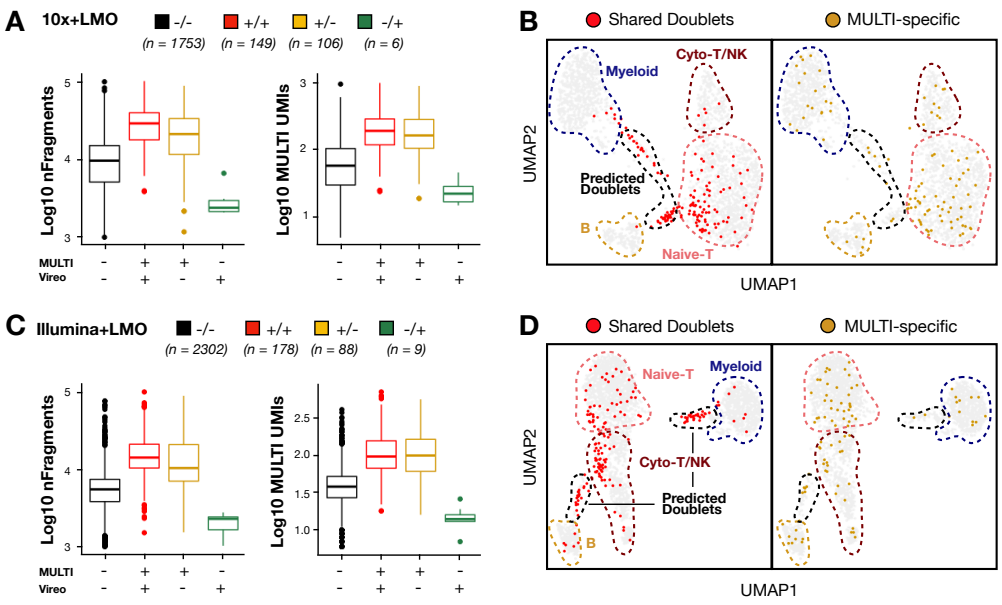

Figure 5-6: MULTI-ATAC-seq identifies Vireo false-positive and false-negative doublet classifications.

(A) Box plots describing the number of snATAC-seq fragments and MULTI-ATAC-seq nUMIs for singlets (S) and doublets (D) called by both MULTI-ATAC-seq and Vireo, MULTI-specific doublets (M), and Vireo-specific doublets (V) in the 10x-Tn5+LMO dataset.

(B) ATAC space for mutual doublets (red) and MULTI-specific doublets (gold) suggest Vireo insensitivity to homotypic doublets.

(C) Same analysis as in Fig. 5-6a and Fig. 5-6b for the Illumina-Tn5+LMO dataset.

We next applied this same analytical workflow to assess the true identity of cells discordantly called by MULTI-ATAC-seq and AMULET. As was observed previously, MULTI-ATAC-seq-specific doublets were enriched for both the number of snATAC-seq fragments and MULTI-ATAC-seq nUMIs, suggesting that these nuclei were AMULET false-negatives (**Fig. 5-7a**). However, AMULET-specific doublets displayed a similar trend, suggesting that were MULTI-

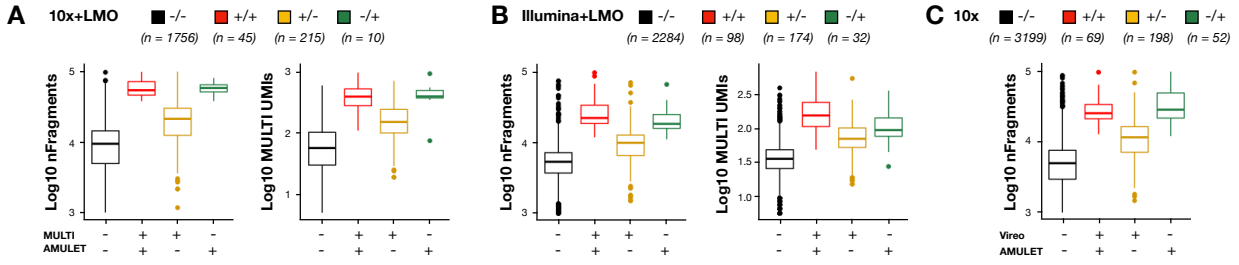

Figure 5-7: MULTI-ATAC-seq and AMULET provide complementary doublet prediction results.

(A) Box plots describing the number of snATAC-seq fragments and MULTI-ATAC-seq nUMIs (right) for singlets (S) and doublets (D) called by both MULTI-ATAC-seq and AMULET, MULTI-ATAC-seq-specific doublets (M), and AMULET-specific doublets (DD) in the 10x-Tn5+LMO dataset.

(B) Same analysis as in Fig. 5-7a except for the Illumina-Tn5+LMO dataset.

(C) Same analysis as in Fig. 5-7a except for the 10x-Tn5 control dataset using Vireo doublet classifications as the reference.

ATAC-seq false-negatives. These results were additionally observed in the Illumina-Tn5+LMO dataset (**Fig. 5-7b**) and the 10x-Tn5 control dataset (**Fig. 5-7c**) when compared to Vireo. Collectively, this benchmarking analysis illustrates the accuracy of MULTI-ATAC-seq doublet predictions, while additionally highlighting AMULET as the preferred computational snATAC-seq doublet prediction tool.

5.3.6 Benchmarking effects of Illumina transposition and LMO labeling on snATAC-seq data.

After validating that MULTI-ATAC-seq successfully demultiplexes PBMC donors, we next sought to assess how LMO labeling and Illumina transposition reagents influence 10x Genomics-based snATAC-seq data quality. TO this end, we first aggregated and read-depth normalized all three snATAC-seq datasets. Mapping sample preparation condition labels onto ATAC space for the 8,074 read-depth-normalized nuclei passing quality-control illustrated that LMO-labeled and control nuclei transposed with 10x Genomics Tn5 are evenly intermixed, while nuclei transposed with Illumina reagents form distinct clusters (**Fig. 5-8a**).

Notably, all PBMC cell types were detected across all sample preparation conditions and exhibited similar chromatin accessibility profiles for literature-supported cell type markers (**Fig. 5-8b**). Differences in fragment size periodicity and transcription start site (TSS) enrichment underlied the differences between 10x- and Illumina-transposed nuclei (**Fig. 5-8c**), with Illumina-transposed nuclei exhibiting increased TSS enrichment and nucleosome-free fragments. Differences in snATAC-seq quality-control metrics are expected when using distinct reagents sources, and the ability to detect PBMC cell types demonstrates the suitability of using alternative transposition reagents for 10x Genomics-based snATAC-seq.

To quantitatively assess the impact of MULTI-ATAC-seq LMO labeling on snATAC-seq data, we used Jensen-Shannon Divergence (JSD) and hierarchical clustering to compute the global dissimilarity between monocytes and naïve T-cells binned by PBMC donor and sample

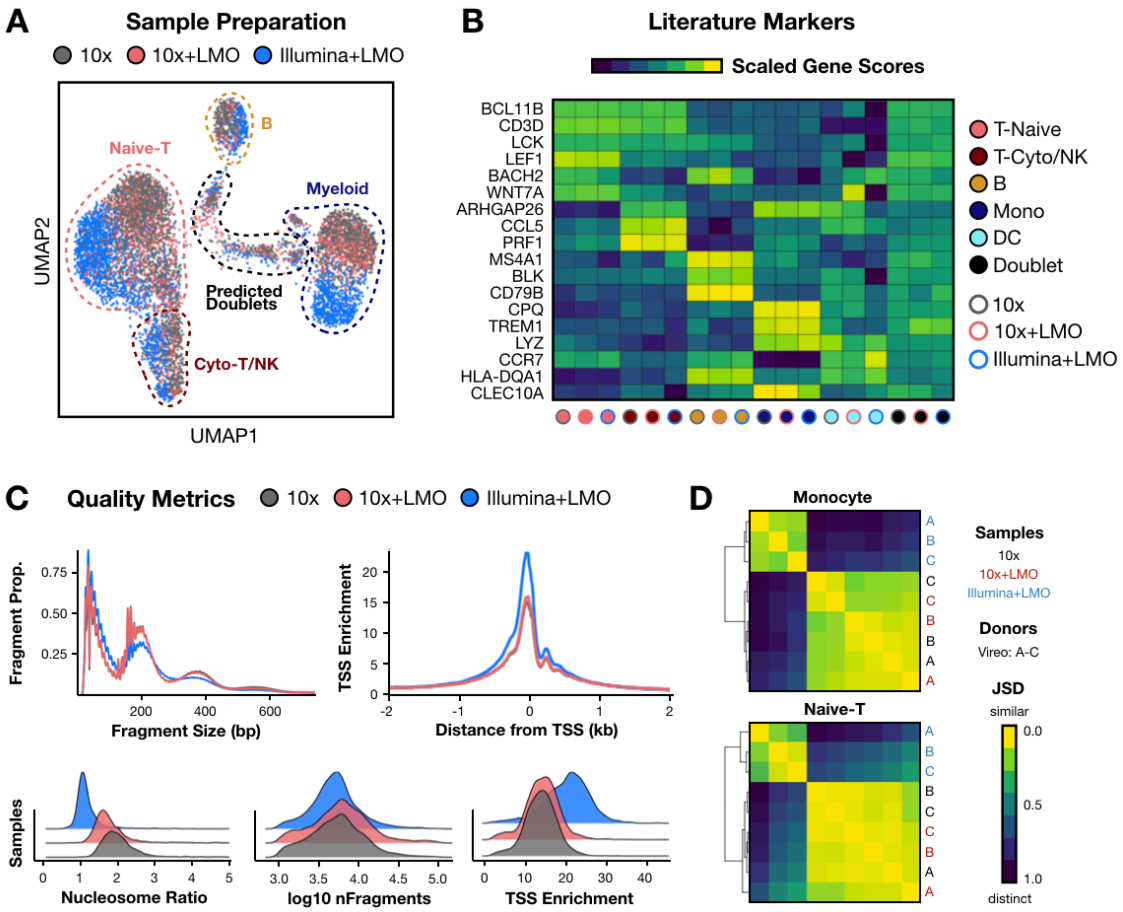

Figure 5-8: MULTI-ATAC-seq LMO labeling does not alter snATAC seq data-quality, while Illumina transposition produces high-quality snATAC-seq data with increased nucleosome-free tagmentation.

(A) ATAC space for all nuclei following read-depth normalization colored by sample preparation (10x-Tn5 control (grey), 10x-Tn5 with MULTI-ATAC-seq LMO labeling (red), Illumina-Tn5 with LMO labeling (blue)). Cell type annotations highlighted with colored dotted lines. n = 8,074 nuclei.

(B) Literature-supported marker gene heatmap with genes arranged as in Fig. 5-4b. Columns correspond to distinct cell types (colored by circle fill) and are split by sample preparation (colored by circle outline).

(C) snATAC-seq quality-control metrics – fragment size distribution and TSS enrichment as a function of distance from aggregated ENCODE-defined TSSs plotted as line plots (top), as well as nucleosome ratio, total number of snATAC-seq fragments, and global TSS enrichment scores plotted as ridge plots (bottom) – grouped according to sample preparation conditions.

(D) Heat map of mean Jensen Shannon Divergence (JSD) between monocytes (top) and naïve T-cells (bottom) grouped according to Vireo-defined donor and sample preparation condition and ordered using hierarchical clustering.

preparation conditions. This analysis revealed that monocyte and naïve T-cell nuclei predominantly cluster by transposition condition (i.e., 10x vs Illumina), matching our previous observations (**Fig. 5-8d**). Moreover, within 10x-transposed monocytes, nuclei clustered according to PBMC donor and not by the presence or absence of LMOs (**Fig. 5-8d**, top). Donor-specific clustering was not as clean for naïve T-cells, although LMO labeling status did not dominate the hierarchical clustering results (**Fig. 5-8d**, bottom). Collectively, these analyses demonstrate that

LMO labeling does not significantly alter snATAC-seq data quality and does not introduce variability in snATAC-seq that is more pronounced than natural inter-individual variability.

5.3.7 MULTI-ATAC-seq epigenetic modifier screen design.

Our proof-of-concept experiment demonstrated how MULTI-ATAC-seq can accurately demultiplex snATAC-seq samples while improving data-quality through doublet detection. However, this experiment did not illustrate how MULTI-ATAC-seq can enable snATAC-seq experimental designs that would otherwise be untenable due to high reagent costs (e.g., chemical compound screens). To this end, we designed a larger MULTI-ATAC-seq experiment where PBMCs were perturbed *ex vivo* in resting or inflammatory conditions (i.e., with anti-CD3/CD28 antibodies) with varying doses of epigenetic modifying agents which have established immunomodulatory capacities [49,50]: the pan-histone deacetylase (HDAC) inhibitor SAHA and the EZH2 methyltransferase inhibitor GSK126 (**Fig. 5-9a**).

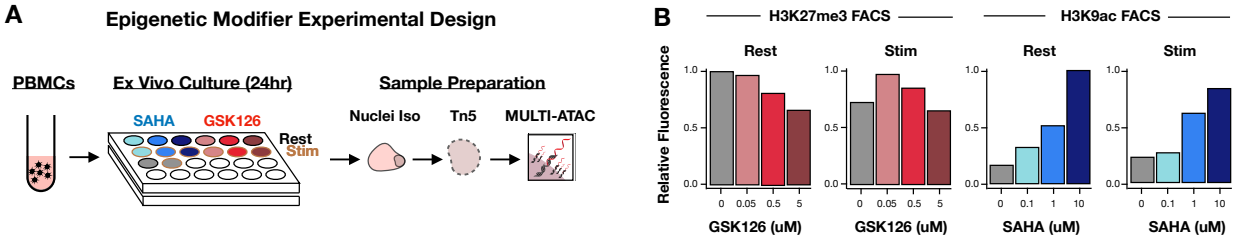

Figure 5-9: MULTI-ATAC-seq epigenetic modifier screen experimental design and dose selection using flow cytometry.

(A) Experimental design schematic overview. PBMCs were cultured *ex vivo* for 24 hours in media supplemented with varying doses of SAHA (blue), GSK126 (red), or DMSO (grey) in the presence (black outline) or absence of anti-CD3/CD28 antibodies (dark gold outline). Following *ex vivo* perturbation, nuclei were isolated, transposed, and LMO labeled prior to snATAC-seq.

(B) Flow cytometry analysis of H3K27me3 marks across varying GSK126 doses (left column, blue) and H3K9ac marks across varying SAHA doses (right column, red) in both resting (top row, black outline) and stimulated conditions (bottom row, gold outline).

Before performing MULTI-ATAC-seq, we first used flow cytometry to identify SAHA and GSK126 doses which induce detectable shifts in relevant epigenetic modifications (e.g., H3K9ac for SAHA, H3K27me3 for GSK126) over 24 hours in PBMCs (**Fig. 5-9b**). Using these conditions, we then employed MULTI-ATAC-seq to multiplex 14 total samples representing resting and

CD3/CD28-stimulated PBMCs treated with three SAHA doses (0.1, 1, and 10 μM), three GSK126 doses (0.05, 0.5, and 5 μM) or vehicle (DMSO) into a single 10x Genomics snATAC-seq microfluidic lane.

5.3.8 MULTI-ATAC-seq identifies shits in immune cell identity and population structure associated with acute inflammation, inflammation-independent responses, and dose-dependent signatures.

Following data pre-processing and quality-control, we analyzed a final snATAC-seq dataset consisting of 5,036 nuclei spanning major PBMC cell types (e.g., T-cells, NK cells, B-cells, and myeloid cells; **Fig. 5-10**, top left). Mapping MULTI-ATAC-seq classifications onto ATAC space revealed that CD3/CD28 stimulation drove unique chromatin accessibility states across all PBMC cell types (**Fig. 5-10**, bottom left). Notably, although anti-CD3/CD28 antibodies only

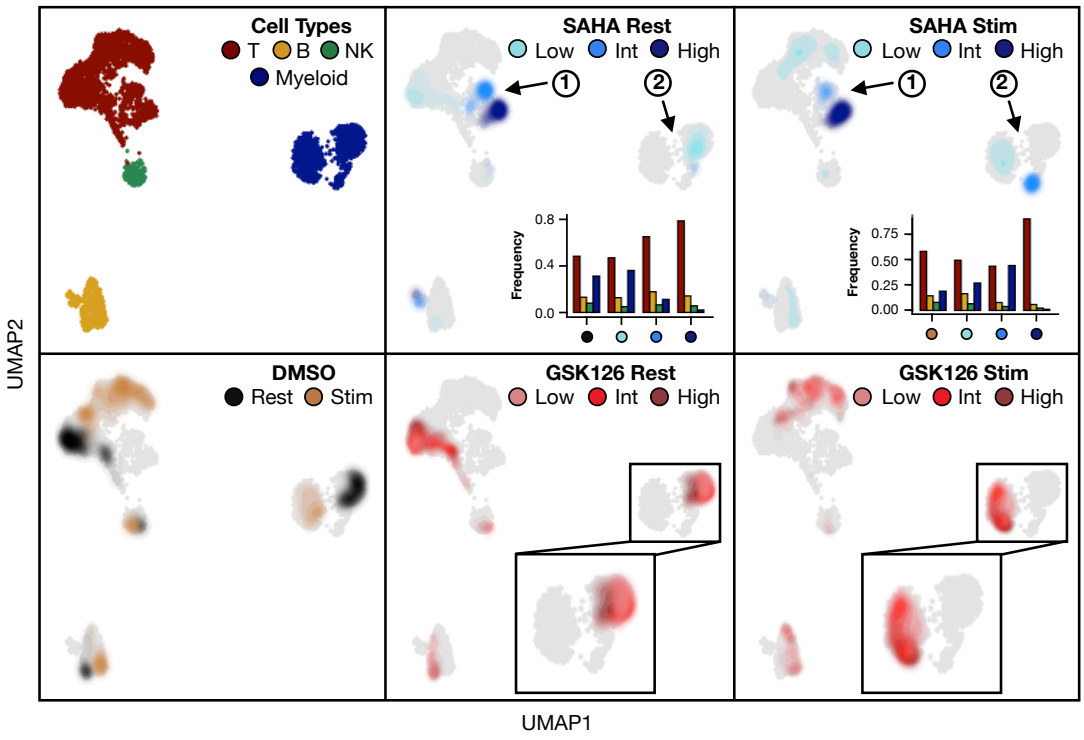

Figure 5-10: Survey of MULTI-ATAC-seq epigenetic modifier screen results highlights PBMC drug response signatures detectable using snATAC-seq.

ATAC space colored by PBMC cell type (top left) and *ex vivo* perturbation conditions grouped by drug (DMSO, GSK126, SAHA) and CD3/CD28-stimulation status (Rest, Stim). Arrows in SAHA-colored ATAC spaces correspond to relevant phenotypes: SAHA dose-specific response independent of CD3/CD28-sitmulation (1) and myeloid toxicity (2). Insets in GSK126-colored ATAC spaces correspond to GSK126 dose-specific myeloid response only resting myeloid cells.

activate T-cells directly through T-cell receptor activation, T-cell activation subsequently leads to the release of cytokines which induce responses in other immune lineages [51]. MULTI-ATAC-seq is able to systematically detect these responses.

Beyond the CD3/CD28-specific responses, MULTI-ATAC-seq additionally identified three patterns of responses to epigenetic modifier perturbation. First, we observed T-cell clusters in ATAC space corresponding to intermediate and high SAHA doses that were robust to CD3/CD28-stimulation (**Fig. 5-10**, top middle/right, arrow #1). Second, we observed myeloid-specific toxicity to the highest SAHA dose (**Fig. 5-10**, top middle/right, arrow #2; inset histograms). Notably, this phenotype would be difficult to discern from global repression of myeloid-specific chromatin domains using bulk measurements. Finally, we observed distinct GSK126 dose-specific trends in both resting and CD3/CD28-stimulated myeloid cells (**Fig. 5-10**, bottom middle/right, inset). Considered together, these results demonstrate how MULTI-ATAC-seq can be used for high-throughput compound screens coupled to snATAC-seq, illustrates how responses to epigenetic modulation can be context (in)-dependent, and highlights how single-cell resolution measurements can deconvolve phenotypes (e.g., cell-type-specific drug toxicity and effects of cell-cell communication) that are obscured by bulk assays.

5.3.9 MULTI-ATAC-seq identifies global de-repressive effect of SAHA treatment on chromatin organization and influence of GSK126 on Th1 differentiation.

After assessing the types of global drug responses that can be detected by MULTI-ATAC-seq, we next sought to determine whether single-cell resolution measurements were sufficiently sensitive to detect the established effect of EZH2 inhibition on CD3/CD28-induced T-cell differentiation. Specifically, EZH2 is known to inhibit Th1 differentiation while promoting the survival of Th1 cells [49]. As an EZH2 inhibitor, we thus expected GSK126 to increase Th1 differentiation while leading to increased Th1 cell death.

To assess this nuanced immunomodulatory phenotype, we first subsetted T-cells from our existing dataset and used literature-supported marker genes to annotate T-cell sub-types (**Fig 5-11a**, **Fig. 5-11b**). In resting T-cells, we identified naïve T-cells associated with state-specific

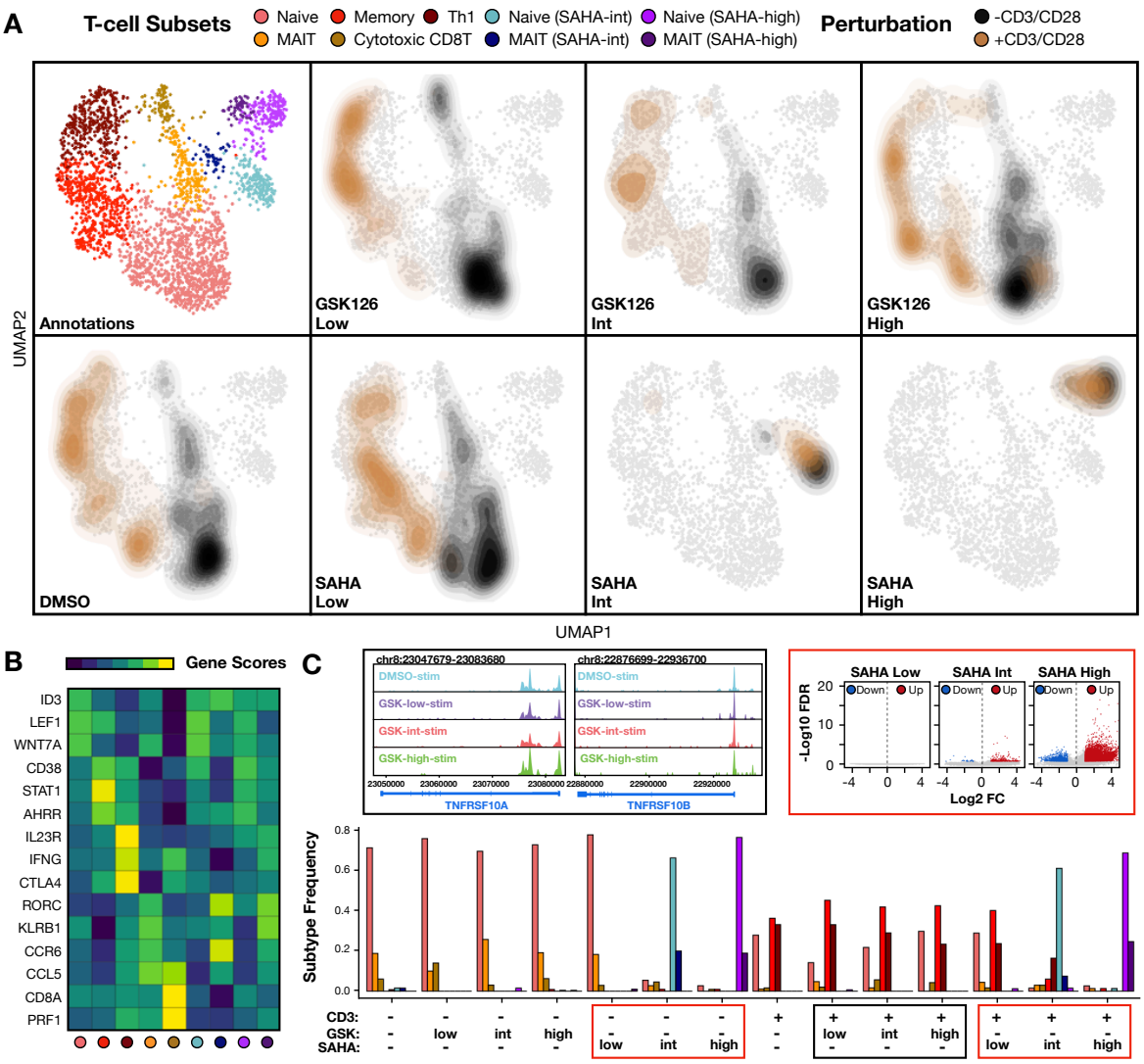

Figure 5-11: T-cell sub-type analysis demonstrates global de-repressive effect of SAHA on chromatin state and influence of GSK126 treatment on Th1 differentiation

(A) ATAC space colored by T-cell sub-type (top left) and perturbation conditions grouped by drug (DMSO, GSK126, SAHA) and colored by CD3/CD28-stimulation status.

