

Dissecting Variant Effects with Multiplexed Multi-omics in Health and Disease.

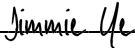
by  
Mary Grace Gordon

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

in  
Biological and Medical Informatics

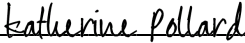
in the  
GRADUATE DIVISION  
of the  
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:  
  
755C0380968F429... Jimmie Ye  
Chair

DocuSigned by:  
  
Nadav Ahituv

DocuSigned by:  
  
David Erle

DocuSigned by:  
  
9B0DDB91029D4F1... Katherine Pollard

---

Committee Members

Copyright 2022

By

Mary Grace Gordon

## **DEDICATION**

I dedicate this work to my steadfast life partner, John Tapp. You moved across the country with me at the beginning of this journey and have been incredibly supportive every step of the way. You have kept me sane throughout this process by making sure I had some semblance of work life balance, from forcing me take naps to vacations and ensuring I didn't survive off frozen food and canned soup.

## ACKNOWLEDGEMENTS

They say it takes a village to raise a child, but I've found this also applies to training PhD students. I feel incredibly fortunate to have had an amazing group of people supporting me throughout this journey. First, I would like to thank my PhD advisors Jimmie Ye and Nadav Ahituv. I am so fortunate to have joined two incredible labs. You have both helped me to develop into an independent scientist. You both entrusted me with so much responsibility from the start, you believed in my scientific abilities before I believed in myself. Thank you to Jimmie for really pushing me to be extraordinarily productive, your excitement for science is infectious. Thank you for always being willing to hop on a zoom call to help answer questions, even if you were enjoying a vacation in Mexico or sipping wine in Sonoma. Thank you Nadav for being so supportive over my PhD. For always being willing to stay late for a meeting, like when you helped me practice for my qualifying exam. You have been incredibly supportive and have always known exactly what to say to make me feel less stressed.

Thank you to my thesis committee members, Dr. Katie Pollard and Dr. David Erle, for all your input and support along the way. You both have provided valuable feedback throughout my PhD and helped me navigate some of the challenges around co-mentorship. Katie, I admire you as a scientist and it's been especially helpful to have a woman role model who handles balancing work and life so well. David, you are an amazing scientist and I want to thank you for always asking thoughtful and challenging questions and pushing my science. Also, thank you for bringing me into the CoLabs community, where I've been able to get amazing feedback from a community of immunologists.

During my PhD I have had the fortune to collaborate with many amazing scientists. First, I'd like to thank all past and current members of the Ye and Ahituv labs for your support, both personally and scientifically. I want to thank Dr. Yang Sun and Dr. Alyssa Ward, you are both amazing

scientists and organizers, I feel so fortunate to have had the chance to work with you both so closely. Thank you to Dr. Meena Subramaniam and Dr. Rachel Gate for setting such great examples for me and helping me get started in the lab. Thank you to Mincheol Kim for always being willing to chat about stats with me. Thank you to Dr. Ryan Ziffra for suffering through long experiments with me and commiserating over failed experiments. To Lindsay Liang, I wouldn't want to debug pipelines with anyone else, I'm fortunate to call you a friend.

Thank you to my friends, both near and far. To Dr. Beatrice Ary, who convinced me to apply to UCSF and has been incredibly supportive, you're like the sister I never had. To Dr. Mackenzie and Branden DuPont, I'm glad you came to SF despite not wanting to move here, I felt fortunate to have friends that feel more like family living so close for a while. To my friends from Davidson College who have almost all managed to find the time to visit SF during my PhD. Thank you all for reminding me there's life outside of lab and for always humoring me trying to explain my research and celebrating my scientific accomplishments (Lexie Winograd, Molly Marshall, Kari Sickles, Taylor MacDonald, Timmy Basista, Marcus Bailey, Dr. Emily Kneble, Margaret Kaufmann, Caroline Queen). To the CCB ladies who welcomed me into their friend group despite my lack of chemistry knowledge: Dr. Kaitlyn Tsai, Dr. Taylor Arhar, Dr. Susanna Elledge, and Dr. Lisa Kirkemo. To my iPQB cohort.

I'd like to also thank the amazing mentors I've had along the way who helped me get to graduate school in the first place. Thank you to Pamela Floodman and Dr. Moyra Smith for taking a chance on a clueless undergraduate student. Thank you to my mentors and Professors at Davidson College. Thank you to Dr. Karen Hales for letting me work your lab and introducing me to cellular and molecular biology and to my undergraduate advisor, Dr. Malcolm Campbell, for encouraging me to seriously explore research as a career path and for introducing me to the field of genomics. Special thanks to Dr. Thomas Markello, I wouldn't have considered doing a PhD without your

encouragement. Your love of learning is contagious, and you are incredibly generous with your time, whether teaching me about a genetic disorder, giving a historic tour of NIH, or jamming out to some Jonathan Colton- you made doing science full time a blast.

Last, I would like to thank my immediate and extended family. Thanks especially to my Mom and Dad (Carrie and Alan Gordon) for their support throughout my PhD, but also through my entire education. You made sure I understood the importance of education from a young age and have been intentional about providing for me so I could have the opportunity to pursue higher education, though I think I might have surprised you (and myself) by choosing to pursue a Doctorate. I also want to thank my late grandfather Stan Gordon, I didn't fully appreciate all the work you did behind the scenes to support my education, I know you'd be so proud of this accomplishment. I'd also like to thank my future in-laws Eveline and RJ Tapp, who have welcomed me into their family as if they were my own parents. You have provided continuous encouragement and I am so fortunate to have had you both living in SF while I've worked on my degree. Of course, I really owe this degree in part to John Tapp, my future husband. I'm not sure I would have made it here without you by my side.

## CONTRIBUTIONS

This thesis contains published and unpublished work resulting from research conducted by M. Grace Gordon at the University of California, San Francisco in pursuit of her PhD. Each publication is analogous to a standard thesis chapter. While these efforts were led by M. Grace Gordon, co-authors of both publications contributed significantly to these works in discussions, feedback, writing and analyses. Additional contributions from collaborators are outlined in detail below.

Chapter 2 of this thesis is a reprint of work published in *Science*<sup>1</sup>, overseen by Dr. Chun J. Ye. All authors of this publication were critical to the completion of this work. Detailed contributions are included in the chapter 2 acknowledgements, section 2.7.

Chapter 3 of this thesis. This work was overseen by Dr. Chun J. Ye in collaboration with Dr. Alyssa Ward, Dr. Yang Sun, Lindsay Liang, Taibo Li, Pooja Kathail, Mincheol Kim, Divyashree Kushnoor, Tara Teed, Dr. Phil De Jagar, Dr. Stephan Sanders, and Dr. Alexis Battle.

Chapter 4 of this thesis is a reprint of materials published in *Nature Protocols*<sup>2</sup>, overseen by Dr. Nadav Ahituv. All co-authors were instrumental in the completion of this work. Detailed contributions are included in the chapter 4 acknowledgements, section 4.10.

Chapter 5 of this thesis was co-supervised by Drs. Nadav Ahituv and Chun J. Ye. This work was completed in collaborations with Dr. Ryan Ziffra, and Dr. Taka Inoue.

1. Perez, R. K. *et al.* Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. *Science* **376**, eabf1970 (2022).
2. Gordon, M. G. *et al.* lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. *Nature Protocols* vol. 15 2387–2412 (2020).

## Dissecting Variant Effects with Multiplexed Multi-omics in Health and Disease.

M. Grace Gordon

### Abstract

**Background:** Recent advancements in Next Generation Sequencing (NGS) and high throughput multiplexed single-cell multi-omics present an unprecedented opportunity to investigate cell composition, gene expression, map associations between single nucleotide polymorphisms and quantitative molecular traits in heterogenous cell populations across hundreds of donors and perform context relevant functional validation studies. Here we present work (i) using multiplexed single-cell RNA sequencing (scRNA-seq) to study cellular and genetic correlates of systemic lupus erythematosus (SLE), (ii) generating an atlas of healthy immune cells from a diverse cohort, using multiplexed multi-omic profiling of RNA and open chromatin. (iii) We streamlined lentiviral Massively Parallel Reporter Assay (lentiMPRA) protocols and established a computational pipeline to support these assays and (iiii) worked to extend this method to single cells (scMPRA).

**Methods:** (i) We profiled 1.2 million cells using multiplexed scRNA-seq (mux-seq) from 261 donors. We evaluated changes in cell composition and gene expression between cases and controls. We mapped cell type specific expression quantitative trait loci. (ii) We profiled over 1 million cells using multiplexed multi-omics from ~400 donors. We evaluated cell composition across a diverse cohort, mapped immune regulatory networks, and investigated genetic architecture of chromatin accessibility and gene expression. (iii) We established a robust protocol for lentiMPRA and developed pipeline using Nextflow for seamless data analysis to functionally validate putative non-coding regulatory sequences. (iv) We have designed an approach for scMPRA and have generated promising preliminary results. **Results and Conclusions:** (i) Our mux-seq platform is robust and scalable (ii) and can be applied to multi-omic datasets. (iii) Genetic associations can be validated using lentiMPRA with our streamlined methods. (iv) scMPRA holds promise as an option for validating genetic associations in distinct cellular environments.



# Table of Contents

<b>CHAPTER 1: INTRODUCTION .....</b>	<b>1</b>
1.1 REFERENCES .....	5
<b>CHAPTER 2: SINGLE-CELL RNA-SEQ REVEALS CELL TYPE–SPECIFIC MOLECULAR AND GENETIC ASSOCIATIONS TO LUPUS. ....</b>	<b>7</b>
2.1 ABSTRACT .....	7
2.2 INTRODUCTION .....	7
2.3 RESULTS.....	8
2.3.1 <i>A census of circulating immune cells in SLE.....</i>	<i>8</i>
2.3.2 <i>Compositional analysis reveals CD4<sup>+</sup> T cell lymphopenia in SLE.....</i>	<i>9</i>
2.3.3 <i>Decrease of circulating naïve CD4<sup>+</sup> T cells in SLE.....</i>	<i>11</i>
2.3.4 <i>Clonal expansion of cytotoxic GZMH<sup>+</sup> T cells in SLE.....</i>	<i>12</i>
2.3.5 <i>Expression changes across 11 peripheral immune cell types in SLE.....</i>	<i>14</i>
2.3.6 <i>Pronounced type-1 interferon response in classical monocytes. ....</i>	<i>15</i>
2.3.7 <i>Expression modules predict CD4<sub>naïve</sub> lymphopenia, disease status, and stratifies SLE patients.....</i>	<i>16</i>
2.3.8 <i>Identification of cell-type-specific cis-eQTLs across eight immune cell types. ....</i>	<i>17</i>
2.3.9 <i>Identification and annotation of cell-type-specific SLE-associated loci. ....</i>	<i>18</i>
2.3.10 <i>Modification of genetic effects on gene expression by interferon activation. ....</i>	<i>20</i>
2.4 DISCUSSION.....	21
2.5 METHODS SUMMARY .....	25
2.6 SUPPLEMENTARY MATERIALS.....	27
2.7 REFERENCES AND NOTES.....	28
2.8 ACKNOWLEDGMENTS: .....	36
<b>CHAPTER 3: THE IMMUNE CELL CENSUS: MULTIPLEXED MULTI-OMICS ENABLES DISCOVERY OF IMMUNE REGULATORY PROGRAMS AND GENETIC ARCHITECTURE OF MOLECULAR TRAITS. ....</b>	<b>49</b>

3.1 ABSTRACT .....	49
3.2 INTRODUCTION .....	49
3.3 RESULTS.....	51
3.3.1 <i>Cell phenotyping of more than one million single cells.</i> .....	51
3.3.2 <i>Cell composition</i> .....	51
3.3.3 <i>Networks</i> .....	52
3.3.4 <i>Genetics</i> .....	53
3.4 DISCUSSION.....	53
3.5 METHODS .....	54
3.5.1 <i>Sample collection</i> .....	54
3.5.2 <i>Genotyping</i> .....	54
3.5.3 <i>Single Cell Sequencing</i> .....	55
3.5.4 <i>Single Cell Analysis</i> .....	55
3.5.5 <i>QTL calling</i> .....	56
3.6 REFERENCES .....	57

**CHAPTER 4: LENTIMPRA & MPRAFLOW FOR HIGH-THROUGHPUT FUNCTIONAL CHARACTERIZATION OF GENE**

<b>REGULATORY ELEMENTS. ....</b>	<b>63</b>
4.1 ABSTRACT .....	63
4.2 INTRODUCTION .....	63
4.3 DEVELOPMENT OF THE PROTOCOL .....	65
4.4 APPLICATIONS OF THE METHOD .....	66
4.5 COMPARISONS WITH OTHER METHODS .....	67
4.6 EXPERIMENTAL DESIGN.....	68
4.6.1 <i>Library design</i> .....	68
4.6.2 <i>Library generation</i> .....	69
4.6.3 <i>Association sequencing</i> .....	70

4.6.4 Lentiviral prep.....	70
4.6.5 Infection and sequencing.....	70
4.6.6 Data processing.....	71
4.6.7 Necessary Expertise.....	72
4.6.8 Limitations.....	72
4.7 ANTICIPATED RESULTS.....	72
4.8 AUTHOR CONTRIBUTIONS STATEMENTS.....	78
4.9 ACKNOWLEDGMENTS.....	78
4.10 REFERENCES.....	79
<b>CHAPTER 5: DEVELOPING SCMPRA TO DISSECT GENE BY ENVIRONMENT INTERACTIONS. ....</b>	<b>86</b>
5.1 ABSTRACT.....	86
5.2 INTRODUCTION.....	86
5.3 METHODS.....	86
5.3.1 Molecular Biology Approach.....	89
5.3.2 Mathematical calibration of scMPRA.....	89
5.4 RESULTS.....	90
5.5 CONCLUSIONS.....	91
5.6 REFERENCES.....	96

## List of Figures

Figure 2.1: Changes in the composition of circulating immune cells in SLE. ....	39
Figure 2.2: Reduction of naïve CD4+ and expansion of cytotoxic CD8+ T cells in SLE.....	41
Figure 2.3: Type-1 interferon response of myeloid cells in SLE .....	43
Figure 2.4: Prediction of disease status and molecular stratification of SLE.....	45
Figure 2.5: Cell-type-specific genetic determinants of gene expression .....	47
Figure 2.6: Interferon modifies cell-type-specific genetic effects on gene expression.....	48
Figure 3.1: Experimental design, cell phenotyping, and composition. ....	60
Figure 3.2: Evaluating Networks of Genes.....	61
Figure 3.3: Molecular trait genetics. Mann-Whitney U test for enrichment of eQTLs .....	62
Figure. 4.1: Schematics of lentiMPRA.....	74
Figure. 4.2: Overview of MPRAflow association utility .....	75
Figure. 4.3: Overview of count utility .....	76
Figure 4.4: Overview of Saturation Mutagenesis Utility .....	77
Figure 5.1: scMPRA design.....	93
Figure 5.2: LCL and HepG2 Feature Barcode pilot results.....	94
Figure 5.3: K562 and HepG2 5' amplicon pilot results.....	95

## List of Tables

Table 4.1: Association Utility options. Blue rows are mandatory and orange are optional. ....	82
Table 4.2: Count utility options. Blue rows are mandatory and orange are optional.....	83
Table 4.3: Saturation Mutagenesis utility options. Blue rows are mandatory and orange are optional.....	83
Table 4.4: Troubleshooting table.....	84

## Chapter 1: Introduction

Deoxyribonucleic acid (DNA) encodes the blueprint for life, where composition of a DNA sequence dictates many aspects of the resulting organism. Understanding the relationship between genotypes and phenotypes and how environmental factors modulate these associations has broad implications for human health<sup>1</sup>. Past research has resulted in the discovery of genetic origins of diseases, enabling increased understanding of mechanisms of disease resulting in more effective therapies<sup>2</sup>. These relationships can be explored through observational studies, where natural genetic variation is observed in large cohorts of individuals, or through experimental studies, where the effect of genetic perturbations can be observed *in vitro* or *in vivo*.

The most common observational studies for complex traits are Genome Wide Association Studies (GWAS), where linear models are used to test for relationships between a single nucleotide polymorphism (SNP) a complex trait (ex. Disease status, Height, etc.)<sup>3,4</sup>. Historically these studies have been conducted for traits measured at the organism level, but recent advancements in next generation sequencing technologies have provided the opportunity to investigate genetic relationships with molecular traits, such as gene expression or chromatin state<sup>5-7</sup>. Investigating genetic associations to functional traits not only furthers understanding of the genetic architecture of these traits but can also be leveraged to functionally annotate GWAS to gain a better understanding of molecular underpinnings of these associations<sup>8</sup>.

In the past, these methods have been conducted in bulk tissue, where a biological sample of interest obtained for each donor, processed using RNA-seq or ATAC-seq and in parallel a second sample from the donor is genotyped. While these studies have resulted in significant findings, these approaches are labor intensive, as each donor must be prepared as a separate sample. These methods are particularly ill-suited for heterogenous tissues such as peripheral mononuclear blood cells (PBMCs), where many unique cell types would need to be sorted and

processed separately to avoid loss of cell type specific signals, making these studies only possible as large consortium efforts<sup>9,10</sup>.

Over the last decade significant strides in single cell sequencing technologies have been achieved, enabling orders of magnitude increases in cell throughput as well as increasing the numbers of modalities that can be assayed simultaneously in the same single cells. Two developments are of note: droplet based single cell sequencing and genetic multiplexing<sup>11-13</sup>. Droplet based single cell sequencing uses microfluidics to encapsulate a single cell in a droplet where its transcriptome is captured and uniquely barcoded. However, to ensure only one cell is captured in each droplet, cells are Poisson loaded leaving many reagents unused. To combat this waste and reduce the cost of single cell sequencing, mux-seq was developed. In short, this technique pools cells from genetically distinct donors, allowing for approximately 5 times the number of cells to be recovered because droplets containing cells from two distinct donors can be removed while the remaining cells can be assigned back to their donor of origin based on the combination of SNPs observed in the transcriptome. These advances have provided the unique opportunity to apply single cell methods to hundreds of donors to better understand cell composition, gene expression, and chromatin accessibility at single cell resolution and to find genetic associations with these traits in heterogenous mixtures of primary cells.

We first applied these technologies to study Systemic lupus erythematosus (SLE), a highly heterogenous autoimmune disease with multiorgan manifestations ranging from mild to life-threatening in severity. Previous studies have identified hallmarks of SLE. Flow cytometry experiments have identified a decrease in lymphocytes, lymphopenia, in SLE cases compared to controls<sup>14</sup>. Bulk transcriptomic analyses of PBMCs have identified expression signatures such as increased expression of interferon stimulated genes (ISGs)<sup>15</sup>. However, these previous methods are limited to evaluating cell composition with a limited set of surface markers or profiling bulk

tissues, which mask cell type specific signals and are confounded by unequal distributions of immune cell types in blood. These limitations make it difficult to annotate the more than 100 loci associated with SLE<sup>16</sup>. We used multiplexed single-cell RNA-seq (mux-seq) to profile over 1.2 million PBMCs from 165 SLE cases and 99 healthy controls. In this work, described in chapter 2, we evaluated composition and cell type specific expression differences between cases and controls. We mapped cell-type-specific cis-eQTLs and used these associations to functionally annotate previous SLE associated loci.

Having demonstrated the scalability of mux-seq, we sought to generate a large-scale healthy reference dataset including samples from a diverse set of donors in a study we call the Immune Cell Census. We profiled over 1 million PBMCs from approximately 400 donors, simultaneously sampling gene expression and chromatin accessibility. We first investigated changes in cell composition. We then dissected gene regulation. Lastly, we investigate genetic architecture of chromatin accessibility and gene expression. This work is described in chapter 3.

While observational studies are powerful tools for nominating associations between SNPs and phenotypes, functional studies must be performed to validate these statistical associations. Popular functional studies involve directly perturbing the DNA of cells and reading out their effects on a molecular measurement<sup>17</sup>. However, while there are CRISPR based technologies to knock out, activate, or inhibit targeted regions of the genome, base editing technologies have not reached maturity yet to be able to assay specific SNP changes and how they influence candidate regulatory elements (CREs). Massively Parallel Reporter Assays (MPRAs) have dramatically increased the throughput of these validation efforts by performing these experiments in a pooled manner, where activity of individual CREs and variants in those CREs can be quantified using a DNA barcode<sup>18</sup>. Despite these advancements, limited computational support exists for these analyses. To address these concerns, we developed lentiMPRA and MPRAflow, a portable



scalable pipeline to process raw MPRA data to quantify measurements of expression for CRS and variants in CRS, which is discussed in chapter 4. Lastly, we have worked to develop a single cell MPRA (scMPRA) to address dissecting variant effects in heterogenous cell types and cell states. Progress towards this goal is discussed in chapter 5.

## 1.1 References

1. Nussbaum, R. L., McInnes, R. R. & Willard, H. F. *Thompson & Thompson genetics in medicine, 8e.* (Elsevier, 2015).
2. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).
3. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
4. Visscher, P. M. *et al.* 10 years of GWAS discovery: Biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
5. Storey, J. D. *et al.* Gene-expression variation within and among human populations. *Am. J. Hum. Genet.* **80**, 502–509 (2007).
6. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
7. Kasowski, M. *et al.* Extensive variation in chromatin states across humans. *Science* **342**, 750–752 (2013).
8. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
9. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

10. Vősa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).
11. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
12. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
13. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
14. Banchereau, J. & Pascual, V. Type I interferon in systemic lupus erythematosus and other autoimmune diseases. *Immunity* **25**, 383–392 (2006).
15. Banchereau, R. *et al.* Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients. *Cell* **165**, 551–565 (2016).
16. Bentham, J. *et al.* Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat. Genet.* **47**, 1457–1464 (2015).
17. Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853–1866.e17 (2016).
18. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**, 159–164 (2015).

## **Chapter 2: Single-cell RNA-seq reveals cell type–specific molecular and genetic associations to lupus.**

### **2.1 Abstract**

Systemic lupus erythematosus (SLE) is a heterogeneous autoimmune disease. Knowledge of circulating immune cell types and states associated with SLE remains incomplete. We profiled over 1.2 million PBMCs (162 cases, 99 controls) with multiplexed single-cell RNA-sequencing (mux-seq). Cases exhibited elevated expression of type-1 interferon-stimulated genes (ISG) in monocytes, reduction of naïve CD4<sup>+</sup> T cells that correlated with monocyte ISG expression, and expansion of repertoire-restricted cytotoxic *GZMH*<sup>+</sup> CD8<sup>+</sup> T cells. Cell-type-specific expression features predicted case-control status and stratified patients into two molecular subtypes. We integrated dense genotyping data to map cell-type-specific *cis*-eQTLs and link SLE-associated variants to cell-type-specific expression. These results demonstrate mux-seq as a systematic approach to characterize cellular composition, identify transcriptional signatures, and annotate genetic variants associated with SLE.

### **2.2 Introduction**

Systemic lupus erythematosus (SLE) is a heterogeneous autoimmune disease affecting multiple organ systems, with elevated prevalence in women (1) and individuals of Asian, African, and Hispanic ancestries (2). Bulk transcriptomic profiling has implicated increased type-1 interferon signaling, dysregulated lymphocyte activation, and failure of apoptotic clearance as hallmarks of disease (3). Many genes participating in these immunological processes are proximal to the ~100 known genetic variants associated with SLE (4). Despite this progress, a comprehensive census of circulating immune cells in SLE remains incomplete and annotating the cell types and cell contexts mediating genetic associations remains challenging.

Historically, different approaches have been used to characterize the role of circulating immune cells in SLE. Flow cytometry analyses, quantifying composition based on known cell surface

markers, reported B and T cell lymphopenia (5). Bulk transcriptomic analyses of peripheral blood mononuclear cells (PBMCs) universally found elevated expression of interferon-stimulated genes (ISGs) and molecularly stratified patients based on expression features (3, 6). However, flow cytometry is biased by its use of a limited set of markers, while bulk transcriptomic profiling is underpowered to detect cell-type-specific expression differences. Bulk transcriptomic analysis of sorted cell populations can identify cell-type-specific expression signatures in SLE (7). However, it does not capture cell type frequencies, obscures heterogeneity within sorted populations, and is challenging to scale to well-powered cohorts for detecting subtle disease-associated differences in gene expression.

Single-cell RNA-sequencing (scRNA-seq) of PBMCs holds potential as a comprehensive and unbiased approach to simultaneously profile the composition and cell-type-specific transcriptional states of circulating immune cells. When integrated with dense genotyping data, there are further opportunities to fine map disease-associated variants and identify the cell types and states where they exert their effects. Despite its potential, application of scRNA-seq to population cohorts has been limited by low sample throughput, high cost, and susceptibility to technical variability. To overcome these limitations, we previously developed multiplexed scRNA-seq (mux-seq) to enable systematic and cost-effective scRNA-seq of population cohorts (8).

## **2.3 Results**

### **2.3.1 A census of circulating immune cells in SLE.**

We used mux-seq (8) to profile over 1.2 million PBMCs from 264 unique samples obtained from the California Lupus Epidemiology Study (CLUES) (9) and the ImmVar Consortium (10–12). The 264 samples corresponded to 162 SLE cases including 19 disease flare cases and 10 matched samples post flare treatment, and 99 healthy controls (**Fig. S1A**). Most samples were from women of European or Asian ancestry. The 264 samples and 91 replicates were profiled in 23 pools

across four batches (**Fig. S1B**). Surface protein expression for cells from processing batches three (155,034 cells) and four (375,261 cells) were also profiled using 16 and 99 DNA-conjugated antibodies respectively. 1,444,450 cells remained after quality control and doublet removal using freemuxlet (8) (mean doublet rate 22.12%, **Fig. S1C**). Following additional removal of doublets using Scrublet (13) (67,969 droplets), contaminating platelets, and red blood cells (112,805 cells), 1,263,676 cells remained in the final dataset (**Fig. S1C**). Genotype based sample demultiplexing resulted in an average of 3,560 singlets (standard deviation: 1,103) assigned to each sample (**Fig. S1D**).

