

**UCSF**

**UC San Francisco Electronic Theses and Dissertations**

**Title**

Implementing and Applying Multiplexed Single Cell RNA-sequencing to Reveal Context-specific Effects in Systemic Lupus Erythematosus

**Permalink**

<https://escholarship.org/uc/item/3mx1354t>

**Author**

Subramaniam, Meena

**Publication Date**

2019

Peer reviewed|Thesis/dissertation

Implementing and Applying Multiplexed Single Cell RNA-sequencing to Reveal  
Context-specific Effects in Systemic Lupus Erythematosus

by  
Meena Subramaniam

DISSERTATION

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

in

Biological and Medical Informatics

in the

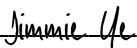
GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:




7CBD3BB0EE2B418...

Jimmie Ye

Chair

DocuSigned by:



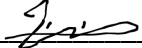
DocuSigned by: 4DB...

Noah Zaitlen



Jonathan Weissman

DocuSigned by: 4C9...



Nir Yosef

1CDCE045B4714BE...

Committee Members



## **Acknowledgements**

The work described in this thesis would not have been possible without significant support from many individuals. I would like to thank my mentors Jimmie Ye and Noah Zaitlen for their constant mentorship and support, particularly in helping me identify and pursue the topics which interested me the most. I would like to thank the Biological and Medical Informatics Program at UCSF for administrative assistance, as well as the UCSF Discovery Fellows Program for their financial as well as mentorship support. Finally, I would also like to thank my thesis committee members Nir Yosef and Jonathan Weissman for their helpful feedback through this process.

The chapter entitled “Multiplexed droplet single-cell RNA-sequencing using natural genetic variation” was published in Nature Biotechnology (PMID: 29227470, doi: 10.1038/nbt.4042).

# **Implementing and Applying Multiplexed Single Cell RNA-sequencing to Reveal Context-specific Effects in Systemic Lupus Erythematosus**

**Meena Subramaniam**

## **Abstract**

Droplet single-cell RNA-sequencing (dscRNA-seq) has enabled rapid, massively parallel profiling of transcriptomes. However, assessing differential expression across multiple individuals has been hampered by inefficient sample processing and technical batch effects. Here we describe a computational tool, demuxlet, that harnesses natural genetic variation to determine the sample identity of each cell and detect droplets containing two cells. These capabilities enable multiplexed dscRNA-seq experiments in which cells from unrelated individuals are pooled and captured at higher throughput than in standard workflows. Using simulated data, we show that 50 SNPs per cell are sufficient to assign 97% of singlets and identify 92% of doublets in pools of up to 64 individuals. Given genotyping data for each of 8 pooled samples, demuxlet correctly recovers the sample identity of >99% of singlets and identifies doublets at rates consistent with previous estimates. We also apply demuxlet to assess cell type-specific changes in gene expression in 8 pooled lupus patient samples treated with IFN- $\beta$  and perform eQTL analysis on 23 pooled samples.

Systemic lupus erythematosus (SLE) is an autoimmune disease defined by a broad range of symptoms that disproportionately affects women. Our knowledge of which immune cells mediate the etiology and pathogenesis of the disease remains incomplete. Identifying pathogenic cells using bulk gene expression analysis is confounded by the functional overlap and frequency

variation of immune cell types. Here, we used multiplexed single-cell RNA-seq (scRNA-seq) to profile ~1 million peripheral blood mononuclear cells from 134 SLE cases and 58 healthy controls. Cases were marked by a reduction of naive CD4<sup>+</sup> T cells, clonal restriction of effector memory CD8<sup>+</sup> T cells, and elevated expression of interferon-stimulated genes in classical monocytes. An additional 15 cases experiencing active disease flares displayed increased expansion of effector memory CD8<sup>+</sup> T cells and the presence of macrophages not seen in managed disease. Although cell-type-specific expression contributed most to inter-individual expression variability across all cells, cell composition accounted for more variability in genes differentially expressed in cases. We integrated dense genotyping data to map thousands of genetic variants, including SLE-associations, whose effects on expression are modified by cell type or interferon activation. Population-scale scRNA-seq analysis reveals changes in cell composition and state associated with SLE, and when integrated with genetic data, ascribes function to disease-associated and disease-modified variants.