(B) Literature-supported marker gene heatmap. Three literature-supported marker genes for each detected T-cell sub-type are arranged in the following column order: Naïve T, memory T, Th1, MAIT, cytotoxic CD8+ T. Averaged scaled gene scores are shown for each annotated cell type (coral = naïve T, red = memory T, dark-red = Th1, orange = MAIT, gold = cytotoxic CD8+ T, light blue = SAHA-int naïve T, dark blue = SAHA-int MAIT, light purple = SAHA-high naïve T, dark purple = SAHA-high MAIT.

(C) T-cell sub-type proportions binned according to CD3/CD28-stimulation status (CD3), drug (GSK or SAHA), and dose (low, intermediate, and high). Black box highlights CD3/CD28-stimulated, GSK126-treated T-cells and corresponds to black inset (top left) showing Genome Browser track for the TNFRSF10A and TNFRSF10B loci. Red box highlights SAHA-treated T-cells and corresponds to red inset (top right) statistically-summarizing differentially-accessible peaks between DMSO-treated naïve T-cells and naïve T-cells treated with varying doses of SAHA.

transcription factors (e.g., ID3 [52]) and genes that are down-regulated after TCR engagement (e.g., LEF1 and WNT7A [53]), cells resembling mucosal-associated invariant T (MAIT) cells associated with known MAIT markers (e.g., KLRB1 [54] and CCR6 [55]), and cytotoxic CD8+ T-cells marked by cytotoxic genes such as PRF1. In CD3/CD28-stimulated T-cells, we identified T-cell sub-types known to arise during Th1 differentiation such as memory T-cells associated with known memory T-cell markers (e.g., CD38 [56]) and genes up-regulated during TCR signaling (e.g., STAT1 [57]), as well as Th1 cells marked by IFNG, IL23R, and CTLA4 [58,59]. Finally, we also observed naïve T-cell and MAIT clusters linked to the intermediate (**Fig 5-11a**, blue) and high SAHA doses (**Fig 5-11a**, purple) in a CD3/CD28-independent fashion.

We next aimed to verify these annotations by comparing the sub-type frequencies associated with each perturbation condition. These analyses revealed that resting PBMCs treated with any dose of GSK126, the lowest SAHA dose, or DMSO were highly-enriched for naïve T-cells, matching expectations due to the lack of TCR engagement (**Fig. 5-11c**). Moreover, comparing the total number of differentially-accessible transcription factor binding motifs between control and SAHA-treated T-cells demonstrated that the magnitude of global chromatin accessibility scaled with SAHA dose (**Fig. 5-11c**, red box). This matches the known de-repressive mechanism of HDAC inhibitors and explains the observed apportionment of T-cells into dose-specific sub-type clusters in ATAC space. Finally, CD3/CD28-stimulated PBMCs were enriched for memory and Th1 cells at the expense of naïve T-cells, matching expectations.

Having established the accuracy of the annotation workflow, we next analyzed the T-cell sub-type frequencies amongst GSK126-treated CD3/CD28-stimulated T-cells to assess whether MULTI-ATAC-seq detected effects of EZH2 inhibition on Th1 differentiation. These analyses revealed that GSK126 treatment was associated with enriched proportions of memory T-cells across all doses (**Fig. 5-11c**). Moreover, we observed that the proportion of naïve T-cells increased with GSK126 dose, while the proportion of Th1 cells showed the opposite trend. Finally,

T-cells treated with the highest GSK126 dose exhibited increased open chromatin proximal to the promoters for the TNFRSF10A/B loci, which are genes with known roles in Th1 apoptosis [60]. Collectively, these results demonstrate that MULTI-ATAC-seq can detect nuanced immunomodulatory responses such as the effect of EZH2 inhibition on Th1 differentiation. Supporting this claim, we observe that GSK126 exposure increased Th1 differentiation (as inferred from increased memory T-cell proportions) and decreased Th1 survival (as inferred from dose-dependent decreases in Th1 proportions and chromatin accessibility proximal to the TNFRSF10A/B loci), matching the documented effects of GSK126 in this context.

5.4 Discussion

Sample multiplexing technologies expand the purview of single-cell genomics experiments and have been effectively applied to generate high-quality, large-scale single-cell transcriptomics datasets. Despite these efforts, analogous methods for multiplexing single-cell epigenomics assays are relatively under-developed. In this study, we present an extension of the MULTI-seq single-cell sample multiplexing technology for snATAC-seq – MULTI-ATAC-seq.

To demonstrate the utility of MULTI-ATAC-seq for snATAC-seq sample multiplexing, we performed a proof-of-concept experiment on PBMCs from three unrelated donors where ground-truth sample-of-origin classifications were obtained from *in silico* genotyping. These analyses demonstrated high concordance between MULTI-ATAC-seq and *in silico* genotyping results, illustrated how MULTI-ATAC-seq does not alter snATAC-seq data quality, and showed how alternative transposition reagents are compatible with snATAC-seq workflows on the 10x Genomics platform. Moreover, we used this proof-of-concept dataset to illustrate snATAC-seq doublet detection algorithms such as AMULET which identify doublets using bi-allelic read-count ratios can outperform community-detection-based doublet detection methods.

To demonstrate the types of large-scale experiments that MULTI-ATAC-seq makes possible, we next multiplexed 14 snATAC-seq samples of resting and CD3/CD28-stimulated PBMCs perturbed with titrating doses of two epigenetic modifying agents (SAHA and GSK126) into a single 10x Genomics microfluidics lane. These analyses illustrated three types of global immunomodulatory responses that can assessed using multiplexed snATAC-seq – namely, perturbation responses that are independent of stimulation status (e.g., SAHA effects on T-cells), dose- and cell-type-specific drug toxicities (e.g., high-dose SAHA effect on myeloid cells), and dose-dependent shifts in chromatin state (e.g., GSK126 effects on resting and CD3/CD28-stimulated myeloid cells). We also demonstrate how coupling perturbation-response experimental designs to high-dimensional single-cell measurements can detect nuanced immunomodulatory phenotypes – specifically, dose-dependent effects of GSK126 on Th1 differentiation and survival.

Collectively, this work represents an effective extension of the MULTI-seq sample multiplexing platform for scATAC-seq experiments. To date, snATAC-seq applications of antibody- and click-chemistry-based multiplexing techniques have not been described, although they are feasible in theory but would similarly suffer from limitations such as species-specificity and sample processing constraints (as summarized previously [13]). In contrast, the recently described multiplexing platform CASB, which localizes biotinylated sample barcodes to cell or nuclear membranes labeled with streptavidin-conjugated concavilin A molecules [45], shares many of the same benefits as MULTI-seq and MULTI-ATAC-seq. However, comparing MULTI-ATAC-seq to CASB highlights strengths and weakness for each method.

One relative strength of MULTI-ATAC-seq is that MULTI-ATAC-seq sample barcodes incorporate a PCR handle that is orthogonal to those introduced during gDNA transposition. The presence of an orthogonal PCR handle enables MULTI-ATAC-seq and snATAC-seq libraries to be prepared independently via selective PCR amplification, which in turn facilitates sequencing of each library at user-defined proportions. In contrast, CASB – in its current implementation –

does not introduce an orthogonal PCR handle for the sample barcode library, and thus requires sample barcode and snATAC-seq libraries to be prepared in a single pool. This can lead to sequencing inefficiencies, as sample barcode libraries require an order of magnitude fewer reads than snATAC-seq libraries. Moreover, as CASB was only used for one documented snATAC-seq experiment, it remains unclear how the performance of the method will vary across different samples or assay types with varying amounts of transposed gDNA. For example, targeted single-cell epigenomics assays such as snCUT&Tag generate less transposed gDNA than in snATAC-seq, which could further compound sequencing inefficiencies when using CASB

In contrast, one strength of CASB relative to MULTI-ATAC-seq is that the biotin-streptavidin interaction between membrane-bound concavilin A and sample barcode oligonucleotides is stable at high temperatures. This allows for nuclei to be tagged and pooled prior to transposition, which in turn minimizes technical difficulty of the downstream workflow, as well as batch effects that may be introduced due to slight variations in transposition reaction conditions. In its current iteration, MULTI-ATAC-seq requires sample tagging after transposition, which limits the scalability of the method. However, strategies for increasing the membrane retention of LMOs at higher temperatures may prove to be sufficient for enabling pooled transposition after MULTI-ATAC-seq labeling.

Looking forward, we anticipate that single-cell epigenomics experiments will be further propelled by interfacing the MULTI-ATAC-seq platform with targeted single-cell epigenomics assays such as snCUT&Tag, as well as paired single-cell transcriptomics and epigenomics assays. Coupling snCUT&Tag to MULTI-ATAC-seq will make it possible to interrogate multiple epigenetic marks, transcription factors, and/or nuclear compartments in large numbers of samples in an economical fashion. Moreover, coupling snCUT&Tag and MULTI-ATAC-seq to a paired scRNA-seq read-out will further improve the interpretability of multiplexed snCUT&Tag datasets, as each layer of the epigenome will be ilnked to a common level of biological information. As a

result, it will become possible to define causal links between the raw read-out of the assay (e.g., genomic loci) to specific regulatory molecules (e.g., transcription factor X or histone modification Y) in a fashion that provides mechanistic insights into how epigenomic states are defined at single-cell resolution.

5.5 Materials and Methods

5.5.1 Design LMOs and sample barcode oligonucleotides.

Anchor and co-anchor LMOs were synthesized as described previously [13]. Briefly, the anchor LMO has a 3' lignoceric acid (LA) modification with two oligonucleotide domains. The 5' end is complimentary to the co-anchor LMO, which bears a 5' palmitic acid (PA), and the 3' end is complimentary to the 5' end of the sample barcode oligonucleotide. The sample barcode oligonucleotide was designed to have five domains: (1) 5' anchor LMO hybridization site, (2) TruSeq R2 PCR handle for barcode amplification and library preparation, (3) 8 base-pair sample barcode with Hamming distance > 3 relative to all other utilized barcodes, (4) 8 base-pair unique molecular identifier [46], and (5) 3' R1 PCR handle used for scATAC-seq bead hybridization and downstream NGS. Identically designed anchor and co-anchor CMOs are conjugated to cholesterol at the 3' or 5' ends via a triethylene glycol (TEG) linker and are commercially available from Integrated DNA Technologies (IDT).

Anchor LMO: 5'-TGGAATTCTCGGGTGCCAAGGGTAACGATCCAGCTGTCACT-{LA}-3'

Co-anchor LMO: 5'-{PA}-AGTGACAGCTGGATCGTTAC-3'

Sample barcode (V1): 5'-CCTTGGCACCCGAGAATTCCAGTGACTGGAGTTCAGACGTGTGC TCTTCCGATCT-{BC}-{UMI}-GACGCTGCCGACGA-3'

Sample barcode (V2): 5'-CCTTGGCACCCGAGAATTCCAGTGACTGGAGTTCAGACGTGTGC

TCTTCCGATCT-{BC}-{UMI}-CTGTCTCTTATACACATCTGACGCTGCCGACGA-3'

5.5.2 Cell culture.

For analytical flow cytometry experiments to compare LMO and CMO labeling kinetics and to perform the LMO titration experiment, HEK293 cells were cultured in DMEM H-21 with 10% FBS and 1% pen/strep on 6-well plates (Corning cat#353046) at 37 °C and 5% CO_2. For all experiments involving PBMCs, cells were obtained from HemaCare and were thawed according to the 'Fresh Frozen Human Peripheral Blood Mononuclear Cells for Single Cell RNA Sequencing' protocol from 10x Genomics. PBMCs were thawed into RPMI 1640 Medium, GlutaMAX™ Supplement, HEPES (ThermoFisher cat#72400047) supplemented with 10% FBS and 1% pen/strep (RPMI-FBS-PS) and plated at 0.5-1x10^7 cells/dish in 10cm ultra-low attachment culture dishes (Corning cat#3262). PBMCs were then rested overnight at 37 °C and 5% CO_2 and suspension PBMCs were transferred to a 15 mL conical tube. Adherent cells were then lifted with TrypLE Express (ThermoFisher cat#12605036) and pooled with the suspension PBMCs. After washing with PBS, trypsinized PBMCs were then subjected to distinct workflows.

For the MULTI-ATAC-seq proof-of-concept experiment, trypsinized PBMCs were immediately used for MULTI-ATAC-seq. For analytical flow cytometry experiments to assess epigenetic modifications across titrating SAHA and GSK126 doses, trypsinized PBMCs were re-plated into ultra-low attachment 6-well plates (Corning cat#3471) at ~1x10^6 cells/well and cultured at 37 °C and 5% CO_2 for 24 hours in RPMI-FBS-PS with or without 25 μL/mL ImmunoCult™ Human CD3/CD28 T Cell Activator (Stemcell Technologies cat# 10971) supplemented with 50 nM, 500 nM, or 5 μM GSK126 (Selleck cat#S7061) or 100 nM, 1 μM, or 10 μM SAHA (Selleck cat#S1047). For the MULTI-ATAC-seq epigenetic modifier screen experiment, trypsinized

PBMCs were re-plated into ultra-low attachment 24-well plates (Corning cat#3473) at ~7x10^5 cells/well and cultured as described for the PBMC analytical flow cytometry experiment.

5.5.3 Analytical flow cytometry.

All analytical flow cytometry data was analyzed using FlowJo and R. For experiments to compare LMO and CMO labeling kinetics and to perform the LMO titration experiment, nuclei were isolated from HEK293 cells according to the "Nuclei Isolation for Single Cell ATAC Sequencing" protocol from 10x Genomics. Briefly, ~5x10^5 cells were pelleted and resuspended in 100 μL ice-cold lysis buffer containing: 10mM Tris-HCl, 10mM NaCl, 3mM $MgCl_2$, 0.1% Tween-20, 0.1% Nonidet P40 substitute, 0.01% digitonin, 1% BSA and 173 μL nuclease-free water. Cells were incubated in lysis buffer for 5' on ice before being diluted with 1mL of ice-cold wash buffer (i.e., lysis buffer without Nonidet P40 substitute or digitonin). Nuclei were then pelleted and resuspended in 50 μL ice-cold 1X nuclei buffer (10x Genomics) and counted.

Following nuclei isolation, 50,000 nuclei were aliquoted into 1.5 mL Eppendorf tubes, pelleted, and resuspended in 50 μL Illumina transposition mix (Illumina, cat# 20034197) containing: 25 μL 2X TD buffer, 16.5 μL 1X PBS, 0.1% Tween-20, 0.01% digitonin, 2.5 μL Tn5 transposase, and 5 μL nuclease-free water. Nuclei were transposed on a thermomixer (Eppendorf) for 1hr at 37 °C without mixing. After transposition, nuclei were labeled with LMOs or CMOs as described previously [8]. For the LMO vs CMO comparison experiment, nuclei were labeled with 100 nM anchor LMO/CMO pre-hybridized to either FAM or AF647-conjugated oligonucleotide probes before labeling with 100 nM co-anchor LMO/CMO. For the LMO titration experiment, nuclei were labeled with 1 nM, 5 nM, 10 nM, 50 nM, or 100 nM anchor LMO pre-hybridized to an AF647-conjugated oligonucleotide probe before labeling with matching concentrations of co-anchor LMO.

LMO/CMO-labeled nuclei labeling reactions were quenched with 1.2mL 1% BSA and pelleted prior to resuspension in ice-cold PBS. For the LMO vs CMO comparison experiment, an aliquot of nuclei from each LMO/CMO and FAM/AF647 condition was immediately analyzed using a BD FACSCalibur instrument, while the remaining LMO- or CMO-labeled nuclei were pooled and incubated on ice for 1hr prior to analysis. For the LMO titration experiment, nuclei were analyzed immediately after LMO labeling, and the number of AF647 oligonucleotide probes present on each nuclear membrane was quantified using AF647 calibration beads (Bangs Laboratories, cat#647).

For analytical flow cytometry experiments to assess epigenetic modifications across titrating SAHA and GSK126 doses, PBMCs were lifted and stained for 30 minutes using a Far Red LIVE/DEAD Fixable Dead Cell Stain (Invitrogen Catalog #L34973). Fixation, permeabilization, and immunostaining were performed according to the "Flow Cytometry, Methanol Permeabilization Protocol for Rabbit Antibodies" protocol from Cell Signaling Technology. Briefly, cells were pelleted, resuspended in 4% PFA, and fixed for 15 minutes at room temperature, followed by two washes with PBS. Fixed cells were permeabilized for 10 minutes on ice in 90% MeOH in PBS, after which MeOH was washed out by pelleting in excess PBS. SAHA-treated PBMCs and corresponding DMSO controls were incubated for 1 hour in 1:200 monoclonal Rabbit anti-H3K9ac (Cell Signaling Technology cat#9649), while GSK126-treated PBMCs and DMSO controls were incubated in 1:200 dilution monoclonal Rabbit anti-H3K27me3 (Cell Signaling Technology cat#9733). Afterwards, cells were washed twice with 0.5% PBS-BSA and then incubated in 1:200 AF488-conjugated Goat anti-Rabbit secondary antibody (Invitrogen cat#A-11008) for 30 minutes at room temperature. Cells were then washed twice with 0.5% PBS-BSA and then resuspended in PBS, followed immediately by analytical flow cytometry using a BD FACSCalibur instrument.

5.5.4 MULTI-ATAC-seq sample preparation.

For the proof-of-concept MULTI-ATAC-seq experiment, cryopreserved PBMCs from three healthy unrelated donors were thawed, cultured, and trypsinized (see section 5.5.2). Nuclei were then purified from trypsinized PBMCs (see section 5.5.3). Three sets of 15,000-nuclei aliquots from each PBMC donor were then subjected to transposition in 1.5mL low-retention tubes for 1 hour at 37 °C in either 10x Genomics or Illumina transposition reagents (see section 5.5.3). After transposition, all samples were immediately put on ice and two sets of nuclei were gently pipette-mixed and labeled with 5 μL of 100 nM anchor LMO pre-hybridized to MULTI-ATAC-seq sample barcode oligonucleotides for 5 minutes. After anchor LMO labeling, nuclei were gently pipette-mixed before 5 μL of 125 nM co-anchor LMO was added to each sample. Nuclei were then incubated on ice for another 5 minutes before all samples were diluted with 1% BSA in PBS and centrifuged at 4 °C and 500xg for 5 minutes. The supernatant was then aspirated, pellets were resuspended in 450 μL 1% BSA in PBS, and nuclei samples from each set were pooled and pelleted. The supernatant was then carefully aspirated before nuclei pellets were resuspended to a total volume of 15 μL in 10x Genomics ATAC buffer. 15 μL aliquots from each nuclei pool were then loaded into three 10x Genomics microfluidics lanes according to the "Chromium Next GEM Single Cell ATAC Reagent Kits v1.1" user-guide from 10x Genomics.

The sample preparation workflow for the MULTI-ATAC-seq epigenetic modifier screen experiment was identical to the proof-of-concept experiment, with the following exceptions. First, Illumina transposition reagents were not used. Second, all samples were labeled with MULTI-ATAC-seq LMOs and sample barcode oligonucleotides, and no 'no-LMO' control samples were used. Third, nuclei were resuspended in 100 μL 1% BSA in PBS instead of 450 μL 1% BSA in PBS after LMO quenching and pelleting to enable pooling of all 14 samples into a single 1.5mL low-retention tube. Fourth, pooled nuclei were resuspended in 30 μL 10x Genomics ATAC buffer,

counted, and diluted to 1,000 nuclei/μL prior to loading a single 10x Genomics microfluidics lane.

5.5.5 MULTI-ATAC-seq library preparation and next-generation sequencing.

Sequencing libraries were prepared using a custom protocol based on the 10X Genomics scATAC-seq and MULTI-seq [13] workflows. Briefly, the 10x workflow was followed until after the post-linear PCR 1.2X SPRI clean-up, where 1 μL of the scATAC-seq library was removed for the MULTI-ATAC-seq library prep PCR (95 °C, 5'; 98 °C, 20"; 67 °C, 30"; 72 °C, 20"; 14 cycles; 72 °C, 1'; 4 °C hold). Each reaction volume was a total of 50 μL containing 26.25 μL 2X KAPA HiFi HotStart master mix (Roche), 2.5 μL of 10μM Illumina TruSeq i7 index (IDT), 2.5 μL of 10μM SI-PCR-B (IDT), 1 μL scATAC-seq library, and 17.75 μL nuclease-free water.

SI-PCR-B: 5'-AATGATACGGCGACCACCGAGA-3'

TruSeq I7: 5'-CAAGCAGAAGACGGCATACGAGAT-{I7}-GTGACTGGAGTTCAGACGTGTGCT CTTCCGATCT-3'

Following the MULTI-ATAC-seq library prep PCR, low molecular weight oligonucleotides were removed via a 2.0X SPRI clean-up, SPRI beads were washed twice with 80% EtOH, and the final MULTI-ATAC-seq library was eluted in 7.5 μL Buffer EB. The remaining 39μL of the scATAC-seq library was then used for the ATAC library prep PCR, where 7.5 μL of 100 μM SI-PCR-B (IDT) was substituted for the standard 7.5 μL of SI-PCR-B (10x Genomics). Excess primers were included in this step in order to avoid deleterious effects of SI-PCR-B competition between transposed DNA and MULTI-ATAC-seq sample barcode oligonucleotides. MULTI-ATAC-seq and snATAC-seq libraries were then pooled and sequenced on NovaSeq SP flow cells according to recommendations from 10x Genomics.