### **2.3.2 Compositional analysis reveals CD4<sup>+</sup> T cell lymphopenia in SLE.**

Louvain clustering (14) of normalized and batch corrected single-cell transcriptomic profiles identified 23 clusters which were assigned to 11 cell types: CD14<sup>+</sup> classical and CD16<sup>+</sup> non-classical monocytes (cM and ncM); conventional and plasmacytoid dendritic cells (cDC and pDC); CD4<sup>+</sup> and CD8<sup>+</sup> T cells (CD4 and CD8); natural killer cells (NK); B cells (B); plasmablasts (PB); proliferating T and NK cells (Prolif); and progenitor cells (Progen) (**Fig. S2A**). Regions of the Uniform Manifold Approximation and Projection (UMAP) (15) were occupied by cells of different cell types (**Fig. 1A**), and to a lesser extent, different case-control status and ethnicity (**Figs. 1B** and **S2B**). Pool and processing batch had no observable effects on the distribution of cells (**Fig. S2C, D**).

We first assessed changes in cellular composition in SLE by comparing the frequencies of 11 cell types between cases and controls of Asian and European ancestry separately. Cell type percentage estimates from mux-seq were reproducible between biological replicates (Median Pearson  $R_{\text{cases}}=0.79$  and  $R_{\text{controls}}=0.85$ ) (**Fig. S2E**) and correlated with estimates obtained from surface protein profiling for batch four (Median Spearman  $R=0.88$ ). Compared with controls, cases were most notably marked by a decrease in CD4 percentage (Weighted Least Squares

(WLS); Asian: -20.4%; European: -10.0%; Fisher's method  $P_{\text{meta:Fisher}} < 5.58 \times 10^{-16}$ ), an increase in cM (Asian: +11.9%; European: +8.8%;  $P_{\text{meta:Fisher}} < 9.75 \times 10^{-7}$ ) and Prolif percentages (Asian: +0.55%; European: +0.38%;  $P_{\text{meta:Fisher}} < 1.93 \times 10^{-3}$ ; **Fig. 1C; Table S1**). While most changes were correlated between ethnicities (Pearson  $R=0.97$ ), Asian cases were marked by a greater reduction in CD4 percentage ( $\text{Log}_2\text{FC}=-0.36$ ,  $P_{\text{WLS}} < 5.60 \times 10^{-5}$ ; **Fig. 1D**). Cases not receiving therapy (N=21) exhibited similar changes in composition compared with cases receiving therapy (Pearson  $R_{\text{Asian}}=0.89$  and  $R_{\text{European}}=0.92$ ; **Fig. S2H**). Compared with cases not receiving oral steroids (OS; N=78), cases treated with OS (N=82) exhibited an increase in CD8 percentage (Asian: +5.2%, European: +3.9%;  $P_{\text{meta:Fisher}} < 4.23 \times 10^{-3}$ ) and a decrease in ncM percentage (Asian: -1.3%, European: -1.0%;  $P_{\text{meta:Fisher}} < 3.54 \times 10^{-3}$ ; **Fig. S2F**). Cases treated with azathioprine (AZ, N=15) had a decrease in NK percentage (Asian: -4.3%, European -7.7%;  $P_{\text{meta:Fisher}} < 6.68 \times 10^{-5}$ ) and an increase in PB percentage (Asian: +0.2%, European: +0.3%;  $P_{\text{meta:Fisher}} < 1.36 \times 10^{-3}$ ; **Fig. S2F**) compared with cases not receiving AZ. Cases treated with mycophenolate mofetil (N=54), hydroxychloroquine (N=113), methotrexate (N=13), or a calcineurin inhibitor (N=10) did not exhibit significant differences in composition compared with cases not receiving each of these therapies. These results suggest that the decrease in CD4<sup>+</sup> T cell and increase in classical monocyte percentages in patients with SLE are not due to therapy.

We next assessed if changes in CD4 and cM percentages were due to changes in the absolute abundance of either population. We analyzed lymphocyte and monocyte abundances reported in the UCSF Electronic Health Record (EHR) Complete Blood Count. Reported abundances in the EHR were highly correlated with the estimated abundances from mux-seq (Pearson  $R_{\text{lympho}}=0.97$  and  $R_{\text{mono}}=0.87$ ; **Fig. S2G**). Comparing an additional 100 cases with 154 controls matched for ethnicity, age, and sex, cases exhibited a significant reduction in lymphocyte abundance (ordinary least squares (OLS); Asians:  $-7.4 \times 10^8$  cells/L,  $P_{\text{OLS}} < 3.46 \times 10^{-9}$ , Europeans:  $-5 \times 10^8$  cells/L,  $P_{\text{OLS}} < 1.07 \times 10^{-6}$ ; **Fig. 1E**) but no difference in monocyte abundance (Asians:  $P_{\text{OLS}}=0.61$ ,

Europeans:  $P_{OLS}=0.98$ ). To assess if a causal relationship exists between lymphocyte decrease and SLE, we performed Generalised Summary-data-based Mendelian Randomizations using summary statistics for genetic associations to immune cell composition (16, 17). The mediation effect of variants associated with lymphocyte abundance ( $\beta_{\text{lympho}\rightarrow\text{SLE}}=-0.39$ ,  $P_{\text{lympho}\rightarrow\text{SLE}}<0.008$ ), but not monocyte abundance ( $\beta_{\text{mono}\rightarrow\text{SLE}}=0.009$ ,  $P_{\text{mono}\rightarrow\text{SLE}}<0.92$ ), was negative on SLE risk. A reverse causation analysis did not show mediation of SLE risk on lymphopenia ( $P_{\text{SLE}\rightarrow\text{lympho}}<0.24$  and  $P_{\text{SLE}\rightarrow\text{mono}}<0.20$ ; **Fig. 1F**) though an alternative explanation of horizontal pleiotropy cannot be excluded.

### 2.3.3 Decrease of circulating naïve CD4<sup>+</sup> T cells in SLE.

Previous studies identified impaired activation of T and B memory cells and elevated expression of ISGs in lymphocytes from patients with SLE (18). To characterize changes in frequencies and transcriptomic profiles of lymphoid populations in SLE, we re-clustered lymphoid cells and assigned the resulting 26 clusters to 14 subpopulations (**Fig. 2A**). Within non-T cells, we identified two NK and four B cell subpopulations. The NK compartment consists of NK<sub>Bright</sub> cells expressing high levels of *GNL1* and moderate levels of *NKG7* and NK<sub>Dim</sub> cells expressing high levels of *NKG7* and *CD16 (FCGR3A)* (**Fig. 2B**). The B cell compartment consists of naïve cells expressing *TCL1A* (B<sub>Naive</sub>), memory cells expressing *BANK1* (B<sub>Mem</sub>), plasma cells expressing *MZB1* (B<sub>Plasma</sub>), and an atypical memory subpopulation expressing *FCRL5*, *CD11c*, *TBX21*, and lacking expression of *CD21* (B<sub>Atypical</sub>; **Fig. 2B**). Atypical B cells may also contain age-associated B cells which share some (*CD11c*<sup>+</sup>, *TBX21*<sup>+</sup>, *CD21*<sup>-</sup>) but not all of the expression markers (*FCRL5*; 19). As a percentage of lymphocytes, neither NK nor B cell subpopulations significantly differed by case-control status.

In the CD4<sup>+</sup> T cell compartment, we identified canonical subpopulations of naïve cells expressing *CCR7* (CD4<sub>Naive</sub>), effector memory cells lacking *CCR7* expression while expressing OX40 receptor



(*TNFRSF4*) and *IL7R* ( $CD4_{EM}$ ), and regulatory cells expressing the canonical transcription factor *FOXP3* and its direct target *RTKN2* (*20*) ( $CD4_{Reg}$ ; **Fig. 2A, B**). Compared with controls, the most pronounced difference in cases was a reduction of  $CD4_{Naive}$  percentage (WLS; Asian: -21.7%; European: -11.8%; Fisher's method  $P_{meta:Fisher} < 8.63 \times 10^{-21}$ ; **Fig. 2C, Table S2**), with Asian cases exhibiting almost a two-fold lower percentage than European cases ( $P_{WLS} < 5.20 \times 10^{-5}$ ). No significant association between  $CD4_{Naive}$  percentage and age (Spearman  $P = 0.76$ ; **Fig. S3A**) or treatment (**Fig. S3B**) was detected. Importantly, cases not on therapy ( $N = 21$ ) exhibited a similar decrease in  $CD4_{Naive}$  percentage compared with controls (Asian: -25.6%, European: -9.7%;  $P_{meta:Fisher} < 2.66 \times 10^{-7}$ , **Fig. S3E**).

#### 2.3.4 Clonal expansion of cytotoxic $GZMH^+$ T cells in SLE.

Within the  $CD8^+$  T cell compartment, we identified naïve cells expressing *CCR7* ( $CD8_{Naive}$ ) and three effector memory subpopulations, including mucosal-associated invariant T cells expressing *KLRB1* and *GZMK* ( $CD8_{MAIT}$ ) and two clusters lacking the expression of *KLRB1* and expressing the chemokine *CCL5*, effector molecule *PRF1*, and exhaustion marker *LAG3* (**Fig. 2A, B**). The two non-MAIT clusters could be distinguished by the expression of granzymes ( $CD8_{GZMH}$ : *GZMH* and *GZMB*,  $CD8_{GZMK}$ : *GZMK*) and mirrored the NK subpopulations ( $NK_{Dim}$ : *GZMH* and *GZMB*;  $NK_{Bright}$ : *GZMK*) (**Figs. 2B and S3C**). Within the  $CD8_{GZMH}$  population, 6% were  $CD4^+CD8^-$  cells based on *CD4* surface expression in the subset of samples also profiled using DNA-conjugated antibodies. Compared with controls, the  $CD8_{GZMH}$  percentage was significantly increased in cases (Asian: +8.6%; European: +6.0%;  $P_{meta:Fisher} < 3.43 \times 10^{-4}$ ; **Fig. 2C, Table S2**) and was observed at similar percentages in flaring and untreated cases (**Fig. S3C-E**). Additionally, we observed a reduction in  $CD8_{MAIT}$  percentage in cases (Asian: -1.1%; European: -0.7%;  $P_{meta:Fisher} < 6.93 \times 10^{-6}$ ; **Fig. 2C, Table S2**).

In addition to increased frequency within lymphocytes, CD8<sub>GZMH</sub> cells were a transcriptionally heterogeneous population with elevated expression of cytotoxic, exhaustion, and ISG signatures in SLE cases compared with controls (**Fig. 2D**). The expression of these signatures was not associated with treatment (**Fig. S3F**). Additionally, only the ISG signature was inversely correlated with age (Pearson  $R=-0.39$ ,  $P<6.57\times 10^{-7}$ ). Across cells, the correlation between cytotoxic and ISG signature genes (mean  $R_{\text{Pearson}}=0.16$ ) and between cytotoxic and exhaustion signature genes (mean  $R_{\text{Pearson}}=0.10$ ) were generally low (**Figs. 2E**). Thus, in cases these pathways are unlikely to be jointly activated in the same cells. This was in stark contrast to the high correlation between signature genes calculated across CD8<sub>GZMH</sub> pseudobulk expression profiles from different individuals, highlighting the limitation of bulk analysis in uncovering additional heterogeneity within a seemingly homogeneous population (**Fig. 2E**).

To further investigate the clonality of the CD8<sub>GZMH</sub> and CD8<sub>GZMK</sub> populations, we amplified and sequenced the CDR3 region of the T cell receptor (TCR), recovering paired *TCRA* and *TCRB* sequences from 10.2% of CD4 and 8.7% of CD8 cells with no differences in the number of unique TCRs detected between cases (N=83) and controls (N=20) ( $P_{\text{wilcoxon}}=0.72$ ). Of the expanded CD8 clones, 59% were from CD8<sub>GZMH</sub> cells and 21% from CD8<sub>GZMK</sub> cells (**Fig. 2F**). Compared with controls, cases exhibited a restricted repertoire in CD8 cells ( $P_{\text{wilcoxon}}<0.01$ ; **Fig. 2G**) but not CD4 cells ( $P_{\text{wilcoxon}}=0.91$ ; **Fig. S3G, H**). Within the CD8<sub>GZMH</sub> subpopulation, cells expressing the cytotoxic signature were expanded at a ~4:1 ratio to cells expressing the ISG signature (44.8% vs 9.7%, **Fig. 2H**). As a positive control, clones expressing the invariant *TRAV1-2* and *TRAJ33* chains were enriched within the CD8<sub>MAIT</sub> cluster compared to other cell types (Tukey's HSD  $P<0.001$ ; **Fig. S3I**).

### 2.3.5 Expression changes across 11 peripheral immune cell types in SLE.

Bulk transcriptomic analyses of PBMCs have consistently reported the association between SLE and elevated expression of ISGs, which is normally observed during acute viral infections (21). Longitudinal bulk analysis of 158 pediatric cases confirmed the elevated expression of ISGs in patients with more severe acute presentations and increased renal and neurological involvement (3). However, bulk analysis has limited power to pinpoint the cell types producing the ISG signature or identify additional cell-type-specific signatures. Recent analysis of 33 pediatric cases demonstrated the potential of scRNA-seq to assign cell-type specificity to previously identified ISGs from bulk analysis (6).

Here, we characterize the transcriptional differences for each of 11 circulating immune cell types between SLE cases and controls. 302 genes were differentially expressed (DE) in at least one cell type between cases and controls of either Asian or European ancestry, not confounded by medication ( $|\log_{2}FC| > 0.5$ ;  $P_{\text{adjusted}} < 0.05$ ; **Table S3**; **Figs. S4A** and **S4G**). Hierarchical clustering of pseudobulk expression profiles of these DE genes across cell types resulted in six modules (**Figs. 3A**). Compared with controls, cases upregulated a module of ISGs across all cell types ( $\text{Pan}_{\text{up}}$ ) and a myeloid-specific module ( $\text{Mye}_{\text{up}}$ ) containing *IFITM1/3*, *IFITM3*, *APOBEC3A*, *RNASE2*, and *IFIT2*. Both modules were enriched for type-I interferon signaling and innate immune pathways (**Fig. 3B**). Additionally, we identified a downregulated module across all cell types enriched for the interaction between lymphoid and non-lymphoid cells ( $\text{Pan}_{\text{down}}$ ), a myeloid-specific downregulated module ( $\text{Mye}_{\text{down}}$ ) enriched for hedgehog signaling, a T cell-specific upregulated module ( $\text{T}_{\text{up}}$ ) enriched for leukocyte activation, and a B cell-specific upregulated module ( $\text{B}_{\text{up}}$ ) enriched for AP-1 transcriptional response and TLR signaling (**Fig. 3B**).

Our results were validated by single-cell transcriptomic analyses of PBMCs activated *in vitro* by recombinant interferon beta (rIFNB1) (8) and from pediatric patients with SLE (6). For each cell

type, particularly myeloid populations, expression fold changes between cases and controls were highly correlated with fold changes between rIFN $\beta$ 1-stimulated and unstimulated cells (**Fig. S4B**). Of the 100 ISGs previously identified from bulk analysis and analyzed in pediatric SLE (6), 64 were DE in at least one cell type and mainly resided in the Pan<sub>up</sub> (46/79) and Mye<sub>up</sub> (8/64) modules. Interestingly, 11 genes were DE only across PBMC pseudobulks, illustrating a likely confounding effect of bulk analysis due to differences in cellular composition between cases and controls (**Table S4**). Importantly, the large sample size of our cohort resulted in the identification of 238 previously undescribed DE genes in adult SLE, 56 of which were myeloid specific.

### 2.3.6 Pronounced type-1 interferon response in classical monocytes.

Myeloid cells exhibited the most DE genes between cases and controls, consisting of known and novel genes associated with SLE. To further investigate their heterogeneity, we re-clustered myeloid cells into six clusters differentiating the monocyte lineage (cM: *CD14*<sup>+</sup> classical, ncM: *FCGR3A*<sup>+</sup> non-classical, ncM<sub>comp</sub>: *C1QA*<sup>+</sup>/*FCGR3A*<sup>+</sup> complement-expressing non-classical) and the dendritic cell lineage (cDC1: *CLEC10A*<sup>+</sup> conventional type-1, cDC2: *CLEC9A*<sup>+</sup> conventional type-2, pDC: *IRF7*<sup>+</sup> plasmacytoid; **Figs. 3C, D and S4C, D**). Although pDCs can derive from either myeloid or lymphoid progenitors, their expression profiles were more similar to, and thus jointly analyzed with, other myeloid populations (22). We also detected *AXL*<sup>+</sup> dendritic cells within both cDC1s and pDCs consistent with their suggested distribution as a transitioning population between cDCs and pDCs (23) (**Fig. S4E**). As a percentage of myeloid cells compared with controls, cases exhibited reduced percentages of pDCs (WLS; Asian: -0.6%; European: -1.8%; Fisher's method  $P_{\text{meta:Fisher}} < 2.33 \times 10^{-24}$ ), cDC1s (Asian: -2.0%; European: -1.9%;  $P_{\text{meta:Fisher}} < 2.65 \times 10^{-14}$ ), and cDC2s (Asian: -0.2%; European: -0.1%;  $P_{\text{meta:Fisher}} < 2.51 \times 10^{-7}$ ) and increased percentages of cMs (Asian: +3.6%; European: +3.7%;  $P_{\text{meta:Fisher}} < 1.78 \times 10^{-5}$ ) and ncM<sub>comp</sub>s (Asian: +0.5%; European: +0.2%;  $P_{\text{meta:Fisher}} < 1.67 \times 10^{-3}$ ; **Fig. 3E, Table S5**).

Next, we used RNA velocity to assess the transcriptional heterogeneity of each myeloid cell type along a trajectory of inferred activation (24, 25). In cMs, ncMs, and ncM<sub>comps</sub>, velocity analysis of DE genes revealed that inferred activation largely reflected the degree of average ISG expression (Mye<sub>up</sub>; **Fig. 3F**) with regions of high activation enriched for cells from SLE cases (**Fig. 3G**). These results were similar in cDC populations (**Fig. S4F**). Ordering cMs along inferred activation showed higher activation from cases with higher SLE Disease Activity Index (SLEDAI) (26) defined using clinical features (T-test; Asian:  $P < 5 \times 10^{-4}$ ; European:  $P < 3.2 \times 10^{-7}$ ; **Fig. 3H**). The average inferred activation was better correlated with SLEDAI in Europeans ( $R_{\text{Pearson}} = 0.66$ ) than Asians ( $R_{\text{Pearson}} = 0.52$ ; **Fig. 3I**). In both ethnicities, a wide range of average inferred activations were observed in patients with lower disease activity (SLEDAI between 0 and 4) suggesting that clinical measures underlying SLEDAI do not fully capture the molecular heterogeneity of SLE.

### **2.3.7 Expression modules predict CD4<sub>naïve</sub> lymphopenia, disease status, and stratifies SLE patients.**

Previous work in mouse models has demonstrated that type-1 interferons upregulate *CD69* thereby inhibiting lymphocyte egress from lymphoid tissue (27). We hypothesized that the pleiotropic effects of type-1 interferons in patients with SLE may underlie the monocyte-dominant expression of ISGs and inhibit CD4<sup>+</sup> T cells from exiting lymphoid tissue, resulting in the observed decrease of circulating naïve CD4<sup>+</sup> T cells. Consistent with this hypothesis, both the Pan<sub>up</sub> and Mye<sub>up</sub> gene module scores were highly correlated with CD4<sub>Naïve</sub> abundance (Asian: Pearson  $R_{\text{Panup}} = -0.52$ , European:  $R_{\text{Panup}} = -0.57$ ,  $P_{\text{meta:Fisher}} < 1.04 \times 10^{-3}$ ; Asian:  $R_{\text{Myeup}} = -0.35$ , European:  $R_{\text{Myeup}} = -0.48$ ,  $P_{\text{meta:Fisher}} < 0.02$ ; **Figs. 4A and S5A**).

One of the diagnostic difficulties of SLE is the extensive heterogeneity in disease manifestations. Consistent with this heterogeneity, individual clinical features weakly correlated with module scores (**Fig. 4B**). We therefore utilized the expression of module genes over pseudobulks of the

relevant cell types as features for clinical prediction and molecular stratification of SLE. While the 302 expression features had good out-of-sample predictive power for case-control status (Area Under the Curve (AUC) = 0.84; **Fig. 4C**), they had only modest predictive power for individual clinical features, reflective of the modest correlation between clinical features and module scores (**Figs. 4D** and **S5B**). To molecularly stratify cases, we performed principal component analysis (PCA) over expression features followed by K-means clustering to identify two clusters that broadly tracked with case-control status (**Fig. 4E**), SLEDAI score (**Fig. 4F**), and along principal component 1 (PC1). Cases in the High cluster had significantly higher inferred activation of monocytes compared with cases in the Low cluster ( $P_{\text{Wilcoxon}} < 6.20 \times 10^{-9}$ ; **Fig. S5C**). PC1 correlated most with genes in the Pan<sub>Up</sub>, Mye<sub>Up</sub>, and B<sub>Up</sub> modules including the myeloid-specific-expression of *IFITM3*, a gene previously described to stratify pediatric SLE cases (3) (**Fig. 4E**). To assess the correspondence between molecular clusters and clinical features, we projected 94 held-out cases each to a molecular cluster based on expression features (**Fig. 4G**). Cases assigned to the High cluster were enriched for disease flare (15/19 flare cases, **Fig. S5D**) and portended over five times the odds of having anti-Smith antibodies ( $P_{\text{adjusted:Fisher}} < 0.05$ ; **Fig. 4H**). These results demonstrate that cell-type-specific expression profiles obtained using mux-seq can be used to link cell-intrinsic states with changes in composition, predict case-control status, and molecularly stratify patients with SLE.

### **2.3.8 Identification of cell-type-specific *cis*-eQTLs across eight immune cell types.**

We next integrated mux-seq data with genotyping data to map cell-type- and cell-context-specific *cis* expression quantitative trait loci (eQTLs) that may mediate SLE disease associations. Across the eight most abundant cell types, linear regression followed by meta-analysis (28, 29) of three cohorts (92 CLUES Europeans, 98 CLUES Asians, 46 ImmVar Europeans) identified 3,331 genes with at least one *cis*-eQTL in a cell type (FDR<0.05), which we termed cell-type-by-cell-type *cis*-eQTLs (CBC-eQTL) (**Table S6**). Analysis of the genetic architecture of gene expression (30)

resulted in estimates of average *cis* heritability ranging from 0.03 to 0.09 per cell type and average *cis* genetic correlations (rG) ranging from 0.25 to 0.75 for pairs of cell types. Since cells were simultaneously processed, we also estimated shared residual effects (rE) between cell types (e.g., shared environmental and *trans* genetic effects) ranging from 0.03 to 0.12. Clustering of rG and rE reflected known lineages between circulating immune cell types (**Fig. 5A**).

The rG and rE estimates suggest that pleiotropic genetic and shared residual effects are common across immune cell types, which may confound the ability to detect cell-type-specific signals among CBC-eQTLs. To account for pleiotropy, we decomposed per cell-type expression profiles into a shared component across all cell types and eight cell-type-specific components, then mapped *cis*-eQTLs associated with each component (31). We identified 535 genes with at least one cell-type-specific *cis*-eQTL (cs-eQTL) (FDR<0.05) and 1,207 shared *cis*-eQTLs (sh-eQTLs) (**Fig. 5B**; **Table S7**). The effect sizes of CBC-, sh-, and cs-eQTLs were correlated between individuals of European and Asian ancestries (**Fig. S6A, B**), which separated by genotype principal components (**Fig. S6C**). Compared to CBC-eQTLs, cs-eQTLs for each cell type were significantly and specifically enriched for regions of chromatin accessibility in the same or closely related cell types (32), suggesting that decomposition analysis is more likely to identify *cis*-eQTLs overlapping cell-type-specific *cis*-regulatory elements (**Fig. 5C**).

### **2.3.9 Identification and annotation of cell-type-specific SLE-associated loci.**

We next integrated GWAS summary statistics from 9 immune-mediated and 7 non-immune-mediated traits/diseases to identify cell types where cs-eQTLs harbored the most GWAS associations. Linkage disequilibrium (LD) score regression (33) revealed enrichment of disease heritability for relevant cell types across autoimmune diseases (**Fig. 5D**). The highest enrichment for SLE variants was in cMs and B cells, consistent with our finding that cMs are the highest

expressers of type-1 ISGs and previous work demonstrating that activated B cells produce autoantibodies and secrete cytokines related to disease pathogenesis (34, 35) (**Fig. 5D**).

We next performed Bayesian genetic colocalization analyses utilizing sh- and cs-eQTLs to fine-map 43 loci associated with SLE (4, 36). Among the five loci colocalized with sh-eQTLs (Posterior probability (PP) > 0.6) was the *UBE2L3* locus. Previously identified *UBE2L3* cis-eQTLs in lymphoblastoid cells lines, B cells, and monocytes were replicated by colocalization analysis utilizing CBC-eQTLs (B, cM, ncM PP>50%). However, analysis utilizing sh- and cs-eQTLs predicted colocalization of the SLE-association and an *UBE2L3* sh-eQTL (PP=88.5%) suggesting that this association is shared across cell types (**Fig. S6D**).

Among the seven SLE-associated loci colocalizing with cs-eQTLs was 17q21, a locus associated with asthma (37), Crohn's disease (38), and type-1 diabetes (39). This locus has been difficult to dissect as it encompasses three genes, *IKZF3*, *GSDMB*, and *ORMDL3* implicated in lymphocyte development (40), pyroptosis (41), and inflammation (42). *ORMDL3* is a regulator of sphingolipid biosynthesis, linked to the autophagy pathway associated with multiple autoimmune diseases (43), and implicated in the development and differentiation of lymphocytes in SLE pathogenesis (44). *ORMDL3* was ubiquitously expressed across cell types with the highest expression in lymphoid populations (**Fig. 5E, F**). Colocalization was predicted between SLE-associations and both *ORMDL3* sh-eQTLs (PP>88%) and cs-eQTLs in Bs, CD8s, and pDCs (PP>96.1%, 92.0%, and 92.1% respectively) (**Fig. 5G**). While *GSDMB* and *IKZF3* were also expressed in most cell types (**Fig. 5F**), neither gene had a cs-eQTL and the highest posterior probability of colocalization was observed between SLE-associations and *GSDMB* sh-eQTLs at 63.8%. Further, conditional analysis (45) confirmed that the SLE associations observed near *IKZF3* (**Fig. 5G**) were independent of the *GSDMB* and *ORMDL3* associations, and that the conditioned SLE-associations still colocalized with the *ORMDL3* cs- and sh-eQTLs. The minor allele (T) of



rs7216389, a tagging variant in the locus associated with asthma and SLE ( $P < 6.09 \times 10^{-7}$ ) (4), conferred an increase of *GSDMB* and *ORMDL3* expression across all cell types, but an additional increase of *ORMDL3* expression in CD8s and Bs suggesting cell-type-specific genetic effects in these cell types that was not observed for *GSDMB* (Fig. 5G). These results are consistent with previous observations in CD8s and Bs where SNPs in high LD with rs7216389 impacted regulatory elements affecting *ORMDL3* expression (46).