## Table of Contents

Chapter 1: Introduction.....	1
References.....	4
Chapter 2: Multiplexed droplet single-cell RNA-sequencing using natural genetic variation.....	6
Introduction.....	7
Results.....	10
Discussion.....	16
Methods.....	17
Figures.....	26
References.....	30
Chapter 3: Multiplexed RNA-sequencing of 1M immune cells reveals the cellular, molecular, and genetic correlates of systemic lupus erythematosus.....	36
Introduction.....	36
Results.....	39
Methods.....	53
Figures.....	58
Discussion.....	67
References.....	70

## List of Figures

Figure 2.1: demultiplexing and doublet identification from single cell data.....	26
Figure 2.2: Performance of demuxlet. ....	27
Figure 2.3: Interindividual variability in IFN- $\beta$ response.....	28
Figure 2.4 – Genetic control over cell type proportion and gene expression (N=23). ....	29
Figure 3.1: Overview and compositional changes in SLE.....	58
Figure 3.2: Bulk expression differences and variance decomposition. ....	60
Figure 3.3: Myeloid changes in SLE .....	62
Figure 3.4: Lymphoid changes in SLE. ....	63
Figure 3.5: cis-eQTL mapping demonstrates cell type specificity and environmental specificity in genetic effects. ....	65
Figure 3.6: SLE flare cohort analysis demonstrates reproducibility of our findings and disease flare specific alterations .....	66



## Chapter 1: Introduction

The annotation and functional interpretation of genetic variants from the human genome has been one of the largest challenges in studying complex traits. Although Genome Wide Association Studies (GWAS) have shed light on genetic loci that may be involved in disease pathology, the function of the individual variants that are associated with complex traits are often unknown and poorly understood in the context of a biological mechanism<sup>1</sup>. To overcome this, recent studies have quantified genetic variant associations with cellular composition, gene expression, chromatin accessibility, and protein expression in tissues of interest which provide more functional context that is relevant to disease states<sup>2-4</sup>. Gene expression and chromatin accessibility have also been used to identify potential biomarkers that are unique to disease states as well as subtype patients into groups where the disease is thought to have multiple functionally distinct mechanisms, suggesting their value in developing better diagnostics and targeted therapies in the future<sup>5,6</sup>.

Despite these advances, gene expression profiling of tissues does not offer clarity when the relevant cell type for a specific disease is unknown. For example, in the case of Systemic Lupus Erythematosus (SLE), an autoimmune disease with highly heterogeneous manifestations that is difficult to diagnose, GWAS studies have pointed to numerous cell types in the blood having involvement in the disease etiology<sup>7</sup>. Likely due to this, studying the peripheral blood in bulk likely does not capture the variability between patients or signatures from the specific cell subtypes that are dysregulated. Several cell subtypes including B cells, CD4 T cell subsets, and CD8 T cells as well as broad clinical phenotypes such as lymphopenia have been studied in the context of SLE, and shown to change in abundance and state through the disease course<sup>8-10</sup>.

Additionally, previous work has shown that in the case of SLE, cell type specific gene expression implicates different signatures in different ancestral backgrounds, suggesting that studying interindividual variation across many cell types may inform more personalized medicine and targeted strategies<sup>11</sup>.

Current bulk gene expression profiling across many cell types is performed by using antibodies for population markers to enrich for each cell type separately, and then prepare individual RNA libraries for each cell type. This procedure is costly, laborious, and prone to confounding effects due to the high number of independent experiments performed. Experimental procedures often take hours to prepare the samples for sequencing, and this impacts the quality and accuracy of the resulting data. Furthermore, the enrichment of specific cell subtypes based on previously identified markers biases the profiling towards known populations and does not lead to the discovery of any novel or unknown cell states that may be relevant to disease. These factors result in biased bulk gene expression profiles that are suboptimal for discovery and studying interindividual variation.

Recent advances in droplet based single cell RNA sequencing have enabled the capture of heterogeneous populations in an unbiased manner<sup>12</sup>. To date, scRNA-seq has been used to characterize heterogeneity in tissues globally, in response to stimulation and knock-down perturbations, and in tumor composition<sup>13-15</sup>. Although these studies have led to the discovery of novel cell populations and shed new light on the dynamics of transcriptional regulation, interindividual variability at single cell resolution remains largely uncharacterized. Previously, single-cell qPCR experiments have shown that the distributional properties of gene expression

may be controlled by genetic variants<sup>16</sup>. This suggests that characterizing interindividual variability across single cells and identifying its genetic basis will aid in the discovery and interpretation of disease-causing genetic polymorphisms. Characterizing molecular quantitative traits in specific cell types from peripheral blood mononuclear cells (PBMCs) will aid in the annotation of SLE-associations and shed new light on the pathogenesis of SLE.