5.5.6 MULTI-ATAC-seq proof-of-concept experiment computational analysis.

snATAC-seq data from the MULTI-ATAC-seq proof-of-concept experiment were pre-processed using cellranger-atac (10x Genomics). Briefly, snATAC-seq FASTQs were pre-processed using cellranger-atac count and aligned to the hg19 reference genome (v1.2.0). Cellranger-atac outputs were then analyzed using the ArchR analysis package [47]. For the 10x-Tn5+LMO and Illumina-Tn5+LMO datasets, nuclei barcodes passing the ArchR quality-control workflow (minTSS = 2, minFrags = 1000) were then used during MULTI-ATAC-seq FASTQ pre-processing and sample classification using the deMULTIplex R package, as described previously [13]. For all three datasets, nuclei barcodes passing the ArchR quality-control workflow were used for Vireo *in silico* genotyping [17] with the following settings: cellSNP: minMAF = 0.1, UMItag = None; vireo: N = 3. After MULTI-ATAC-seq and Vireo sample classification, datasets were parsed to their final form by selecting nuclei associated with ATAC profiles matching literature-supported marker genes that could be assigned to their sample-of-origin using MULTI-ATAC-seq. Default ArchR parameters were then used for dimensionality reduction using iterative LSI and UMAP, Louvain clustering, and data imputation.

These cleaned datasets were used to benchmark doublet predictions for MULTI-ATAC-seq, Vireo, AMULET [48] and ArchR. To generate AMULET doublet predictions, AMULET was run on the cellranger-atac 'possorted_bam.bam' and 'singlecell.csv' outputs using human autosomes masked for repetitive elements in the hg19 reference genome. AMULET doublets were thresholded using a p-value cut-off of 0.05. For ArchR, 'default' doublet predictions were generated using the 'addDoubletScores' function (k = 10, knnMethod = 'UMAP', LSIMethod = 1) and thresholded according to doublet formation rate estimates (10x Genomics) relative to the total number of detected nuclei (e.g., 10x-Tn5 = 125, 10x-Tn5+LMO = 60, Illumina-Tn5+LMO = 80).

To optimize ArchR doublet detection parameters, the following workflow was performed. First, doublet score vectors spanning different 'knnMethod' inputs (e.g., 'UMAP' and 'LSI') and 'k'

inputs (e.g., 0.1%-25% of the total number of nuclei in each dataset) were computed using the 'addDoubletScores' function. Second, logistic regression models were fit for each doublet score vector using the 'glm' R function (family = 'binomial', link = 'logit'). Third, the predictive capacity of each model was computed on the MULTI-ATAC-seq or Vireo doublet classifications using ROC analysis as implemented in the 'ROCR' and 'pROC' R packages [61,62]. Optimal ArchR parameters were identified as those resulting in the lowest AUC value.

To assess how Illumina transposition reagents and the presence of MULTI-ATAC-seq barcodes influence snATAC-seq data quality, .h5 outputs for all three datasets generated by cellranger-atac count were read-depth normalized using cellranger-atac aggr. The resulting aggregated dataset was then parsed to include high-quality cells identified using the previously-described analysis workflow. Naïve T-cells and classical monocytes were then subsetted and JSD analysis for quantifying global differences in ATAC space between PBMC donors in each experimental condition was performed using the following workflow. First, UMAP embeddings were computed for subsetted naïve T-cells and classical monocytes. Second, UMAP coordinates for each donor and experimental condition were used to compute 2-dimensional kernel density estimations with the 'kde2d' function in the 'MASS' R package [63]. Third, kernel density estimations were fed into the 'JSD' function in the 'philentropy' R package [64] to generate a JSD matrix representing the global dissimilarity between each donor in each experimental condition. Finally, JSD scores were scaled from 0-1, and the scaled matrices were clustered using hierarchical clustering and plotted as heatmaps using the ComplexHeatMap R package [65]

5.5.7 MULTI-ATAC-seq epigenetic modifier screen experiment computational analysis.

For the epigenetic modifier screen experiment, ATAC-seq and MULTI-ATAC-seq data were pre-processed, quality-controlled, classified into MULTI-ATAC-seq samples, and submitted for downstream dimensionality reduction, unsupervised clustering, and data imputation using the

same workflow described for the proof-of-concept experiment. To maximize nuclei counts for immune cell sub-type annotation, nuclei which exhibited ATAC profiles matching literature-supported marker genes but which could not be assigned to their MULTI-ATAC-seq sample-of-origin were retained but excluded from all documented comparisons between samples. Moreover, default ArchR parameters were used for pseudo-bulking T-cell sub-types, peak calling (with Macs2), and motif analysis.

5.6 References

1. Ramsköld D, Luo S, Wang YC, Li R, Deng Q, Faridani OR, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nature Biotechnology. 2012; 30(8): 777-82.
2. Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Reports. 2012; 2(3): 666-73.
3. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015; 161(5): 1202-14.
4. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015; 161(5): 1187-1201.
5. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, et al. Massively parallel digital transcriptional profiling of single cells. Nature Communications. 2017; 8: 14049.
6. Habib N, Avraham-Davidi I, Basu A, Burks T, Shekhar K, Hofree M, et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. Nature Methods. 2017; 14(10): 955-8.
7. Gierahn TM, Wadsworth MH 2nd, Hughes TK, Bryson BD, Butler A, Satija R, et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Nature Methods. 2017; 14(4): 395-8.
8. Hughes TK, Wadsworth MH 2nd, Gierahn TM, Do T, Weiss D, Andrade PR, et al. Second-Strand Synthesis-Based Massively Parallel scRNA-Seq Reveals Cellular States and Molecular Features of Human Inflammatory Skin Pathologies. Immunity. 2020; 53(4): 878-94.
9. Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. Science. 2017; 357(6352): 661-7.

10. Rosenberg AB, Roco CM, Muscat RA, Kuchina A, Sample P, Yao Z, et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science. 2018; 360(6385): 176-82.
11. Stoeckius M, Zheng S, Houck-Loomis B, Hao S, Yeung BZ, Mauck WM 3rd, et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biology. 2018; 19(1): 224.
12. Gaublomme JT, Li B, McCabe C, Knecht A, Yang Y, Drokhlyanksy E, et al. Nuclei multiplexing with barcoded antibodies for single-nucleus genomics. Nature Communications. 2019; 10(1): 2907.
13. McGinnis CS, Patterson DM, Winkler J, Conrad DN, Hein MY, Srivastava V, et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nature Methods. 2019; 16(7), 619-26.
14. Gehring J, Park JH, Chen S, Thomson M, Pachter L. Highly multiplexed single-cell RNA-seq by DNA oligonucleotide tagging of cellular proteins. Nature Biotechnology. 2019; 38(1): 35-8.
15. Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nature Biotechnology. 2017; 36(1): 89-94.
16. Heaton H, Talman AM, Knights A, Imaz M, Gaffney DJ, Durbin R, et al. Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. Nature Methods. 2020; 17(6): 615-20.
17. Huang Y, McCarthy DJ, Stegle O. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. Genome Biology. 2019; 20(1): 273.
18. Xu J, Falconer C, Nguyen Q, Crawford J, McKinnon BD, Mortlock S, et al. Genotype-free demultiplexing of pooled single-cell RNA-seq. Genome Biology. 2019; 20(1): 290.

19. Guo C, Kong W, Kamimoto K, Rivera-Gonzalez GC, Yang X, Kirita Y, Morris SA. CellTag Indexing: genetic barcode-based sample multiplexing for single-cell genomics. Genome Biology. 2019; 20(1): 90.
20. Shin D, Lee W, Lee JH, Bang D. Multiplexed single-cell RNA-seq via transient barcoding for simultaneous expression profiling of various drug perturbations. Science Advances. 2019; 5(5): eaav2249.
21. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. Cell Systems. 2019; 8(4): 329-37.e4.
22. Wolock SL, Lopez R, Klein AM. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. Cell Systems. 2019; 8(4): 281-91.e9.
23. Cook DP, Vanderhyden BC. Context specificity of the EMT transcriptional response. Nature Communications. 2020; 11(1): 2142.
24. Srivatsan SR, McFaline-Figueroa JL, Ramani V, Saunders L, Cao J, Packer J, et al. Massively multiplex chemical transcriptomics at single-cell resolution. Science. 2020; 367(6473): 45-51.
25. McFarland JM, Paolella BR, Warren A, Geiger-Schuller K, Shibue T, Rothberg M, et al. Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. Nature Communications. 2020; 11(1): 4296.
26. Miyamoto M, Kannan S, Uosaki H, Kakani T, Murphy S, Andersen P, Kwon C. Cardiac progenitors auto-regulate second heart field cell fate via Wnt secretion. bioRxiv. 2021. doi: 10.1101/2021.01.31.428968.
27. Hurskainen M, Mižíková I, Cook DP, Andersson N, Cyr-Depauw C, Lesage F, et al. Single cell transcriptomic analysis of murine lung development on hyperoxia-induced damage. Nature Communications. 2021; 12(1):1565.

28. Rifes P, Isaksson M, Rathore GS, Aldrin-Kirk P, Møller OK, Barzaghi G, et al. Modeling neural tube development by differentiation of human embryonic stem cells in a microfluidic WNT gradient. Nature Biotechnology. 2020; 38(11): 1265-73.
29. Jong H. Park, Tiffany Tsou, Paul Rivaud, Matt Thomson, Sisi Chen. Designing signaling environments to steer transcriptional diversity in neural progenitor cell populations. bioRxiv. 2020. doi: 10.1101/2019.12.30.890087.
30. Fawkner-Corbett D, Antanaviciute A, Parikh K, Jagielowicz M, Gerós AS, Gupta T, et al. Spatiotemporal analysis of human intestinal development at single-cell resolution. Cell. 2021; 184(3): 810-26.
31. Hu KH, Eichorst JP, McGinnis CS, Patterson DM, Chow ED, Kersten K, et al. ZipSeq: barcoding for real-time mapping of single cell transcriptomes. Nature Methods. 2020; 17(8): 833-43.
32. Lareau CA, Duarte FM, Chew JG, Kartha VK, Burkett ZD, Kohlway AS, et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. Nature Biotechnology. 2019; 37(8): 916-24.
33. Satpathy AT, Granja JM, Yost KE, Qi Y, Meschi FM, McDermott GP, et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nature Biotechnology. 2019; 37(8): 925-36.
34. Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. Science. 2015; 348(6237): 910-4.
35. Wu SJ, Furlan SN, Mihalas AB, Kaya-Okud HS, Feroze AH, Emerson SN, et al. Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression. Nature Biotechnology. 2021. doi: 10.1038/s41587-021-00865-z.

36. Bartosovic M, Kabbe M, Castelo-Branco G. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. Nature Biotechnology. 2021. doi: 10.1038/s41587-021-00869-9.
37. Grosselin K, Durand A, Marsolier J, Poitou A, Marangoni E, Nemati F, et al. High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. Nature Genetics. 2019; 51(6): 1060-6.
38. Zhu C, Zhang Y, Eric Li YE, Lucero J, Behrens MM, Ren B. Joint profiling of histone modifications and transcriptome in single cells from mouse brain. Nature Methods. 2021; 18(3): 283-92.
39. Chen S, Lake BB, Zhang K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nature Biotechnology. 2019; 37(12): 1452-7.
40. Cao JC, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, Hill AJ, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science. 2018; 361(6409): 1380-5.
41. Zhu C, Yu M, Huang H, Juric I, Abnousi A, Hu R, Lucero J, et al. An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. Nature Structural and Molecular Biology. 2019; 26(11): 1063-70.
42. Yost KE, Satpathy AT, Wells DK, Qi Y, Wang C, Kageyama R, et al. clonal replacement of tumor-specific t cells following pd-1 blockade. Nature Medicine. 2019; 25(8): 1251-9.
43. Ludwig LS, Lareau CA, Ulirsch JC, Christian E, Muus C, Li LH, et al. Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. Cell. 2019; 176(6): 1325-39.
44. Shema E, Bernstein B, Buenrostro JD. Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution. Nature Genetics. 2019; 51(1): 19-25.

45. Fang L, Li G, Sun Z, ZHu Q, Cui H, Li Y, et al. CASB: a concanavalin A-based sample barcoding strategy for single-cell sequencing. Molecular & Systems Biolgoy. 2021; 17(4): e10060.
46. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Maria Kasper M, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nature Methods. 2014; 11(2): 163-6.
47. Granja JM, Corces MR, Pierce SE, Bagdatli T, Choudhry H, Chang HY, Greenleaf WJ. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nature Genetics. 2021; 53(3): 403-11.
48. Thibodeau A, Eroglu A, McGinnis CS, Lawlor N, Nehar-Belaid D, Kursawe R, et al. AMULET: a novel read count-based method for effective multiplet detection from single nucleus ATAC-seq data. Genome Biology. 2021; *in press*.
49. Karantanos T, Christofides A, Bardhan K, Li L, Boussiotis VA. Regulation of T Cell Differentiation and Function by EZH2. Frontiers in Immunology. 2016; 7: 172.
50. Licciardi PV, Ververis K, Tang ML, El-Osta A, Karagiannis TC. Immunomodulatory effects of histone deacetylase inhibitors. Current Molecular Medicine. 2013; 13(4): 640-7.
51. Lawlor N, Nehar-Belaid D, Grassmann JDS, Stoeckius M, Smibert P, Stitzel ML, et al. Single Cell Analysis of Blood Mononuclear Cells Stimulated Through Either LPS or Anti-CD3 and Anti-CD28. Frontiers in Immunology. 2021; 12: 636720.
52. Miyazaki M, Rivera RR, Miyazaki K, Lin YC, Agata Y, Murre C. The opposing roles of the transcription factor E2A and its antagonist Id3 that orchestrate and enforce the naive fate of T cells. Nature Immunology. 2011; 12(10): 992-1001.
53. Willinger T, Freeman T, Herbert M, Hasegawa H, McMichael AJ, Callan MFC. Human naive CD8 T cells down-regulate expression of the WNT pathway transcription factors lymphoid enhancer binding factor 1 and transcription factor 7 (T cell factor-1) following antigen encounter in vitro and in vivo. Journal of Immunology. 2006; 176(3): 1439-46.

54. Parrot T, Gorin J, Ponzetta A, Maleki KT, Kammann T, Emgård J, et al. MAIT cell activation and dynamics associated with COVID-19 disease severity. Science Immunology. 2020; 5(51): eabe1670.
55. Chen P, Deng W, Li D, Zeng T, Huang L, Wang Q, et al. Circulating Mucosal-Associated Invariant T Cells in a Large Cohort of Healthy Chinese Individuals From Newborn to Elderly. Frontiers in Immunology. 2019; 10: 260.
56. Sandoval-Montes C, Santos-Argumedo L. CD38 is expressed selectively during the activation of a subset of mature T cells with reduced proliferation but improved potential to produce cytokines. Journal of Leukocyte Biology. 2005; 77(4): 513-21.
57. Gamero AM, Larner AC. Signaling via the T cell antigen receptor induces phosphorylation of Stat1 on serine 727. Journal of Biological Chemistry. 2000; 275(22): 16574-8.
58. Feng J, Hu Y, Song Z, Liu Y, Guo X, Jie Z. Interleukin-23 facilitates Th1 and Th2 cell differentiation in vitro following respiratory syncytial virus infection. Journal of Medical Virology. 2015; 87(4): 708-15.
59. Tagar TN, Turley DM, Padilla J, Karandikar NJ, Tan L, Bluestone JA, Miller SD. CTLA-4 regulates expansion and differentiation of Th1 cells following induction of peripheral T cell tolerance. Journal of Immunology. 2004; 172(12): 7442-50.
60. Falschlehner C, Schaefer U, Walczak H. Following TRAIL's path in the immune system. Immunology. 2009; 127(2): 145-54.
61. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005; 21: 3940-1. 37.
62. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011; 12: 77.
63. Venables WN, Ripley BD. Modern Applied Statistics with S, Fourth edition. 2002.

64. Dorst H. Philentropy: Information Theory and Distance Quantification with R. Journal of Open Source Software. 2018; 3: 765.
65. Gu Z, Ellis R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016; 32(18): 2847-9.

Chapter 6: Single-cell 'screen-by-sequencing' with peripheral blood mononuclear cells reveals immunomodulation trajectories, off-target drug activities, and novel effects on immune population homeostasis

6.1 Abstract

High-throughput chemical screens are traditionally performed on cell line systems and are limited to low-dimensional read-outs due to measurement scalability concerns. In contrast, bulk 'screen-by-sequencing' approaches generate high-dimensional measurements of drug-induced gene expression patterns using bulk RNA-seq, but are similarly limited to simple experimental systems due to the aggregative nature of the approach. Single-cell RNA-sequencing sample multiplexing approaches can couple chemical perturbation screens to high-dimensional, single-cell measurements which, in turn, facilitates the use of complex *in vitro* systems which better recapitulate *in vivo* disease biology. To demonstrate the potential of single-cell 'screen-by-sequencing', we use MULTI-seq to generate single-cell gene expression response profiles of peripheral blood mononuclear cells to >500 drug compounds in resting and/or inflammatory conditions. These analyses reveal the ability of single-cell screen-by-sequencing to detect the primary modes of immunomodulation in lymphoid and myeloid cells, the phenotypic consequences of off-target drug activities, the effect of a subset of non-steroidal anti-inflammatory drugs on immune population homeostasis, and the utility of target-enrichment strategies for improving single-cell screen-by-sequencing scalability in the future.

6.2 Introduction

Cell-based chemical screening faces a tradeoff: to maximize the number of tested compounds, the amount of information collected must be minimized. For example, screens to discover cancer chemotherapeutics typically measure a single parameter across thousands of candidate drugs – usually cell growth or viability – and thereby necessarily discover drugs which limit the growth or viability of cancer cells [1]. As a result, therapies aiming to induce other clinically-relevant phenotypes (e.g., blocking tumor metastasis and therapy resistance, inducing systemic anti-tumor immune responses, etc.) are lacking, as such complex phenotypes are difficult to assess in a high-throughput fashion. 'Screen-by-sequencing' approaches [2-7] represent a potential solution to this limitation, as bulk RNA sequencing data can identify chemical perturbations which induce phenotypes (vis-à-vis gene expression signatures) that are ordinarily obscured by low-dimensional read-outs. However, bulk screen-by-sequencing studies fail to capture two critical features of human disease: intercellular heterogeneity and communication.

The advent of single-cell genomics sample multiplexing technologies [8-12] enables 'single-cell screen-by-sequencing' studies, wherein the effects of large numbers of chemical perturbations can be deeply interrogated without obscuring intercellular heterogeneity. For example, pioneering methods such as sci-Plex [13] and MIX-seq [14] have demonstrated how single-cell screen-by-sequencing can uncover variable responses to drug treatment within ostensibly homogenous cell lines, disentangle biological covariates (e.g., drug-induced expression signatures vs toxicity [14]), and uncover previously-unknown drug activities (e.g., metabolic effects of histone deacetylase (HDAC) inhibitor treatment due to acetyl-CoA depletion [13]). Building on these efforts, Zhao and colleagues recently described the use of patient-derived glioblastoma slice cultures for scRNA-seq-coupled chemical screens [15]. Although the authors tested only 6 compounds in this proof-of-principle study, their work highlights how single-cell

screen-by-sequencing can be expanded beyond cell lines to systems which better recapitulate *in vivo* disease biology, while additionally uncovering how drugs alter intercellular communication networks (e.g., glioblastoma tumor-immune interactions).

To further demonstrate the promise of single-cell screen-by-sequencing, we used MULTI-seq [8] to profile the transcriptional impact of >500 immunomodulatory compounds in >1.5M resting and stimulated peripheral blood mononuclear cells (PBMCs) at single-cell resolution using scRNA-seq. We selected PBMCs for this study because they represent a complex *in vitro* system with heterogenous and interacting cell types [16] which mirror *in vivo* peripheral immune responses such as inflammation [17], viral infection [18], and chemical perturbation [19]. This effort represents the largest per-sample single-cell screen-by-sequencing dataset generated to date, and yielded key insights that we hope will inspire future single-cell screen-by-sequencing endeavors. These insights will be presented in four vignettes.

First, we provide a broad survey of the scRNA-seq data where we highlight immune cell sub-types, differentiation trajectories, and other biological processes that are recapitulated in our *in vitro* system. We then use this framework to characterize the primary modes of drug-induced patterns of immunomodulation across T-cells and myeloid cells. Second, we discuss how divergent responses to tyrosine kinase inhibitors (TKIs) with identical primary molecular targets – as well as convergent responses to TKIs with distinct targets – illustrate how single-cell screen-by-sequencing can capture off-target drug activities. Third, we describe a previously-undescribed effect of a subset of non-steroidal anti-inflammatory drugs (NSAIDs) on T-cell-mediated macrophage apoptosis which demonstrates the types of insights that can be gained by screening on complex *in vitro* systems. And fourth, we document how targeted transcript enrichment strategies can be used to maximize information content while minimizing next-generation sequencing costs for future studies.

6.3 Results

6.3.1 Study design.

Standard screen-by-sequencing practices involve characterizing individual compounds across a wide range of doses with biological and technical replicates [20]. Due to the high costs associated with single-cell genomics experiments, however, we first sought to perform a pilot screen which maximized the number of compounds tested at the expense of other experimental parameters. To this end, we treated PBMCs isolated from a single healthy donor with 516 unique chemical compounds sourced from commercially-available immunomodulatory and FDA-approved drug libraries (Selleck Chem) at a single-dose (1 μM) and time-point (24 hours) without replicates. Collectively, these compounds span >30 major drug classes and >150 unique primary molecular targets. Since many of the screened compounds are clinically used to treat diverse inflammatory diseases, we concurrently treated PBMCs with drugs and soluble anti-CD3/CD28 tetrameric antibodies, which mimic acute inflammation by activating T-cells through T-cell receptor (TCR) and co-receptor engagement. Drug perturbations were additionally performed on resting PBMCs (i.e., without anti-CD3/CD28 antibody exposure) for 279/516 of the compounds. Perturbations were performed in ten batches of 96 (with vehicle-only controls), and were labeled with condition-specific MULTI-seq sample barcodes prior to pooling and scRNA-seq using the 10x Genomics platform (**Fig. 6-1**; Methods).

6.3.2 Vignette I: PBMC single-cell screen-by-sequencing data survey – PopAlign identifies high-impact perturbations and 'broad' or 'local' patterns of immunomodulation.

Following sample demultiplexing and quality-control filtering (Methods), we analyzed a final scRNA-seq dataset comprised of 995,173 single-cell transcriptomes representing 801 unique experimental conditions. Focusing on a subset of CD3/CD28-activated cells, we identified

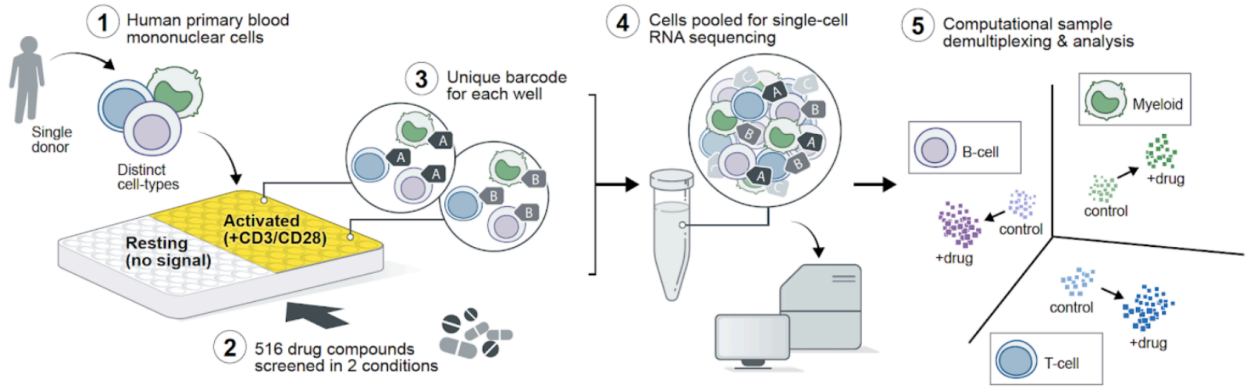

Figure 6-1: Schematic overview of PBMC scRNA-seq pilot screen.