We further used expression decomposition to perform a modified transcriptome-wide association study (TWAS) using CONTENT (47). Across SLE, Crohn's disease, and rheumatoid arthritis, joint modeling of shared and cell-type-specific gene expression identified 93 genes associated with SLE (73 novel), more than twice the number identified by CBC approaches (Fig. 5H). Results were significantly enriched for known SLE associations where 51% of candidate genes, defined as the most proximal gene to each SLE association (6), were replicated in the TWAS with  $p$ -values  $< 0.05$  ( $P_{\text{Enrichment}} < 1.2 \times 10^{-24}$ ). Importantly, both the joint and CBC analyses enabled by mux-seq significantly outperformed a standard TWAS using pseudobulk PBMC transcriptomic profiles. These analyses highlight the advantage of leveraging cell-type-specific *cis*-eQTLs to annotate GWAS associations, detangle GWAS signals in gene dense loci, and power TWAS analysis to identify novel associations.

### **2.3.10 Modification of genetic effects on gene expression by interferon activation.**

We next assessed if variable type-1 interferon activation observed in patients with SLE could modify genetic effects on gene expression *in vivo*, consistent with our previous *in vitro* work (11, 48). In SLE cases, we identified 35 genes with a *cis*-eQTL interacting with the Pan<sub>up</sub> ISG signature, a proxy for type-1 interferon activation, which we call IFN-eQTL (FDR < 0.1). IFN-eQTL effect size estimates correlated between samples of Asian and European ancestries (Fig. S7).

Previous interferon response *cis*-eQTLs (reQTLs) identified in monocyte-derived dendritic cells *in vitro* (48) were significant in cMs, but not other cell types (**Fig. 6A**).

Among the IFN-eQTLs, we replicated rs11080327 (A>G) as an IFN-eQTL for *SLFN5* in myeloid (cM:  $P < 2.5 \times 10^{-10}$ , ncM:  $P < 0.001$ ) and B cells ( $P < 5.8 \times 10^{-6}$ ) but not in NK or T cells (**Fig. 6B**). These results are consistent with the identification of rs11080327 as a *cis*-eQTL in lymphoblastoid cell lines (49) and as a *cis*-reQTL in monocyte-derived dendritic cells stimulated with rIFNB1 (11). We then performed multiplexed single-cell ATAC-seq of PBMCs from 5 healthy donors either unstimulated or stimulated with rIFNB1. In most cell types, we observed less accessibility in genomic regions near rs11080327 at baseline and a genotype dependent increase of accessibility after stimulation (**Fig. 6C**). This was most pronounced in cMs, where the strongest IFN-eQTL was observed. These results are consistent with luciferase reporter assays demonstrating the region overlapping rs11080327 harboring a *cis*-regulatory element that is activated in response to type-1 interferon (11). Overall, our findings illustrate that variability in cell activation *in vivo* could modify genetic effects on gene expression suggesting that genetic differences may not only predispose individuals to SLE but also affect individual's response to a disease state.

## 2.4 Discussion

SLE remains a challenging autoimmune disease to diagnose and treat. The paucity of targeted therapies, in conjunction with the heterogeneity of disease manifestations and treatment response, highlight the need for improved molecular characterization. In a large multiethnic cohort, we demonstrate the use of mux-seq as a systematic approach to characterize changes in cell-type composition and cell-type-specific gene expression in adult SLE. We further show how integration of population genetics with single-cell RNA-sequencing could be utilized to annotate genetic variants with cell-type-specific effects on gene expression associated with SLE and other autoimmune diseases.

Using mux-seq, we linked compositional changes to variation in immune cell transcriptional states in SLE. Compositionally, the decrease of naïve CD4<sup>+</sup> T cells in cases, particularly those of Asian ancestry, appears to explain the known lymphopenia observed in patients with SLE and importantly was not associated with immunosuppressant treatment, consistent with reports suggesting mycophenolate mofetil, hydroxychloroquine, and steroids have either no or transient effects on the composition of white blood cells (50). Transcriptionally, cMs and ncMs produced the most prominent type-1 ISG signature, including genes specific to myeloid cells, consistent with observations in pediatric SLE (6). This finding justifies further investigation into the heterogeneity of type-1 interferon response across leukocyte subsets, particularly in SLE patients being treated with antagonists against the type-1 interferon receptors that have shown mixed results in clinical trials (51). While both cDCs and pDCs also express ISGs, their scarcity in circulation limited their contribution to the overall ISG signature. We did not detect *IFNB1* or *IFNA* transcripts in pDCs or other myeloid cell types and thus the source of type-1 interferons in SLE remains elusive and is likely not among circulating immune cells (52). The inverse correlation between naïve CD4<sup>+</sup> T cell abundance and monocyte ISG expression suggests the following model of the pleiotropic effects of type-1 interferons *in vivo*: ISG production through the interferon signaling cascade and sequestration of T cells in sites of inflammation through the regulation of *CD69* and *S1PR1* (27). While age was inversely correlated with the ISG signature consistent with previous reports, naïve CD4 T cell abundance was not correlated with age and remains inversely correlated the ISG signature after adjusting for age (53). Thus, age is likely not a primary factor for causing SLE, consistent with healthy female first-degree relatives showing similar inverse correlation between age and serum IFN $\alpha$  (7). Matched profiling of cells from disease-damaged tissue and blood in cases could further shed light on both the source of type-1 interferons and confirm the role of lymphocyte trafficking in SLE.

A striking observation from our data is the expansion of *GZMH*<sup>+</sup> but not *GZMK*<sup>+</sup> cytotoxic CD8<sup>+</sup> T cells in SLE, in some cases consisting of ~50% of all lymphocytes. While two cytotoxic CD8<sup>+</sup> T cell populations were also observed in pediatric SLE (6), the frequency of *GZMH*<sup>+</sup> CD8<sup>+</sup> T cells was not reported to be significantly increased despite elevated expression of *GZMB* and *PRF1*, which may originate from both *GZMH*<sup>+</sup> CD8<sup>+</sup> T and NK<sub>dim</sub> cells. While *GZMB* and *PRF1* have been described as markers for CD8<sup>+</sup> T cell subsets enriched in SLE (54), *GZMH* was higher expressed, more ubiquitous, and more differentially expressed between cases and controls. The function of granzyme-H is not well characterized, but previous work demonstrated its divergent roles in initiating caspase-dependent apoptosis in T cells while initiating caspase-independent apoptosis in NK cells (55, 56). The significant clonal expansion of *GZMH*<sup>+</sup> CD8<sup>+</sup> T cells, specifically the cytotoxic subpopulation, suggests a pathogenic role for these cells in SLE and are consistent with independent work (54). One model for the initiation and exacerbation of SLE suggested by these results is an adaptive immune response initiated by foreign and autoantigens followed by chronic exposure to antigens in damaged tissue resulting in "epitope spreading", where new autoantigens are introduced to the immune system and become future targets of the autoimmune response (57). Analysis of immune repertoires of both B and T cells and matching analysis of their antigenic specificity of SLE patients longitudinally would be instructive for deciphering the role of cell-mediated immunity in pathogenesis.

Integrating measurements of cellular composition and cell-type-specific expression with genotyping provided an opportunity to assess the genetic determinants of cell-type- and cell-context-specific gene expression and ascribe functionality to SLE-associated variants. In the presence of pleiotropic effects, mux-seq enabled the decomposition of gene expression into shared and cell-type-specific components and mapping of *cis*-eQTLs associated with these components. Enrichment analyses of orthogonal functional genomic datasets supported the annotation of cell-type-specific *cis*-eQTLs. Integrated analysis of GWAS data and cell-type-

specific *cis*-eQTLs provided insight into immune cell types that mediate disease associations and for individual loci, enabled the fine-mapping and annotation of disease-associated variants. Using decomposed expression components also significantly improved our ability to identify novel disease-associated genes using TWAS compared to using pseudobulk expression profiles over PBMCs or individual cell types. Finally, using quantitative measures of interferon activation from mux-seq, we identified *cis*-eQTLs whose effects on gene expression could be modified by elevated interferon levels, a critical disease environment in SLE. These results highlight the importance of cellular context for the interpretation of genetic variants associated with disease risk and perhaps disease heterogeneity.

Mux-seq is a cost-effective and systematic approach for enabling cellular phenotyping of large population cohorts. Genetic analysis of cohorts across populations are important for understanding the differences in SLE risk between ancestries and the involvement of environmental triggers. Longitudinal profiling of SLE cases, particularly patients in remission or active flare, could reveal new insights into the initiation of disease, variation in disease activity, new homeostatic states in patients, and response to treatment. While we examined and controlled for treatment associated differences in cellular composition and cell-type-specific expression between SLE and healthy controls, we did observe notable effects of treatment including the depletion of NK cells in patients treated with azathioprine. Since mux-seq leverages natural genetic variation as sample barcodes, it is compatible with multimodal single-cell profiling of chromatin state and cell-surface protein abundance. The integration of richer epigenetic and cellular phenotypes along with improvements to current transcriptomic workflows will undoubtedly improve molecular sub-phenotyping of SLE, the power to detect cell-type-specific and cell-context-specific molecular QTLs, and the resolution for annotating SLE associations.

## 2.5 Methods Summary

Detailed materials and methods can be found in the Supplementary Materials. Briefly, we collected PBMCs from SLE cases in the California Lupus Epidemiological Study (CLUES) cohort, matching healthy controls from the UCSF Rheumatology Clinic, and additional controls from the Immune Variation Project (ImmVar). Presence of clinical features important to SLE were collected.

Antibody stained or unstained PBMCs were pooled and profiled using 10x Genomics' Chromium Single Cell 3' V2 chemistry and processed using the 10x Cell Ranger pipeline. Freemuxlet was used to assign cells to their donor of origin and, along with Scrublet (13), remove doublets. Platelets, Megakaryocytes, Red blood cells (RBC) were removed using gene markers. Technical variation was removed using COMBAT and regressing out nUMIs, and mitochondrial percent. Standard approaches in Scanpy version 1.6 were used to filter cells, perform dimensionality reduction, cluster using Louvain, and project cells using UMAP (58). Cell types were annotated using canonical marker genes and confirmed in cells with antibody staining.

For each cell type, percentage is calculated as the number of cells divided by the total number of cells assigned to the sample. Differences in percentages were compared using weighted least squares. UCSF electronic health record queries compared individuals with multiple healthy encounters and cases with a M32.\* ICD-10 code. Mendelian randomization was performed using the GSMR package version 1.91.5beta on UK Biobank cell count QTLs and a separate SLE study (4). To examine changes in expression, pseudobulk expression profiles were computed for each cell type and individual using EdgeR. EdgeR was used to perform differential expression analysis (59).  $CD8_{GZMH}$  signature scores were calculated using Scanpy `score_genes` on canonical markers (see Supplement). Module scores per individual were calculated by the mean pseudobulk expression for genes in each module. Co-expression analysis was performed on the

top 300 DE genes, and clustered by Spearman correlation. Expression modules were recovered by hierarchical clustering of DE genes, revealing 6 modules. ToppGene was used to find enrichment of modules in pathways (60). Molecular clusters were defined using PCA. RNA velocity was performed on cM using the scVelo package. Sklearn's Logistic Regression function was used for all prediction models.

TCR sequencing was performed by amplifying TRA and TRB CDR3 sequences from cDNA and processed with the Cell Ranger pipeline. Only cells with paired TRA and TRB were used. TCRs were analyzed with the singleTCR package. Expanded clonotypes defined as a TCR sequence detected in at least two cells, were identified using Normalized Shannon's entropy.

Samples collected at UCSF were genotyped using the Affymetrix World LAT array. ImmVar samples were genotyped on the OmniExpressExome54 chip. Data was processed using Axiom Best Practices or by previously published methods for the ImmVar cohort. Samples were evaluated for call rate, missingness, and heterozygosity then imputed using the Michigan Imputation Server with the Haplotype Reference Consortium version 1.1 reference set. Only SNPs with  $R_{sq} > 0.3$  and minor allele frequency  $> 10\%$  were retained. Heritability was calculated with the GCTA package's Bivariate GREML function. *Cis*-eQTLs were mapped  $\pm 100$ kb of each gene using the MatrixEQTL package accounting for Genotype PCs, Expression PCs, age, sex, SLE status, batch as covariates in the linear model. Cell type specific eQTLs were mapped using the fastGxC method (31). CLUES Asian, CLUES European, and ImmVar samples were analyzed separately then meta-analyzed using the METASOFT package. Empirical p-values and FDRs were calculated with the qvalue package. LocusZoom was used to visualize loci. SLE cases were analyzed for reQTLs with MatrixEQTL using the ISG score as an interaction term and accounting for genotype PCs, age, sex, and batch.

ATAC-seq enrichment was calculated using a Mann-Whitney test and previously published ATACseq peaks from sorted cell types. GWAS enrichment was calculated using LDscore regression (33). TWAS analyses were performed using CONTENT (47). Colocalization analyses were performed with COLOC (36).

10X Chromium scATAC-seq kit was used to process PBMCs from 5 healthy individuals incubated for 8 hours with IFNB or culture media alone. Sequencing data was processed with CellRanger and demultiplexed with Freemuxlet. The ArchR package and Scanpy were used for downstream processing (61).

## **2.6 Supplementary Materials**

Materials and Methods

Figs. S1 to S8

Tables S1 to S12



## 2.7 References and Notes

1. E. E. Carter, S. G. Barr, A. E. Clarke, The global burden of SLE: prevalence, health disparities and socioeconomic impact. *Nat. Rev. Rheumatol.* **12**, 605–620 (2016).
2. A. Kaul et al., Systemic lupus erythematosus. *Nat. Rev. Dis. Primers.* **2**, 16039 (2016).
3. R. Banchereau et al., Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients. *Cell.* **165**, 551–565 (2016).
4. J. Bentham et al., Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat. Genet.* **47**, 1457–1464 (2015).
5. J. Banchereau, V. Pascual, Type I interferon in systemic lupus erythematosus and other autoimmune diseases. *Immunity.* **25**, 383–392 (2006).
6. D. Nehar-Belaid et al., Mapping systemic lupus erythematosus heterogeneity at the single-cell level. *Nat. Immunol.* **21**, 1094–1106 (2020).
7. S. Sharma et al., Widely divergent transcriptional patterns between SLE patients of different ancestral backgrounds in sorted immune cell populations. *J. Autoimmun.* **60**, 51–58 (2015).
8. H. M. Kang et al., Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
9. C. M. Lanata et al., Genetic contributions to lupus nephritis in a multi-ethnic cohort of systemic lupus erythematosus patients. *PLoS One.* **13**, e0199003 (2018).
10. T. Raj et al., Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science.* **344**, 519–523 (2014).

11. M. N. Lee et al. , Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science*. **343**, 1246980 (2014).
12. C. J. Ye et al., Intersection of population variation and autoimmunity genetics in human T cell activation. *Science*. **345**, 1254665 (2014).
13. S. L. Wolock, R. Lopez, A. M. Klein, Scrublet: Computational identification of cell Doublets in Single-cell transcriptomic data. *Cell Syst*. **8**, 281-291.e9 (2019).
14. V. A. Traag, L. Waltman, N. J. van Eck, From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
15. L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018), (available at <http://arxiv.org/abs/1802.03426>).
16. Z. Zhu et al., Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* **9**, 224 (2018).
17. C. Bycroft et al. , The UK Biobank resource with deep phenotyping and genomic data. *Nature*. **562**, 203–209 (2018).
18. V. R. Moulton et al., Pathogenesis of human systemic lupus erythematosus: A cellular perspective. *Trends Mol. Med.* **23**, 615–635 (2017).
19. K. Rubtsova, A. V. Rubtsov, M. P. Cancro, P. Marrack, Age-Associated B Cells: A T-bet-Dependent Effector with Roles in Protective and Pathogenic Immunity. *J. Immunol.* **195**, 1933–1937 (2015).
20. A. Ferraro et al., Interindividual variation in human T regulatory cells. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E11111-20 (2014).

21. Y. Kotliarov et al., Broad immune activation underlies shared set point signatures for vaccine responsiveness in healthy individuals and disease activity in patients with lupus. *Nat. Med.* **26**, 618–629 (2020).
22. H. Shigematsu et al., Plasmacytoid dendritic cells activate lymphoid-specific genetic programs irrespective of their cellular origin. *Immunity*. **21**, 43–53 (2004).
23. A.-C. Villani et al., Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*. **356** (2017), doi:10.1126/science.aah4573.
24. V. Bergen, M. Lange, S. Peidli, F. A. Wolf, F. J. Theis, Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* (2020), doi:10.1038/s41587-020-0591-3.
25. G. La Manno et al., RNA velocity of single cells. *Nature*. **560**, 494–498 (2018).
26. J. P. Buyon et al., The effect of combined estrogen and progesterone hormone replacement therapy on disease activity in systemic lupus erythematosus: a randomized trial. *Ann. Intern. Med.* **142**, 953–962 (2005).
27. L. R. Shiow et al. , CD69 acts downstream of interferon- $\alpha/\beta$  to inhibit S1P1 and lymphocyte egress from lymphoid organs. *Nature*. **440**, 540–544 (2006).
28. B. Han, E. Eskin, Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am. J. Hum. Genet.* **88**, 586–598 (2011).
29. A. A. Shabalín, Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*. **28**, 1353–1358 (2012).

30. J. Yang, S. H. Lee, M. E. Goddard, P. M. Visscher, GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
31. A. Lu et al., Fast and powerful statistical method for context-specific QTL mapping in multi-context genomic studies. *bioRxiv* (2021), p. 2021.06.17.448889, , doi:10.1101/2021.06.17.448889.
32. D. Calderon et al., Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nat. Genet.* **51**, 1494–1505 (2019).
33. H. K. Finucane et al., Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
34. E. Nashi, Y. Wang, B. Diamond, The role of B cells in lupus pathogenesis. *Int. J. Biochem. Cell Biol.* **42**, 543–550 (2010).
35. X. Hu et al. , Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. *Am. J. Hum. Genet.* **89**, 496–506 (2011).
36. C. Giambartolomei et al., CommonMind Consortium, B. Pasaniuc, P. Roussos, A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics.* **34**, 2538–2545 (2018).
37. M. F. Moffatt et al., Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature.* **448**, 470–473 (2007).
38. L. Jostins et al., Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature.* **491**, 119–124 (2012).

39. J. C. Barrett et al., Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* **41**, 703–707 (2009).
40. B. Morgan et al., Aiolos, a lymphoid restricted transcription factor that interacts with Ikaros to regulate lymphocyte differentiation. *EMBO J.* **16**, 2004–2013 (1997).
41. L. Li, Y. Li, Y. Bai, Role of GSDMB in Pyroptosis and Cancer. *Cancer Manag. Res.* **12**, 3033–3043 (2020).
42. Y. Zhang et al., The ORMDL3 Asthma Gene Regulates ICAM1 and Has Multiple Effects on Cellular Inflammation. *Am. J. Respir. Crit. Care Med.* **199**, 478–488 (2019).
43. B. James, S. Milstien, S. Spiegel, ORMDL3 and allergic asthma: From physiology to pathology. *J. Allergy Clin. Immunol.* **144**, 634–640 (2019).
44. J. Dang et al. , ORMDL3 facilitates the survival of splenic B cells via an ATF6 $\alpha$ -endoplasmic reticulum stress-Beclin1 autophagy regulatory pathway. *J. Immunol.* **199**, 1647–1659 (2017).
45. J. Yang et al., Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–75, S1-3 (2012).
46. B. J. Schmiedel et al., 17q21 asthma-risk variants switch CTCF binding and regulate IL-2 production by T cells. *Nat. Commun.* **7**, 13426 (2016).
47. M. Thompson et al., Multi-context genetic modeling of transcriptional regulation resolves novel disease loci. *bioRxiv* (2021), , doi:10.1101/2021.09.23.461579.
48. C. J. Ye et al., Genetic analysis of isoform usage in the human anti-viral response reveals influenza-specific regulation of ERAP2 transcripts under balancing selection. *Genome Res.* **28**, 1812–1825 (2018).

49. J. F. Degner et al. , DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*. **482**, 390–394 (2012).
50. S. J. Rivero, E. Díaz-Jouanen, D. Alarcón-Segovia, Lymphopenia in systemic lupus erythematosus. *Arthritis Rheum*. **21**, 295–305 (1978).
51. E. F. Morand et al., Trial of anifrolumab in active systemic lupus erythematosus. *N. Engl. J. Med*. **382**, 211–221 (2020).
52. M. Cella et al., Plasmacytoid monocytes migrate to inflamed lymph nodes and produce large amounts of type I interferon. *Nat. Med*. **5**, 919–923 (1999).
53. T. B. Niewold et al., Age- and sex-related patterns of serum interferon-alpha activity in lupus families. *Arthritis Rheum*. **58**, 2113–2119 (2008).
54. P. Blanco et al., Increase in activated CD8+ T lymphocytes expressing perforin and granzyme B correlates with disease activity in patients with systemic lupus erythematosus. *Arthritis Rheum*. **52**, 201–211 (2005).
55. L. Casciola-Rosen, F. Andrade, D. Ulanet, W. B. Wong, A. Rosen, Cleavage by granzyme B is strongly predictive of autoantigen status: implications for initiation of autoimmunity. *J. Exp. Med*. **190**, 815–826 (1999).
56. M. Faroudi et al., Lytic versus stimulatory synapse in cytotoxic T lymphocyte/target cell interaction: manifestation of a dual activation threshold. *Proc. Natl. Acad. Sci. U. S. A*. **100**, 14145–14150 (2003).
57. C. L. Vanderlugt, S. D. Miller, Epitope spreading in immune-mediated diseases: implications for immunotherapy. *Nat. Rev. Immunol*. **2**, 85–95 (2002).

58. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
59. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* **26**, 139–140 (2010).
60. J. Chen, E. E. Bardes, B. J. Aronow, A. G. Jegga, ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* **37**, W305-11 (2009).
61. J. M. Granja et al., Author Correction: ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* (2021), doi:10.1038/s41588-021-00850-x.
62. R. K. Perez, et al., Multiplexed scRNA-seq reveals the cellular and genetic correlates of systemic lupus erythematosus Analysis Code (2021), , doi:10.5281/ZENODO.4724043.
63. W. E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.* **8**, 118–127 (2007).
64. N. Rappoport et al., Comparing Ethnicity-Specific Reference Intervals for Clinical Laboratory Tests from EHR Data. *J Appl Lab Med.* **3**, 366–377 (2018).
65. P. Virtanen et al., SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods.* **17**, 261–272 (2020).
66. D. Zemmour et al., Single-cell gene expression reveals a landscape of regulatory T cell phenotypes shaped by the TCR. *Nat. Immunol.* **19**, 291–301 (2018).

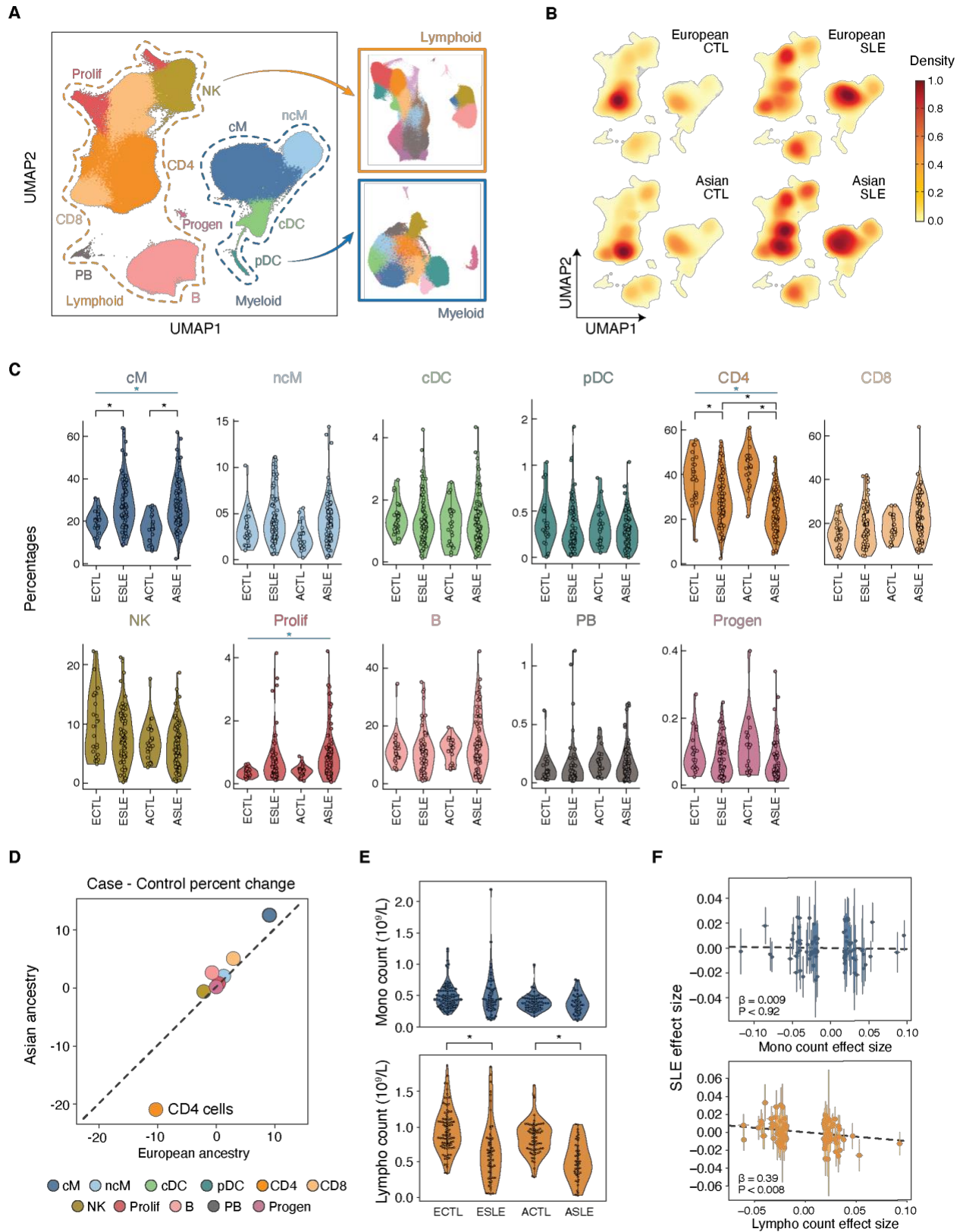
67. M. Aringer et al., 2019 European League Against Rheumatism/American College of Rheumatology classification criteria for systemic lupus erythematosus. *Arthritis rheumatol.* 71, 1400–1412 (2019).
68. D. D. Gladman, D. Ibañez, M. B. Urowitz, Systemic lupus erythematosus disease activity index 2000. *J. Rheumatol.* 29, 288–291 (2002).
69. S. Purcell et al., PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575 (2007).
70. S. Das et al., Next-generation genotype imputation service and methods. *Nat. Genet.* 48, 1284–1287 (2016).
71. R. J. Pruim et al., LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics.* 26, 2336–2337 (2010).
72. N. Mancuso et al., Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. *Am. J. Hum. Genet.* 100, 473–487 (2017).
73. C. Wallace, Statistical testing of shared genetic control for potentially related traits. *Genet. Epidemiol.* 37, 802–813 (2013).
74. 1000 Genomes Project Consortium et al., A global reference for human genetic variation. *Nature.* 526, 68–74 (2015).