PBMCs from a single healthy donor comprised of diverse immune cell types (1) were cultured in 96-well plates prior to drug treatment in the presence or absence of anti-CD3/CD28 tetrameric antibodies (2). After 24 hours, PBMCs were tagged with MULTI-seq barcodes marking each specific condition (3), pooled, and subjected to scRNA-seq and next-generation sequencing (4). Experimental conditions were then computationally demultiplexed according to MULTI-seq barcode counts and analyzed to identify drug- and cell-type-specific transcriptional responses in major immune cell types such as myeloid lineage cells and B and T lymphocytes (5).

major PBMC cell types including T-cells (CD3D+), monocytes/macrophages (LYZ+), B-cells (MS4A1+), natural killer cells (CD3D-NKG7+), and dendritic cells (CD3D-CCR7+). We additionally identified low-abundant cell populations such as granulocytes (GATA2+), plasma cells (MZB1+), and a cluster consisting of cells treated with the proteasome inhibitor bortezomib which expressed high levels of protein homeostasis genes (e.g., HSPA1A) and localized separately from all other cell types in gene expression space (**Fig. 6-2a**).

Using these cell type annotations, we then used PopAlign [21] to quantify the magnitude and cell-type-specificity of perturbation responses. After stringent significance thresholding, PopAlign identified a number of experimental conditions which deviated significantly from controls. Among the PopAlign hits, we identified compounds which induced broad responses across three major PBMC cell types (T-cells, B-cells, and myeloid cells) as well as drugs which prompted localized responses in individual or pairs of cell types (**Fig. 6-2b**). While both broad and local immunomodulatory compounds were comprised of drugs with myriad molecular targets, many of these classifications were supported by the existing literature and could be discerned via differential localization in gene expression space (**Fig. 6-2c**). For example, drugs inhibiting the phosphatase calcineurin were classified as broad modulators in CD3/CD28-activated conditions.

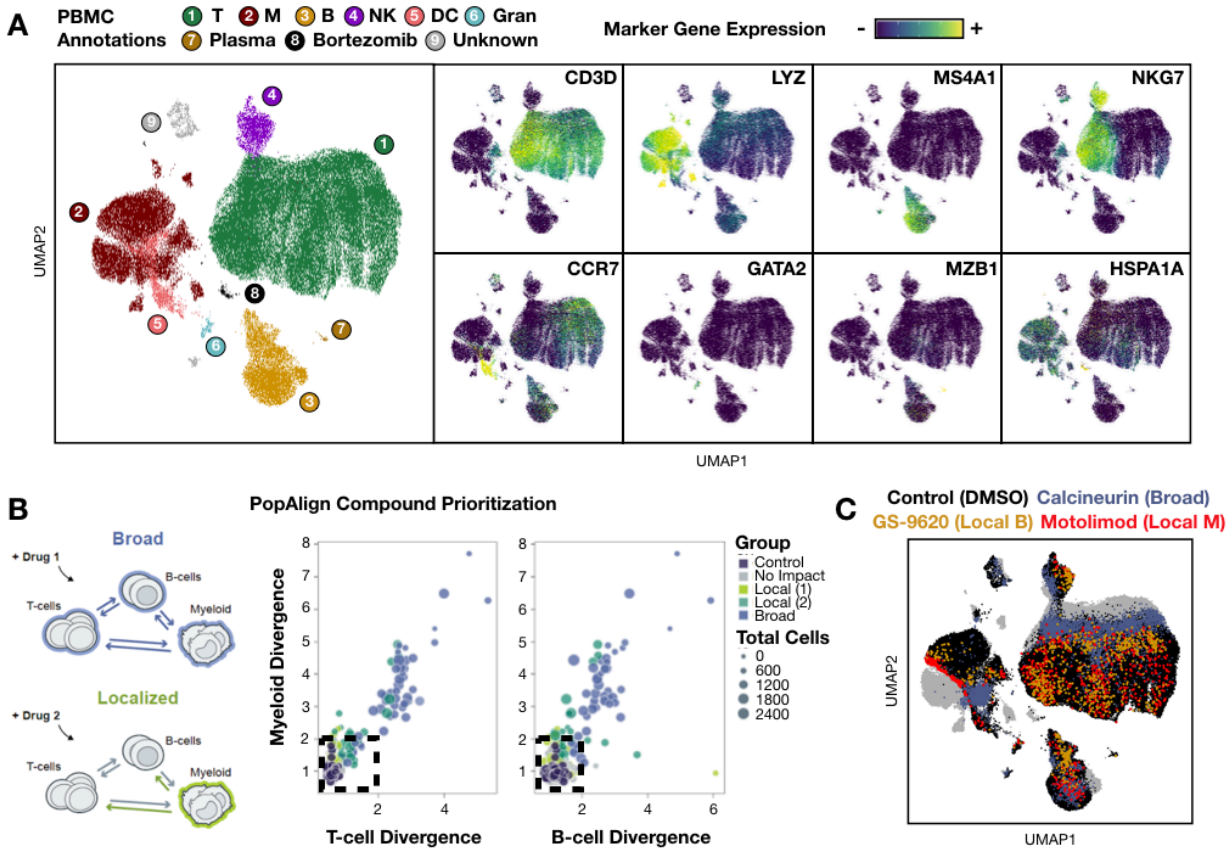

Figure 6-2: PopAlign analytical framework identifies global trends across PBMC scRNA-seq drug screen data.

(A) Gene expression space for PBMCs co-treated with PopAlign-defined high-impact compounds and anti-CD3/CD28 antibodies (along with CD3/CD28-stimulated controls). Cells colored according to cell type annotation workflow (left) with literature-supported marker genes (right). n=143,872 cells

(B) Scatter plots describing the gene expression divergence between each experimental condition and vehicle-alone controls in myeloid cells and T and B lymphocytes (right). Point size corresponds to the number of cells assigned to each scRNA-seq sample. Point color corresponds to PopAlign classifications of immune cell type modulation patterns (see descriptive schematic (left)). High-impact compounds exhibit significant divergence values for any PBMC cell type (demarcated with black box).

(C) PBMC gene expression space colored to highlight PopAlign prioritization results. PopAlign-defined broad modulators (e.g., calcineurin inhibitors; blue) localize separately from control PBMCs (black) in myeloid cells and T and B lymphocytes. Local modulators for B-cells (e.g., GS-9620, goldenrod) and myeloid cells (e.g., Motolimod, red) exhibit differential localization only in specific immune cell type clusters.

This classification reflects the role of calcineurin in the TCR intracellular signaling cascade which, in turn, alters intercellular signaling patterns between T-cells and other immune cell types [22]. In contrast, the TLR7 agonist motolimod and TLR8 agonist GS-9620 were classified as 'local' myeloid and B-cell modulators, respectively, which matches the cell-type-restricted expression of TLR7/8 [23]. Taken together, these analyses illustrate how PopAlign provides an abstracted view of single-cell screen-by-sequencing datasets, highlighting high-impact perturbations and compounds eliciting phenotypes only in therapeutically-relevant cell populations.

6.3.3 Vignette I: PBMC single-cell screen-by-sequencing data survey – Immune sub-type annotation and identification of drug-specific gene expression programs.

Although PopAlign enables rapid and facile parsing of single-cell screen-by-sequencing datasets, uncovering the fine-grained details of transcriptional responses requires further analysis. To this end, we next analyzed scRNA-seq data subsets representing the top perturbations (n = 72 drugs) in CD3/CD28-stimulated T-cells, B-cells, and myeloid cells.

Among the 91,110 T-cells in our subsetted data, we observed a variety of sub-types including FOXP3+ regulatory T-cells (Tregs) and three groups representing previously-described CD4+ and CD8+ T-cell responses to TCR engagement (**Fig. 6-3a**). Specifically, we identified a CD4+ T-cell differentiation group comprised of naïve (IL7R+), central memory (TCM; SELL+ CCR7+) and effector-memory T-cells (TEM; IL2RA+) resembling T helper 1 (Th1) cells [24]. Visual inspection of control cells in gene expression space demonstrated that naïve T-cells were depleted from CD3/CD28-alone controls (**Fig. 6-3a**, inset), matching expectations. Amongst CD8+ T-cells, we identified two differentiation trajectories leading to memory precursor effector cells (MPECs; CD27+ XCL1+) or short-lived effector cells (SLECs; CCL5+) expressing limited (SLEC-1, SLEC-St) or high (SLEC-2, SLEC-3) levels of cytotoxicity genes (e.g., GNLY, GZMB) [25]. Finally, we observed drug-specific sub-types including steroid-treated CD8T+ SLECs expressing glucocorticoid receptor targets (e.g., TXNIP, TSC22D3) [26], as well as HDAC inhibitor-treated T-cells expressing HIST1H1C/D which did not resemble other T-cell sub-types.

B-cell sub-types manifested in a similar fashion as CD4+ T-cells, as we observed naïve (TCL1A+ CXCR4+), memory (IGHA+ CD27+ CXCR4+), and activated B-cells (MIR155HG+ CXCR4-) among the 14,308 B-cells in our subsetted data (**Fig. 6-3b**) [27-29]. Analogous corticosteroid and HDAC-treated B-cell sub-types were also identified (**Fig. 6-3b**, inset), which collectively demonstrates the consistency of T- and B-lymphocyte phenotypes in the context of CD3/CD28-induced acute inflammation.

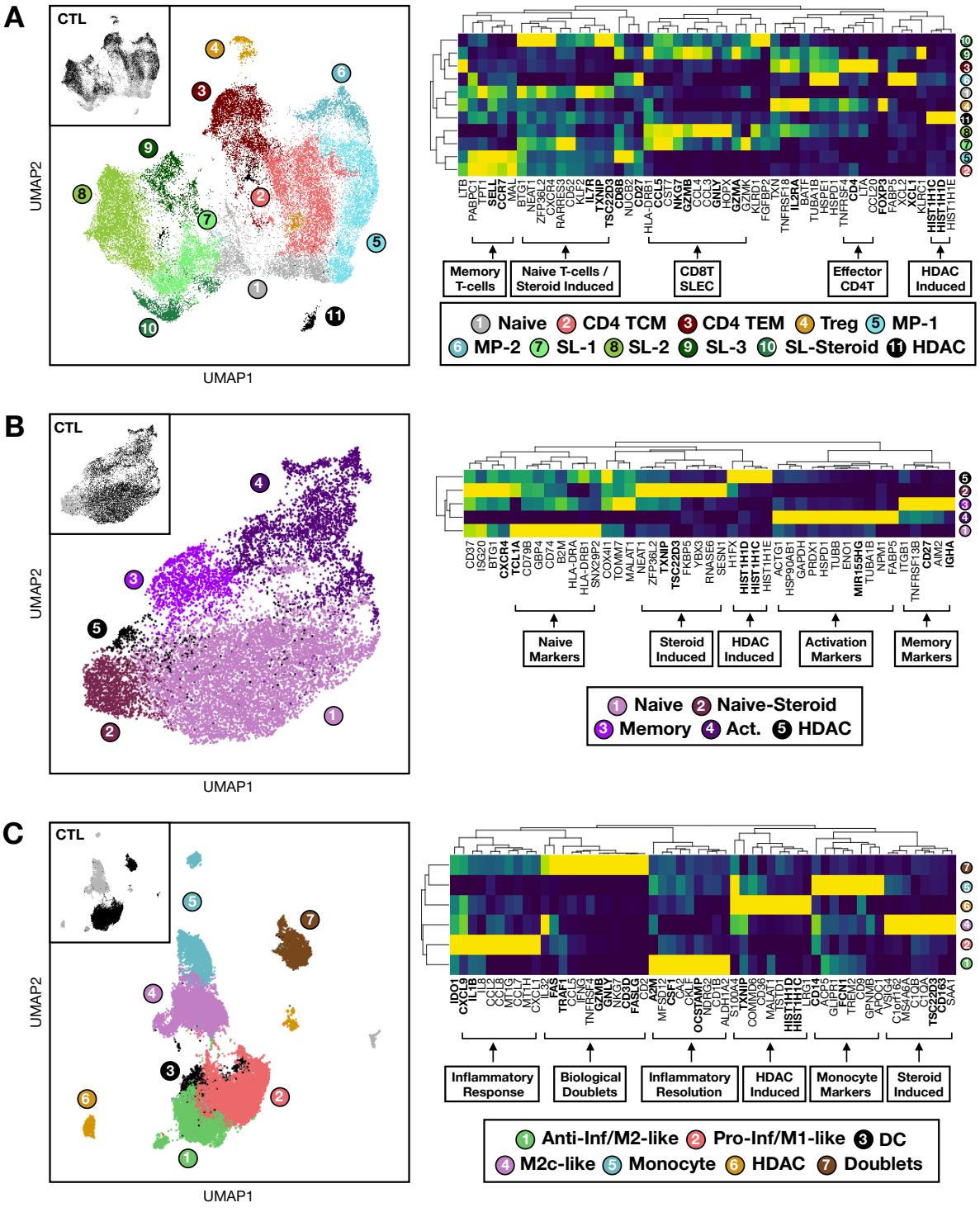

Figure 6-3: PBMC scRNA-seq drug screen data sub-type annotations.

Gene expression space colored by sub-type annotation (left) with associated marker gene heatmaps (right) for T-cells (A), B-cells (B), and monocytes/macrophages (C) co-treated with anti-CD3/CD28 antibodies and high-impact chemical perturbations. Heatmap fill reflects the scaled average expression of normalized marker gene counts across sub-types. Sub-type and gene order was defined using hierarchical clustering on the scaled matrix. Bolded genes were used for literature-informed sub-type annotation, while non-bolded genes were identified through differential expression analysis. Gene expression space insets show the distribution of vehicle-alone control cells and thereby highlight non-overlapping drug-induced sub-types. TCM: CD4+ central memory T-cells; TEM: CD4+ effector memory T-cells; Treg: regulatory T-cells; MP-1/2: CD8+ memory precursor effector cell sub-types; SL-1/2/3/Steroid: CD8+ short-lived effector cell sub-types; HDAC: histone deacetylase inhibitor-treated T-cells, B-cells, or myeloid cells; DC: dendritic cells; Act.: activated B-cells.

Among the 27,443 monocytes and macrophages in our subsetted data, we observed a number of sub-types reflecting different stages of polarization (**Fig. 6-3c**). For example, we observed macrophages expressing the polarization regulator IDO1 [30] along with variable levels of pro-inflammatory (e.g., CXCL9 and IL1B) or anti-inflammatory genes (e.g., CSF1, A2M, and OCSTAMP) [31]. These macrophage sub-types exist along a continuum and are analogous to 'M1' and 'M2' macrophage polarization states that are observed *in vivo* at the onset and resolution of an acute inflammatory response, respectively [32]. Interestingly, we also identified unpolarized (IDO1-) monocytes expressing canonical monocyte markers such as CD14 and M-ficolin (FCN1) [33] which did not overlap with control macrophages in gene expression space (**Fig. 6-3c**, inset). Similar to the depletion of naïve T-cells following CD3/CD28 stimulation, the lack of un-polarized monocytes in control samples matches expectations since polarization is induced by cytokine signaling (e.g., IFNG) from differentiated effector T-cells [32]. Beyond un-polarized monocytes, we also observed a number of drug-specific (i.e., non-overlapping with control samples) cell clusters corresponding in part to HDAC inhibitor (HIST1H1C/D+ macrophages) or corticosteroid exposure (CD163+ TSC22D3+ 'M2c-like' macrophages) which, together with the associated T- and B-cell counterparts, demonstrate the global immunomodulatory effects of these drug classes.

Finally, we annotated a cluster co-expressing markers of pro-inflammatory macrophages (e.g., IDO1 and CXCL9) and cytotoxic CD8+ SLECs (e.g., CD3D, CD8B, CCL5, and GZMB) which also exhibited higher numbers of RNA UMIs and enriched expression of apoptosis genes (e.g., FAS, FASLG, and TRAF1). Interestingly, these cells were uniformly classified as singlets according to MULTI-seq barcode counts despite the co-expression of lymphoid and myeloid markers and were only detected in CD3/CD28-stimulated PBMC conditions. Based on these observations, we annotated these cells as 'biological' doublets – i.e., doublets formed independently of Poisson droplet filling due to a *bona fide* biological process [34]. Macrophage/T-cell biological doublets were observed previously in scRNA-seq data of PBMCs stimulated *in vitro*

with anti-CD3/CD28 antibodies [25], and are likely produced in part through a FAS/FASLG-dependent T-cell-mediated macrophage apoptosis pathway that is active during inflammation [35-37], matching our observations.

Considered collectively, our fine-grained data annotation workflow demonstrates the ability of PBMC-based single-cell screen-by-sequencing to detect immune cell sub-types (e.g., naïve T-cells, Tregs, monocytes, etc.) and biological processes (e.g., macrophage polarization, CD8+ SLEC/MPEC bifurcation, T-cell mediated macrophage apoptosis, etc.) which mimic *in vivo* biology and perturbation responses. Moreover, these analyses identified a number of perturbations which induce demonstrably distinct gene expression programs relative to controls (e.g., HDAC inhibitors, bortezomib, and corticosteroids). However, the majority of perturbations identified as PopAlign high-impact 'hits' did not induce dramatic transcriptional reprogramming in PBMCs, as determined by their significant overlap with control samples in gene expression space. This observation suggests that the immunomodulatory capacity of these compounds manifests more at the level of population structure (i.e., the relative frequencies of cell sub-types) than of immune cell identity (i.e., differentially-expressed genes).

6.3.4 Vignette I: PBMC single-cell screen-by-sequencing data survey – Population response clustering reveals primary modes of T-cell and myeloid cells immunomodulation.

To explore the possibility that immunomodulatory compounds perturb immune population structure, we performed hierarchical clustering on an immune cell-type frequency matrix binned according to perturbation condition (Methods). In T-cells, this analysis identified four types of responses which reflect varying degrees of inhibition of the T-cell activation response to CD3/CD28 stimulation (**Fig. 6-4**). For example, population response clustering identified a 'control response' group where T-cell activation remained intact (**Fig. 6-4**, black labels). This group was depleted of naïve T-cells and enriched for differentiated T-cell sub-types at relatively consistent

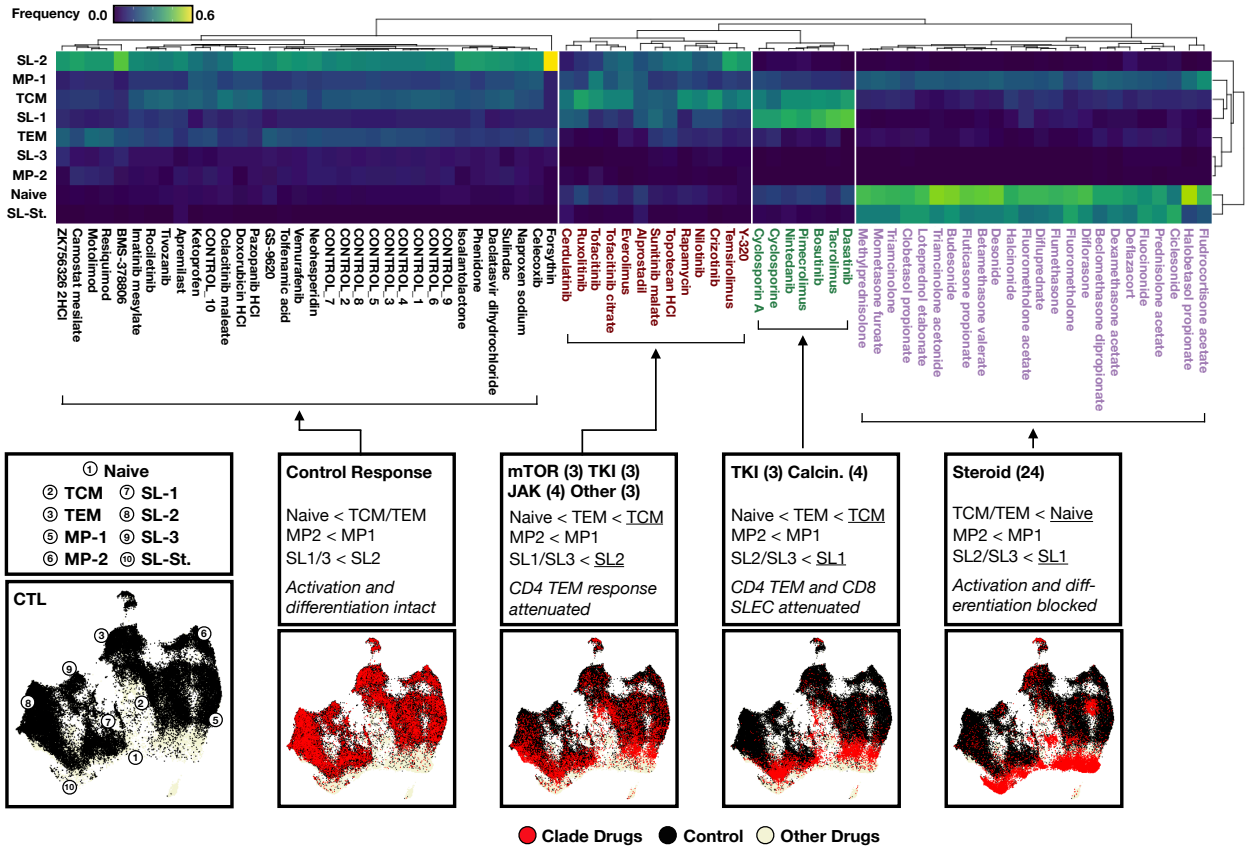

Figure 6-4: Population response clustering reveals primary modes of T-cell immunomodulation after CD3/CD28 stimulation. T-cell sub-type frequency heatmap ordered by hierarchical clustering of sub-types (rows) and perturbation conditions (columns; top). Condition labels are colored according to clade assignment and are associated with distinct legends and T-cell gene expression space embeddings (bottom). Clade-specific legends include information about drug classes, effects on frequencies of CD4+ T-cell, CD8+ SLEC, and CD8+ MPEC lineages (major effects relative to control are underlined), and a brief summary of the population-level response. Reference distribution of control samples in T-cell gene expression space provided for comparison (bottom left). 'Control Response' embeddings only highlight non-control samples. TCM: CD4+ central memory T-cells; TEM: CD4+ effector memory T-cells; MP-1/2: CD8+ memory precursor effector cell sub-types; SL-1/2/3/St.: CD8+ short-lived effector cell sub-types; CTL: controls; TKI: tyrosine kinase inhibitor; Calcin: calcineurin inhibitors.

frequencies which reflect the multi-lineage differentiation 'set-point' after 24 hours of CD3/CD28 activation. At the other extreme, population response clustering grouped corticosteroids together due to their strong enrichment for naïve T-cells and steroid-induced non-cytotoxic CD8+ SLECs (**Fig. 6-4**, purple labels). This phenotype suggests a complete block on T-cell activation, which reflects the ability of corticosteroids to inhibit T-cell activation through myriad mechanisms [38,39].

In addition to the control and corticosteroid response groups, we additionally observed two groups of drugs exhibiting marginal enrichment for naïve T-cells, which suggests an inhibitory effect on T-cell activation. However, this modest T-cell activation inhibition was linked to distinct

effects on downstream T-cell differentiation trajectories in these two drug groups. Specifically, we observed a population response group (**Fig. 6-4**, red labels) including inhibitors of JAK (e.g., tofacitinib, cerdulatinib, and ruxolitinib) and mTOR (e.g., rapamycin, everolimus, and temosirolimus) where a subset of cytotoxic CD8+ SLECs were detected while CD4+ TEM differentiation was specifically diminished (**Fig. 6-5**, top). Notably, these shifts in population structure were supported by significance testing using the single-cell proportion test [40] (**Table 6-1**) and mTOR and JAK inhibitors have well-documented attenuative effects on Th1 differentiation [41-43], supporting the suitability of our analytical approach.

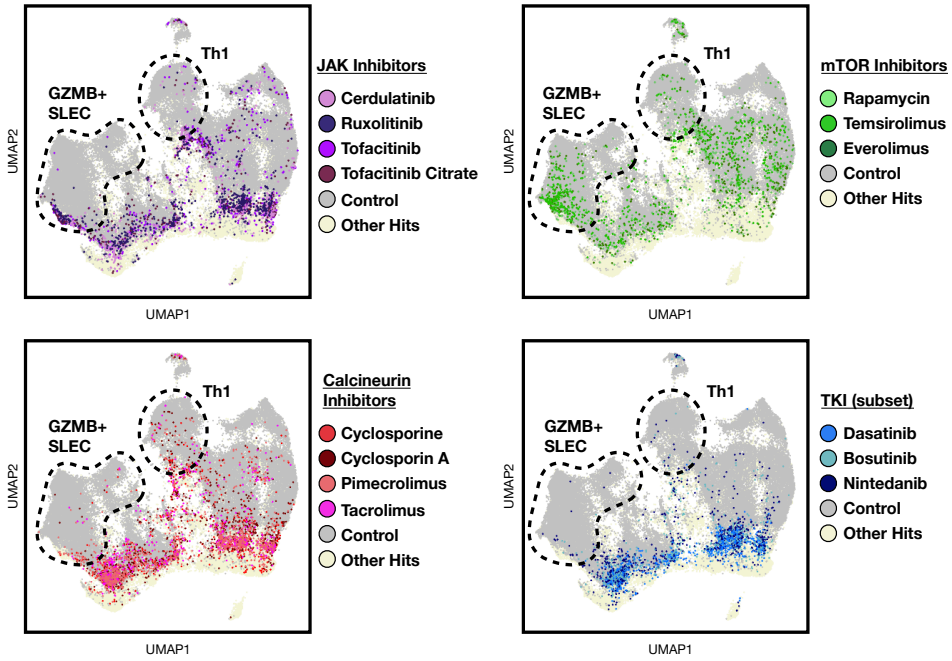

Figure 6-5: Differential effects of distinct drug classes on CD4+ TEM and cytotoxic CD8+ SLEC activation trajectories.
T-cell gene expression space colored by drug annotation for JAK inhibitors (top left, purple), mTOR inhibitors (top right, green), calcineurin inhibitors (bottom left, red), and a subset of TKIs (bottom right, blue). Control cells shown in grey while other drugs are shown in beige. Clusters in gene expression space annotated as Th1 cells and cytotoxic CD8+ SLECs are denoted with dotted lines.

In contrast to the mTOR/JAK inhibitor phenotype, we also observe a population response group (**Fig. 6-4**, green labels) including calcineurin inhibitors and a subset of tyrosine kinase inhibitors (TKIs; dasatinib, bosutinib, and nintedanib) which are depleted of both CD4+ TEMs and cytotoxic CD8+ SLECs (**Fig. 6-5**, bottom) in a statistically-significant fashion (**Table 6-1**).

Calcineurin is one of many components of the TCR signal transduction cascade, and calcineurin inhibitors attenuate T-cell activation primarily by blocking NFATc-induced gene expression [22]. Interestingly, dasatinib and nintedanib do not have any known direct effects on calcineurin activity, but instead are known to suppress TCR-mediated signal transduction by inhibiting LCK phosphorylation [44,45].

Table 6-1: T-cell sub-type fold-changes for immunomodulatory drugs relative to control samples.

Fold-change in T-cell sub-type proportions for JAK inhibitors, mTOR inhibitors, calcineurin inhibitors, and a subset of TKI relative to CD3/CD28-alone control samples. Positive and negative fold-changes passing significance testing are shown in bolded red and blue, respectively. Positive and negative fold-changes not passing significance testing are shown in italicized light red and light blue, respectively. TCM: CD4+ central memory T-cells; TEM: CD4+ effector memory T-cells; MPEC-1/2: CD8+ memory precursor effector cell sub-types; SLEC-1/2/3.: CD8+ short-lived effector cell sub-types.

Drugs	Naive	TCM	TEM	SLEC-1	SLEC-2	SLEC-3	MPEC-1	MPEC-2
Ruxoltinib	**4.8**	*1.8*	**-7.9**	**2.4**	**-2.2**	**-11.6**	*1.1*	**-3.7**
Cerdulatinib	**3.7**	*1.5*	**-3.6**	**2.9**	**-2.9**	**-15.9**	*1.4*	*-1.9*
Tofacitinib	**3.1**	*1.6*	**-5.8**	*1.2*	*-1.8*	**-17.4**	*2.0*	*-1.4*
Tofacitinib citrate	**3.5**	*1.5*	**-7.6**	*1.7*	*-1.2*	**-9.2**	*1.4*	**-3.3**
Rapamycin	**2.3**	*1.5*	*-1.5*	*1.0*	*-1.1*	**-7.6**	*1.2*	**-6.5**
Everolimus	**2.5**	*1.4*	**-2.5**	*1.2*	*-1.1*	**-4.8**	*1.7*	**-13**
Temsirolimus	*1.3*	*1.3*	**-2.1**	*1.5*	*1.3*	**-2.6**	*1.1*	**-10.3**
Cyclosporine	**4.4**	*1.2*	*-1.6*	**4**	**-7.7**	**-8.6**	*1.2*	**-36.8**
Cyclosporin A	**4.1**	*1.3*	-1.9	**3.9**	**-12.8**	**-29.9**	*1.2*	**-64.3**
Pimecrolimus	**4.3**	*1.5*	**-3.2**	**4.1**	**-5.8**	**-19.9**	*1.0*	**-21.4**
Tacrolimus	**3.1**	*1.5*	**-2.3**	**4.7**	**-12.4**	**-7.1**	*-1.2*	**-15.3**
Dasatinib	**5**	*1.7*	**-125.2**	**5**	**-101.6**	**-Inf**	*-1.3*	**-53.9**
Bosutinib	**4.4**	*1.6*	**-3.9**	**4.5**	**-12.9**	**-Inf**	*-1.2*	**-58.4**
Nintedanib	**3.8**	*1.5*	**-4.1**	**4.4**	**-7.3**	**-24.4**	*1.0*	**-52.4**

Collectively, these results illustrate the main modes of T-cell immunomodulation in our experimental system – namely, varying degrees of T-cell activation inhibition resulting in differential abundances of effector CD4+ T-cells and cytotoxic CD8+ SLECs. Notably, many of these phenotypes are supported by decades of clinical and basic immunology research (e.g., mTOR and JAK inhibitor effects on Th1 differentiation, complete-inhibition effect of corticosteroids, etc.). However, these types of drug response signatures would be particularly difficult to assess systematically using bulk screen-by-sequencing approaches (or even differential expression analysis of scRNA-seq data binned by drug), as the relative frequencies of immune cell sub-types are difficult to deconvolve. Moreover, these results illustrate the

convergent effects of diverse drug classes on T-cell activation (e.g., calcineurin inhibitors and LCK-targeting TKIs blocking Th1 and cytotoxic SLEC differentiation) which would be challenging to observe using traditional screening methods which measure limited numbers of biomolecules.

Having identified the primary modes of T-cell immunomodulation using population response clustering, we next applied the same analytical workflow to myeloid cells. Myeloid population response clustering identified six types of responses which generally reflect different stages of macrophage polarization (**Fig. 6-6**). For example, we identified a 'control-response' group which was slightly enriched for pro-inflammatory over anti-inflammatory macrophages and was depleted of un-polarized monocytes and corticosteroid-specific M2c-like macrophages, matching expectations. We also identified two population response groups which mapped cleanly onto the corresponding T-cell responses, namely a corticosteroid-specific group (**Fig. 6-6**, purple

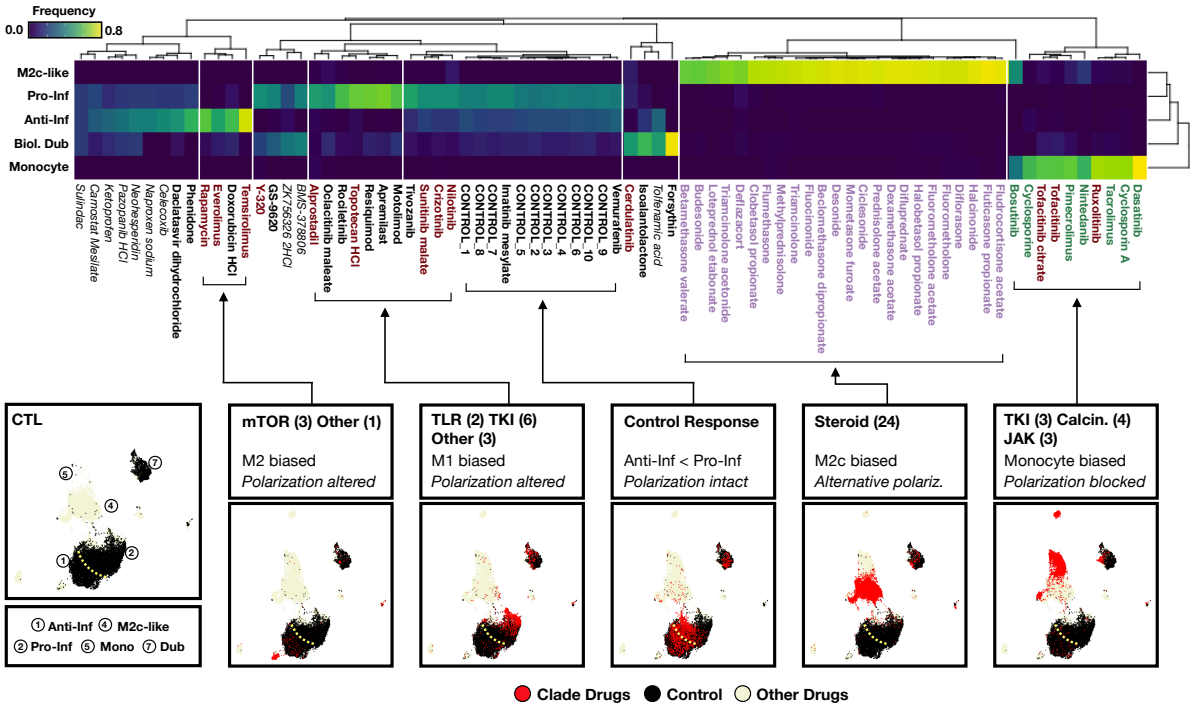

Figure 6-6: Population response clustering reveals primary modes of myeloid cell immunomodulation.

Myeloid sub-type frequency heatmap ordered by hierarchical clustering of sub-types (rows) and perturbation conditions (columns; top). Condition labels are colored according to clade assignments from T-cell analysis (**Fig. 6-4**) and are associated with distinct legends and myeloid gene expression space embeddings (bottom). Reference distribution of control samples in myeloid gene expression space provided for comparison (bottom left), and pro- and anti-inflammatory macrophages are delineated with a dotted yellow line in all embeddings. 'Control Response' embeddings only highlight non-control samples. Italicized drug labels denote drugs with significant myeloid cell depletion relative to control samples. CTL: controls; Biol. Dub: biological doublets; TLR: toll-like receptor; TKI: tyrosine kinase inhibitor; Mono: monocyte.

labels) enriched for M2c-like macrophages, and a group including T-cell activation inhibitors (**Fig. 6-6**, green labels) which were enriched for un-polarized monocytes (**Fig. 6-7**, top).

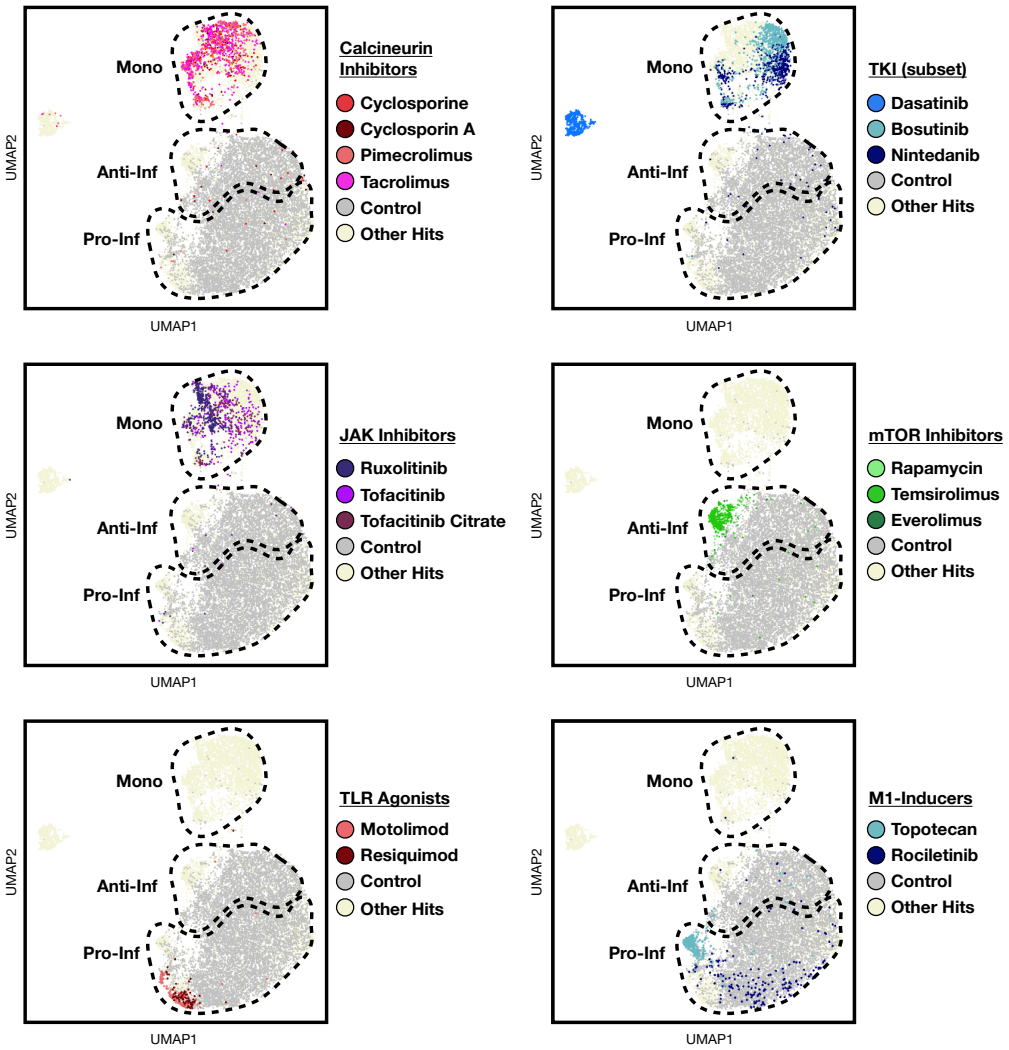

Figure 6-7: Differential effects of distinct drug classes on macrophage polarization.

Myeloid cell gene expression space colored by drug annotation for calcineurin inhibitors (top left, red), a subset of TKIs (top right, blue), JAK inhibitors (middle left, purple), mTOR inhibitors (middle right, green), and M1-biasing agents such as the TLR agonists motolimod and resiquimod (bottom left, red) and topotecan and rociletinib (bottom right, blue). Control and other drugs are colored as in Fig. 6-5. Clusters in gene expression space annotated as monocytes, pro-inflammatory macrophages, and anti-inflammatory macrophages are denoted with dotted lines. DCs, M2c-like macrophages, HDAC-treated macrophages, and biological doublets were omitted from this embedding for visualization purposes.

In contrast to these consistent connections between T-cell and myeloid responses, we additionally observed that T-cell 'control-response' drugs (**Fig. 6-6**, black labels) and compounds linked to impaired Th1 differentiation (**Fig. 6-6**, red labels) were split across a variety of myeloid population response groups. For instance, JAK inhibitors clustered with T-cell activation inhibitors

due to enrichment for un-polarized monocytes, while mTOR inhibitors were associated with anti-inflammatory macrophages (**Fig. 6-7**, middle), despite the observed similarity between mTOR and JAK T-cell responses. In an analogous vein, drugs associated with altered Th1 differentiation (e.g., the type-I topoisomerase inhibitor topotecan) clustered with a subset of T-cell 'control-response' drugs (e.g., the EGFR inhibitor rociletinib) in a clade marked by enrichment for pro-inflammatory over anti-inflammatory macrophages relative to control samples (**Fig. 6-7**, bottom). Notably, the TLR7 against motolimod and TLR7/8 agonist resiquimod were also members in this clade, which reinforces the accuracy of this classification due to the known pro-inflammatory effect of these drugs on macrophages [46]. Finally, many T-cell 'control-response' drugs exhibited cell-type-specific toxicity in myeloid cells, causing >90% depletion in total myeloid cell counts relative to controls (**Fig. 6-6**, italicized labels), which are the focus of later analyses on perturbations to immune population homeostasis mechanisms (see Section 6.3.6).

Considered together, these results complement our T-cell analyses by outlining the main modes of myeloid immunomodulation in our system – namely, blocking/altering macrophage activation and biasing polarization towards pro- or anti-inflammatory phenotypes. When mapping between these immunomodulatory modes, many drug perturbations elicited responses that match expectations – e.g., drugs blocking T-cell activation also block macrophage polarization, TLR7 agonists do not alter T-cells but induce pro-inflammatory macrophages, etc. However, other perturbations defied these expectations by inducing cell state transitions in a cell-type-specific fashion – e.g., JAK and mTOR inhibitors induce similar population-level responses in T-cells but distinct responses in myeloid cells. As a result, these results reinforce the need to assess the immunomodulatory effects of candidate drug compounds across the entire immunological milieu, and position single-cell screen-by-sequencing as an ideal approach to address this need.

6.3.5 Vignette II: Off-target activities of TKIs against T-cell activation and macrophage polarization regulators dictate phenotypic responses.

Immune cell sub-type annotation and population response clustering workflows delineated the primary modes of immunomodulation in our single-cell screen-by-sequencing dataset. Moreover, many of the observed drug-induced immunomodulatory effects were consistent with existing knowledge of the function of these drugs' primary molecular targets during inflammation. One notable exception, however, relates to the TKIs dasatinib, bosutinib, imatinib, and nilotinib, which are dedicated Bcr-Abl transgene inhibitors currently used in the clinic to treat chronic myelogenous leukemia [47]. Dasatinib and bosutinib functioned as macrophage polarization and T-cell activation inhibitors in our data, while these process remained intact following exposure to nilotinib and imatinib. Moreover, TKIs whose primary targets are not Bcr-Abl including nintedanib (primary target: VEGFR/PDGFR), crizotinib (ALK/MET), and sunitinib malate (KIT/VEGFR/FLT3) were similarly split across these two stereotyped immune responses (**Fig. 6-8**), which demonstrates that TKI primary targets did not predict phenotypic response.

Notably, the Bcr-Abl-inhibiting TKIs analyzed in our dataset are known to interact with the Bcr-Abl fusion protein via distinct mechanisms. Specifically, imatinib and nilotinib inhibit Bcr-Abl by binding and stabilizing the inactive conformation of the enzyme, while bosutinib and dasatinib bind Bcr-Abl in its active conformation and limit kinase activity [47]. This fact, considered along with the overlapping immunomodulatory effects of TKIs whose primary proposed targets are independent of Bcr-Abl suggest that off-target drug activities contributed to immune phenotypic responses in our experimental system.

To investigate this hypothesis, we analyzed publicly-available data describing the dissociation constants (K_d) of 72 distinct TKIs and 442 purified wild-type or mutant kinases using multiplexed competition binding assays [48]. Specifically, we parsed these data to identify kinases with associated K_d values less than the drug dose used for this study (1 μM) across all Group I

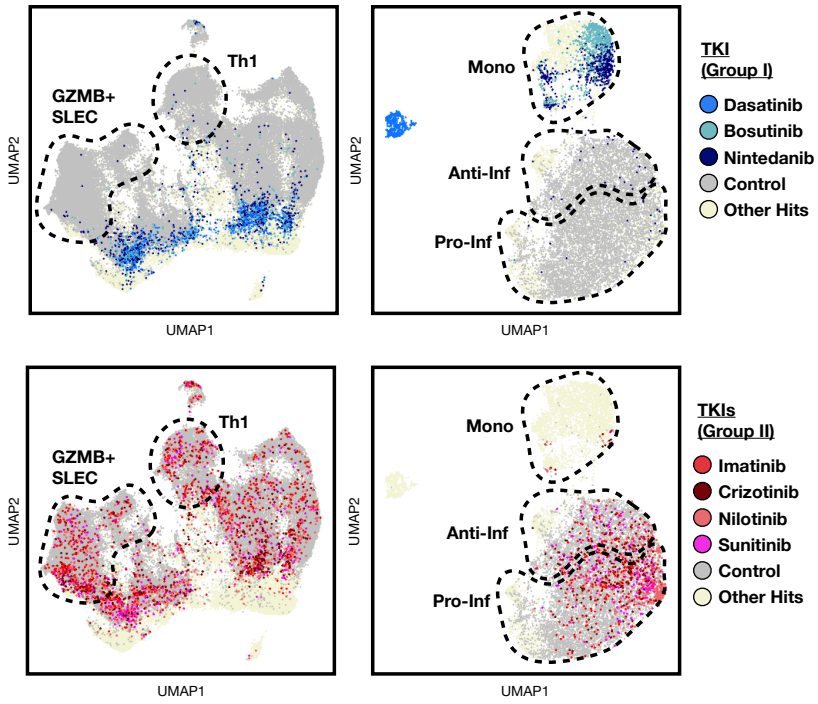

Figure 6-8: Divergent myeloid and T-cell responses to TKIs with overlapping and divergent primary molecular targets.
Gene expression space for T-cells (left column) and myeloid cells (right column) colored by drug annotation for two subsets of TKIs inducing distinct immune cell responses (split by row) as in Fig. 6-5. Clusters in gene expression space annotated as monocytes, pro-/anti-inflammatory macrophages, Th1 effector cells, and cytotoxic CD8+ SLECs are denoted with dotted lines.

TKIs (i.e., dasatinib, bosutinib, and nintedanib) and/or Group II TKIs (i.e., imatinib, nilotinib, crizotinib, and sunitinib malate). We then checked this parsed list for kinases exhibiting uniformly lower K_d values amongst Group I relative to Group II TKIs, or vice versa. This analysis revealed that Group I TKIs have unique affinities for the kinases BTK, MAP2K5, YES1, SIK1, and LCK compared to Group II TKIs (**Table 6-2**). Notably, BTK and Src family kinases (e.g., LCK and YES1) are known regulators of T-cell activation and macrophage polarization [49-51], while SIK1 inhibition is linked to decreased pro-inflammatory macrophage polarization (reviewed in [52]). Moreover, a subset of these predicted off-target drug activities have been empirically validated at the level of immune phenotypes (e.g., dasatinib and nintedanib limit T-cell activation via LCK inhibition [44,45], dasatinib and bosutinib limit pro-inflammatory activity in myeloid cells [53], etc.).