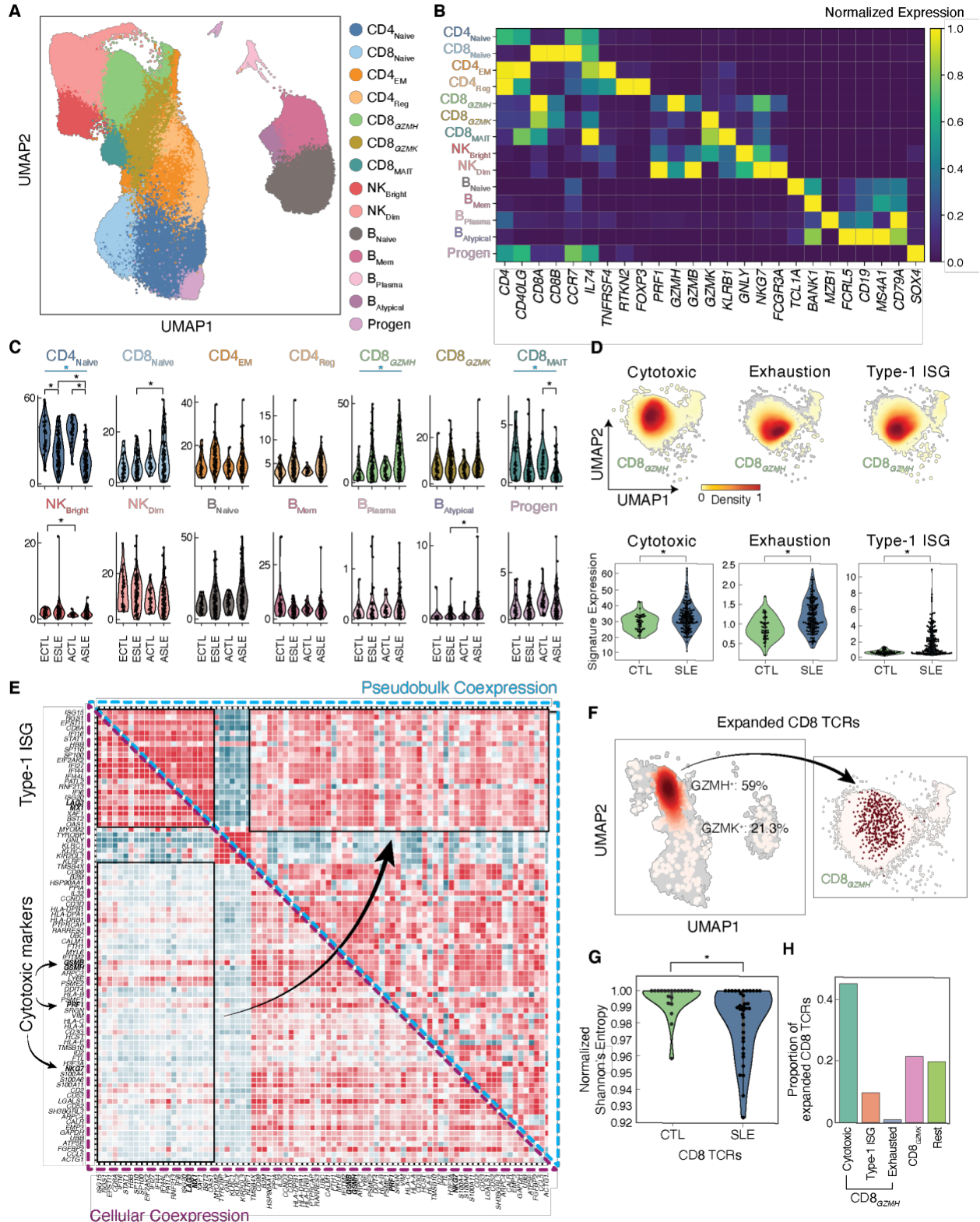
## 2.8 Acknowledgments:

We thank all members of the Ye lab for discussions. **Funding:** C.J.Y, L.A.C, and J.Y. are supported by NIH P30AR070155. C.J.Y is supported by NIH R01AR071522, U01HG012192, R21AI133337, and CZI P0535277. L.A.C and J.Y are supported by CDC U01DP005120, and L.A.C is supported by the Lupus Research Alliance. N.Z. is supported by NIH K25HL121295, U01HG009080, R01HG006399, R01CA227237, R03DE025665, and DoD W81XWH-16-2-0018. L.F. is supported by NIH R01CA194511 and R01CA223484. A.R. is supported by the Manton Foundation, Klarman Cell Observatory and Howard Hughes Medical Institute. M.G.G is supported by NIH 1F31HG011007. G.C.H. was supported by NSF under GRFP 1650113. M.T. Is supported by NIH T32HG002536. **Author Contributions:** M.S, G.H, S.T, and L.M performed all experiments. R.K.P, M.G.G., M.S, and C.J.Y. wrote the manuscript. R.K.P., M.G.G., and C.J.Y. revised the manuscript. R.K.P and C.J.Y performed all preprocessing, cell-type annotations, single-cell analysis, pseudobulk, DE analysis, and clinical predictions. M.C.K performed the trajectory analysis and RNA velocity. N.R. performed the UCSF EHR database queries. M.G.G performed the mendelian randomization analysis, and eQTL analysis. S.K and T.L performed the TCR sequencing experiments and R.K.P. performed the TCR analysis. M.M, M.C, L.T, C.L, M.D, J.Y, and L.A.C provided access to CLUES samples and all patient information. M.G.G., B.B., A.L., M.T., N.Z. developed and implemented the decomposition and modified TWAS methods. A.D. aided heritability and subtype analyses. M.S, J.W, D.D, and O.R performed the Immvar sequencing experiments. M.D, A.R, J.Y, L.A.C, and N.A.Z provided critical edits and feedback to the manuscript. **Competing interests:** A.R. is a co-founder and equity holder of Celsius Therapeutics, an equity holder in Immunitas, and was an SAB member of ThermoFisher Scientific, Syros Pharmaceuticals, Neogene Therapeutics and Asimov until July 31, 2020. From August 1, 2020, A.R. is an employee of Genentech. ORR is a co-inventor on patent applications filed at the Broad related to single cell genomics. ORR has given numerous lectures about single cell genomics to a wide variety of audiences and, in some cases, has received remuneration to

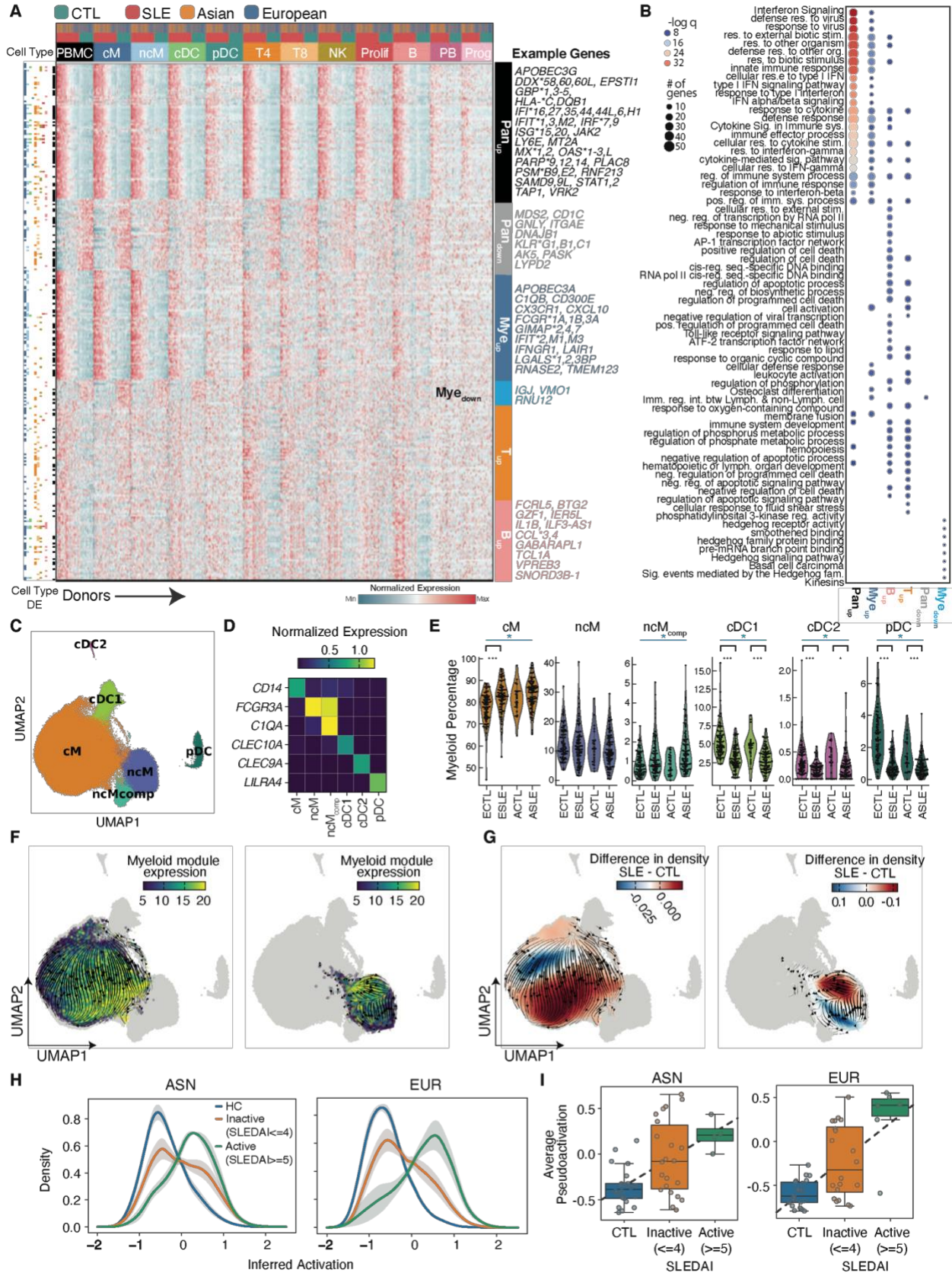
cover time and costs. C.J.Y. is a SAB member for and hold equity in Related Sciences and ImmunAI, a consultant for and hold equity in Maze Therapeutics, and a consultant for Trex Bio. C.J.Y. has received research support from Chan Zuckerberg Initiative, Chan Zuckerberg Biohub, and Genentech. **Data and materials availability:** All data is available in the Human Cell Atlas Data Coordination Platform and at the GEO Accession Number GSE174188. Code is available at [10.5281/zenodo.4724043](https://doi.org/10.5281/zenodo.4724043) (62)



**Figure 2.1: Changes in the composition of circulating immune cells in SLE.** A) UMAP and assignment of 1.2M cells to 11 cell types: classical and non-classical monocytes (cM and ncM); conventional and plasmacytoid dendritic cells (cDC and pDC); CD4+ and CD8+ T cells (CD4 and CD8); natural killer cells (NK); B cells (B); plasmablasts (PB); proliferating lymphocytes (Prolif); CD34+ progenitors (Progen). Sub-clustering of lymphoid (orange box) and myeloid (blue box) populations. B) Cell density plots of cases and controls separated by ethnicity. C) Percentage (y-axis) vs cases-control status (x-axis) for each cell type separated by ethnicity. Cell types with significant percentage changes between cases and controls are highlighted (black bar and star \*: WLS Padjusted < 0.05; blue bar and star indicate significant meta-analysis by Fisher's method). D) Correlation in percentage change versus controls between European (x-axis) and Asian (y-axis) cases. E) Monocyte (top) and lymphocyte (bottom) abundances (y-axes) vs case-control status (x-axis) from the UCSF EHR. Significant differences between cases and controls are highlighted (\*: OLS Padjusted < 0.05). F) Scatter plot of effect sizes on SLE status (y-axis) vs effect sizes on monocyte (top) or lymphocyte (bottom) abundance (x-axes) for genetic variants associated with both traits reported (4, 17). ECTL: European control; ESLE: European case; ACTL: Asian control; ASLE: Asian case.

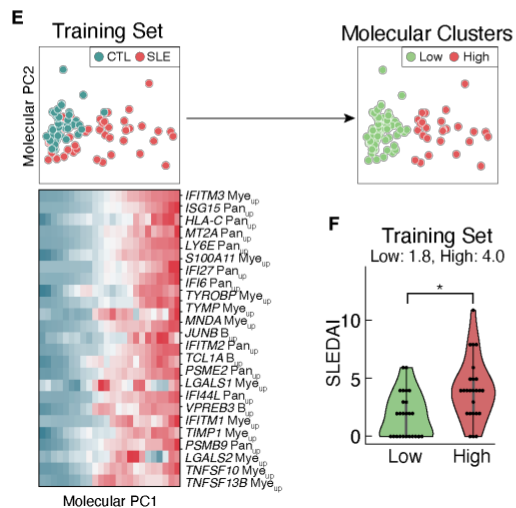
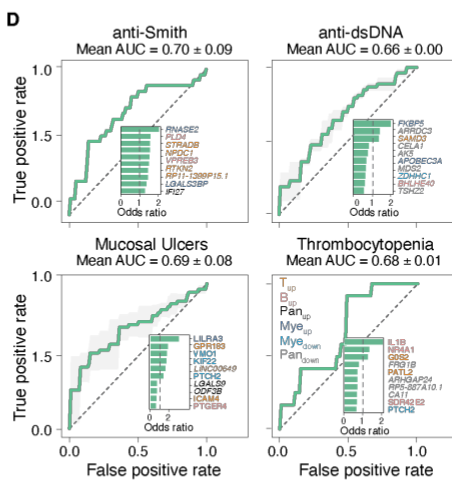
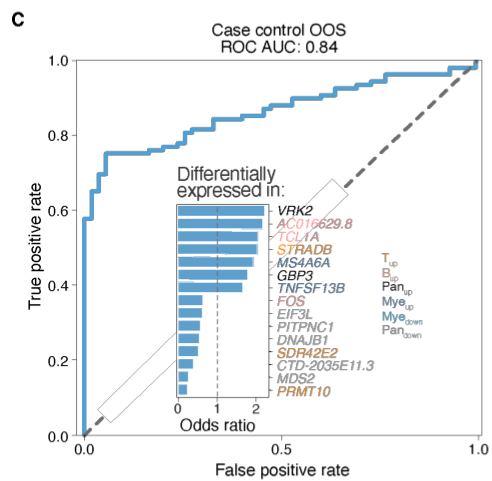
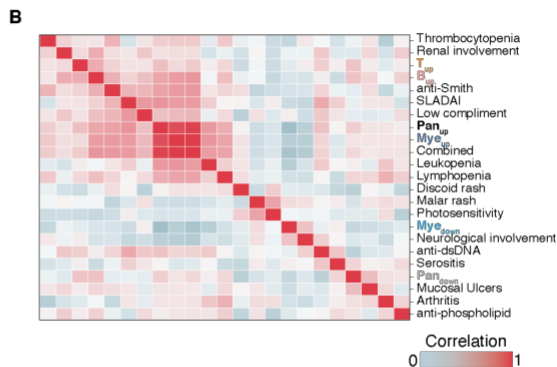
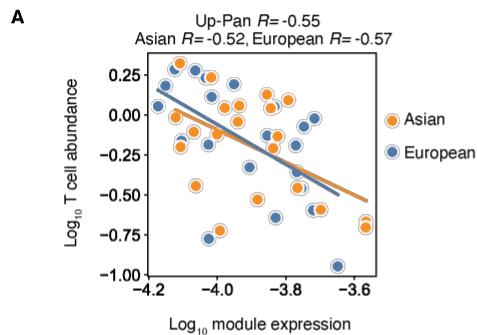


**Figure 2.2: Reduction of naïve CD4+ and expansion of cytotoxic CD8+ T cells in SLE. A)** UMAP of lymphoid cells re-clustered into 14 subpopulations: naïve, effector memory and regulatory CD4<sup>+</sup> T cells (CD4<sub>Naive</sub>, CD4<sub>EM</sub>, CD4<sub>Reg</sub>); naïve, *GZMH*<sup>+</sup> cytotoxic, *GZMK*<sup>+</sup> cytotoxic, and mucosal-associated invariant CD8<sup>+</sup> T cells (CD8<sub>Naive</sub>, CD8<sub>GZMH</sub>, CD8<sub>GZMK</sub>, CD8<sub>MAIT</sub>); CD56<sup>bright</sup> and CD56<sup>dim</sup> natural killer cells (NK<sub>bright</sub>, NK<sub>dim</sub>); naïve, memory, plasma and atypical B cells (B<sub>Naive</sub>, B<sub>Mem</sub>, B<sub>Plasma</sub>, B<sub>Atypical</sub>); Progen: CD34<sup>+</sup> progenitors. **B)** Expression of marker genes (columns) used to annotate each subpopulation (rows) colored by normalized expression levels. **C)** Percentage (y-axis) vs case-control status (x-axis) for each lymphoid subpopulation separated by ethnicity. Subpopulations with significant percentage changes between cases and controls are highlighted (black bar and star \*: WLS  $P_{\text{adjusted}} < 0.05$ ; blue bar and star indicate significant meta-analysis by Fisher's method). **D)** Density plot showing average expression of cytotoxic, exhaustion, and type-1 interferon stimulated gene (ISG) signatures in CD8<sub>GZMH</sub> cells (top) and across individuals (bottom) separated by case-control status and ethnicity (black bar and star \*: WLS  $P < 0.05$ ). **E)** Co-expression of top 300 differentially expressed genes between cases and controls in CD8<sub>GZMH</sub> cells computed across single cells (lower triangular matrix) or across donor-specific pseudobulk expression profiles (upper triangular matrix). **F)** All (light pink) and expanded (red) TCR sequences detected shown on UMAP of all cells (left) and *GZMH*<sup>+</sup> cells (right). **G)** Normalized Shannon's Entropies of CD8<sup>+</sup> TCR repertoire diversity (y-axis) in cases and controls (x-axis) (black bar and star \*: WLS  $P < 0.05$ ). **H)** Percentage of expanded CD8<sup>+</sup> TCRs identified as *GZMH*<sup>+</sup> cells expressing cytotoxic, ISG, and exhaustion signatures, *GZMK*<sup>+</sup> cells (GZMK), and all other cells (Rest). ECTL: European control; ESLE: European case; ACTL: Asian control; ASLE: Asian case.



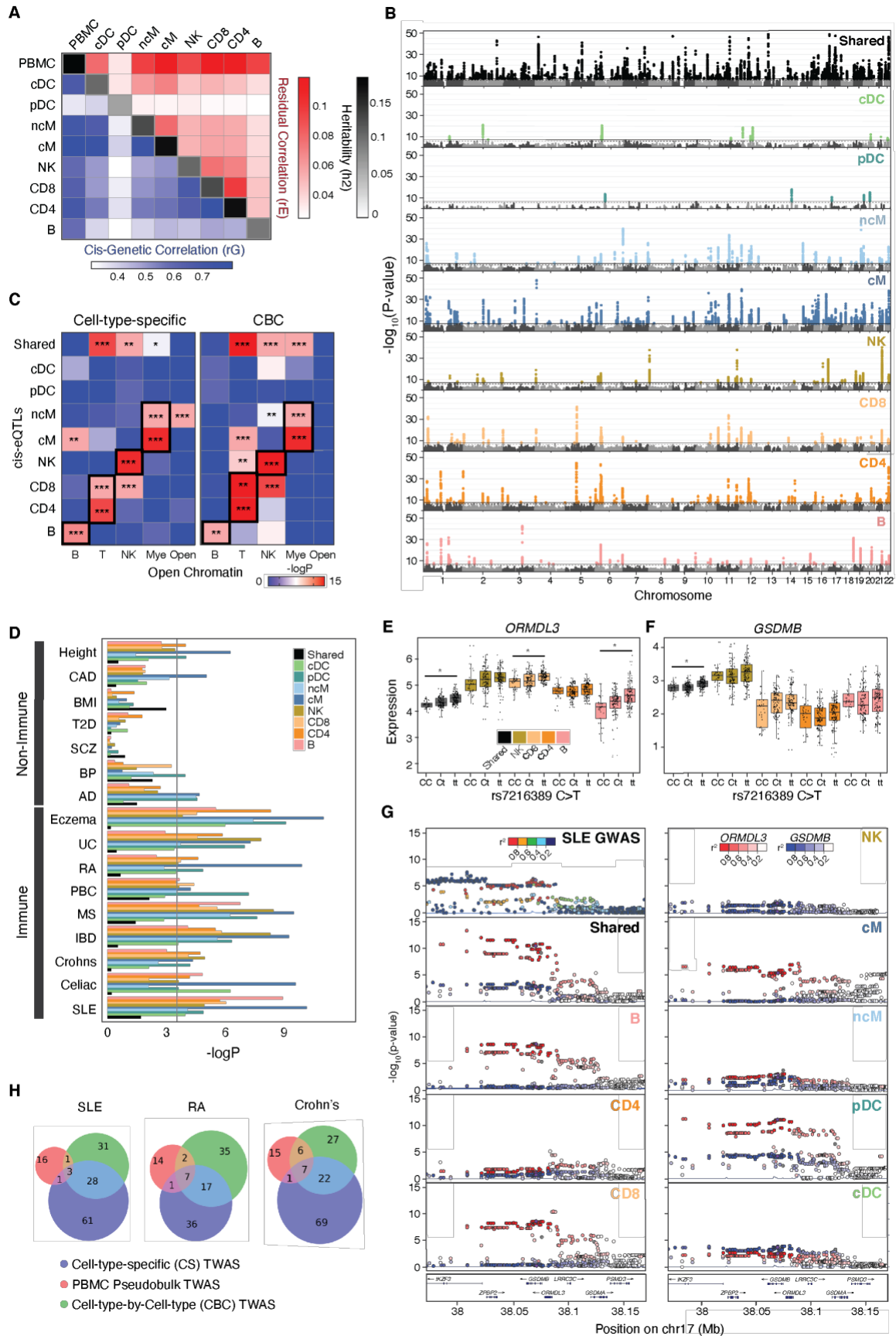
**Figure 2.3: Type-1 interferon response of myeloid cells in SLE.** **A)** Heatmap of pseudobulk gene expression profiles of 302 differentially expressed genes detected in at least one of 11 cell types. For each gene, colored row bars indicate cell types it was differentially expressed in. Colored columns indicate cell type, case-control status, and ethnicity. Labeled modules were identified using hierarchical clustering. **B)** Top GSEA pathway enrichment results for each module. Each dot color represents the  $-\log q$  value and the size represents the number of genes overlapping with the gene ontology. **C)** Identification of six myeloid cell types including classical, non-classical, and complement-expressing non-classical monocytes (cM, ncM, ncM<sub>comp</sub>), conventional type 1, conventional type 2, and plasmacytoid dendritic cells (cDC1, cDC2, pDC). **D)** Marker genes used for annotating each cell type. **E)** Percentages of myeloid cells (y-axis) vs case-control status and ethnicity (x-axis) for each myeloid subpopulation. Myeloid subpopulations with significant percentage changes between cases and controls are highlighted (black bar and star \*: WLS  $P < 0.01^*$ ,  $0.001^{**}$ ,  $0.0001^{***}$ ; blue bar and star indicate significant meta-analysis by Fisher's method). RNA velocity stream plots for cM (right UMAP), ncM and ncM<sub>comp</sub> (left UMAP) subpopulations colored by **F)** the average expression of  $\text{Mye}_{\text{up}}$  genes enriched for type-1 ISGs and **G)** the relative density of cells from SLE cases vs healthy controls. **H)** Distribution of the degree of inferred activation for individuals across disease activities (HC: healthy controls, Inactive: SLEDAI between 0 and 4, Active: SLEDAI greater than or equal to 5). **I)** The average inferred activation across cells per sample (y-axis) vs disease activity (x-axis) for Asians (left) and Europeans (right) separately. ECTL: European control; ESLE: European case; ACTL: Asian control; ASLE: Asian case.