Table 6-2: Off-target drug activities for TKIs inducing divergent T-cell and myeloid phenotypes.

Table repurposed from [47] and subsetted to only include kinases with K_d values lower than 1 μM across all three Group I TKIs (blue headings) and/or all four Group II TKIs (red headings). Numerical values are in nM. Cell fill corresponds to recorded K_d values (dark red: K_d < 10 nM; red: K_d < 100 nM; light red: K_d < 1 μM; grey: K_d < 10 μM). Bolded row names correspond to genes where K_d values are uniformly higher for all Group I TKIs relative to Group II TKIs.

	Dasatinib	Nintedanib	Bosutinib	Crizotinib	Imatinib	Nilotinib	Sunitinib
BTK	1.4	310	4.8	>1e4	>1e5	>1e5	>1e4
LCK	0.2	6.2	0.59	30	40	47	230
MAP2K5	3.3	1.8	8.2	>1e5	>1e5	190	46
SIK1	3.9	670	30	>1e5	>1e5	>1e5	>1e4
YES1	0.3	80	4	770	>1e5	>1e4	120
ABL1	0.016	52	0.037	97	44	15	120
BLK	0.21	380	3.3	110	520	500	65
CSF1R	0.58	48	380	210	11	45	2.5
DDR1	0.69	12	120	510	0.7	1.1	>1e4
DDR2	3.2	42	140	>1e5	15	33	>1e4
EPHB1	0.45	550	33	120	>1e5	>1e4	480
FGR	0.5	300	6.3	670	>1e4	320	270
FYN	0.79	630	11	>1e4	>1e4	>1e4	520
KIT	0.57	2.7	520	>1e5	14	22	1.3
LYN	0.57	940	4.2	940	890	100	270
MAP3K2	140	9.4	30	72	>1e5	>1e5	57
MAP3K3	280	34	54	110	>1e5	>1e5	220
MAP4K1	980	35	15	39	>1e5	890	16
MAP4K3	640	290	5.1	75	>1e5	>1e5	180
MAP4K5	45	390	0.5	79	>1e5	>1e5	41
MINK1	430	82	3.2	>1e4	>1e5	>1e5	29
PDGFR	0.63	15	200	>1e5	14	73	0.075
SIK2	6.4	280	29	200	>1e5	>1e5	580
SLK	720	51	4.7	18	>1e5	>1e5	56
SRC	0.21	580	1	560	>1e5	>1e4	>1e4
STK35	770	810	2	610	>1e5	>1e5	>1e4
TXK	2.1	860	72	850	>1e5	>1e5	>1e5
YSK4	79	5.2	16	980	>1e5	>1e5	17

Collectively, these results demonstrate how off-target drug activities can dictate immune responses in a fashion that is independent of a given drug's primary mechanism of action. Such off-target drug activity is thought to be ubiquitous amongst therapies currently used in the clinic (with both positive and negative consequences), and predicting the poly-pharmacology of drugs is an active area of research [54]. Our data illustrate how cell-based screens coupled to scRNA-seq enable researchers to empirically measure the cumulative effects of off-target drug activities at the level of immune cell phenotypes in the presence of active intercellular communication networks. We anticipate that this feature of single-cell screen-by-sequencing data will prove to be an invaluable approach in the context of cancer immunology as traditional chemotherapies and immunotherapies are used in concert to perturb complex tumor microenvironments.

6.3.6 Vignette III: Macrophage-depleting NSAID response associated with enhanced T-cell-mediated macrophage apoptosis

Comparing population-level responses to drug perturbation in T-cells and myeloid cells revealed diverse cell-type-specific effects on T-cell activation and macrophage polarization. Beyond these primary immunomodulatory modes, however, we additionally observed a subset of drugs associated with significant macrophage-specific toxicity (**Fig. 6-6**, italicized labels). Interestingly, this effect was dependent on CD3/CD28-stimulation (**Fig. 6-9a**), and among the top 10 macrophage-depleting compounds, 50% belonged to a single class – non-steroidal anti-inflammatory drugs (NSAIDs). Comparing the NSAIDs in our dataset which did and did not exhibit the macrophage-depleting phenotype revealed no correlations with cyclooxygenase (COX)-1/2 specificity, chemical structure, or published IC50 values (data not shown). Moreover, fluorescence imaging of PBMCs stimulated for 24 hours with anti-CD3/CD28 antibodies and treated with either a macrophage-depleting NSAID (naproxen sodium) or two non-depleting NSAIDs (zaltoprofen

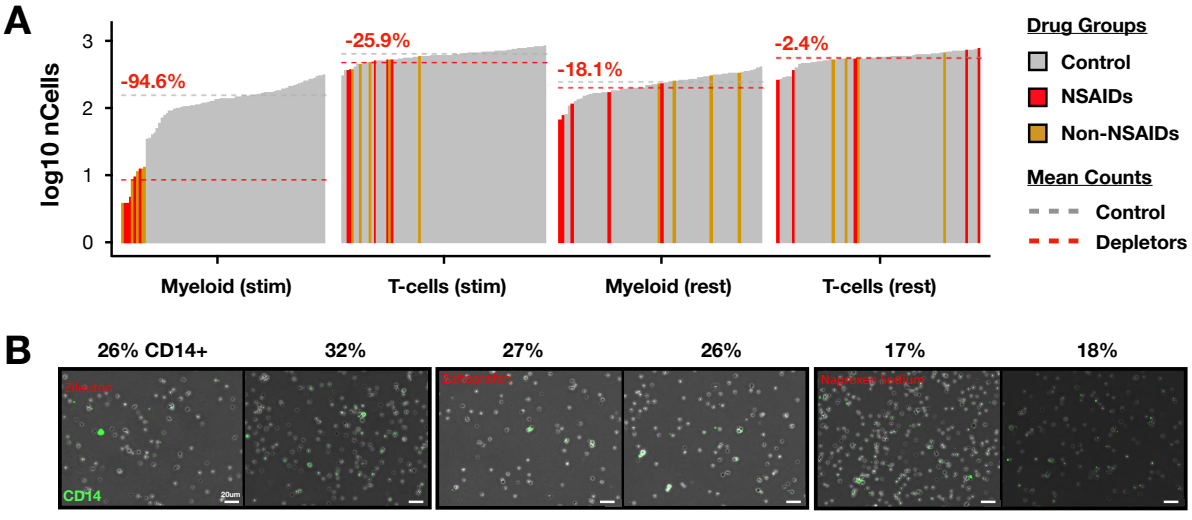

Figure 6-9: Context-specific macrophage depletion is dependent on CD3/CD28-stimulation.

(A) Bar plots showing the number of myeloid and T-cell counts for drugs exhibiting a macrophage-depletion phenotype including NSAIDs (red) and non-NSAIDs (gold) along with control samples (grey) in CD3/CD28-stimulated and resting conditions. Mean counts for each cell type and condition are demarcated with dotted lines for controls (grey) and macrophage-depleting compounds (red). Differences between mean counts are displayed as percentages in red text.

(B) Fluorescence imaging of PBMCs stained for CD14 (green) following stimulation for 24 hours with anti-CD3/CD28 antibodies and treatment with two non-macrophage-depleting NSAIDs (zileuton and zaltoprofen) and one macrophage-depleting NSAID (naproxen sodium) in duplicate. Proportion of CD14+ cells displayed as percentages on top of representative images.

and zileuton) recapitulated these observations, as naproxen sodium caused significant reductions in CD14+ immune cells in this context (**Fig. 6-9b**).

Considering these results along with the previous observation of T-cell/macrophage biological doublets in our dataset, we hypothesized that the mechanism of macrophage-depletion was related to an increase in the rate of T-cell-mediated macrophage apoptosis. To test this hypothesis, we performed a MULTI-seq experiment where PBMCs were stimulated with anti-CD3/CD28 antibodies and perturbed with titrating doses (10 nM, 32 nM, 100 nM, 320 nM, 1μM, 3.2 μM, and 10 μM) of the macrophage-depleting NSAID naproxen sodium and the non-depleting NSAID etodolac for 24 hours. We additionally included titrating doses of the mTOR inhibitors rapamycin as a positive control for a *bona fide* perturbation of both T-cells and macrophages.

Following MULTI-seq sample classification and scRNA-seq data quality-control, we proceeded with an analysis of 11,537 PBMCs. Similar to our previous dataset, we identified clusters in gene expression space corresponding to the major PBMC cell types that were comprised of cells from all drug conditions (**Fig. 6-10a**). Moreover, we identified T-cell sub-types including CD4+ naïve/TCMs and TEMs and CD8+ MPECs and SLECs (**Fig. 6-10b**), as well as macrophage sub-types including anti- and pro-inflammatory macrophages and T-cell/ macrophage biological doublets (**Fig. 6-10c**). Notably, we observed two subsets of biological doublets in these data which were primarily distinguished by the total number of detected RNA UMIs, which perhaps represents different stages of apoptosis.

To interrogate how T-cells and macrophages respond to (non-)macrophage-depleting NSAIDs, we first computed the relative proportions of immune cell sub-types in each drug condition relative to CD3/CD28-stimulated DMSO controls. In rapamycin-perturbed cells, we observed differences recapitulating our previous results, namely enrichment for naïve/TCM T-cells (**Fig. 6-11a**) and anti-inflammatory macrophages (**Fig. 6-11b**) at the expense of Th1 cells

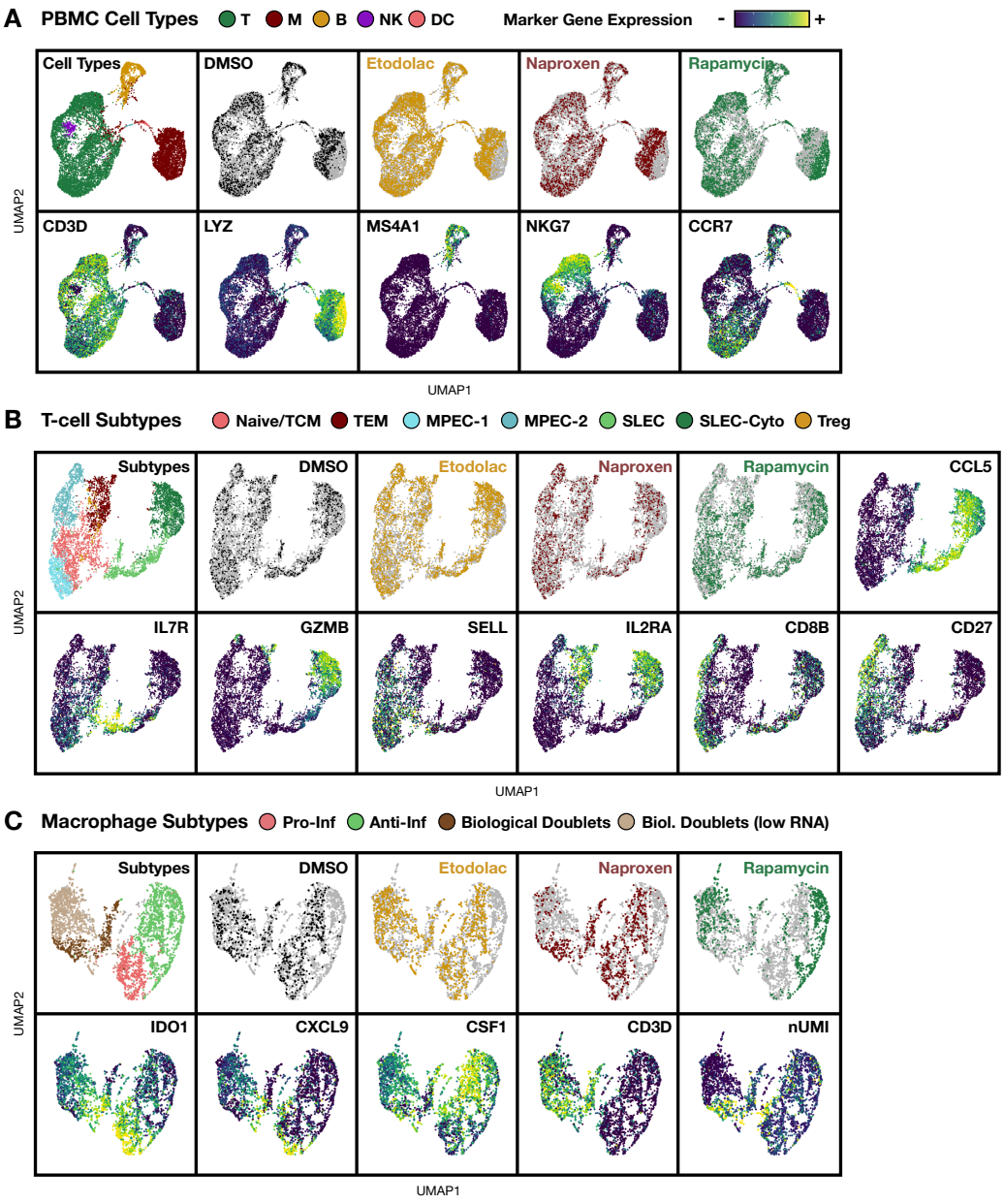

Figure 6-10: PBMC cell type and T-cell and myeloid sub-type annotations in the naproxen sodium, etodolac, and rapamycin dose-response scRNA-seq dataset.

Gene expression space for PBMCs (A), T-cells (B), and myeloid cells (C) colored according to cell type annotations or sub-type annotations, classified drug perturbation, and the expression of informative marker genes. NK: natural killer cell; DC: dendritic cell; TCM: CD4+ central memory T-cells; TEM: CD4+ effector memory T-cells; MPEC-1/2: CD8+ memory precursor effector cell sub-types; SLEC: non-cytotoxic CD8+ short-lived effector cells; SLEC-cyto: cytotoxic CD8+ short-lived effector cells; Treg: regulatory T-cell.

and pro-inflammatory macrophages. Although the shifts in T-cell sub-type proportions did not pass significance testing using the single-cell proportion test [40], we interpreted these results as being generally supportive of our analytical workflow.

In cells treated with naproxen sodium, we observed a marginal decrease in early and cytotoxic CD8+ SLECs (**Fig. 6-11a**) and a significant increase in high-UMI T-cell/macrophage biological doublets at the expense of low-UMI biological doublets compared to both DMSO controls and etodolac (**Fig. 6-11b**). Moreover, comparing the numbers of cells assigned to etodolac and naproxen sodium treatment groups revealed a 52% reduction in total PBMC cell counts (etodolac: 599 cells/sample; naproxen sodium: 289 cells/sample), despite each titration series being seeded with the same number of cells and culture wells. Considered together, these observations support the hypothesis that naproxen sodium increases the tendency of macrophages and cytotoxic CD8+ SLECs to form biological doublets and likely engage in T-cell-mediated macrophage apoptosis. Moreover, the proportion of low- or high-UMI biological doublets amongst etodolac-treated macrophages was not significantly different from DMSO controls,

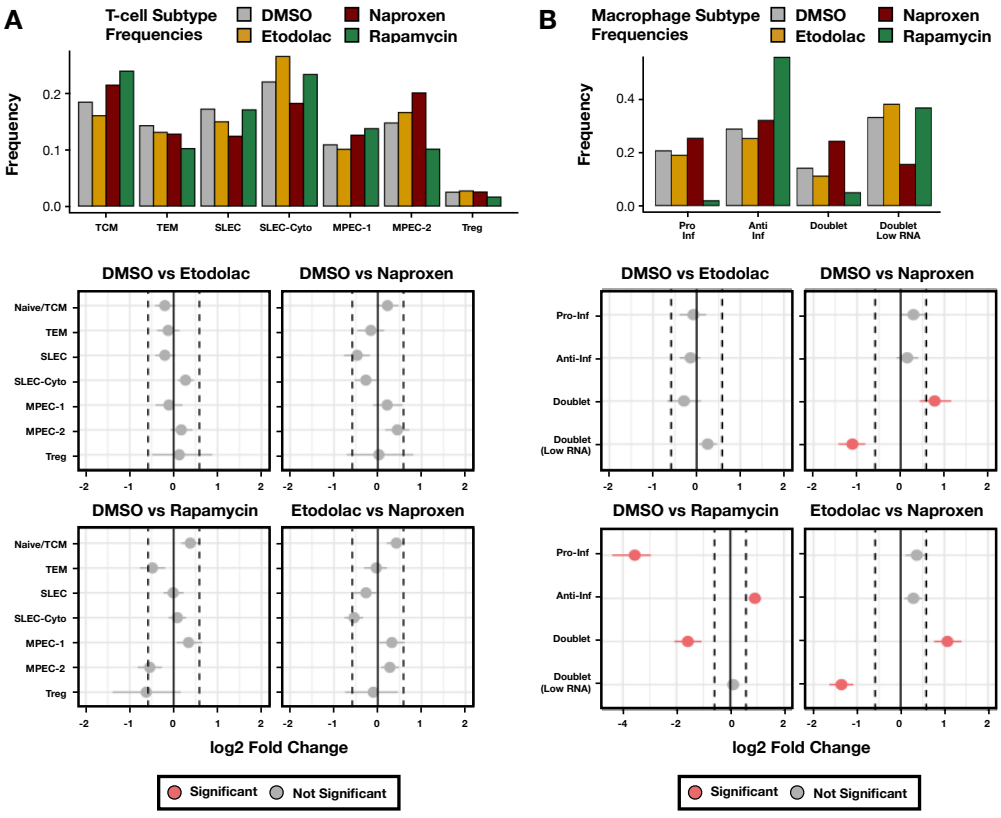

Figure 6-11: Naproxen sodium induces T-cell/macrophage biological doublet formation.

Bar plots showing immune sub-type frequencies across drug conditions (top) and significance testing results from the single-cell proportion test (bottom) for T-cells (A) and macrophages (B). Significance thresholds: Fold-change > 1.5, FDR < 0.05.

suggesting that the cellular response to naproxen sodium which underlies this phenotype is independent of the drug's inhibitory effects on COX1/2.

To better refine our understanding of how naproxen sodium enhances T-cell-mediated macrophage apoptosis, we performed differential gene expression analysis using cells grouped by immune sub-types and drug treatment. While many statistically-significant differentially-expressed genes (DEGs) were detected between rapamycin and DMSO control conditions, no genes passed significance testing for any immune cell sub-type when comparing naproxen sodium to etodolac or DMSO controls (data not shown). The same pattern was observed when binning cells by both immune sub-type and drug-dose, nor did we observe overt differences in sub-type proportions across drug doses (**Fig. 6-12**).

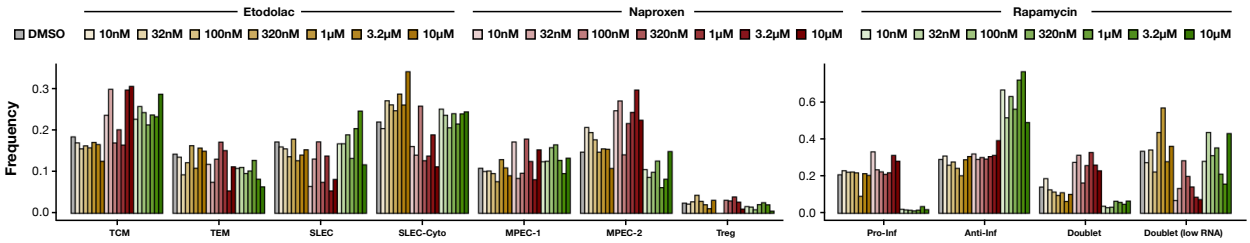

Figure 6-12: Immune sub-type frequencies do not show overt dose-sensitive trends.
Bar plots showing immune cell sub-type frequencies across drugs binned by dose for T-cells (left) and macrophages (right).

Collectively, this MULTI-seq dose-response experiment revealed that compared to the non-macrophage-depleting NSAID etodolac, naproxen sodium treatment was associated with decreased PBMC cell counts, decreased proportions of cytotoxic CD8+ SLECs, and increased proportions of pre-apoptotic T-cell/macrophage biological doublets with high numbers of RNA UMIs. Interestingly, naproxen sodium-treated samples were depleted of low-RNA biological doublets, which perhaps reflects sub-optimal MULTI-seq labeling performance due to the increased number of apoptotic cells in these conditions. As a result, we conclude that naproxen sodium – and likely the other macrophage-depleting NSAIDs observed in our single-cell PBMC

screen-by-sequencing dataset – reduce macrophage cell counts by enhancing T-cell-mediated macrophage apoptosis during CD3/CD28 stimulation. Unfortunately, differential gene expression and immune sub-type proportion analyses did not reveal additional insights into the mechanism by which this phenotype manifests.

However, in addition to demonstrating how single-cell screen-by-sequencing offers the requisite resolution for uncovering novel drug activities, these results illustrate the importance of screening drug activities in complex cellular systems. Put simply, the macrophage-depleting capacity of these NSAIDs would not have been uncovered had T-cell or myeloid cell lines been employed for this study, as T-cell-mediated macrophage apoptosis requires the presence of both T-cells and macrophages. Thus, this vignette epitomizes how the ability to incorporate complex *in vitro* systems into single-cell screen-by-sequencing pipelines holds the potential to redefine our understanding of how existing and novel drug compounds influence the immune system.

6.3.7 Vignette IV: Targeted transcript enrichment maintains scRNA-seq data information content while minimizing next-generation sequencing costs.

scRNA-seq sample multiplexing technologies dramatically lower the cost of large-scale scRNA-seq experiments and make studies that were previously economically-unfeasible possible. For example, using MULTI-seq to generate this PBMC single-cell screen-by-sequencing dataset in a pooled-sample format resulted in a reduction of >$900,000 in droplet microfluidics reagents compared to parallel sample processing. However, next-generation sequencing requirements for highly-multiplexed experimental designs which require large numbers of cells remain incredibly high. For instance, sequencing a 1,500,000 cell dataset at a depth of 40,000 reads per cell (RPC) translates to >$100,000 in next-generation sequencing costs alone.

One potential solution to this issue is to use cDNA target enrichment strategies [54,55] to limit next-generation sequencing read allotment to genes associated with informative levels of

biological information [56]. Such approaches have been successfully applied in other scRNA-seq experiments [57], but to the best of our knowledge, have not been tested on PBMCs following chemical perturbation. To determine whether target enrichment is suitable for our PBMC single-cell screen-by-sequencing platform, we performed a follow-up experiment wherein PBMCs were co-cultured for 24 hours with anti-CD3/CD28 antibodies and immunomodulatory drugs that were found to induce distinct gene expression responses in our initial screen (e.g., corticosteroids, mTOR inhibitors, and TKIs). We also included stimulated and resting control samples that were cultured without drugs in the presence or absence of anti-CD3/CD28 antibodies, respectively. Samples from each condition were tagged with MULTI-seq sample barcodes and pooled prior to scRNA-seq using two 10x Genomics microfluidic lanes. One cDNA library was then submitted for full-transcriptome sequencing, while the other was subjected to target enrichment using the Human Immunology Panel from 10x Genomics. Specifically, one final cDNA library was concentrated and reconstituted in the presence of biotinylated 'bait' oligonucleotides which feature complementarity to distinct mRNA domains of 1,056 genes with established immunological functions. Following bait hybridization, targeted transcripts are then purified onto streptavidin-conjugated magnetic beads while the remaining cDNA library is discarded. The target-enriched library is then subjected to a final library preparation PCR amplification step prior to sequencing.