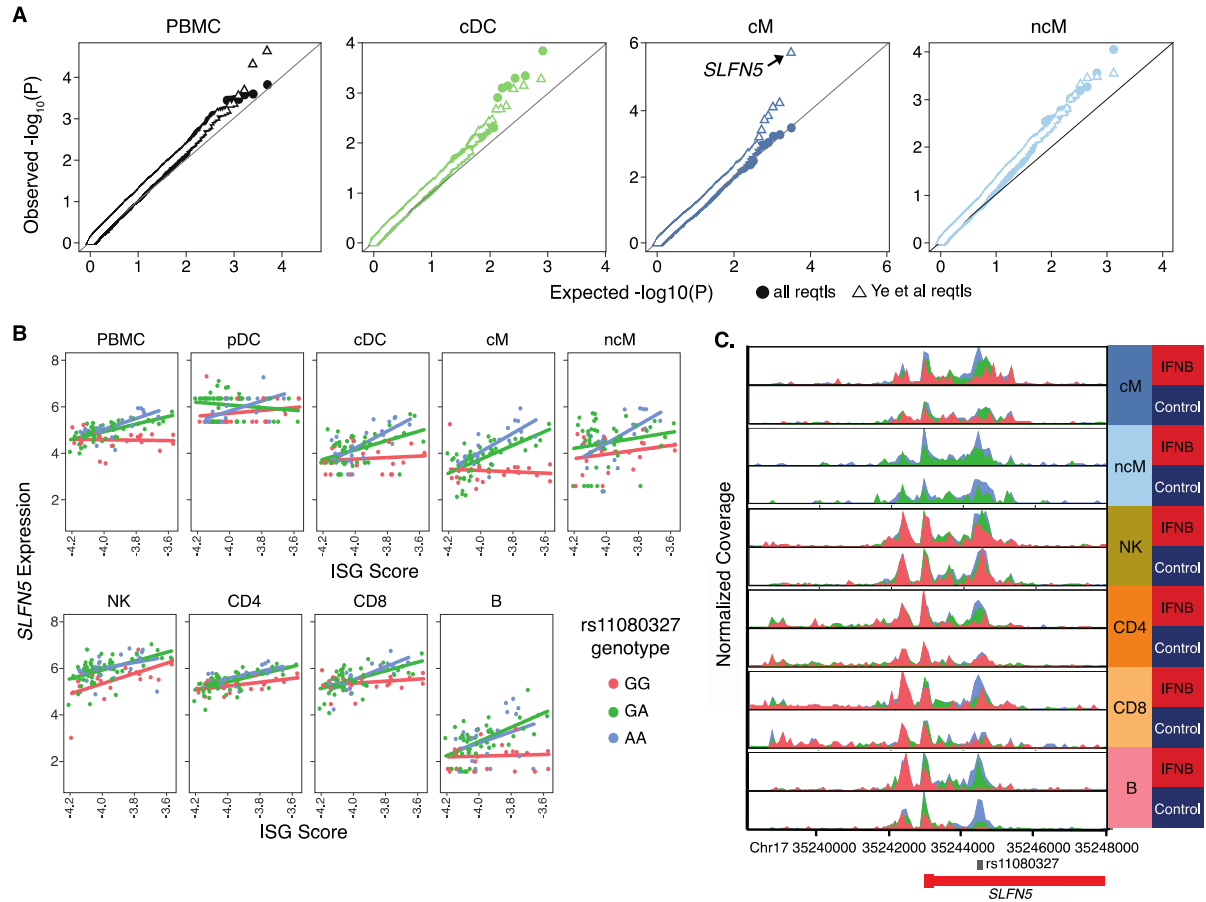


Clinical Feature	+		-	
	High	Low	High	Low
	21	29	5	38

**Figure 2.4: Prediction of disease status and molecular stratification of SLE.** **A)** Correlation between  $\log_{10}$  expression of  $\text{Pan}_{\text{up}}$  (x-axis) and  $\log_{10}$  abundance of  $\text{CD4}_{\text{naive}}$  cells in processing batch 4 cases only. **B)** Correlation matrix between average expression of each of six gene module and clinical feature. Receiver operating curve for out-of-sample prediction of **C)** case-control status and **D)** individual clinical variables using a logistic regression model trained on 302 expression features. Inset depicts the most important molecular features inferred by the model colored by the module each feature belongs to. **E)** Principal component analysis of training set based on 302 expression features. Green: Control, Red: Case. Heatmap shows the top 25 most correlated expression features to molecular PC1. Expression was binned and averaged across 24 equal steps across molecular PC1. K-means clustering of samples based on principal components into two molecular sub-phenotypes (Low, High). **F)** Distribution of SLEDAI scores (y-axis) for each molecular sub-phenotype (x-axis) in the training data (Wilcoxon rank-sums  $P < 0.05$ ). **G)** Projection of out-of-sample test set onto molecular PC1 and molecular PC2 and colored by case-control status (left) and molecular cluster membership (right). Heatmap shows the top 25 most correlated expression features to molecular PC1 in the test set. **H)** Odds ratio of having a clinical feature given membership to the High molecular cluster versus the Low molecular cluster.



**Figure 2.5: Cell-type-specific genetic determinants of gene expression.** **A)** *Cis* genetic correlation ( $r_G$ : lower triangular plot), shared residual correlation ( $r_E$ : upper triangular plot), and heritability ( $h^2$ : diagonal) of eight cell types and PBMCs. *Cis* is defined 100 kb within the TSS. **B)** Manhattan plots of shared- (sh-eQTL; black) and cell-type-specific-*cis*-eQTLs (cs-eQTL; colored) determined by mapping *cis*-eQTLs associated with shared and cell-type-specific expression components from decomposition analysis. Associations are reported as  $-\log_{10}(P\text{-value})$  (y-axis) ordered by chromosomes (x-axis). **C)** Enrichment of cs-eQTLs (left) and CBC-eQTLs (right) for disjoint sets of cell-type-specific regions of open chromatin. Mann-Whitney test  $P < 0.01^*$ ,  $0.001^{**}$ ,  $0.0001^{***}$ . **D)** Enrichment of shared or cs-eQTLs among GWAS associations for seven non-immune (CAD: coronary artery disease, BMI: body mass index; T2D: type-2 diabetes; SCZ: schizophrenia; BP: bipolar disease; AD: Alzheimer's disease) and nine immune-mediated diseases/traits (UC: ulcerative colitis; RA: rheumatoid arthritis; PBC: primary biliary cirrhosis; MS: multiple sclerosis; IBD: inflammatory bowel disease; SLE: systemic lupus erythematosus). Bonferroni corrected significance threshold shown as black line. Boxplots of decomposed shared- and cell-type-specific expression of **E)** *ORMDL3* and **F)** *GSDMB* in all individuals grouped by genotype for rs7216389 (\* COLOC Posterior Probability > 0.7). **G)** LocusZoom plots of SLE GWAS, sh-eQTLs, and cs-eQTLs associated with *ORMDL3* (red) and *GSDMB* (blue) expression. **H)** Number of associations identified by a modified transcriptome wide association analysis (TWAS) using decomposed shared and cell-type-specific expression matrices (blue), cell-type by cell-type expression matrices (green) or pseudobulk PBMCs (red).



**Figure 2.6: Interferon modifies cell-type-specific genetic effects on gene expression. A)** Quantile-quantile plot of expected  $-\log_{10}(P\text{-value})$  (x-axis) vs observed  $-\log_{10}(P\text{-value})$  (y-axis) of *cis*-IFN-QTLs (filled). Previously identified response-QTLs (reQTLs) from monocyte derived dendritic cells highlighted (unfilled). **B)** Normalized expression of *SLFN5* expression (y-axis) versus ISG score (x-axis) separated by rs11080327 genotype (color). Line indicates best linear regression fit for each genotype. **C)** Gene locus plot of *SLFN5* scATAC-seq peaks for six peripheral immune cell types in unstimulated and rIFN $\beta$ 1 stimulated conditions, separated by genotype. Location of rs11080327 is indicated.

## **Chapter 3: The Immune Cell Census: Multiplexed Multi-omics enables discovery of immune regulatory programs and genetic architecture of molecular traits.**

### **3.1 Abstract**

Over the last decade significant strides in single cell sequencing technologies have been achieved, enabling orders of magnitude increases in cell throughput as well as increasing the numbers of modalities that can be assayed simultaneously. We applied single cell multi-omics (ATAC+RNA) sequencing to profile over one million PBMCs across 400 diverse individuals. We investigated cell composition as well as cell type specific chromatin accessibility and gene expression. Further, we mapped cell type specific QTLs for chromatin accessibility and gene expression to assess the genetic architecture of these two molecular traits. The resulting data gives insight into immune regulatory programs and provides a reference dataset for the scientific community.

### **3.2 Introduction**

Recent advancements in throughput of single cell technologies have presented the opportunity to profile millions cells across many tissues, diseases, environmental contexts, and hundreds to thousands of donors<sup>1-4</sup>. Immunologists have made significant strides towards this goal, generating comprehensive cell atlases of blood by profiling peripheral mononuclear blood cells (PBMCs) with single cell RNA sequencing (scRNA-seq)<sup>5-7</sup>. These large cohorts have enabled population genetics to be applied to single cell genomics, resulting in mapping expression quantitative trait loci (eQTLs) in a variety of cell type and cell states.<sup>8</sup> These context specific eQTLs hold promise for better functional annotation of genetic loci associated with disease, which will help elucidate the cell types, cell states, and genes underlying disease signals<sup>9,10</sup>.

However, the number of modalities that can be simultaneously measured in a single cell continues to increase providing new opportunities to further our understanding of cell regulation circuitry. These include methods to simultaneously capture RNA and surface proteins, RNA and Assay for

Transposase-Accessible Chromatin (ATAC), ATAC and surface proteins, as well as all three modalities<sup>11-14</sup>. Methods for simultaneous capture of ATAC-seq and RNA-seq have recently been commercialized to perform robustly at scale, making them ideal for atlas efforts. This combination of modalities holds great potential to better dissect regulatory programs and link putative regulatory elements with their target genes. It also presents a unique opportunity to investigate the genetic architecture of chromatin accessibility, gene expression, and relationships between these two modalities.

One major shortcoming of many of these atlas efforts to date, is a lack of diversity reflective of the human population. Like many genetic studies, previous efforts have focused on individuals of European descent. These practices have potential to widen gaps in health disparities between populations as these references will be used for biomedical research, drug target discovery, and therapeutic development<sup>15</sup>. Additionally, many naturally occurring genetic variants will not be sampled by limiting these studies to small homogenous fractions of populations, leading to holes in our understanding of variant effects and genetic diseases<sup>16,17</sup>.

Here we present the Human Immune Cell Census. A cross-sectional cohort of ~400 donors of African (AFR), East Asian (EAS), European (EUR), and Latinx (AMR) descent. We profiled over 1 million single cells, simultaneously capturing measures of gene expression and chromatin accessibility with paired dense genotyping data. We used this data to evaluate cell composition, dissect regulatory programs, and investigate genetic architecture underlying gene expression and chromatin accessibility.

### **3.3 Results**

#### **3.3.1 Cell phenotyping of more than one million single cells.**

Multiplexed single cell multi-ome data was generated from approximately 400 donors and genotyped in parallel. Genotyping data was used to identify and remove droplets containing multiple cells, then to assign the remaining single cells to their donors of origin (Fig1A). Standard workflows were applied to remove low quality cells based on both the ATAC and RNA fractions of the data. Dimensionality reduction was performed on each modality separately, as well as jointly. Cell types were identified based on Leiden clustering, marker genes, and motifs. While all three projections had sufficient resolution to capture the major cell groups, the joint projection maximized distances between lineages while retaining resolution in closely related cell types (Fig B-D).

#### **3.3.2 Cell composition**

Cell counts and proportions are often used as a diagnostic tool in the clinic. However, these tools are often limited to major cell classifications such as B, T, and Myeloid cells. Flow sorting and CyTOF can be used to quantify more fine-grained cell types, however the number of surface proteins that can be profiled are limited to tens or one hundred respectively<sup>18,19</sup>. The throughput of these methods is limited, as each sample must be run as a separate experiment, where single cell sequencing methods are parallelizable. Here we quantified composition differences in 8 cell types (B, CD4T, CD8T, NK, cM, ncM, cDC, and pDC). Across the donors, significant variation in composition was observed (Fig1E). To evaluate factors that may influence compositional changes we first visualized changes in composition across the sampled populations. Subtle decreases in CD4T cells in EAS and B cells in EUR are observed (Fig1F) when compared to the other 3 groups together, however they are confounded by technical effects. To quantify the factors underlying variance in composition, principal components (PCs) of cell composition per donor were calculated and a multivariate linear model was fit to evaluate the effect of meta information of donors as well as technical factors from the single cell experiments. Across the first 4 PCs, very



little variance was explained by any one of these factors, with less than 20% of the variance explained by the sum of all factors evaluated (Fig1G). This indicates that while there may be trends in differences of proportions between groups, it does not explain much if the variance indicating intraindividual differences contribute to the vast majority variance in cell composition.

### **3.3.3 Networks**

Gene expression is controlled by gene regulatory networks, where chromatin remodelers open regions of chromatin around regulatory regions and their target genes, transcription factors bind to these regions and promote the transcription of the target gene. To investigate these networks, measures of open chromatin and gene expression in single cells can be leveraged to discover novel, cell type specific cis regulatory elements and to better predict the genes they act on. First, over 100K peaks were called. Most peaks fall into the intronic regions of the genome, but appreciable numbers are also located in promoters and distal regions (Fig2A). This is to be expected as regulatory elements are in non-coding regions of the genome. As regulatory elements often are cell type specific, all called peaks were evaluated for cell type specificity using differential accessibility analysis of each cell type against the rest. Many of the peaks were found to be differentially accessible ( $\log_2FC > 0.5$ ,  $FDR < 0.05$ ) in the cell types examined, while many shared peaks were only shared with closely related cell types, for example classical and non-classical monocytes (Fig2B). To evaluate which genes putative regulatory elements act on, correlation between peaks and gene expression were calculated. These correlated peaks were mostly located in intronic regions of the gene (blue), or upstream reflecting promoters or putative enhancers which was observed in most cell types, here shown in NK cells (Fig2C). Peak correlations to the NK marker gene SMAD7 in NK cells were very high ( $> 0.8$ ) for four regions around the gene. These regions highlighted in the user track, overlapped H3K27ac and clusters of ENCODE annotated cis-regulatory elements suggesting that these region house important regulatory elements for SMAD7 in NK cells (Fig2D).

### 3.3.4 Genetics

While GWAS have associated thousands of loci with hundreds of complex traits, most of these loci are in non-coding regions. This has posed a challenge in the field, as functional mechanisms are difficult to untangle when they do not directly change protein sequences. Additionally, GWAS do not provide any insight into cell contexts in which these loci are important<sup>20</sup>. Understanding the genetic architecture of molecular traits has potential to elucidate the context and genes through which these loci may act, which can help determine mechanisms underlying complex diseases. Here we mapped eQTLs and atacQTLs in 8 cell types (B, CD4+T, CD8+T, NK, cM, ncM, cDC, and pDC). While all peaks identified were tested, most of the significant atacQTLs ( $p < 5 \times 10^{-8}$ ) were between non-coding SNPs and non-coding peaks, though an appreciable number of peaks did intersect with exons (~26-35%). First enrichment of atacQTLs and eQTLs were calculated in cell type specific peaks (Fig3A-B). atacQTLs were not enriched in their cell type specific peaks, where eQTLs were. While the atacQTL result was unexpected, it could suggest that SNPs which control chromatin accessibility may not always reside in peaks that are unique to the specific cell type and may rather fall into shared peaks. The eQTL results suggest that SNPs have the largest effects on expression when they reside in open chromatin. Further follow up analyses are required to better understand this result. Additionally, to evaluate which cell types are most important in immunological diseases, LD score regression was used to find enrichment of GWAS traits for immunological diseases (Ulcerative Colitis, Rheumatoid Arthritis, Primary Biliary Cirrhosis, IBD, Crohn's, Celiac, and Lupus) in cell type specific ATAC peaks (Fig3C). Particularly interesting signals from these results include Primary Biliary Cirrhosis, B cell peaks had the highest and most significant enrichment ( $p\text{-value} < 0.001$ ). Multiple studies in mice have shown that B cells play an important role in managing inflammation in the disease<sup>21</sup>.

### 3.4 Discussion

Single cell sequencing of population scale cohorts holds great promise to further our understanding of molecular mechanisms and how they vary between donors and fine-grained cell

types in health and disease. Profiling healthy PBMCs across large diverse populations is critical to further our understanding of baseline activity of immune cells. Here, we profiled over 1 million cells across four diverse human populations. We found that cell composition did not differ significantly across populations, but significant variation was observed across donors. We identified thousands of cell type specific peaks that were in regions of the genome that house regulatory elements. We leveraged multi-omic measurements to pair putative regulatory elements with their target genes and observed that in SMAD7, a NK cell marker, associated peaks overlapped relevant enhancer and promoter markers. Lastly, we mapped atacQTLs and eQTLs, finding relevant eQTLs were enriched in marker peaks and found GWAS enrichment in disease relevant cell type specific peaks. This rich dataset has potential for deeper analysis and will serve as a reference for many future studies.

### **3.5 Methods**

#### **3.5.1 Sample collection**

Primary blood samples were collected from healthy donors at two collection sites: UCSF and CUMC. Inclusion criteria for participants included no history of immunological disease or cancer. No sample was collected until 3 weeks after any vaccination or acute illness. Participants were between 18 and 55 years of age and were part of four self-reported ethnicities: AFR, EAS, EUR, AMR. 50mL of blood was collected and PBMCs were isolated with SepMate-50 tubes following manufacturer's user guide then cryopreserved.

#### **3.5.2 Genotyping**

Genotyping was performed on Illumina's MEGA (multi-ethnic global array) chip in three batches and processed at the Berkeley Genomics Core. Plink 1.9 was applied to merge and filter genotype data. For each sample 1,748,250 sites were genotyped and mapped to grch38. Individuals with more than 5% missing SNP calls were removed. Resulting genotypes were filtered by Missingness ( $< 0.02$ ), minor allele frequency (MAF  $> 0.01$ ), and Hardy-Weinberg equilibrium ( $p >$

1e-4). KING was applied to estimate cryptic relatedness between resulting individuals where one of each pairs of individuals with third-degree relatedness or higher was removed (KING coefficient > 0.0442). The resulting 894294 variants from 347 individuals were used for imputation with the TOPMed Imputation Server using the TOPMed r2 reference panel. Eagle 2.4 was applied to phase imputed genomes. Sites with an imputation  $R^2 > 0.3$  were retained for downstream analyses.

### **3.5.3 Single Cell Sequencing**

For each processing pool,  $10^6$  cells from 20 donors each were first used for nuclei isolation (if you don't have a nuclei isolation step, we need to write more).  $10^5$  nuclei from 20 donors each were pooled and loaded into 4 wells of 10X Genomics Single Cell Multiome ATAC + Gene Expression kit. Briefly, cells from each donor were thawed at 37C water bath, then cell counts and viability were determined with Cellaca MX High-Throughput Automated Cell Counter. Donors with cell viabilities less than 65% were excluded. Nuclei for each donor were isolated and counted again on the Cellaca. Nuclei were equally pooled between donors, filtered using Corning 40um cell filter, then single nuclei were loaded into 4 wells at  $5 \times 10^4$  nuclei per well and run on the Chromium controller for single nuclei capture. Resulting libraries were sequenced at UCSF's Center for Advanced Technology (CAT) core at a depth of  $4 \times 10^4$  reads for ATAC and  $2 \times 10^4$  reads for RNA using S4 flowcells on Illumina's Novaseq Platform. 50x10x24x90 cycles used for Read 1:i7:i5:Read 2 in the RNA fraction, while the ATAC fraction used 50x10x24x90 for Read 1:i7:i5:Read 2.

### **3.5.4 Single Cell Analysis**

Single cell data was processed using cellranger-arc 2.0, then demultiplexed using demuxlet on the ATAC fraction of the data. Downstream processing of the RNA fraction was completed with Scanpy and processing of the ATAC fraction was completed with ArchR. Cobolt was used for joint projections. Leiden clustering was applied to the data then marker genes were utilized to determine cell types in the clusters.

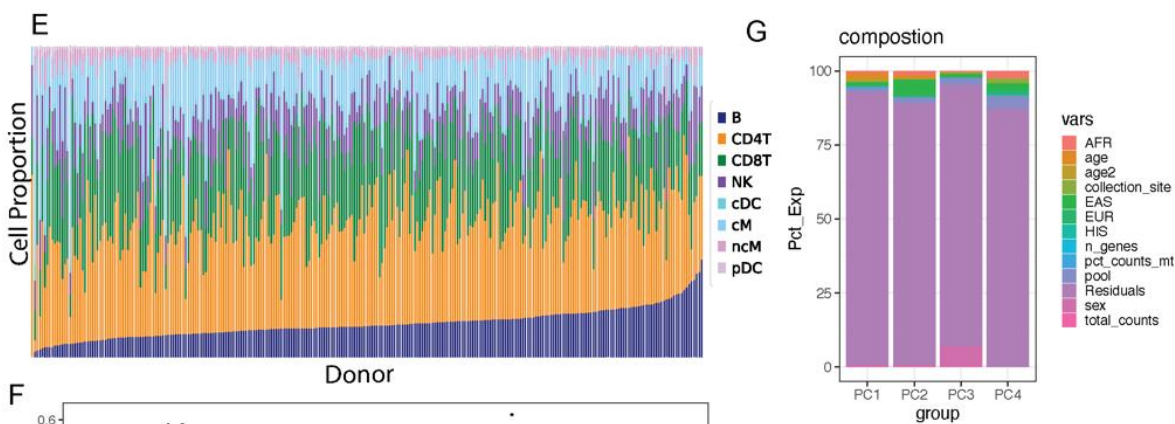
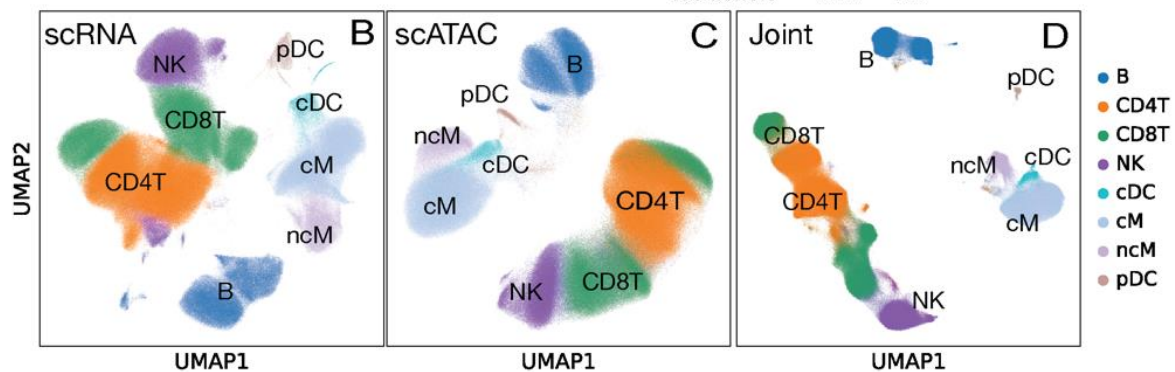
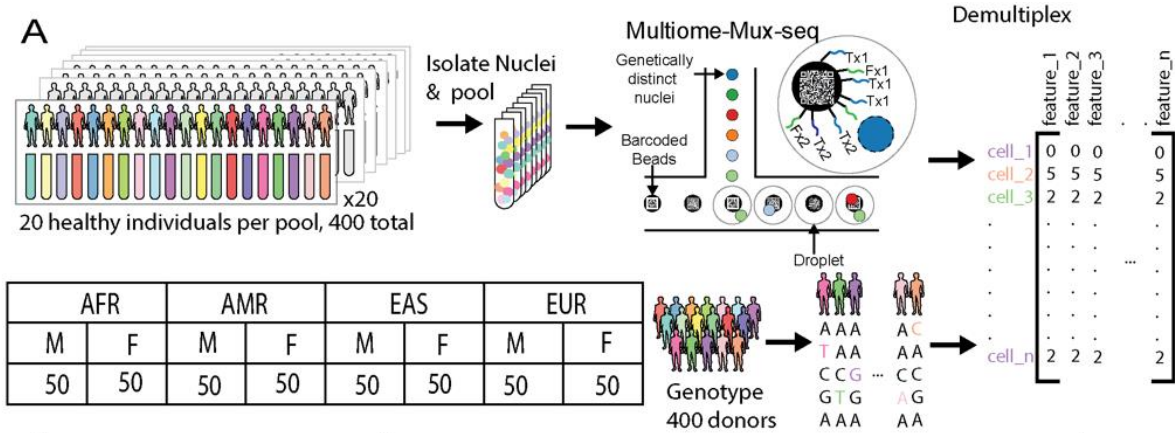
### **3.5.5 QTL calling**

For each cell type profiled, pseudobulk matrices were generated from the single cell ATAC and RNA data turning a cell by feature matrix into a donor by feature pseudo-count matrix for each cell type. Log counts per million were calculated then standardized for each Pseudo-count matrix with EdgeR. For each feature a linear model was for expression of each feature using SNPs within 100kb of the genomic feature tested using matrixeQTL. For each model pool, age, sex, 10 genotype PCs and 10 phenotype PCs were included. Each population was processed separately, then meta-analyzed using Metasoft to control for population structure.

### 3.6 References

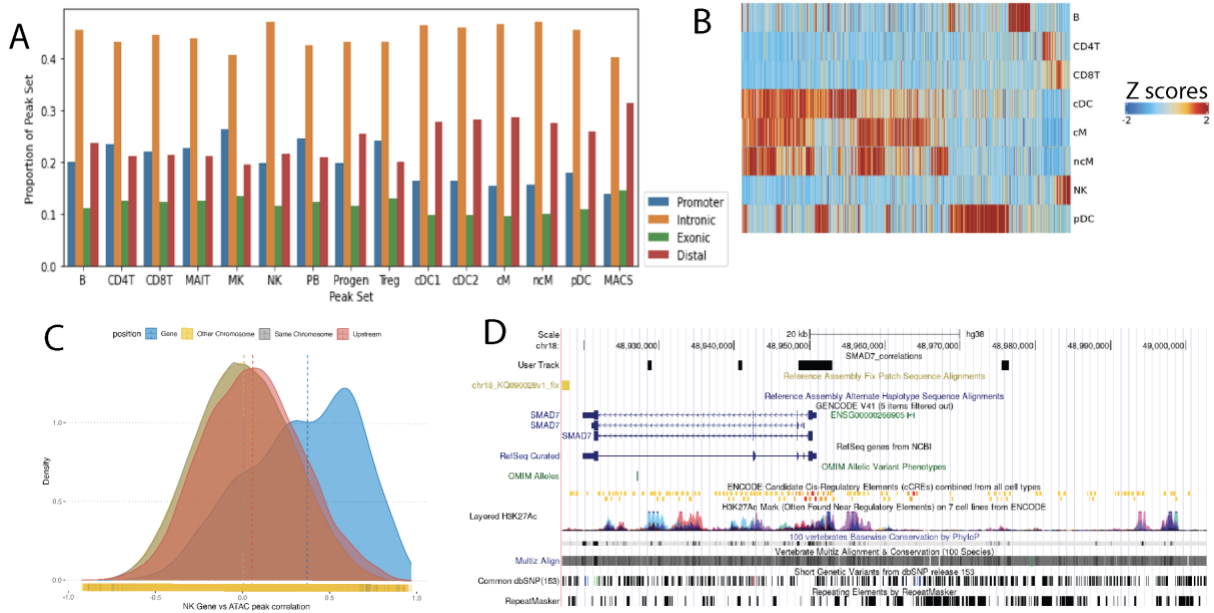
1. Rozenblatt-Rosen, O. *et al.* Building a high-quality Human Cell Atlas. *Nat. Biotechnol.* **39**, 149–153 (2021).
2. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
3. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
4. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
5. Yazar, S. *et al.* Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease. *Science* **376**, eabf3041 (2022).
6. Perez, R. K. *et al.* Single-cell RNA-seq reveals cell type–specific molecular and genetic associations to lupus. *Science* **376**, eabf1970 (2022).
7. Oelen, R. *et al.* Single-cell RNA-sequencing of peripheral blood mononuclear cells reveals widespread, context-specific gene expression regulation upon pathogenic exposure. *Nat. Commun.* **13**, 3267 (2022).
8. Sumida, T. S. & Hafler, D. A. Population genetics meets single-cell sequencing. *Science (New York, N.Y.)* vol. 376 134–135 (2022).
9. Westra, H.-J. & Franke, L. From genome to function by studying eQTLs. *Biochim. Biophys. Acta* **1842**, 1896–1902 (2014).
10. van der Wijst, M. *et al.* The single-cell eQTLGen consortium. *Elife* **9**, (2020).
11. Gaublotte, J. T. *et al.* Nuclei multiplexing with barcoded antibodies for single-nucleus genomics. *Nat. Commun.* **10**, 2907 (2019).
12. Lyu, M. *et al.* TEAseq-based identification of 35,696 Dissociation insertional mutations facilitates functional genomic studies in maize. *J. Genet. Genomics* **48**, 961–971 (2021).