Following MULTI-seq sample classification and quality-control filtering, we proceeded with an analysis of 7,148 and 14,986 PBMCs from the full-transcriptome and target-enriched scRNA-seq datasets, respectively. We identified major PBMC cell types in both datasets, including T-cells, B-cells, myeloid cells, DCs, and NK cells (**Fig. 6-13a**) despite the expected decrease in the number of detected genes and UMIs in the target-enriched data (**Fig. 6-13b**). Moreover, T-cell and myeloid subtypes including CD4+ lineage cells (naïve, TCM, and TEM), CD8+ MPECs and SLECs, M2c-like, pro-, and anti-inflammatory macrophages, monocytes, and T-cell/macrophage biological doublets were also identifiable in both scRNA-seq datasets (**Fig. 6-14**). Collectively,

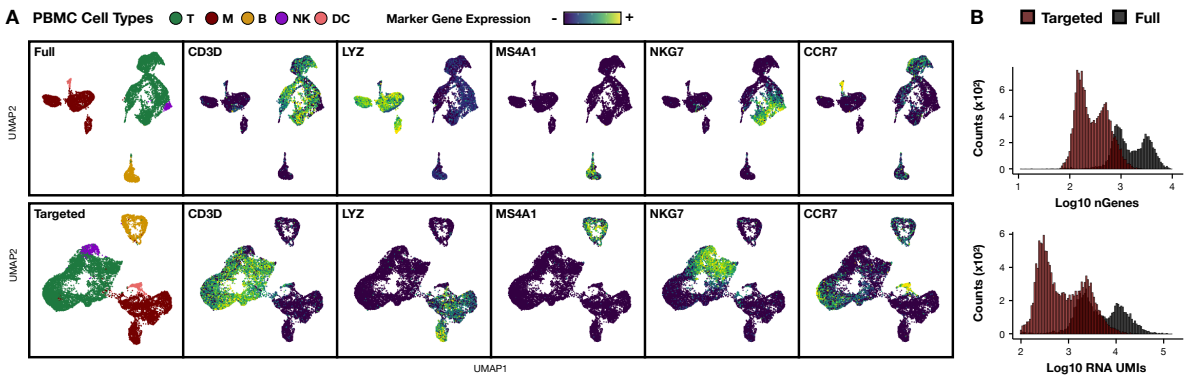

Figure 6-13: PBMC cell type annotation in full-transcriptome and target-enriched scRNA-seq data.

(A) Gene expression space for PBMCs in full-transcriptome (top) and target-enriched (bottom) scRNA-seq datasets colored according to cell type annotations (left) and informative marker gene expression

(B) Histograms describing the number of detected genes (top) and RNA UMIs (bottom) in the full-transcriptome (black) and target-enriched (dark red) scRNA-seq datasets.

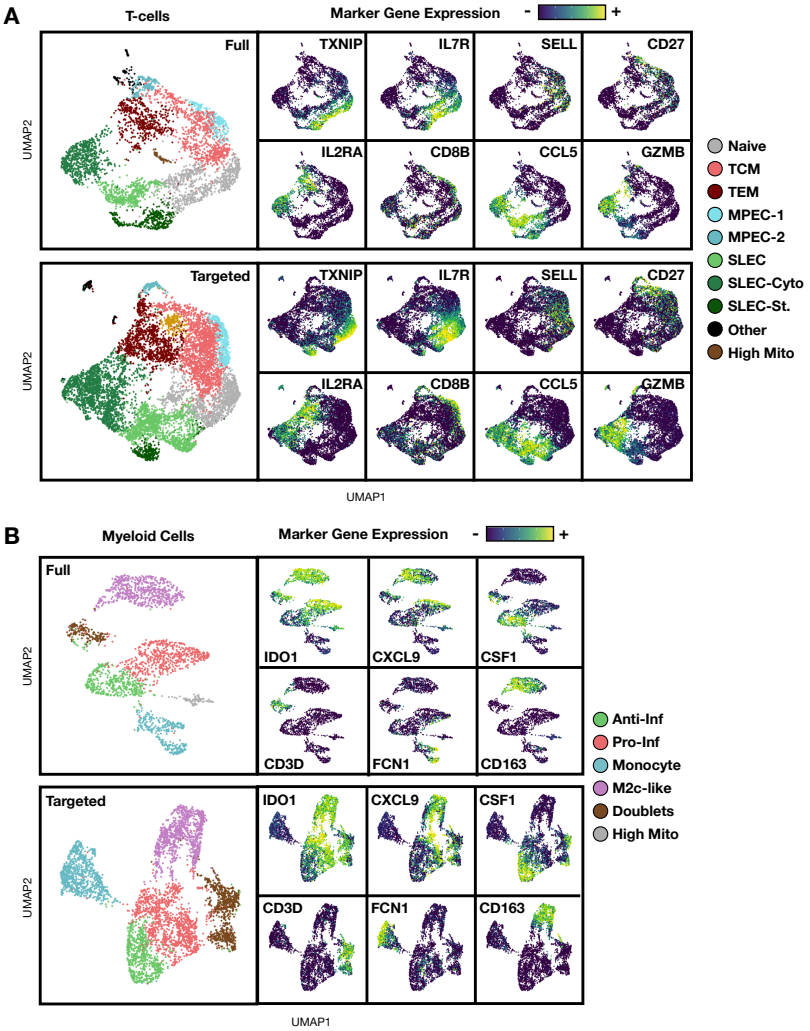

Figure 6-14: T-cell and myeloid sub-type annotations in full-transcriptome and target-enriched scRNA-seq data.

Gene expression space for T-cells (A) and myeloid cells (B) in full-transcriptome (top) and target-enriched (bottom) scRNA-seq datasets colored according to sub-type annotations (left) and informative marker gene expression.

these results demonstrate the preservation of biological information needed for immune cell sub-type annotation following target-enrichment in drug-perturbed PBMC scRNA-seq data.

Beyond maintaining the accuracy of immune annotation workflows, these data additionally highlighted how next-generation sequencing costs can be minimized using targeted transcript enrichment strategies, as our target-enriched library was sequenced to 47% of the depth of the full-transcriptome library (e.g., 12,634 RPC vs 27,091 RPC). However, optimizing the economics of future single-cell screen-by-sequencing experiments would benefit from identifying the minimum sequencing depth needed to resolve immune cell sub-types and drug responses. In this vein, we next characterized the distribution of immune cell sub-type and drug perturbation labels in gene expression space following iterative down-sampling of reads to ~1,000 RPC and ~2,500 RPC for the target-enriched and full-transcriptome libraries, respectively. We additionally compared the number of detected DEGs between immune cell sub-types and drug vs control conditions, as well sub-type/drug group information purity (vis-à-vis the entropy-based metric, ROGUE [58]) for each down-sampled dataset.

Qualitative interrogation of how T-cell and myeloid sub-types were distributed in gene expression space for progressively down-sampled datasets demonstrates that sub-types were visually-discernible with as few as 1,000 RPC in the target-enriched data (**Fig. 6-15a**). For example, monocytes and anti-inflammatory macrophages were non-overlapping in gene expression space in both the down-sampled target-enriched data, as well as the full-transcriptome dataset after down-sampling to 2,500 RPC (**Fig. 6-15a**, right). In contrast, many T-cell subsets (e.g., naïve T-cells, CD4+ TCMs, and immature CD8+ MPECs) were intermixed in gene expression space for the down-sampled full-transcriptome data (**Fig. 6-15a**, left, panel #7), while these subsets were resolvable in the 1,000 RPC target-enriched data (**Fig. 6-15a**, left).

These qualitative observations were generally reflected in subsequent quantitative comparisons of immune sub-type label information purity (i.e., ROGUE score) and the number of

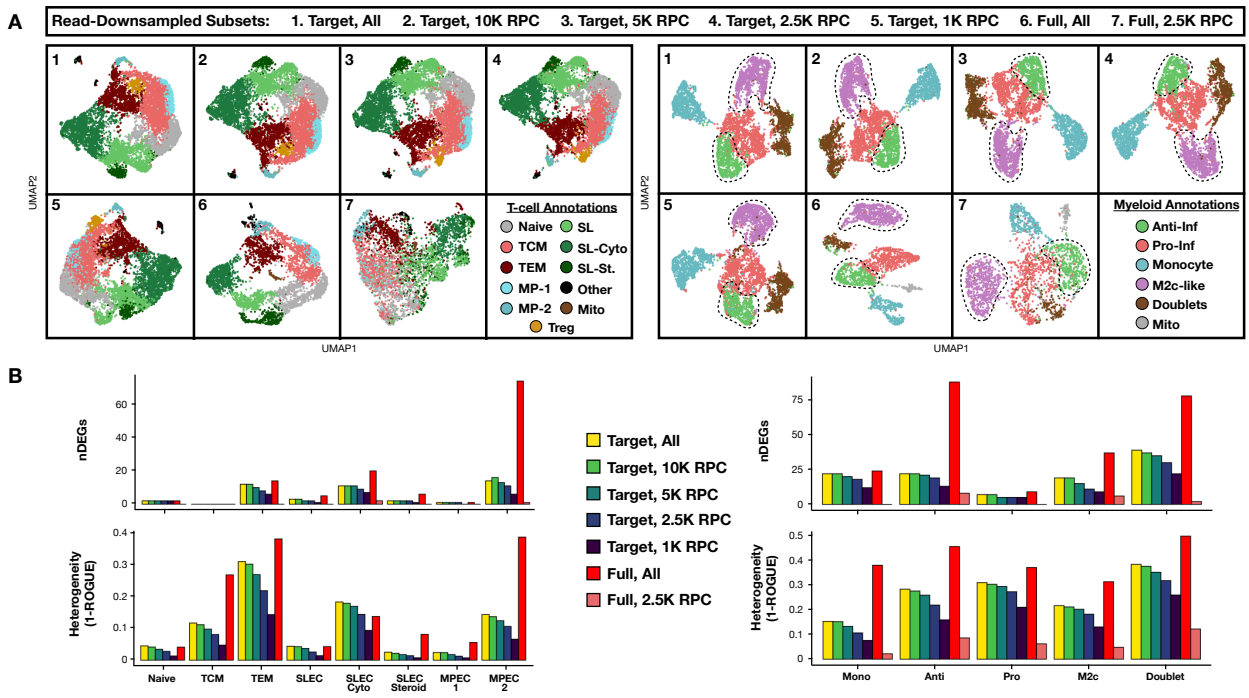

Figure 6-15: Qualitative and quantitative comparisons of immune cell sub-types after iterative down-sampling of next-generation sequencing reads in target-enriched and full-transcriptome scRNA-seq data.

(A) Gene expression space for T-cells (left) and myeloid cells (right) across target-enriched and full-transcriptome scRNA-seq data following iterative down-sampling of sequencing reads (denoted with UMAP panel numbers). Cells are colored by sub-type annotation as in Fig. 6-15. Dotted lines highlight anti-inflammatory macrophages and monocytes.

(B) Bar plots describing the number of detected differentially-expressed genes (top) and preserved heterogeneity (bottom) for T-cell (left) and myeloid cell sub-types (right). Bars are colored according to dataset (target-enriched using the 'viridis' palette; full-transcriptome in red) and extent of read down-sampling.

detected sub-type-specific DEGs. Specifically, we observed that the number of DEGs and magnitude of preserved heterogeneity were depleted across all immune sub-types in the 2,500 RPC full-transcriptome dataset (**Fig. 6-15b**, light coral). In contrast, both metrics were elevated in the 1,000 RPC target-enriched dataset (**Fig. 6-15b**, dark blue) relative to the 2,500 RPC full-transcriptome dataset, and for the full-transcriptome dataset (**Fig. 6-15b**, red) relative to all other datasets. These observations illustrate that some biological signal is lost during target-enrichment (as expected), and that target-enrichment maintains biological information at shallow sequencing depths to a greater extent than full-transcriptome sequencing. Moreover, amongst the down-sampled target-enriched datasets, both metrics decreased to varying extents in different immune sub-types and did not exhibit a clear inflection point, making identification of an 'optimal' minimum sequencing depth for cell sub-type annotation difficult.

We next repeated this analytical workflow using drug perturbation labels instead of immune sub-type labels. Mirroring our previous qualitative observations, most condition-specific shifts in gene expression space were cleanly resolved in the target-enriched dataset. Specifically, resting and CD3/CD28-stimulated control samples (**Fig. 6-16a**, top) were distinguished according to enrichment for monocytes and polarized macrophages in myeloid cell gene expression space (**Fig. 6-16a**, top right), and by enrichment for naïve and differentiated T-cells in T-cell gene expression space (**Fig. 6-16a**, top left), respectively. Moreover, resting and stimulated controls

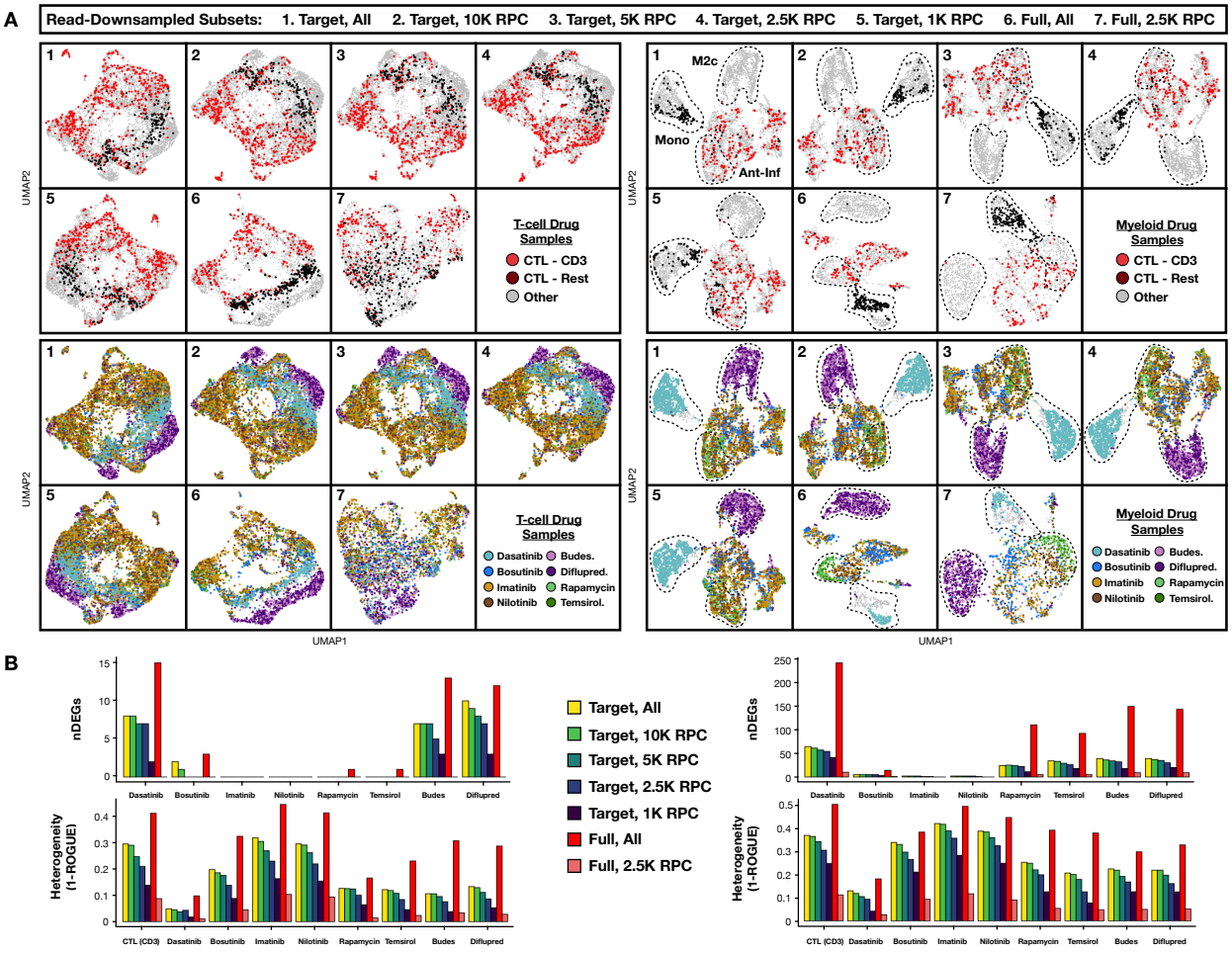

Figure 6-16: Qualitative and quantitative comparisons of immune drug responses after iterative down-sampling of next-generation sequencing reads in target-enriched and full-transcriptome scRNA-seq data.

(A) Gene expression space for T-cells (left) and myeloid cells (right) across target-enriched and full-transcriptome scRNA-seq data following iterative down-sampling of sequencing reads (denoted with UMAP panel numbers). Cells are colored by resting and CD3/CD28-stimulated controls (top) or non-control drug perturbations (bottom). Dotted lines denote monocytes, M2c-like, and anti-inflammatory macrophages and highlight drug perturbations which drive enrichment for those myeloid sub-types.

(B) Bar plots describing the number of differentially-expressed genes relative to CD3/CD28-stimulated controls (top) and preserved heterogeneity (bottom) for T-cells (left) and myeloid cells (right) grouped by drug perturbation. Bars are colored as in Fig. 6-16.

exhibited significant overlap in T-cell gene expression space for the 2,500 RPC full-transcriptome dataset(**Fig. 6-16a**, top left, panel #7), as was observed during the T-cell sub-type analyses. Beyond control samples, drug-specific shifts such as corticosteroid-specific enrichment for M2c-like macrophages and naïve T-cells (**Fig. 6-16a**, bottom, purple) and dasatinib-specific enrichment for monocytes and naïve T-cells (**Fig. 6-16a**, bottom, teal) were also retained in the target-enriched dataset. Notably, dasatinib-treated T-cells were not resolvable in the 2,500 RPC full-transcriptome data (**Fig. 6-16a**, bottom left, panel #7), while macrophages treated with the mTOR inhibitors rapamycin and temsirolimus overlapped with control anti-inflammatory macrophages in all target-enriched datasets (**Fig. 6-16a**, bottom left, panels #1-5) unlike the full-transcriptome datasets (**Fig. 6-16a**, bottoms left, panels #6-7).

Beyond these qualitative observations, quantitative comparisons of drug-specific DEGs and information preservation across decreasing sequencing depth reveal trends similar to those observed for our immune sub-type annotation analyses. Specifically, both metrics were lowest for the 2,500 RPC full-transcriptome dataset, both metrics were highest for the original full-transcriptome dataset, and both metrics decreased variably across different drug sub-sets for the target-enriched datasets (**Fig. 6-16b**). Considered collectively, these analyses demonstrate that existing target-enrichment panels are well-suited for detecting immune sub-types and drug responses at as low as 1,000 RPC. However, detecting the nuances of some drug responses (e.g., mTOR inhibitor-treated anti-inflammatory macrophages) may require further panel optimization and/or deeper sequencing.

6.4 Discussion

Single-cell genomics sample multiplexing technologies make it possible to generate high-dimensional single-cell measurements across large numbers of experimental conditions. In the

context of high-throughput compound screening, these technologies can be employed to perform single-cell screen-by-sequencing studies using complex *in vitro* systems comprised of diverse interacting cell types which better recapitulate *in vivo* biology. Moreover, scRNA-seq can provide more nuanced views on how cells respond to chemical perturbation, which we anticipate will improve clinical trial success rates and open new therapeutic avenues beyond metrics that can be readily assessed using low-dimensional read-outs and/or bulk transcriptomics.

In this descriptive study, we outline the key insights gained from the largest per-sample single-cell screen-by-sequencing experiment performed to date, which produced a final scRNA-seq dataset of ~1,000,000 resting or CD3/CD28-stimulated PBMCs spanning 801 perturbation conditions. First, we describe how PopAlign can rapidly parse single-cell screen-by-sequencing datasets to identify high-impact perturbations and classify compounds according to their cell-type-specific effects on PBMCs. We then dig deeper into the high-impact perturbations and use population-response clustering to describe the primary modes of immunomodulation in T-cells and myeloid cells. Specifically, we identify drugs which block T-cell activation but have diverse effects on CD4+ T-cell and CD8+ SLEC differentiation, as well as drugs which block or bias macrophage polarization towards pro- or anti-inflammatory states. Considering the role of the immune system in battling cancer and the fact that many of the drugs identified in this vignette are commonly used in the clinic as chemotherapeutics (e.g., Bcr-Abl inhibitors, mTOR inhibitors, topoisomerase inhibitors, etc.), we believe that these results illustrate the importance of using single-cell screen-by-sequencing to interrogate and thereby learn to leverage (or avoid) the immunomodulatory capacities of existing treatments in certain disease contexts.

After identifying the primary modes of immunomodulation in our *in vitro* PBMC model of acute inflammation, we next focused on two confounding results from our screen. First, we explored the divergent immune cell responses to Bcr-Abl targeting TKIs in order to understand why the primary molecular targets of these drugs were not predictive of immune cell phenotype

following perturbation. After incorporating publicly-available TKI binding affinity data, these analyses revealed that off-target effects were predictive of immune response, and demonstrated how single-cell screen-by-sequencing can capture the cumulative effects of off-target drug activities. Second, we performed a follow-up MULTI-seq experiment to understand how treatment with a subset of NSAIDs resulted in >90% depletion of macrophages specifically in CD3/CD28-stimulated PBMCs. These analyses revealed that at least one macrophage-depleting NSAID, naproxen sodium, was associated with increased formation of T-cell/macrophage biological doublets compared to DMSO controls and the non-macrophage-depleting NSAID etodolac, which suggests that macrophage-depleting NSAIDs enhance T-cell-mediated macrophage apoptosis through a currently-unknown mechanism.

Finally, we performed a final MULTI-seq experiment which incorporated target-enrichment strategies to demonstrate how next-generation sequencing costs can be limited in future single-cell screen-by-sequencing studies. Specifically, we observed the immune cell sub-type detection was similarly-accurate between target-enriched and full-transcriptome scRNA-seq data, and that sub-type resolution was maintained with 1,000 RPC shallow sequencing in the target-enriched data but was lost at 2,500 RPC in the full-transcriptome data. Moreover, we observed that many drug-specific transcriptional responses were recovered at 1,000 RPC shallow sequencing in the target-enriched data, but the effect of the mTOR inhibitor rapamycin on macrophages was not well-resolved at any sequencing depth. This suggests that further refinement of PBMC-tailored target-enrichment panels is necessary for future single-cell screen-by-sequencing efforts.

Notably, while the majority of the analyses in this study were centered on CD3/CD28-stimulated T-cells and monocytes/macrophages, we see these data as a rich resource for future hypothesis generation regarding the context-specificity of drug responses (e.g., CD3/CD28-stimulated vs resting) as well as how compounds influence other immune cell types (e.g., B-cells, NK cells, and DCs). Moreover, for dynamic experimental systems such as CD3/CD28-stimulated

PBMCs, we anticipate that much can be gained by using time-course experimental designs (along with the more traditional dose-responses studies) to validate drug phenotypes observed in initial single-cell screen-by-sequencing experiments. Time-course experiments will be especially insightful in contexts where perturbations appear to limit different stages of a given trajectory (e.g., JAK and mTOR inhibitors which enable CD4+ TCM differentiation but not Th1 effector T-cell formation; drugs likes rociletinib/topotecan and mTOR inhibitors which bias macrophage polarization towards pro- or anti-inflammatory fates, respectively), as these tests can shed light on the causal nature of the drug responses – e.g., whether the main trajectory is intact but the transition kinetics are slower vs the main trajectory is incommensurably altered.