13. Mimitou, E. P. *et al.* Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nat. Methods* **16**, 409–412 (2019).
14. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **37**, 1452–1457 (2019).
15. Oh, S. S. *et al.* Diversity in clinical and biomedical research: A promise yet to be fulfilled. *PLoS Med.* **12**, e1001918 (2015).
16. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell* **177**, 1080 (2019).
17. Sherman, R. M. *et al.* Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* **51**, 30–35 (2019).
18. Cheung, R. K. & Utz, P. J. Screening: CyTOF-the next generation of cell detection. *Nat. Rev. Rheumatol.* **7**, 502–503 (2011).
19. Kalisky, T., Blainey, P. & Quake, S. R. Genomic analysis at the single-cell level. *Annu. Rev. Genet.* **45**, 431–445 (2011).
20. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).
21. Moritoki, Y. *et al.* B cells suppress the inflammatory response in a mouse model of primary biliary cirrhosis. *Gastroenterology* **136**, 1037–1047 (2009).

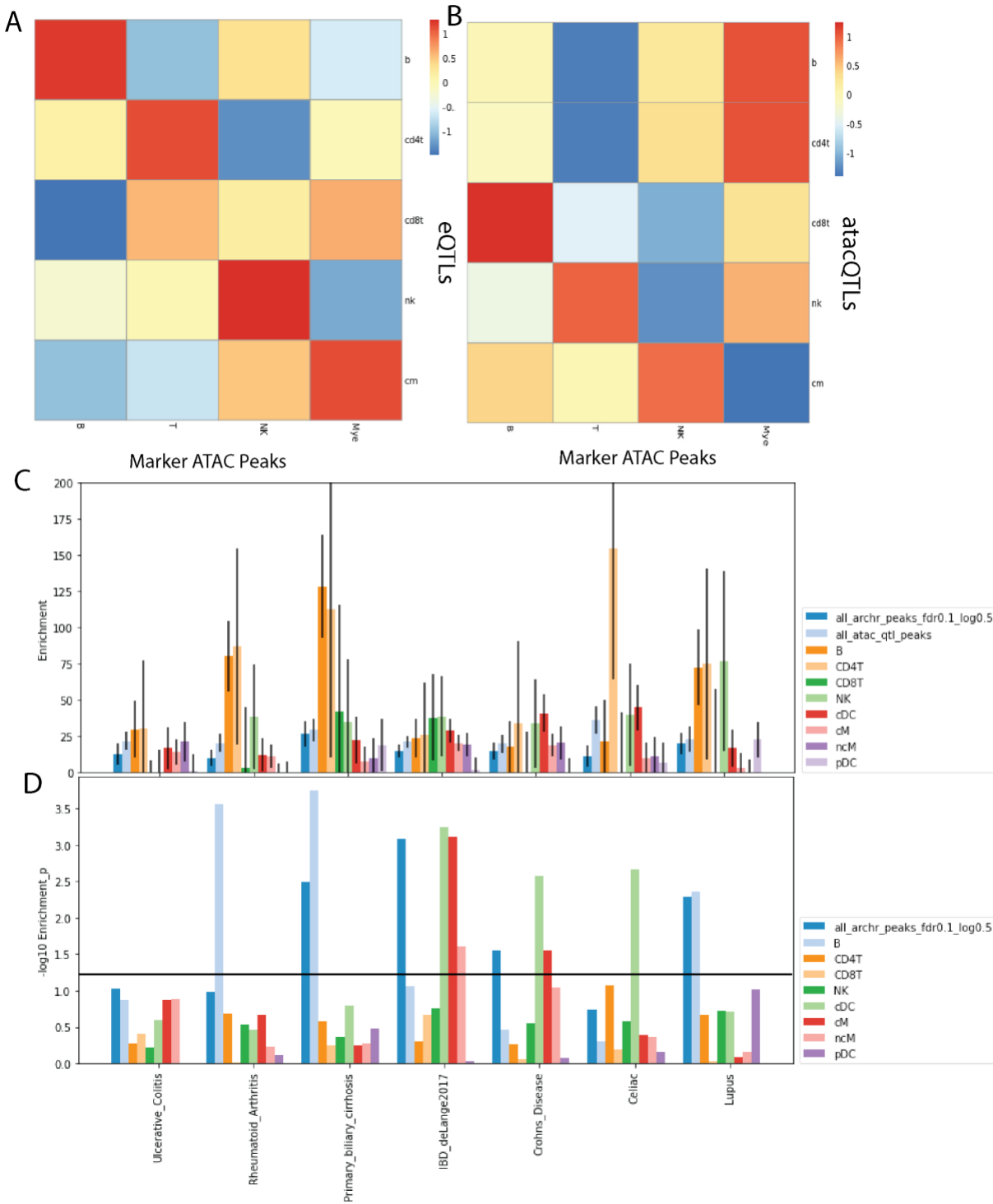




**Figure 3.1: Experimental design, cell phenotyping, and composition.** A) Schematic of experimental design. UMAP projections calculated from snRNA data (B), snATAC data (C) and jointly (D) across both simultaneously assayed modalities. Marker genes from snRNA seq were used to call cell types annotated in these UMAPs. (E) Proportion of each cell type per donor. (F) Box plots of composition distribution per self-reported ethnicity. (G) Percent variance explained for each cell composition PC of composition for measured meta data information.



**Figure 3.2: Evaluating Networks of Genes.** (A) Proportion of peaks in different regions of the genome. (B) Cell type specific peaks. (C) Distribution of distance between gene and peak links colored by genome annotation; in gene (blue), upstream (red), same chromosome (grey), other chromosomes (yellow). (D) SMAD7 locus. User track shows peaks significantly correlated to expression of SMAD7. Encode Tracks and Histone enhancer marks displayed.



**Figure 3.3: Molecular trait genetics. Mann-Whitney U test for enrichment of eQTLs (A) and atacQTLs (B) in simplified cell type specific peaks. (C) LDSC enrichment of GWAS signal in cell type specific peaks. (D) LDSC enrichment p-values in cell type specific peaks. Black line shows  $p=0.05$ .**

## **Chapter 4: lentiMPRA & MPRAflow for high-throughput functional characterization of gene regulatory elements.**

### **4.1 Abstract**

Massively Parallel Reporter Assays (MPRAs) can simultaneously measure the function of thousands of candidate regulatory sequences (CRSs) in a quantitative manner. In this method, CRSs are cloned upstream of a minimal promoter and reporter gene alongside a unique barcode and introduced into cells. If the CRS is a functional regulatory element, it will lead to the transcription of the barcode sequence, which is measured via RNA sequencing and normalized for cellular integration via DNA sequencing of the barcode. This technology has been used to test thousands of sequences and their variants for regulatory activity, decipher the regulatory code and its evolution, and for the development of genetic switches. Lentivirus-based MPRA (lentiMPRA) produces 'in genome' readouts and allows the use of this technique in hard to transfect cells. Here, we provide a detailed protocol for lentiMPRA along with a user-friendly Nextflow-based computational pipeline, MPRAflow, for quantifying CRS activity from different MPRA designs. The lentiMPRA protocol takes approximately two months, which includes sequencing turnaround time and data processing with MPRAflow.

### **4.2 Introduction**

Gene regulatory elements control a gene's transcription. These include sequences that activate transcription such as promoters and enhancers, silencers that repress a gene, or insulators that restrict genes from interacting with certain regulatory elements. Nucleotide variation in these elements can have a major effect on phenotype. Mutations within them have been shown to be a major cause of human disease<sup>1</sup>. For example, over 90% of all human disease genome-wide association studies (GWAS) have shown associations with noncoding variants<sup>2</sup> and colocalize

with potential gene regulatory elements<sup>3</sup>. In addition, gene regulatory elements can be major drivers of evolutionary speciation, driving differences between species such as morphology, diet, and behavior<sup>4</sup>. These sequences can also be used as genetic switches to tune transgenes to specific levels in certain cell types or tissues.

In this protocol, we focus on gene activation associated regulatory elements, promoters and enhancers. These sequences can be identified in a genome-wide manner by biochemical methods such as chromatin immunoprecipitation followed by sequencing (ChIP-seq<sup>5</sup>), DNase I hypersensitive sites sequencing (DNase-seq<sup>6,7</sup>), assay for transposase-accessible chromatin using sequencing (ATAC-seq<sup>8</sup>), cleavage under targets and release using nuclease (CUT&RUN<sup>9</sup>), Hi-C<sup>10</sup> and others. However, these methods only help annotate candidate regulatory sequences (CRSs), and additional experimental assays must be performed in order to validate their predicted activity. Reporter assays are commonly used to characterize CRS. In this assay, the CRS is placed either upstream of a reporter gene (i.e., in the case of testing promoters) or upstream of a minimal promoter followed by a reporter gene (i.e., in the case of testing enhancers). If the sequence is an activating regulatory element, it will turn on the reporter gene, providing a measurable output. However, these assays are primarily done on an individual basis and as such cannot assess the thousands of CRSs and their variants that have been identified via the aforementioned biochemical assays. Massively parallel reporter assays (MPRAs) overcome this hurdle, providing the ability to test hundreds of thousands of sequences and their variants in parallel for their regulatory function<sup>11</sup>. This is done either by measuring RNA expression driven by the CRS by pairing it to a transcribed barcode, or by using the CRS itself as a barcode, as done in the self-transcribing active regulatory region sequencing (STARR-seq<sup>12</sup>) assay.

Here, we describe both a lentivirus-based MPRA (lentiMPRA) and MPRAflow, a computational tool for MPRA analysis based on the Nextflow framework<sup>13</sup> (**Fig. 1a**). lentiMPRA can be used in any cell type that can be efficiently infected via lentivirus, providing the ability to carry out MPRA in a broad range of cell types and tissues. In addition, due to the viruses' inherent genomic integration, it provides an 'in genome' readout, which we have shown to provide more robust results that can be better predicted by both biochemical and sequence-based features compared to episomal-based MPRA<sup>14</sup>. MPRAflow is a user-friendly computational pipeline that is compatible with a broad range of MPRA experiments.

### **4.3 Development of the protocol**

We developed lentiMPRA to overcome the following limitations: 1) Descriptive assays that detect potential regulatory elements (such as CHIP-seq, DNase-seq, ATAC-seq, CUT&RUN and Hi-C) identify candidate sequences within chromatin, yet most MPRAs analyze sequences in an episomal context; 2) episomal-based MPRA is limited to cells that can easily be transfected. Lentivirus-based assays overcome both these limitations. Lentiviruses integrate into the genome, providing an 'in genome' readout. In addition, they can infect a large number of cells and tissue types, providing a more diverse range of cellular environments for MPRA. In this protocol, we further develop lentiMPRA by placing a barcode in the 5' UTR of the reporter gene. This 5' UTR barcoding method minimizes the distance between the CRS and barcode (102 bp) than previous 3' UTR barcoding method (801 bp), reducing the risk of CRS-barcode swapping<sup>15</sup>. In addition, unlike previous lentiMPRA where each CRS is synthesized together with multiple barcodes in a custom array, the 5' UTR barcoding strategy adds barcodes via the PCR primer. This allows the ability to clone and test hundreds of thousands of CRSs using lentiMPRA.

To subsequently analyze MPRA results, there are several home-brewed MPRA computational analysis pipelines tailored to each lab and MPRA technique. However, these tools are not transferable between labs because of the large variability in MPRA designs, lack of documentation, complicated input files and lack of parameterization of these tools. We thus developed MPRAflow, which provides a user-friendly, flexible, parallelized tool for quantifying CRS activity from a variety of MPRA experimental designs, including lentiMPRA, episomal-based MPRA and saturation mutagenesis designs, with easily interpretable visualizations that can be readily adopted by users regardless of their computational level. In addition to providing normalized fold change per CRS, MPRAflow can generate input files for MPRAalyze<sup>16</sup>, a tool that calculates a transcription rate for each tested CRS by fitting a generalized linear model with DNA and RNA counts. This pipeline allows for the entire analysis to be completed in two commands on a terminal, greatly simplifying the computational tasks associated with MPRA and therefore increasing usability of this protocol.

#### **4.4 Applications of the method**

lentiMPRA can be used for numerous research purposes, such as analyzing hundreds of thousands of different candidate enhancers and their variants (e.g. rare and common GWAS-associated SNPs, evolutionary variants) in the genome, decoding the regulatory code, how it evolved in other species and generating specific genetic switches. It provides the ability to carry out these experiments in hard-to-transfect cells (e.g. primary cells, neurons, and many others) and integrates into the nucleus providing an 'in genome' readout which we have shown is more reproducible and is more predictive of functionality than both biochemical annotations and sequence-based models<sup>14</sup>.

MPRAflow utilizes the pipelining tool Nextflow<sup>13</sup>, which automatically runs MPRA processing code (written in Python, Bash, and R), manages all necessary packages and environments with Anaconda<sup>17</sup>, and is compatible with a multitude of computational architectures including a variety of High Performance Compute (HPC) clusters and cloud computing systems. Additionally, technical replicates and experimental conditions are parallelized through these HPC systems. As MPRAflow is a package that allows non-bioinformatic researchers to easily analyze MPRA data, it can greatly increase the usability of this method in labs that do not have in-house bioinformaticians. Additionally, MPRAflow provides easily interpretable graphics and produces files correctly formatted for readily available tools for further in-depth bioinformatic analysis such as MPRAanalyze<sup>16</sup>.

#### **4.5 Comparisons with other methods**

There are several different varieties of MPRA, such as episomal barcode-based MPRA, STARR-seq, and others<sup>11</sup>. lentiMPRA differs from these methods as it provides an 'in-genome' readout in a wider range of cell types. In STARR-seq, the CRS itself acts as the barcode. This attribute can potentially impact results due to the binding of RNA-associated factors and RNA stability of the assayed sequence<sup>15</sup>. Using on average over fifty 15 bp barcodes per CRS in lentiMPRA reduces this impediment. CRSs are usually generated via oligo synthesis, but can also be produced by other processes, such as PCR or DNA-capture based methods. Barcodes can be either added as part of the synthesis or via PCR, providing flexibility in cloning design. As lentiviruses integrate throughout the genome, we introduced anti-repressors on either side of the virus that together with having over 50 barcodes per assayed sequence assist in overcoming differences due to varying genomic integration sites.



Previous MPRA processing tools have mainly focused on CRS library design or determination of CRS activity from count matrices, overlooking the computationally expensive task of processing sequencing data. MPRAflow is based on computational methods used in our previous MPRA work<sup>14,15,18-20</sup> and contains three utilities: association, count, and saturation mutagenesis. The association utility processes demultiplexed FASTQ files and assigns barcodes to the CRS that they are cloned with in the random pairing design. Sensitive alignment of merged paired-end reads provides robustness against sequencing and synthesis errors without strict read filters, even when CRS libraries contain sequences that differ by only one nucleotide. The count utility processes demultiplexed FASTQ files to perform QC across replicates, normalizes barcode count tables per CRS, and quantifies  $\log_2(\text{RNA/DNA})$  ratios per CRS. MPRAanalyze inputs can also be produced using the count utility. Saturation mutagenesis dissolves multiple variants per CRS into single variant ratios by applying a multivariate linear model and it can be combined with the count utility. Each utility is executed with a single command on a terminal and all utilities provide easily interpretable visualizations of all analyses performed.

## **4.6 Experimental Design**

### **4.6.1 Library design**

CRSs can be identified using many of the aforementioned biochemical assays (ChIP-seq, DNase-seq, ATAC-seq, CUT&RUN, GWAS, Hi-C and others). Variants of interest within these CRSs can be identified via GWAS, GTEx, various genomic websites such as Genome Aggregation Database (gnomAD<sup>21</sup>), comparative genomics and many other databases. The CRSs and variants tested ultimately depends on the goal of the study. Negative and positive controls should be included in the lentiMPRA library. For negative controls, sequences that could be used are those that are known not to be active in the assayed cell, having silencing marks such as H3K27me3 within this tissue, or scrambled CRSs that are randomly selected from the library. For positive controls, sequences that are known to function as promoters/enhancers in

this cell/type or tissue could be used. If such data does not exist, one can characterize CRSs from the cell where the lentiMPRA will be done via the aforementioned biochemical assays. These controls should be present within every technical and biological condition that will be tested. Tools such as MPRAator<sup>22</sup> or MPRA Design Tools<sup>23</sup> can assist in choosing regions to test via MPRA and assembling the FASTA files required to order the libraries. Libraries can contain up to hundreds of thousands of sequences, depending on the infection efficiency of the cells (see **Supplementary Table 1**). The length of these sequences can also vary (as long as the combined length is not over 10 kb, the optimal packaging capacity of lentivirus), depending on how the CRSs are generated (i.e., oligo synthesis, PCR, or capture).

#### **4.6.2 Library generation**

For this protocol, we will focus on oligo synthesis as it is currently the most cost-effective way to generate fixed-length CRSs. Here, the synthesized oligo pool of the CRSs is amplified via two rounds of PCR, first to add the minimal promoter, and then to add the barcode. The amplified fragments are cloned via Gibson assembly into the *SbfI/AgeI* site of pLS-Scel vector (Addgene 137725) to construct the library. The resulting library is digested with I-Scel to remove any vector that did not receive an insert. The recombination products are then electroporated into competent cells and plated onto Ampicillin plates. Sanger sequencing of 16 colonies is then used to confirm the proper assembly of the library. The number of plates will dictate the number of barcodes each CRS will have on average. The number of colonies required for plasmid extraction will depend on the number of CRSs tested and the desired number of barcodes per CRS. Generally, it is ideal to have at least 50 barcodes per CRS and the total number of colonies should roughly equal the desired library complexity. We recommend limiting the complexity of the library due to the finite nature of the multiplicity of infection (MOI) and the associated increase in sequencing costs.

The complexity recommended in this protocol is 0.5-12 million. The library should then be midi-prepped to extract the final plasmid library.

#### **4.6.3 Association sequencing**

To associate the barcode to the CRS, PCR is performed on the plasmid library to add flowcell sequences and sample index to the CRS-barcode pairs. The PCR product is then gel extracted at the appropriate insert size (~471 bp for a 200 bp CRS) and sent for paired end sequencing with an index read for barcode sequence, using custom primers provided in this protocol.

#### **4.6.4 Lentiviral prep**

The next step is to generate a lentivirus library. This is done by transfecting 293T cells with the plasmid library. Following 2 days in culture with titer boost reagent, the virus is collected and concentrated. To titrate the lentivirus, the cell type of interest is plated into 8 wells of a 24 well plate and infected with varying volumes of the virus (0, 1, 2, 4, 8, 16, 32, 64  $\mu$ L) in each well. Cells are monitored for viability throughout this time in order to determine whether certain concentrations are toxic to them. Following a three-day incubation (to reduce non-integrating lentivirus), genomic DNA is extracted from each well. qPCR is then carried out for each condition using primers against genomic DNA, integrated viral DNA, and plasmid backbone DNA. The MOI is calculated for each viral concentration (**Supplementary Table 2**). These values are then plotted against the viral volume to calculate the viral titer. Conditions need to be adjusted if cells are not viable.

#### **4.6.5 Infection and sequencing**

The lentiMPRA library is then infected into the cells of interest and incubated for three days. The number of cells required is determined based on the library complexity and the highest MOI that the cells can be infected with that is not toxic to the cells. It is highly recommended to carry out three technical replicates for each biological condition tested in order to assess reproducibility. The cells are then washed to reduce for non-integrating lentivirus and DNA and RNA are

simultaneously extracted. RNA is treated with DNase and reverse transcription is done using construct-specific primers that contains P7 flowcell sequences and unique molecular identifiers (UMIs), to preserve the true counts of molecules through the amplification process. PCR is carried out on the DNA and RNA samples to amplify barcodes, adding P5 flowcell sequence and sample index upstream, and P7 flowcell sequence and UMI to the barcode. The sequencing libraries are then pooled and sent for paired-end sequencing with a UMI and sample index read.

#### **4.6.6 Data processing**

We built a computational tool, MPRAflow, to easily process demultiplexed FASTA data resulting from lentiMPRA and other MPRA experiments. If the barcodes are randomly paired with the CRS, the association utility can be run to assign barcodes to the appropriate CRS. We provide a workflow tailored to testing distinct CRSs, using Burrows-Wheeler Aligner (BWA<sup>24</sup>) to align sequences to the ordered oligo pool and a workflow for libraries containing single nucleotide variants of the same CRS, using Bowtie2<sup>25</sup> and a list of the expected positions of the variants. The resulting pairing is then used in the count utility, which processes the barcode sequencing of the DNA and RNA, to create normalized  $\log_2(\text{RNA}/\text{DNA})$  ratios for transcriptional activity of each CRS tested along with easy to interpret visualizations. If more robust statistical analyses are desired, we provide the option to generate input files for MPRAnalyze<sup>16</sup>, a generalized linear model approach. Additionally, we provide an alternative workflow for quantifying expression of CRS libraries produced with saturation mutagenesis. It processes data into a matrix of RNA count, DNA count, and  $N$  binary columns indicating whether a specific sequence variant was associated with the barcode ( $T$ ), which are used to fit a multiple linear regression model of  $\log_2(\text{RNA}_j) \sim \log_2(\text{DNA}_j) + N + \text{offset}$  ( $j \in T$ ) and report the coefficients of  $N$  as effects for each variant. The utility processes multiple replicates and conditions in parallel if a high-performance computing (HPC) cluster is available, but can also be run locally. This code is freely available on GitHub (<https://github.com/shendurelab/MPRAflow>).

#### 4.6.7 Necessary Expertise

Basic molecular biology and cell culture skills are required to perform lentiMPRA. For MPRAflow, a basic familiarity with command line tools is needed.

#### 4.6.8 Limitations

There are several limitations for lentiMPRA. This includes a limitation in the number of CRSs that can be tested in cells that are not amenable to high lentivirus concentrations, though that can be amended by using an increased number of cells. The use of oligosynthesis to generate the CRS library can also limit the number of sequences that can be tested including their length. Improvements in DNA synthesis can ultimately overcome this limitation as well as PCR or DNA capture-based methods. Techniques that allow for multiplex pairwise assembly of oligos<sup>26</sup> could also be a way to increase CRS size by patching together specific oligonucleotides.

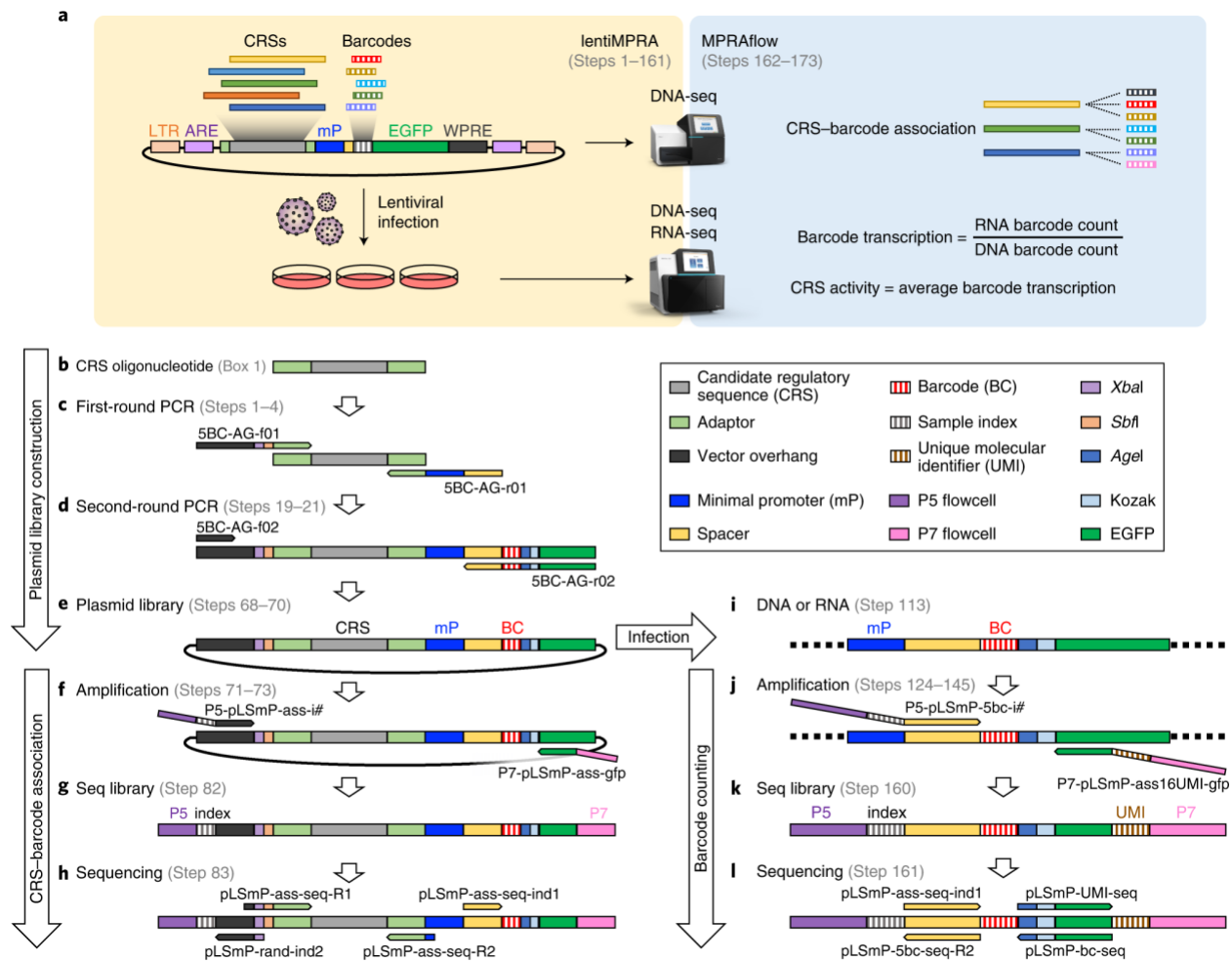
As for MPRAflow, while this tool is applicable to many types of MPRA, it does not support STARR-seq workflows as it does not include functionality for peak calling.