6.5 Methods

6.5.1. PBMC sample preparation, scRNA-seq library preparation, and next-generation sequencing.

For all single-cell screen-by-sequencing experiments, PBMCs from a single healthy donor obtained from HemaCare were thawed according to the 'Fresh Frozen Human Peripheral Blood Mononuclear Cells for Single Cell RNA Sequencing' protocol from 10x Genomics. PBMCs were thawed into RPMI 1640 Medium, GlutaMAX™ Supplement, HEPES (ThermoFisher cat#72400047) supplemented with 10% FBS and 1% pen/strep (RPMI-FBS-PS) and plated at 0.5-1x10^7 cells/dish in 10cm ultra-low attachment culture dishes (Corning cat#3262). PBMCs were then rested overnight at 37 °C and 5% CO_2, lifted with TrypLE Express (ThermoFisher cat#12605036), washed, and re-plated at 2x10^5 cells/well of 96-well flat-bottom ultra-low attachment culture plates (Corning cat#3474). Upon re-plating, PBMCs were cultured at 37 °C and 5% CO_2 for 24 hours in RPMI-FBS-PS with or without 25 μL/mL ImmunoCult™ Human

CD3/CD28 T Cell Activator (Stemcell Technologies cat# 10971) supplemented with materials specific to each described experiment.

For the primary single-cell screen-by-sequencing experiment, PBMCs were cultured with 1 μM of a single chemical compound from the Selleckchem Immunology/Inflammation Compound Library (Selleckchem cat#L4100) or FDA-approved Drug Library (Selleckchem cat#L1300) or equal volumes of DMSO in a total volume of 200 μL. For the dose-response experiment, PBMCs were cultured with 10 nM, 32 nM, 100 nM, 320 nM, 1 μM, 3.2 μM, or 10 μM of naproxen sodium, etodolac, or rapamycin obtained from the Immunology/Inflammation Compound Library. For the target-enrichment experiment, PBMCs were cultured with 1 μM of dasatinib, bosutinib, imatinib, or nintedanib obtained from the FDA-approved Drug Library or budesonide, difluprednate, rapamycin, or temsirolimus from the Immunology/Inflammation Compound Library.

After 24 hours of drug and/or anti-CD3/CD28 antibody stimulation, PBMCs were prepared for scRNA-seq on the 10x Genomics V3 platform. Specifically, suspension PBMCs were transferred to unique wells of a 96-well conical-bottom (Corning cat#249935) or round-bottom (Corning cat#3799) culture plate and placed on ice as adherent PBMCs were lifted using TrypLE Express. Once lifted, trypsinization reactions were quenched with RPMI-FBS-PS and pooled with suspension cells in corresponding wells of the 96-well plate on ice. Cells were then centrifuged for 3 minutes at 400xg, the supernatant was aspirated, and cells were resuspended in ice-cold PBS. This washing procedure was repeated once before resuspension in 180 μL PBS. PBMCs were then MULTI-seq barcoded, quenched with 1% BSA in PBS, washed 1-2 times with 1% BSA in PBS, and pooled as described previously [8]. After MULTI-seq labeling, cells were loaded into 10x Genomics microfluidics lanes with distinct strategies for each described experiment.

For the primary single-cell screen-by-sequencing experiment, PBMCs were loaded into 6-7 microfluidics lanes per 96-well plate at a density targeting 20,000 yielded cells/lane. For the

dose-response experiment, PBMCs were loaded into a single microfluidic lane at a density targeting 24,000 cells. For the target-enrichment experiment, PBMCs were loaded into two microfluidic lanes at two densities targeting either 24,000 yielded cells (target-enriched library) or 8,000 yielded cells (full-transcriptome library). Following droplet microfluidic emulsion, cDNA libraries were prepared according to either the "Chromium Single Cell 3′ Reagent Kits v3" or "Chromium Single Cell 3′ Reagent Kits v3.1" user guides from 10x Genomics. MULTI-seq libraries were prepared as described previously [8]. For the target-enrichment experiment, transcript enrichment was performed using the Human Immunology Panel (10x Genomics cat#1000246) as described in the "Targeted Gene Expression - Single Cell" user guide from 10x Genomics.

After PBMC library preparation, cDNA and MULTI-seq libraries were pooled and subjected to next-generation sequencing using the following formats. For each 96-well plate of the primary single-cell screen-by-sequencing experiment, cDNA and MULTI-seq libraries were sequenced on two lanes of a NovaSeq S4 flow cell (28x8x0x91) at intended read-depths of 45,000 RPC (cDNA) and 2,000 RPC (MULTI). For the dose-response experiment, cDNA and MULTI-seq libraries were sequenced on two lanes of a NovaSeq S1 flow cell (28x8x0x94) at intended read-depths of 34,000 RPC (cDNA) and 2,000 RPC (MULTI). For the target-enrichment experiment, cDNA and MULTI-seq libraries were sequenced on two lanes of a NovaSeq SP flow cell (151x12x12x151) at intended read-depths of 30,000 RPC (full-transcriptome), 12,500 RPC (target-enriched) and 2,000 RPC (MULTI).

6.5.2 scRNA-seq data pre-processing, MULTI-seq sample classification, and quality-control

scRNA-seq FASTQs from the primary single-cell screen-by-sequencing, dose-response, and target-enrichment experiments were pre-processed using Cellranger (10x Genomics) and aligned to the GRCh38 reference transcriptome. For the primary single-cell screen-by-sequencing experiment, raw Cellranger outputs were subjected to the scVI [59] quality-control and batch-

normalization workflow with default parameters. Cell barcodes passing scVI quality-control were then used for MULTI-seq sample classification by thresholding MULTI-seq barcode count distributions using Otsu's method [60].

For the dose-response and target-enrichment experiments, raw Cellranger outputs were read into R and cell-containing droplets were roughly defined by manual thresholding of the RNA UMI count distribution. These parsed datasets were then normalized using 'SCTransform' [61] prior to unsupervised clustering and dimensionality reduction using PCA and UMAP as implemented in the Seurat R package [62]. Low-quality cells selected via membership in clusters associated with low total RNA UMIs and/or high proportions of mitochondrial gene expression were then removed. Cell barcodes passing this quality-control workflow were then used for MULTI-seq sample classification as described previously [8].

6.5.3 PopAlign hit classification

scRNA-seq data from the primary single-cell screen-by-sequencing experiment pre-processed using the previously-described PopAlign workflow [21]. The impact of drug perturbations was quantified for each major PBMC cell type by computing the similarity to unperturbed controls using the minimum Jeffreys' divergence metric. High-impact perturbations were defined as drugs with FDR-adjusted p-values less than 0.05 in T-cells, B-cells, and/or myeloid cells using the minimum Jeffreys' divergence outputs.

6.5.4 Population response clustering

Population response clustering on T-cells and myeloid sub-types was performed using the following workflow. First, immune cell sub-types were annotated according to enriched expression for these literature-supported marker genes. For T-cells, we used naïve T-cell (IL7R+), TCM (CCR7+ CD27+ CD8B-), TEM (CD4+ IL2RA+), MPEC-1 (CCR7+ CD27+ CD8B+), MPEC-2

(CCR7- CD27+ CD8B+), SLEC-1 (CCL5+ CD52+ GNLY-), SLEC-2 (CCL5+ GZMB+ GNLY+), SLEC-3 (CCL5+ GZMB+ GNLY-), and SLEC-steroid (CCL5+ CD52+ TSC22D3+) annotations. Treg and HDAC-treated T-cell clusters were excluded from this analysis. For myeloid cells, we used pro-inflammatory macrophage (IDO1+ CXCL9+ IL1B +/-), anti-inflammatory macrophage (IDO1+ A2M+ OCSTAMP+ CSF1+), M2c-like macrophage (IDO1+ CD163+), monocyte (IDO1- CD9+ FCN1+), and T-cell/macrophage biological doublet (IDO1+ CD3D+ GZMB+) annotations. DCs and HDAC-treated myeloid clusters were excluded from this analysis. Second, the sub-type proportions of T-cell or myeloid sub-types for each drug were converted into a drug x sub-type frequency matrix. Third, frequency matrices were visualized and clustered using hierarchical clustering as implemented in the ComplexHeatmap R package [63]. Statistically-significant (fold-change > 1.5, FDR < 0.05) shifts in immune sub-type frequencies were determined using the single-cell proportion test [40], which computes p-values using a permutation test and confidence intervals via bootstrapping.

6.5.5 Next-generation sequencing read down-sampling comparative analysis of target-enriched and full-transcriptome scRNA-seq data

To determine the extent to which gene expression signatures associated with immune cell sub-types and drug responses degrade at decreasing sequencing depths in the target-enriched and full-transcriptome scRNA-seq datasets, we performed the following workflow. First, FASTQ reads were randomly down-sampled to the intended proportions, and these data were processed using Cellranger and Seurat as described in section 6.5.2. Cell barcodes and metadata initialized in the original datasets were used for the down-sampled data. Second, differential gene expression analysis was then performed for immune cell sub-types as implemented in the Seurat 'FindAllMarkers' function (logfc.threshold = log(1.5), only.pos = T) and for drug responses relative to CD3/CD28-stimulated DMSO control samples as implemented in the 'FindMarkers' function

(logfc.threshold = log(1.5), only.pos = T). Third, the amount of between-sub-type or between-perturbation information retained at different sequencing depths was computed using the entropy-based ROGUE metric using default parameters [58]. Notably, ROGUE calculations were not successful for comparing T-cell sub-type information preservation in the 2,500 RPC down-sampled full-transcriptome data.

6.6 References

1. Nair NU, Greninger P, Friedman A, Amzallag A, Cortez E, Sahu AD, et al. A landscape of synergistic drug combinations in non-small-cell lung cancer. bioRxiv. 2021. doi: 10.1101/2021.06.03.447011.
2. Lamb J, Crawford ED, Peck DD, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006; 313(5795): 1929-35.
3. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. 2017; 171(6): 1437-52.
4. Fattahi F, Steinbeck JA, Kriks S, Tchieu J, Zimmer B, Kishinevsky S, et al. Deriving human ENS lineages for cell therapy and drug discovery in Hirschsprung disease. Nature. 2016; 531(7592): 105-9.
5. Chen B, Ma L, Paik H, Sirota M, Wei W, Chua M, So S, Butte AJ. Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. Nature Communications. 2017; 8: 16022.
6. Le BL, Andreoletti G, Oskotsky T, Vallejo-Gracia A, Rosales R, Yu K, et al. Transcriptomics-based drug repositioning pipeline identifies therapeutic candidates for COVID-19. Scientific Reports. 2021; 11(1): 12310.
7. Ye C, Ho DJ, Neri M, Yang C, Kulkarni T, Randhawa R, et al. DRUG-seq for miniaturized high-throughput transcriptome profiling in drug discovery. Nature Communcations. 2018; 9(1): 4307.
8. McGinnis CS, Patterson DM, Winkler J, Conrad DN, Hein MY, Srivastavan V, et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nature Methods. 2019; 16: 619-26.

9. Shin D, Lee W, Lee JH, Bang D. Multiplexed single-cell RNA-seq via transient barcoding for simultaneous expression profiling of various drug perturbations. Science Advances. 2019; 5(5): eaav2249.
10. Stoeckius M, Zheng S, Houck-Loomis B, Hao S, Yeung BZ, Mauck WM 3rd, et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biology. 2018; 19(1): 224.
11. Gehring J, Park JH, Chen S, Thomson M, Pachter L. Highly multiplexed single-cell RNA-seq by DNA oligonucleotide tagging of cellular proteins. Nature Biotechnology. 2019; 38(1): 35-8.
12. Gaublomme JT, Li B, McCabe C, Knecht A, Yang Y, Drokhlyanksy E, et al. Nuclei multiplexing with barcoded antibodies for single-nucleus genomics. Nature Communications. 2019; 10(1): 2907.
13. Srivatsan SR, McFaline-Figueroa JL, Ramani V, Saunders L, Cao J, et al. Massively multiplex chemical transcriptomics at single-cell resolution. Science. 2020; 367(6473): 45-51.
14. McFarland JM, Paolella BR, Warren A, Geiger-Schuller K, Shibue T, Rothberg M, et al. Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. Nature Communications. 2020; 11(1): 4296.
15. Zhao W, Dovas A, Spinazzi EF, Levitin HM, Banu MA, Upadhyayula P, et al. Deconvolution of cell type-specific drug responses in human tumor tissue with single-cell RNA-seq. Genome Medicine. 2021; 13(1): 82.
16. Longo DM, Louie B, Putta S, Evensen E, Ptacek J, Cordeiro J, et al. Single-Cell Network Profiling of Peripheral Blood Mononuclear Cells from Healthy Donors Reveals Age- and Race-Associated Differences in Immune Signaling Pathway Activation. Journal of Immunology. 2012; 188(4): 1717-25.
17. Jenny M, Klieber M, Zaknun D, Schroecksnadel S, Kurz K, Ledochowski M, Schennach H, Fuchs D. In vitro testing for anti-inflammatory properties of compounds employing peripheral

blood mononuclear cells freshly isolated from healthy donors. Inflammation Research. 2011; 60(2): 127-35.

18. Vijayakumar S, John SF, Nusbaum RJ, Ferguson MR, Cirillo JD, Olaleye O, Endsley JJ. In vitro model of mycobacteria and HIV-1 co-infection for drug discovery. Tuberculosis. 2013; 93: S66-70.
19. Yilmaz S, Boffito M, Collot-Teixeira S, De Lorenzo F, Waters L, Fletcher C, et al. Investigation of low-dose ritonavir on human peripheral blood mononuclear cells using gene expression whole genome microarrays. Genomics. 2010; 96(1): 57-65.
20. Harrill J, Shah I, Setzer RW, Haggard D, Auerbach S, Judson R, Thomas RS. Considerations for Strategic Use of High-Throughput Transcriptomics Chemical Screening Data in Regulatory Decisions. Current Opinions in Toxicology. 2019; 15: 64-75.
21. Chen S, Rivaud P, Park JH, Tsou T, Charles E, Haliburton JR, et al. Dissecting heterogeneous cell populations across drug and disease conditions with PopAlign. PNAS. 2020; 117(46): 28784-94.
22. Tsuda K, Yamanaka K, Kitagawa H, Akeda T, Naka M, Niwa K, et al. Calcineurin inhibitors suppress cytokine production from memory T cells and differentiation of naïve T cells into cytokine-producing mature T cells. PLoS One. 7(2): e31465.
23. Petes C, Odoardi N, Gee K. The Toll for Trafficking: Toll-Like Receptor 7 Delivery to the Endosome. Frontiers in Immunology. 2017; 8: 1075.
24. Chen Y, Zander R, Khatun A, Schauder DM, Cui W. Transcriptional and Epigenetic Regulation of Effector and Memory CD8 T Cell Differentiation. Frontiers in Immunology. 2018; 9: 2826.
25. Lawlor N, Nehar-Belaid D, Grassmann JDS, Stoeckius M, Smibert P, Stitzel ML, et al. Single Cell Analysis of Blood Mononuclear Cells Stimulated Through Either LPS or Anti-CD3 and Anti-CD28. Frontiers in Immunology. 2021; 12: 636720.

26. Breen MS, Bierer LM, Daskalakis NP, Bader HN, Makotkine I, Chattopadhyay M, et al. Differential transcriptional response following glucocorticoid activation in cultured blood immune cells: a novel approach to PTSD biomarker development. Translational Psychiatry. 2019; 9(1): 201.
27. Horns F, Dekker CL, Quake SR. Memory B Cell Activation, Broad Anti-influenza Antibodies, and Bystander Activation Revealed by Single-Cell Transcriptomics. Cell Reports. 2020; 30(3): 905-13.
28. Wen W, Su W, Tang H, Le W, Zhang X, Zheng Y, et al. Immune cell profiling of COVID-19 patients in the recovery stage by single-cell sequencing. Cell Discovery. 2020; 6: 31.
29. Vigorito E, Perks KL, Abreu-Goodger C, Bunting S, Xiang Z, Kohlhaas S, et al. microRNA-155 Regulates the Generation of Immunoglobulin Class-Switched Plasma Cells. Immunity. 2007; 27(6): 847-59.
30. Wang X, Wang H, Wang H, Zhang F, Wang K, Guo Q, et al. The role of indoleamine 2,3-dioxygenase (IDO) in immune tolerance: focus on macrophage polarization of THP-1 cells. Cellular Immunology. 2014; 289(1-2): 42-8.
31. Mantovani A, Sica A, Sozzani S, Allavena P, Vecchi A, Locati M. The chemokine system in diverse forms of macrophage activation and polarization. Trends in Immunology. 2004; 25(12): 677-86.
32. Martinez FO, Gordon S. The M1 and M2 paradigm of macrophage activation: time for reassessment. F1000Prime Reports. 2014; 6: 13.
33. Kapellos TS, Bonaguro L, Gemünd I, Reusch N, Saglam A, Hinkley ER, Schultze JL. Human Monocyte Subsets and Phenotypes in Major Chronic Inflammatory Diseases. Frontiers in Immunology. 2019; 10: 2035.

34. Halpern KB, Shenhav R, Massalha H, Toth B, Egozi A, Massasa EE, et al. Paired-cell sequencing enables spatial gene expression mapping of liver endothelial cells. Nature Biotechnology. 2018; 36(10): 962-70.
35. Jagger AL, Evans HG, Walter GJ, Gullick NJ, Menon B, Ballantine LE, et al. FAS/FAS-L dependent killing of activated human monocytes and macrophages by CD4+CD25- responder T cells, but not CD4+CD25+ regulatory T cells. Journal of Autoimmunity. 2012; 38(1): 29-38.
36. Kaplan MJ, Ray D, Mo RR, Yung RL, Richardson BC. TRAIL (Apo2 ligand) and TWEAK (Apo3 ligand) mediate CD4+ T cell killing of antigen-presenting macrophages. Journal of Immunology. 2000; 164(6): 2897-904.
37. Martinet W, Coornaert I, Puylaert P, De Meyer GRY. Macrophage Death as a Pharmacological Target in Atherosclerosis. Frontiers in Pharmacology. 2019; 10: 306.
38. Van Laethem F, Baus E, Smyth LA, Andris F, Bex F, Urbain J, et al. Glucocorticoids Attenuate T Cell Receptor Signaling. Journal of Experimental Medicine. 2001; 193(7): 803-14.
39. Davis TE, Kis-Toth K, Szanto A, Tsokos GC. Glucocorticoids suppress T cell function by upregulating microRNA 98. Arthritis & Rheumatology. 2013; 65(7): 1882-90.
40. Policastro R, Miller SA. Single Cell Proportion Test. GitHub Repository. 2020. https://github.com/rpolicastro/scProportionTest.
41. Chi H. Regulation and function of mTOR signalling in T cell fate decision. Nature Reviews Immunology. 2012; 12(5): 325-38.
42. Yajnanarayana SP, Stübig T, Cornez I, Alchalby H, Schönberg K, Rudolph J, et al. JAK1/2 inhibition impairs T cell function in vitro and in patients with myeloproliferative neoplasms. British Journal of Haematology. 2015; 169(6): 824-33.
43. Acharya S, Timilshina M, Jiang L, Neupane S, Choi D, Park SW, et al. Amelioration of Experimental autoimmune encephalomyelitis and DSS induced colitis by NTG-A-009 through the inhibition of Th1 and Th17 cells differentiation. Scientific Reports. 2018; 8(1): 7799.

44. Ubieta K, Thomas MJ, Wollin L. The Effect of Nintedanib on T-Cell Activation, Subsets and Functions. Drug Design, Development and Therapy. 2021; 15: 997-1011.
45. Schade AE, Schieven GL, Townsend R, Jackowska AM, Susulic V, Zhang R, et al. Dasatinib, a small-molecule protein tyrosine kinase inhibitor, inhibits T-cell activation and proliferation. Blood. 2008; 111(3): 1366-77.
46. Rodell CB, Arlauckas SP, Cuccarese MF, Garris CS, Li R, Ahmed MS, et al. TLR7/8-agonist-loaded nanoparticles promote the polarization of tumour-associated macrophages to enhance cancer immunotherapy. Nature Biomedical Engineering. 2018; 2(8): 578-88.
47. Lee H, Basso IN, Kim DDH. Target spectrum of the BCR-ABL tyrosine kinase inhibitors in chronic myeloid leukemia. International Journal of Hematology. 2021; 113(5): 632-41.
48. Davis MI, Hunt JP, Herrgard S, Ciceri P, Wodicka LM, Pallares G, et al. Comprehensive analysis of kinase inhibitor selectivity. Nature Biotechnology. 2011; 29(11): 1046-51.
49. Xia S, Liu X, Cao X, Xu S. T-cell expression of Bruton's tyrosine kinase promotes autoreactive T-cell activation and exacerbates aplastic anemia. Cellular & Molecular Immunology. 2020; 17(10): 1042-52.
50. Gabhann JN, Hams E, Smith S, Wynne C, Byrne JC, Brennan K, et al. Btk Regulates Macrophage Polarization in Response to Lipopolysaccharide. PLoS One. 2014; 9(1): e85834.
51. Zarrin AA, Bao K, Lupardus P, Vucic D. Kinase inhibition in autoimmunity and inflammation. Nature Reviews Drug Discovery. 2021; 20(1): 39-63.
52. Sun Z, Jiang Q, Li J, Guo J. The potent roles of salt-inducible kinases (SIKs) in metabolic homeostasis and tumorigenesis. Signal Transduction and Targeted Therapy. 2020; 5(1): 150.
53. Sundberg TB, Choi HG, Song J, Russell CN, Hussain MM, Graham DB, et al. Small-molecule screening identifies inhibition of salt-inducible kinases as a therapeutic strategy to enhance immunoregulatory functions of dendritic cells. PNAS. 2014; 111(34): 12468-73.

54. Vallejo AF, Davies J, Grover A, Tsai C, Jepras R, Polak ME, West J. Resolving cellular systems by ultra-sensitive and economical single-cell transcriptome filtering. iScience. 2021; 24(3): 102147.
55. Saikia M, Burnham P, Keshavjee SH, Wang MFZ, Heyang M, Moral-Lopez P, et al. Simultaneous multiplexed amplicon sequencing and transcriptome profiling in single cells. Nature Methods. 2019; 16(1): 59-62.
56. Cleary B, Cong L, Cheung A, Lander ES, Regev A. Efficient Generation of Transcriptomic Profiles by Random Composite Measurements. Cell. 2017; 171)6: 1424-36.e18.
57. Replogle JM, Norman TM, Xu A, Hussmann JA, Chen J, Cogan JZ, et al. Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. Nature Biotechnology. 2020; 38(8): 954-61.
58. Liu B, Li C, Li Z, Wang D, Ren X, Zhang Z. An entropy-based metric for assessing the purity of single cell populations. Nature Communications. 2020; 11(1): 3155.
59. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nature Methods. 2018; 15(12): 1053-8.
60. Otsu N. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics. 1979; 9(1): 62-6.
61. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology. 2018; 36: 411-20.
62. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biology. 2019; 23: 296.
63. Gu Z, Ellis R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016; 32(18): 2847-9.

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:
Chris McGinnis
0E0C32168B2D40A...
Author Signature

8/26/2021
Date