#### 4.7 Anticipated results

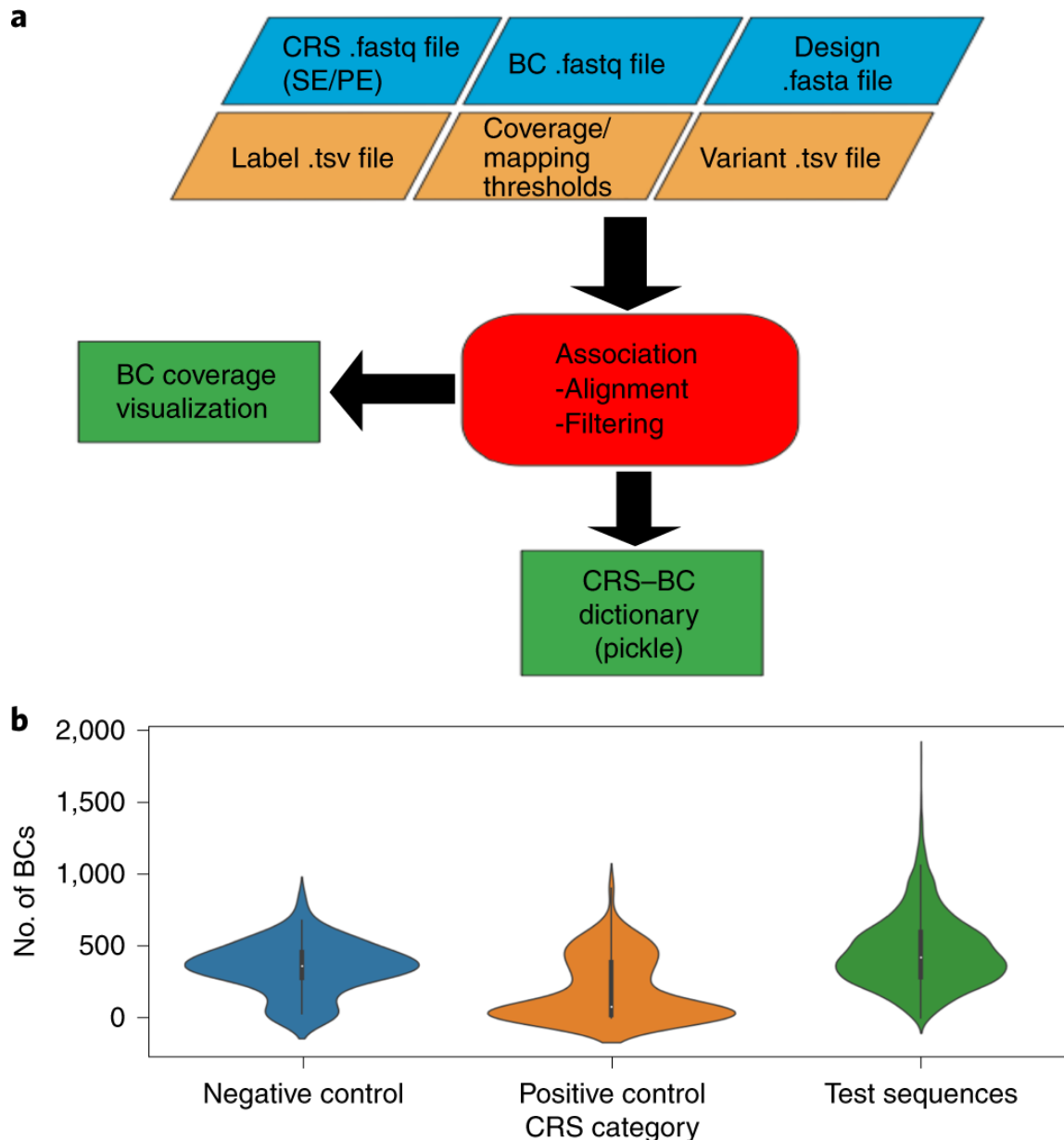
The output of a lentiMPRA experiment will consist of two sets of data: association sequencing, and DNA/RNA barcode sequencing. Success of association sequencing preparation can be assessed by the size of the band (419 bp) observed during library preparation. The association sequencing should contain paired-end reads that cover the CRS (200 bp) and an index read to cover the barcode (15 bp). The recommended sequencing depth will vary significantly with the complexity of the library being tested, but we generally suggest 10 reads per unique barcode expected. MPRAflow's association utility should be run on the example dataset (GSE142696) to determine the number of barcodes per CRS (**Fig. 2b**). Generally, we aim for 50-200 barcodes per CRS, libraries with more than 600 barcodes per CRS should be cloned again

since integration and sequencing will limit the coverage of the library and the sensitivity of the experiment.

The quality of the preparation of the DNA and RNA barcode sequencing library can be assessed by the size of the band (162 bp). The sequencing results should contain paired-end reads that cover the barcode (15 bp) and an index read for a UMI (16 bp). These files should be demultiplexed and run through MPRAflow's count utility. This will return normalized count tables for all experimental conditions and replicates tested as well as a final table of activity of each CRS normalized across replicates. A broad overview of activity can be seen by user-defined categories (**Fig. 3b-c**), allowing for assessment of control sequences. Averaged observed barcodes per CRS can be checked through histograms to verify coverage of the barcodes (**Fig. 3d**). Additionally, the correlation between technical replicates are shown for DNA count, RNA count, and  $\log_2(\text{RNA/DNA})$  (**Fig. 3e**). A successful experiment will allow the user to determine which CRSs increase transcriptional activity and which do not. To determine the active sequences, we can compare our test sequences with scrambled controls. These scrambled sequences provide a null distribution which can be used for robust statistical testing.

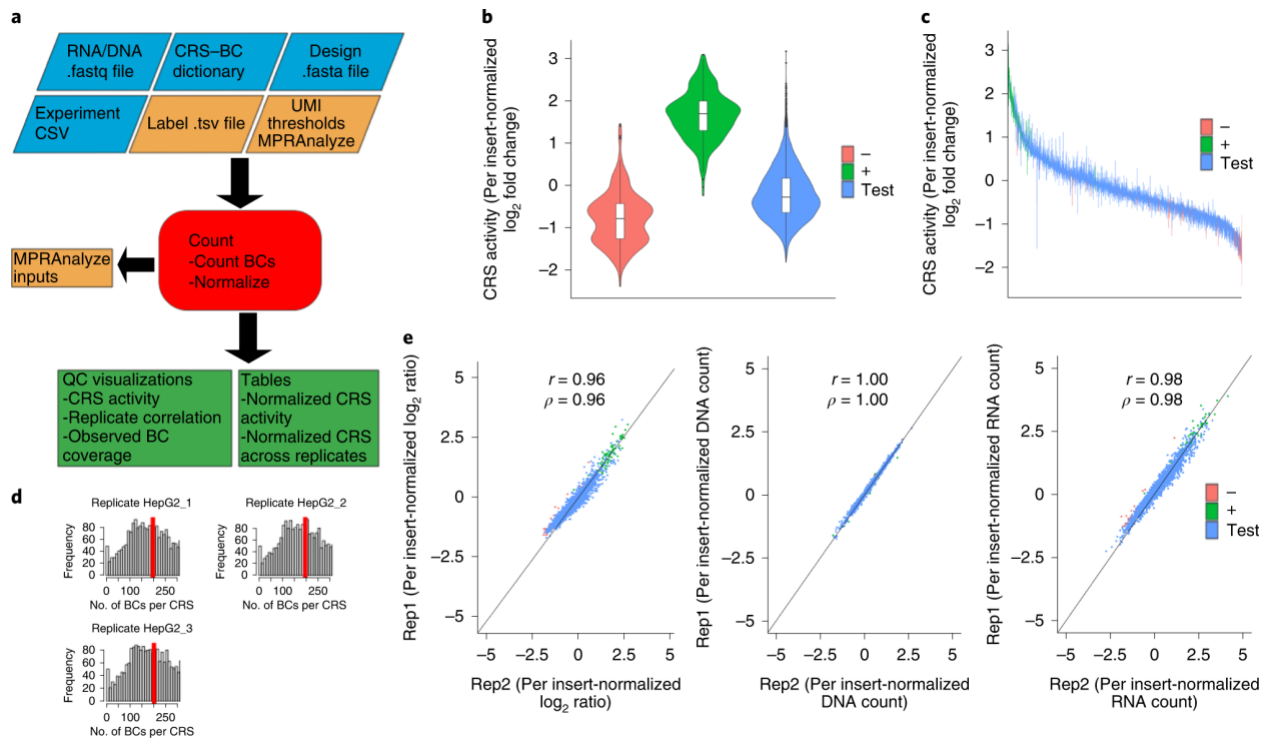


**Figure 4.1: Schematics of lentiMPRA.** **a**, Summary of lentiMPRA and MPRAflow. The lentiMPRA library is sequenced to associate between CRS and barcodes and infected into cells using three replicates. DNA and RNA from the cells is sequenced to obtain barcode transcription and CRS activity. LTR, long terminal repeat; ARE, anti-repressor element; WPRE, Woodchuck hepatitis virus posttranscriptional regulatory element. **b**, CRS oligo. 200-base CRS (grey) is flanked by PCR adaptor sequences (light green). **c**, First round PCR. PCR primers add sequences that are complementary to the vector (black) to the upstream and minimal promoter (mP, blue) and spacer sequences (yellow) downstream of the CRS oligo. **d**, Second round PCR. Reverse primer adds the barcodes (red stripe) and GFP complementary sequences (green). **e**, Plasmid construct. **f**, Amplification for CRS–barcode association. Primers adding P5 (purple) and sample index (grey stripe) upstream and P7 (pink) downstream. **g**, Sequencing library structure. **h**, Sequencing reaction. Paired-end reads specify the CRS sequence, with index read 1 providing the barcode and index read 2 reading the sample index for multiplexing. **i**, Integrated DNA and expressed RNA in infected cells. **j**, Amplification for barcode counting. Primers add P5 and sample index upstream and P7 and unique molecular identifier (UMI, brown stripe) downstream. **k**, Sequencing library structure. **l**, Sequencing reaction. Paired-end reads give barcode, index read 1 gives UMI and index read 2 provides sample index for multiplexing.

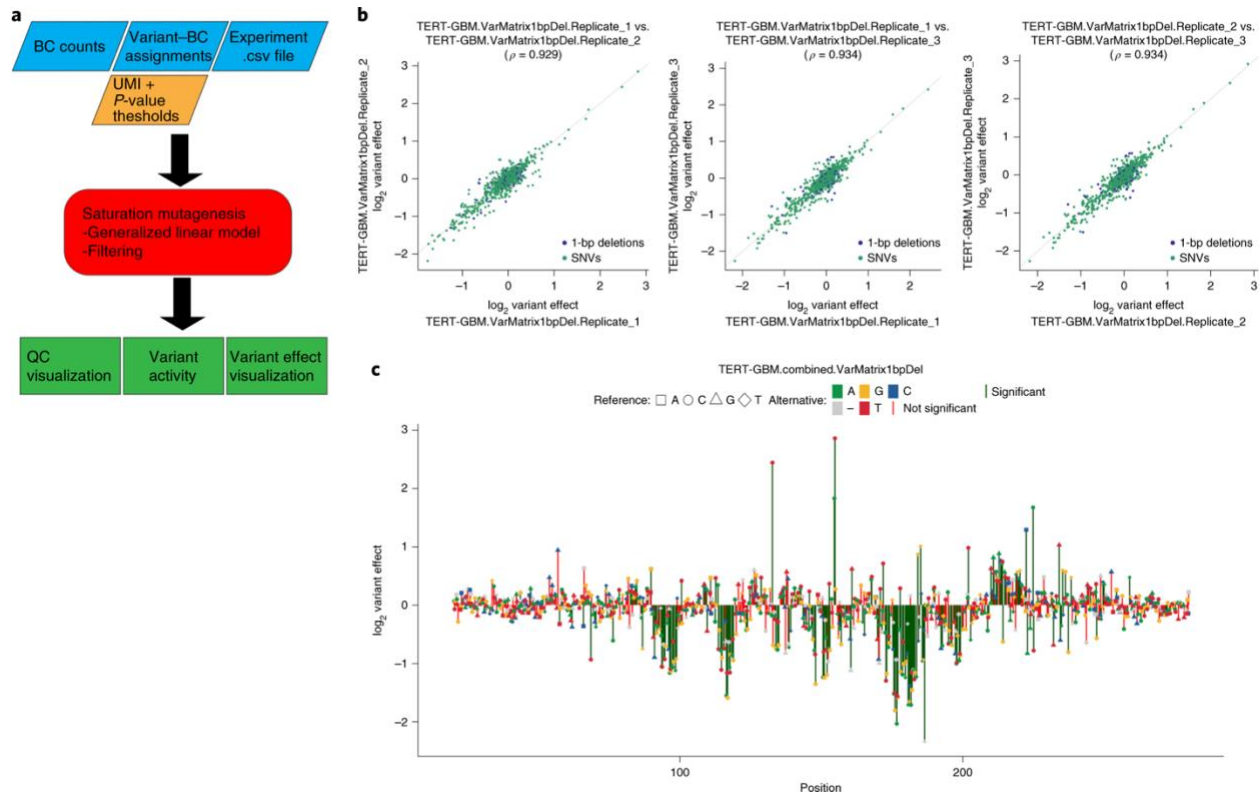


**Figure 4.2: Overview of MPRAFLOW association utility.** **a**, Mandatory inputs (blue), optional flags (orange), output files (green) and utility (red). The program requires FASTQ files for the insert, either single-end (SE) or paired-end (PE) reads and a design file, which is a FASTA file containing the synthesized oligos. The user can also specify a tab delimited file with a mapping of CRS names given in the design file and a grouping, such as control category (e.g. positive or negative control), a tab-separated values file (TSV) of variants in the ordered oligo pool to be used for a tailored alignment strategy, and can accept various parameters for filtering the pairing based on mapping qualities and number of observed barcodes mapping to the CRS. The program outputs a Python dictionary in pickle format, mapping barcodes to their CRS. **b**, A violin plot of barcode coverage for each enhancer, grouped by labels provided in the label TSV. The violin plot features a kernel density, showing the underlying distribution of the data and a boxplot. In the boxplot the white dot is the median, the box represents the interquartile range (IQR), and the whiskers are  $1.5 \times \text{IQR}$ . Outliers are represented as points.





**Figure 4.3: Overview of count utility.** **a**, Mandatory inputs (blue), optional flags and outputs (orange), output files (green) and utility (red). The user must specify the directory containing all FASTQ files for the RNA and DNA sequencing, the CRS-barcode dictionary from the Association Utility, a design file (FASTA file containing the synthesized oligos), and an experimental comma-separated file (CSV) outlining the number of replicates and conditions used. The user can also specify a tab delimited file with a mapping of CRS names given in the design file and a grouping, such as control category (e.g. positive or negative control) and tune parameters such as specifying if a unique molecular identifier (UMI) was used, or if the user would like to generate the input files for MPRAnalyze. The program will produce normalized activity of each CRS from each replicate as well as across replicates along with several visualizations (**b-e**). **b**, CRS activity normalized by insert and grouped by label determined in the label file. The violin plot features a kernel density, showing the underlying distribution of the data and a boxplot. In the boxplot the white dot is the median, the box represents the interquartile range (IQR), and the whiskers are  $1.5 \times \text{IQR}$ . Outliers are represented as points. **c**, Normalized activity of each CRS across replicates colored by label represented as a boxplot across replicates, where the box represents the interquartile range (IQR), and the whiskers are  $1.5 \times \text{IQR}$ . Outliers are represented as points. **d**, Distribution of observed barcode coverage per replicate. **e**, Correlation of normalized  $\log_2(\text{RNA/DNA})$ , DNA counts and RNA counts colored by label.



**Figure 4.4: Overview of Saturation Mutagenesis Utility.** **a**, Mandatory inputs (blue) optional flags and outputs (orange) output files (green) utility (red). The user must specify the directory containing all barcode count files, including DNA and RNA counts, the variant to barcode assignment file, and an experimental comma separated file outlining the number of replicates and conditions used. The user can also set UMI and p-value thresholds that will be used for filtering variants and distinguishing between significant and not-significant variant effects. The program will produce log<sub>2</sub> variant effects together, p-values and a visual output of correlation and a saturation mutagenesis variant effect plot of the region. **b**, Correlation between replicates. Here the correlation between three replicates of the TERT promoter in a glioblastoma cell-line from Kircher et.al. 2019 (Ref. 20) is shown. rho<sub>p</sub> defines the pearson correlation between two samples (model with 1 bp indels). Only variants with a minimum number of 10 barcodes are shown. **c**, Saturation mutagenesis effect plot of the combined model from three replicates of the TERT promoter in a glioblastoma cell-line from Kircher et.al. 2019 (including 1bp indels). Position refers to the variant position of the original target insert. Only variants with a minimum number of 10 barcodes are shown. Significance level is p-value < 1-e5.

## Reporting Summary

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

## Data availability

A 5' lentiMPRA dataset conducted in HepG2 cells<sup>15</sup> has been deposited into the NCBI Sequence Read Archive (SRA) under accession no. GSE142696

## Code availability

The source code is freely available at <https://github.com/shendurelab/MPRAflow>.

## 4.8 Author contributions statements

F.I. and B.M. developed lentiMPRA, R.Z. assisted in developing lentiMPRA, M.G.G., M.S., V.A., S.W., S.F., J.Z., T.A., A.K., I.G.S., N.Y., C.Y., K.S.P., M.K., J.S., N.A. assisted in developing MPRAflow, and all authors contributed to writing the manuscript.

## 4.9 Acknowledgments

This work was supported by the National Human Genome Research Institute grant number 1UM1HG009408 (N.A. and J.S.) and 1R21HG010065 and 1R21HG010683 (N.A.) and Ruth L. Kirschstein Predoctoral Individual National Research Service Award 1F31HG011007 (M.G.G.), NRSA NIH fellowship 5T32HL007093 (V.A.), National Institute of Mental Health grant numbers 1R01MH109907 and 1U01MH116438 (N.A. and K.S.P.), and the Uehara Memorial Foundation (F.I.). J.S. is an investigator of the Howard Hughes Medical Institute.

## Competing interests

The authors declare no competing interests.

#### 4.10 REFERENCES

1. Chatterjee, S. & Ahituv, N. Gene regulatory elements, major drivers of human disease. *Annu. Rev. Genomics Hum. Genet.* **18**, 45–63 (2017).
2. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
3. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
4. Carroll, S. B. Evolution at two levels: on genes and form. *PLoS Biol.* **3**, e245 (2005).
5. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein- DNA interactions. *Science* **316**, 1497–1502 (2007).
6. Crawford, G. E. *et al.* Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 992–997 (2004).
7. Sabo, P. J. *et al.* Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 4537–4542 (2004).
8. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
9. Skene, P. J., Henikoff, J. G. & Henikoff, S. Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nat. Protoc.* **13**, 1006–1019 (2018).
10. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
11. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**, 159–164 (2015).
12. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).

13. Di Tommaso, P. Nextflow enables reproducible computational workflows. *Nat Biotechnol* **35**, 316–319 (2017).
14. Inoue, F. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res* **27**, 38–52 (2017).
15. Klein, J. *et al.* A systematic evaluation of the design, orientation, and sequence context dependencies of massively parallel reporter assays. *Research Square* (2020) doi:10.21203/rs.3.pex-1065/v1.
16. Ashuach, T. *et al.* MPRAnalyze: statistical framework for massively parallel reporter assays. *Genome Biol.* **20**, (2019).
17. *Anaconda Software Distribution.* (2016).
18. Inoue, F., Kreimer, A., Ashuach, T., Ahituv, N. & Yosef, N. Identification and Massively Parallel Characterization of Regulatory Elements Driving Neural Induction. *Cell Stem Cell* **25**, 713–727 (2017).
19. Ryu, H. Massively parallel dissection of human accelerated regions in human and chimpanzee neural progenitors. *bioRxiv* (2018).
20. Kircher, M. *et al.* Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.* **10**, 3583 (2019).
21. Karczewski, K. J. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* (2019).
22. Georgakopoulos-Soares, I., Jain, N., Gray, J. M. & Hemberg, M. MPRAnator: a web-based tool for the design of massively parallel reporter assay experiments. *Bioinformatics* **22**, 137–138 (2006).
23. Ghazi, A. R. Design tools for MPRA experiments. *Bioinformatics* **34**, 2682–2683 (2018).
24. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

25. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
26. Klein, J. C. Multiplex pairwise assembly of array-derived DNA oligonucleotides. *Nucleic Acids Res* **8**, (2015).

## Tables

**Table 4.1: Association Utility options. Blue rows are mandatory and orange are optional.**

Options	Description	
--fastq-insert	Full path to library association fastq for insert (must be surrounded with quotes)	Blue
--variants	tsv with reference_name variant_positions ref_bases alt_bases, only input for variant analyses workflow	Blue
--fastq-bc	Full path to library association fastq for bc (must be surrounded with quotes)	Blue
--design	Full path to fasta of ordered oligo sequences (must be surrounded with quotes)	Orange
--fastq-insertPE	Full path to library association fastq for read2 if the library is paired end (must be surrounded with quotes)	Orange
--min-cov	minimum coverage of bc to count it (default 3)	Orange
--min-frac	minimum fraction of bc map to single insert (default 0.5)	Orange
--mapq	map quality (default 30)	Orange
--baseq	base quality (default 30)	Orange
--cigar	require exact match ex: 200M (default none)	Orange
--outdir	The output directory where the results will be saved and what will be used as a prefix (default outs)	Orange
-w	specific name for work directory (default: work)	Orange
-with-timeline	Create html file showing processing times	Orange
--split	number read entries per fastq chunk for faster processing (default: 2000000)	Orange
--labels	tsv with the oligo pool fasta and a group label (ex: positive_control) if no labels desired a file will be automatically generated	Orange
--h, --help	help message	Orange

**Table 4.2: Count utility options. Blue rows are mandatory and orange are optional.**

Options	Description	
--dir	fasta directory (must be surrounded with quotes)	Blue
--association	pickle dictionary from library association process	Blue
--design	fasta of ordered insert sequences	Blue
--e, -- experiment- file	experiment csv file	Blue
--labels	tsv with the oligo pool fasta and a group label (ex: positive_control), a single label will be applied if a file is not specified	Orange
--outdir	The output directory where the results will be saved (default outs)	Orange
--bc-length	Length of barcode (default 15)	Orange
--umi-length	Length of umi when given (default 10)	Orange
--no-umi	Flag if no umi was used in the experiment	Orange
--merge- intersect	Only retain barcodes in RNA and DNA fraction (TRUE/FALSE, default: FALSE)	Orange
--mpranalyze	Flag to only generate MPRAnalyze outputs	Orange
--thresh	minimum number of observed barcodes to retain insert (default 10)	Orange
-w	specific name for work directory (default: work)	Orange
-with-timeline	Create html file showing processing times	Orange
--h, --help	help message	Orange

**Table 4.3: Saturation Mutagenesis utility options. Blue rows are mandatory and orange are optional.**

Options	Description	
--dir	Directory of count files (must be surrounded with quotes)	Blue
--assignment	Variant assignment file	Blue
--e,-- experiment	experiment csv file	Blue
--outdir	The output directory where the results will be saved (default outs)	Orange
--thresh	minimum number of observed barcodes to retain insert (default 10)	Orange
--pvalue	minimum p-value for significant variant effects (default 1e-5)	Orange
-w	specific name for work directory (default: work)	Orange
-with-timeline	Create html file showing processing times	Orange
--h, --help	help message	Orange



**Table 4.4: Troubleshooting table.**

<b>Step</b>	<b>Problem</b>	<b>Possible reason</b>	<b>Solution</b>
Step 5 of Box 1	Low infection efficiency	Polybrene concentration may not be appropriate	Optimization of polybrene concentration may be required. Seed cells in a 24-well plate and infect control virus along with different amount of polybrene (e.g. 0, 2, 4, 8, 16, 32 $\mu\text{g}/\text{mL}$ at a final concentration), and observe cell death and GFP expression. In our experience, 8 $\mu\text{g}/\text{mL}$ works well for most cell types including HepG2, K562, H1 hESCs, and WTC11 iPSCs. Polybrene kills neural cell types, including neural progenitors and should be avoided when using those cells.
Step 29	Low DNA yield. At least 250 ng insert DNA is required for the recombination reaction.	DNA amplification was not enough. DNA loss during gel extraction.	Multiply the PCR reaction or increase the number of cycles of the second round PCR up to 15 cycles. More cycles (>15 cycles) can decrease the library complexity.
Step 37	Uncut vector DNA appear on the gel.	Insufficient restriction enzyme reaction.	Perform restriction digestion twice or three times (step 30-36).
Step 61	Contamination with empty vectors.	Vector linearization and/or I-SceI digestion were not sufficient.	Lower rate of empty vector contamination (less than 10%, one or two out of 16 colonies) is acceptable. Proceed with the protocol. If higher rate, redo vector linearization with longer incubation time and make sure you have complete linearization using an agarose gel. Perform I-SceI digestion with longer incubation time.

Step	Problem	Possible reason	Solution
Step 61	Mutation and indels observed in the plasmids.	These can be derived from synthesis/PCR/sequencing errors.	As these errors are unavoidable, we usually observe >50% of sequences contain mutations and/or deletions. Proceed with the protocol, and these erroneous sequences should be ruled out during the analysis step. Synthesis error rates might be improved by ordering oligos that are high-fidelity synthesized from the manufacturer.

## **Chapter 5: Developing scMPRA to dissect gene by environment interactions.**

### **5.1 Abstract**

Gene by environment interactions (GxE) are defined as the alteration of a genetic effect in response to environmental variation and may account for missing heritability of complex traits. Current studies have had limited success, as it is difficult to prioritize environmental factors to quantify. Here we describe our efforts towards developing a novel assay, single-cell Massively Parallel Reporter Assay (scMPRA), to characterize how cellular environments (e.g. cell type, cell state, donor) modulate the relationship between gene regulatory elements and gene expression. Future applications of this technology may elucidate how variation in gene expression between cells and individuals result from GxE interactions and increase our understanding of how GxE interactions influence complex traits such as human disease.

### **5.2 Introduction**

Gene regulatory programs are precisely orchestrated as cells differentiate to their terminal states or respond to stimuli in their environment. Regulatory elements, such as enhancers, promoters, and repressors, tune these responses and ensure genes are expressed in the correct cell type and state. Large efforts such as ENCODE, have worked to map these elements in many cell lines and primary cells; however, these experiments require each cell type or cell state to be sorted and profiled separately<sup>1</sup>. These efforts can be arduous, expensive, or impossible in heterogenous tissues, especially where reliable surface markers are unavailable for sorting or when studying developmental trajectories where intermediate cell states fall on a continuum rather than a specific category.

In addition to mapping regulatory regions specific to cell type and state, it is important to understand how genetic variants can disrupt or enhance these regulatory programs. Gene by environment interactions (GxE) occur when genetic effects on trait variation can be modified by

environmental exposures. These interactions are hypothesized to be one source of missing heritability, or the observation that common genetic variants additively explain little of the total trait heritability<sup>2</sup>. This hypothesis is supported by simulation studies showing that increased trait variability due to GxE reduces the power to map associated variants and that including GxE interaction terms in genetic models is essential<sup>3</sup>. While genetic models that include interaction terms have been proposed, they have had little power to map GxE effects due difficulties in accurately quantifying environmental exposures, which can be temporal in nature and may act in combination to produce synergistic effects on phenotype<sup>4</sup>. There is an enormous opportunity to map interactions between genotypes and cellular environments (e.g. cell type or activation state) affecting molecular traits such as gene expression. This is because cellular environments can be experimentally controlled and more accurately measured in vitro, interaction effects acting on molecular traits will likely have larger effect sizes, and candidate GxE effects produce directly testable hypotheses of biological mechanisms that control gene regulation. Dissecting these cellular GxE interactions is critical for our understanding of variation in gene expression between cells and individuals and to shed light on GxE interactions on complex traits such as disease.

Studies profiling transcriptomes in bulk across many cellular environments have already provided compelling evidence of GxE effects. These include the identification of genetic variants whose effects are modified by specific tissues<sup>5</sup> or in response to extracellular stimulation<sup>6</sup>. However, bulk transcriptomic sequencing fails to capture the additional heterogeneity within a tissue. For example, across immune cell types in the peripheral blood, genetic effects on gene expression are less correlated ( $r_G$ ) between cell types distantly related by lineage, suggesting that cellular environments across cell types differentially modify genetic effects<sup>7</sup>. Based on these observations, we hypothesize that the cellular environment differentially modifies genetic effects on gene expression by modulating the activity of regulatory elements harboring the causal variants. Single cell genomics and synthetic biology can be leveraged to simultaneously measure

the cellular environment and test its effects on the expression of synthetic gene constructs harboring various regulatory elements. Currently there are no high-throughput methods available to functionally validate variants across different cellular environments.

Massively Parallel Reporter Assays (MPRAs) are routinely used to functionally annotate regulatory elements in different cell types by linking a synthetic sequence to a transcribed barcode<sup>8</sup>. This technology is highly scalable as many regulatory elements in the genome can be tested simultaneously and has been used to validate thousands of disease and expression associated variants<sup>9</sup>. While MPRAs can nominate candidate regulatory sequences, gene expression is controlled by a complex network of trans activators that may be differentially regulated in across distinct cellular environments<sup>10</sup> specified by cell cycle, cell type, extracellular activation, and donor variability<sup>11</sup>. While it is likely these factors also influence bulk MPRA results, current approaches average the signal of each regulatory element over many cells, losing important information about heterogeneity in cellular environments. Recent advances in single cell RNA-sequencing have shown that gene expression is volatile and heterogeneity across cells can influence the expression of individual transcripts<sup>12,13</sup>. Here we develop a novel method called single-cell Massively Parallel Reporter Assay (scMPRA) that integrates MPRAs and single-cell RNA-seq (scRNA-seq) to map interaction effects between the cellular environment and causal variants that affect gene expression. This technique will enable the simultaneous estimate of cellular environment, testing of the effects of hundreds of sequences harboring expression-associated nucleotide variants, and how those effects can be modified by cellular environments due to individual donor variability, cell type identity or response to stimulation. This method holds potential to be instrumental in understanding how genetic variation interacts with the cellular environment to affect gene expression and to better understand how this relationship contributes to downstream phenotypes such as disease.

## 5.3 Methods

### 5.3.1 Molecular Biology Approach

We have assembled two designs for scMPRA, both utilizing lentiviral based MPRA (lentiMPRA) to test putative regulatory sequences and read out changes as lentiMPRA offers a distinct advantage over episomal based approaches as the regulatory activity is tested via genomic integration, producing more consistent results over 'non-integrating' approaches<sup>8</sup>. The first design uses the 10x Genomics feature barcoding 3' single cell sequencing kit, while the alternative design uses the 5' 10x Genomics kit and amplicon sequencing.

The 10x Genomics feature barcoding 3' kit is designed to read out specific RNAs from single-cell RNA-seq experiments, reserving ~20% of the capture sequences on a gel bead for a feature barcode, while the rest of the capture sequences contain a poly(dT) sequence. The plasmid packaged into the lentiviral vector contains a variable enhancer sequence followed by a minimal promoter, an enhancer barcode, complementary sequence to the feature capture sequence, and a reporter gene (Fig. 1a). Therefore, the cell's transcriptome will be captured with the poly(dT) sequence and the expression of the MPRA construct will be captured with the feature capture sequence. Since each capture sequence contains a cell barcode and a unique molecular identifier (UMI), it is possible to quantify the number of RNA molecules associated with each MPRA construct (Fig 1a).

Alternatively, the 5' approach captures both the MPRA barcodes and transcriptome using a TSO using the same construct utilized in our standard workflow<sup>14</sup>. MPRA barcodes can subsequently be amplified out of the total transcriptome pool to increase the capture of these sequences. Barcodes for each candidate regulatory sequence (CRS), will be assigned based on either association sequencing if barcodes are randomly paired, or will be directly paired through synthesis. Current MPRA methods perform DNA and RNA sequencing, enabling quantification of the change in expression ( $\log_2(\text{RNA counts} / \text{DNA counts})$ ). Due to the inability of scRNA-seq to

capture DNA and RNA in single cells, we will titrate the experiment to ensure the DNA count is equal to one to quantify expression.

### 5.3.2 Mathematical calibration of scMPRA.

A major obstacle to this approach is that current single-cell sequencing technologies do not permit high throughput sequencing of both DNA and RNA in single cells. To circumvent this, carefully titration of the experiment is necessary to ensure that each sequence present in a single cell is unique, allowing us to calculate enhancer activity assuming the DNA count is equal to one. This problem maps to the birthday paradox from probability theory, which defines that if the population size ( $L$ , MPRA library) is large enough, we can sample  $K$  number of sequences, while being sure all sequences are unique<sup>15</sup>. This is defined as  $\prod_{n=1}^{k-1} \frac{L-n}{L}$ , assuming all samples in a library have equal representation. The bigger the library the less often collisions occur (Fig. 1b). However, this technology requires a balance between the collision rate and the cost of sequencing enough cells to cover the full library of tested sequences. The average number of cells needed to be sequenced can be modeled by the batched coupon collector problem. This problem defines the average number of cells one must sequence to cover the entire MPRA library ( $L$ ), given each cell has a MOI of  $K$  MPRA constructs (where the probability of each construct is equal). This is defined as  $\frac{L}{K} \sum_{n=1}^L \left(\frac{1}{n}\right)$ . The larger the library, the more cells you must sequence (Fig. 1c). To overcome these tradeoffs, it is possible to tag each putative regulatory sequence with many barcodes, decreasing the likelihood of collisions while maximizing the coverage of each enhancer. Assuming there are 1000 total sequences one wishes to test, and each sequence is tagged with 40 barcodes, the total library size will be 40,000. By sequencing just 10,000 cells, each enhancer in the library should be covered an average of 200 times, assuming a MOI of 10. Overall, this design allows for efficient sequencing of multiple donors in multiple environmental perturbations profiling thousands of variants of interest in candidate regulatory regions.

## 5.4 Results

To test our designs, we first designed a pilot mixing experiment of LCLs and HepG2 cells, where each of these cell lines will be transduced using the MPRA library described above at a MOI of 10. The premise of this experiment is that enhancers specific to each cell type should only be expressed in their respective cell line. For this experiment we assembled a MPRA library comprised of previously characterized enhancers in LCLs and HepG2 along with negative controls and measure activity using the 10x Genomics Chromium platform. Specifically this library contained 200 sequences were designed including 50 enhancers that are active only in LCLs, 50 enhancers that are active only in HepG2, 30 active in both cell lines, 30 inactive in both cell lines, 10 of the lowest expression in LCLs, 20 synthetic enhancers active in HepG2s, and 10 synthetic enhancers inactive in HepG2, which have all been tested in previous MPRAs<sup>9,16</sup>. We transduced and cultured these cells separately following their respective culturing protocols aiming for a MOI of 10. Seventy-two hours post transduction, we mixed the two populations of cells in a 1:1 ratio and performed scRNA-seq using the 10x Genomics 3' feature barcoding kit followed by paired end sequencing on the Illumina's Nova seq. In parallel we performed bulk DNA and RNA sequencing on a subset of this mixed population using our standard lentiMPRA approach to demonstrate how fine-grained environmental information can improve profiling of regulatory elements<sup>14</sup>.

We successfully produced high quality scRNAseq data during this first experiment and were confidently able to classify LCLs and HepG2 cells (Fig 2a). While globally our positive controls were more highly expressed than negative controls, capture of these constructs were very sparsely captured despite deep sequencing depth of our FBC library (Fig 2b). We postulated that the polyDT capture was out competing capture with the feature barcode, since our constructs also have a polyA tail. To evaluate this hypothesis, we performed a quantitative polymerase chain reaction (qPCR) experiment to quantify abundance of our reporter gene, GFP, in the

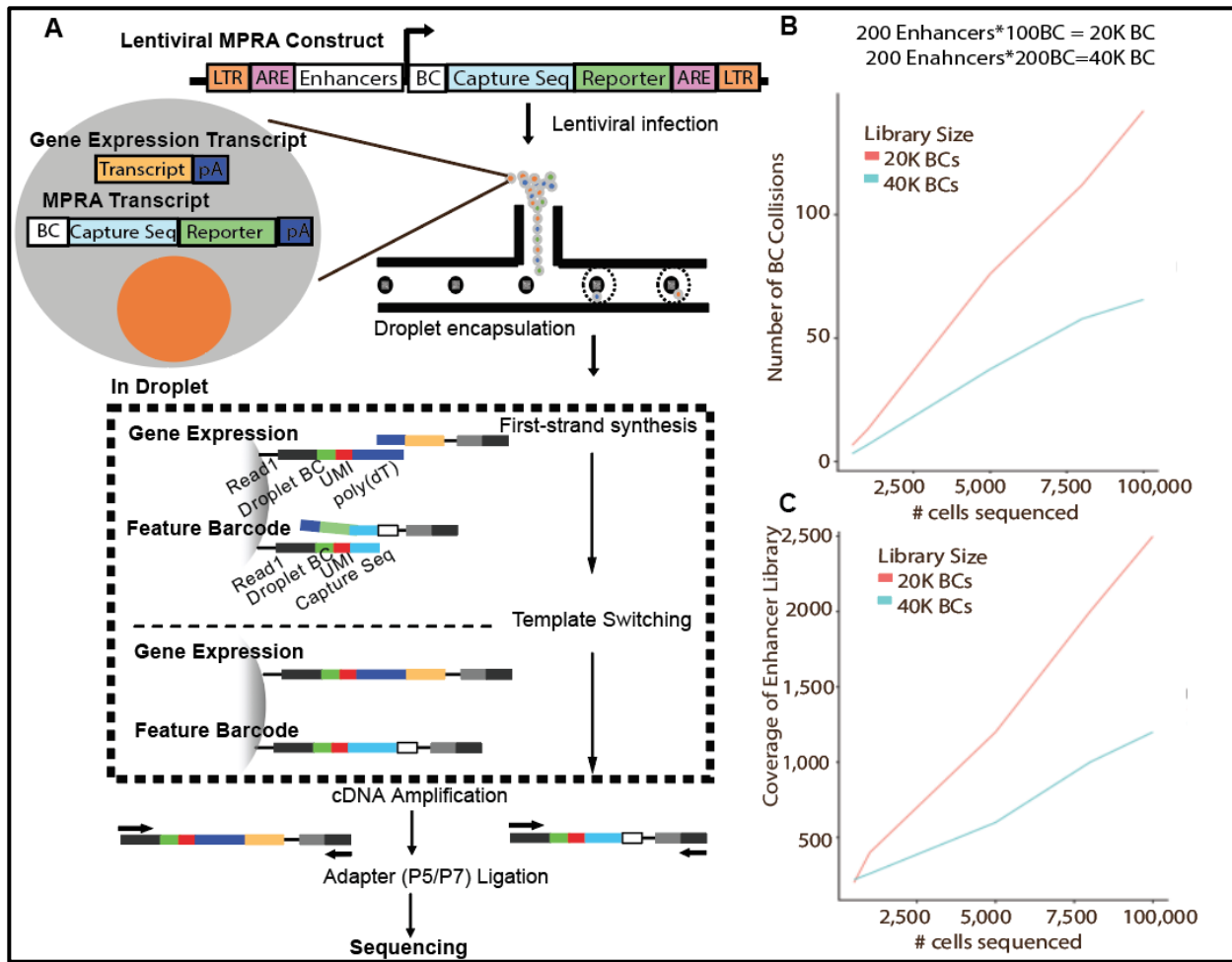


transcriptome library (Fig 2c). High levels of GFP were detected in the transcriptome library, indicating that the feature barcoding approach may not be optimal if the feature includes a polyA.

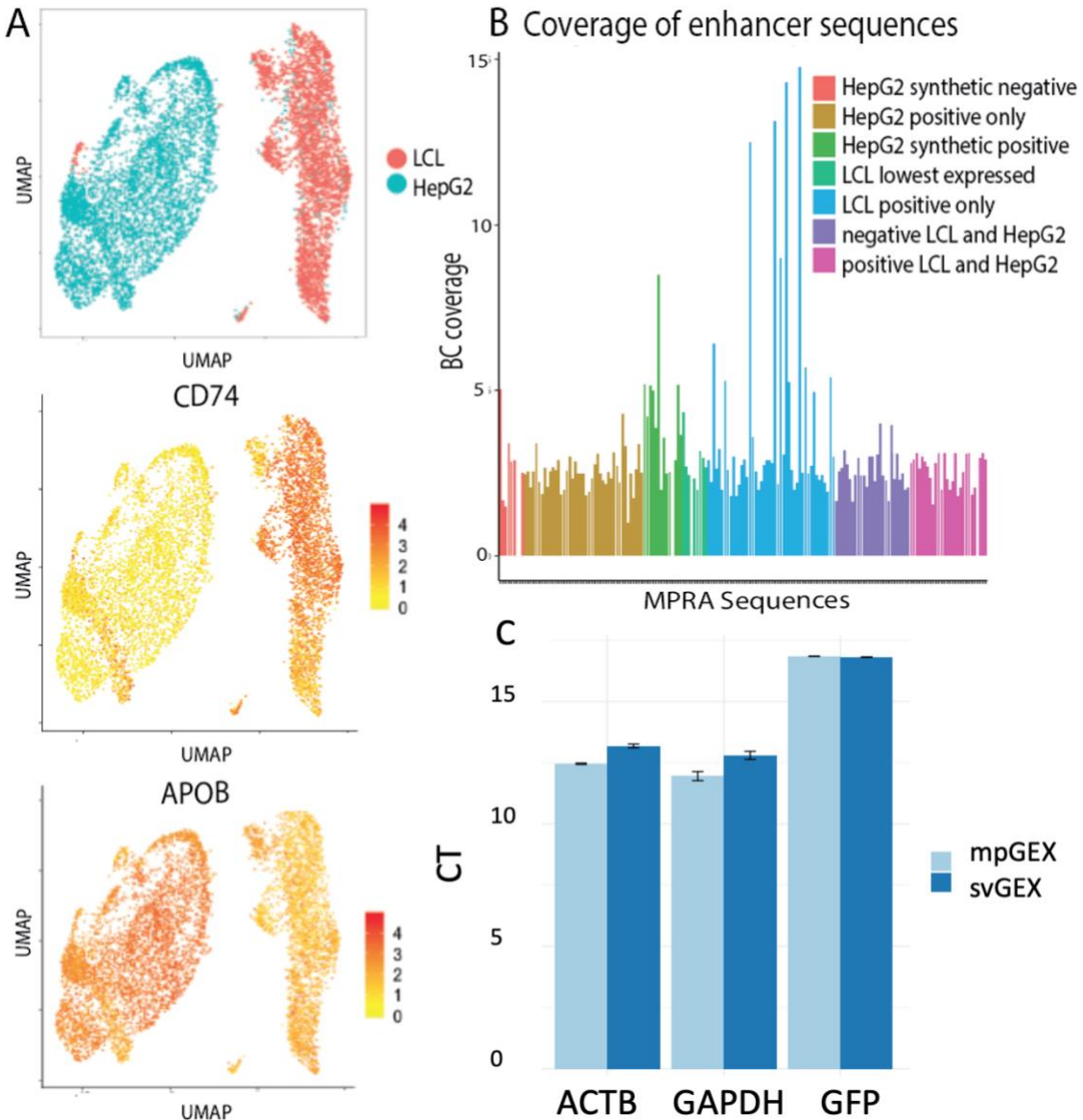
To address this concern, we tested our alternate design using the 5' 10x genomics kit. We synthesized a new pilot library including 220 sequences expected to be active HepG2 and K562 controls that have all been previously tested in the lab<sup>16,17</sup>. Specifically, this library contained 200 sequences were designed including 52 enhancers that are active only in HepG2, 52 enhancers that are active only in K562, 30 active in both cell lines, 30 inactive in both cell lines, 10 of the lowest expression in LCLs, 20 synthetic enhancers active in HepG2s, and 10 synthetic enhancers inactive in HepG2 and 20 sequences scrambled to act as additional negative controls. The parameters for infection and cell culture from the first pilot were used again in this experiment. Once again, high quality scRNA-seq data was produced, where HepG2 and K562 cells were robustly classified (Fig 3a-c). While detection of MPRA constructs improved, the data was sparse. Despite this sparsity, we were able to see higher expression of positive controls compared to the negative controls and the scrambled controls (Fig 3d-f).

## **5.5 Conclusions**

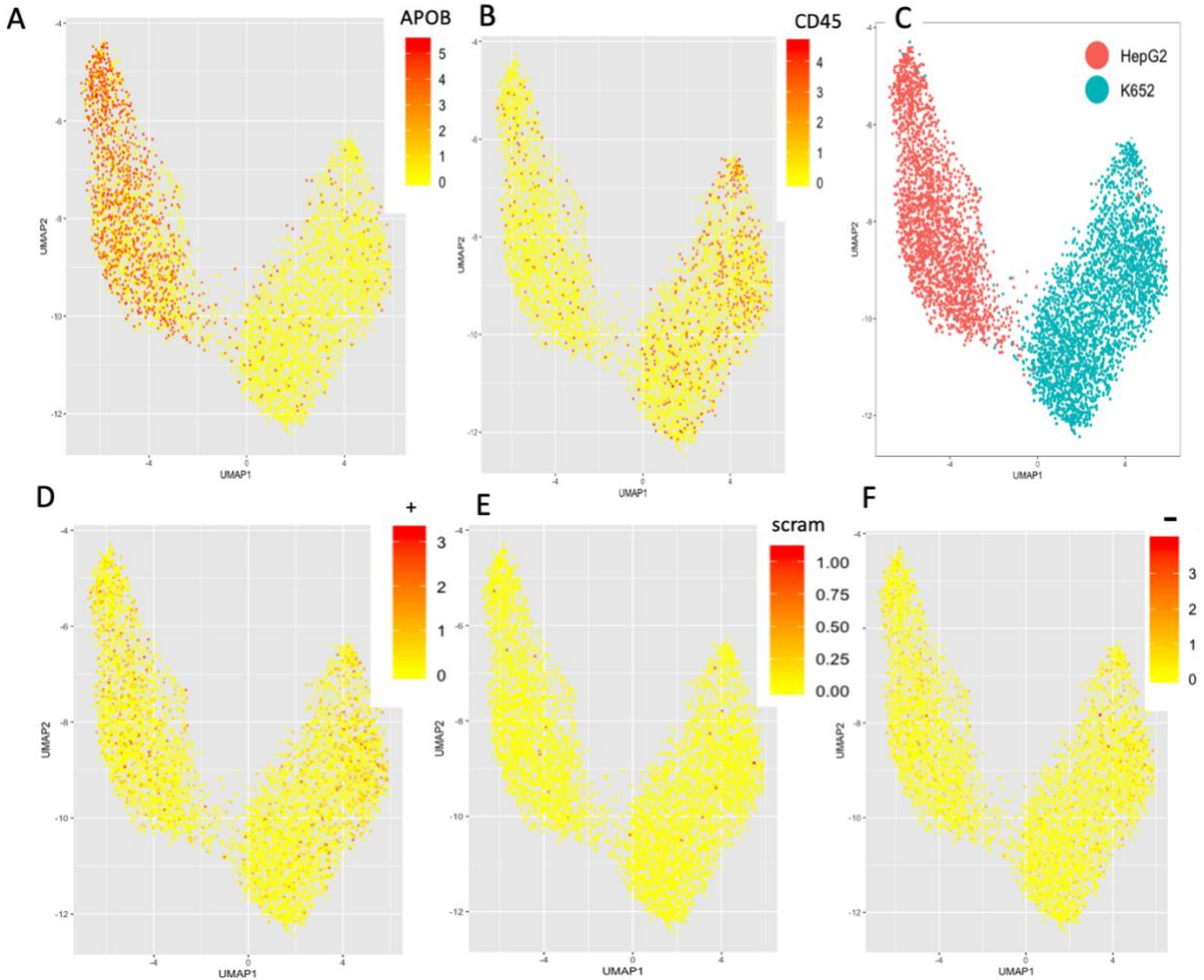
ScMPRA holds great potential for dissecting GxE in single cells. Here we presented work showing a novel design for scMPRA and two pilot experiments. While pilot experiments show constructs can be detected and signal between positive and negative controls reflect the expected outcomes when aggregated, the capture of these constructs is still very sparse. These initial experiments show promise that this method can work, significant optimization is still necessary for robust implementation. To this end we are continuing experiments to maximize capture of constructs including synthesizing barcodes with CRS to reduce loss due to losses in random barcoding due to the association process.



**Figure 5.1: scMPRA design.** A) Design to integrate MPRA and scRNA-seq using the 10x feature barcode technology. B) Experimental design to ensure minimal collisions and maximize coverage of a 200-sequence enhancer library.



**Figure 5.2: LCL and HepG2 Feature Barcode pilot results.** A) Seurat clustering with 15 principal components produces 11 clusters colored by cell type determined by marker genes. LCLs marked by CD74, a component of HLA-II. HepG2 marked by APOB, a lipoprotein made in the liver. B) Unmapped enhancers found in the experiment. C) qPCR cycle thresholds for GFP and control genes: ACTB and GAPDH in scMPRA constructs using a minimal promoter (mP) or SV40 promoter.



**Figure 5.3: K562 and HepG2 5' amplicon pilot results.** (A) Expression of APOB and (B) CD45 (C) specifically mark HepG2 and K562 cell populations. (D) Expression of positive controls is higher than expression of scrambled (E) or previously verified negative controls (F).

## 5.6 References

1. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
2. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
3. Marigorta, U. M. & Gibson, G. A simulation study of gene-by-environment interactions in GWAS implies ample hidden effects. *Front. Genet.* **5**, 225 (2014).
4. McAllister, K. *et al.* Current Challenges and New Opportunities for Gene-Environment Interaction Studies of Complex Diseases. *Am. J. Epidemiol.* **186**, 753–761 (2017).
5. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
6. Ye, C. J. *et al.* Intersection of population variation and autoimmunity genetics in human T cell activation. *Science* **345**, 1254665 (2014).
7. Perez, R. K. *et al.* Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. *Science* **376**, eabf1970 (2022).
8. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**, 159–164 (2015).
9. Tewhey, R. *et al.* Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165**, 1519–1529 (2016).
10. Pope, S. D. & Medzhitov, R. Emerging Principles of Gene Expression Programs and Their Regulation. *Mol. Cell* **71**, 389–397 (2018).
11. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33**, 155–160 (2015).
12. Papalexi, E. & Satija, R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.* **18**, 35–45 (2018).

13. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498 (2015).
14. Gordon, M. G. *et al.* lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. *Nature Protocols* vol. 15 2387–2412 (2020).
15. Mckinney, E. Generalised birthday problem. *American Mathematical Monthly* **73**, 385–387 (1966).
16. Smith, R. P. *et al.* Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat. Genet.* **45**, 1021–1028 (2013).
17. Klein, J. C. *et al.* A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat. Methods* **17**, 1083–1091 (2020).

## 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:

*M Grace Gordon*

BDF4D7B8B8A144E...

Author Signature

8/17/2022

Date