In this work we developed a multiplexed scRNA-seq experimental workflow that significantly decreases the cost and labor time to perform scRNA-seq experiments in large-scale cohorts. We show a proof of concept that our algorithm, *demuxlet*, performs with up to 99% accuracy and are able to replicate biological findings across demultiplexed data. We then applied our workflow to sequence 1M cells across 120 patients with SLE cases and 46 healthy controls, and identified cell composition as well as gene expression features that distinguish cases from controls, and show that cell type specific features are more predictive of clinical criteria for SLE than bulk features. We also performed the first genome-wide single-cell derived eQTL study in patients with SLE and healthy control to discover genetic variants that influence gene expression of different cell subtypes in the immune system.

## References

1. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
2. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
3. Gate, R. E. *et al.* Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nat. Genet.* **50**, 1140–1150 (2018).
4. Battle, A. *et al.* Genomic variation. Impact of regulatory variation from RNA to protein. *Science* **347**, 664–667 (2015).
5. Banchereau, R. *et al.* Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients. *Cell* **165**, 551–565 (2016).
6. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362**, (2018).
7. Guerra, S. G., Vyse, T. J. & Cunninghame Graham, D. S. The genetics of lupus: a functional perspective. *Arthritis Res. Ther.* **14**, 211 (2012).
8. Faddah, S., Elwakd, M., Aboelenein, A. & Hussein, M. Lymphopenia and systemic lupus erythematosus, a preliminary study: Correlation with clinical manifestations, disease activity and damage indices. *The Egyptian Rheumatologist* **36**, 125–130 (2014).
9. Matsushita, M. *et al.* Changes of CD4/CD8 ratio and interleukin-16 in systemic lupus erythematosus. *Clin. Rheumatol.* **19**, 270–274 (2000).
10. Sanz, I. & Lee, F. E.-H. B cells as therapeutic targets in SLE. *Nat. Rev. Rheumatol.* **6**, 326–337 (2010).

11. Sharma, S. *et al.* Widely divergent transcriptional patterns between SLE patients of different ancestral backgrounds in sorted immune cell populations. *J. Autoimmun.* **60**, 51–58 (2015).
12. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).
13. Byrnes, L. E. *et al.* Lineage dynamics of murine pancreatic development at single-cell resolution. *Nat. Commun.* **9**, 3922 (2018).
14. 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).
15. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
16. Wills, Q. F. *et al.* Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat. Biotechnol.* **31**, 748–752 (2013).

## **Chapter 2: Multiplexed droplet single-cell RNA-sequencing using natural genetic variation**

Hyun Min Kang\*<sup>1</sup>, Meena Subramaniam<sup>2-6</sup>, Sasha Targ<sup>2-6,11</sup>, Michelle Nguyen<sup>7-9</sup>, Lenka Maliskova<sup>3,10</sup>, Elizabeth McCarthy<sup>11</sup>, Eunice Wan<sup>3</sup>, Simon Wong<sup>3</sup>, Lauren Byrnes<sup>12</sup>, Cristina Lanata<sup>13,14</sup>, Rachel Gate<sup>2-6</sup>, Sara Mostafavi<sup>15</sup>, Alexander Marson<sup>7-9,16,17</sup>, Noah Zaitlen<sup>3,13,18</sup>, Lindsey A Criswell<sup>3,13,14,19</sup>, Chun Jimmie Ye<sup>3-6\*</sup>

1. Department of Biostatistics and Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, Michigan, United States of America

2. Biological and Medical Informatics Graduate Program, University of California, San Francisco, California, USA

3. Institute for Human Genetics (IHG), University of California San Francisco, California, USA

4. Institute for Computational Health Sciences, University of California San Francisco, California, USA

5. Department of Epidemiology and Biostatistics, University of California San Francisco, California, USA

6. Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, California, USA

7. Department of Microbiology and Immunology, University of California, San Francisco, California, USA

8. Diabetes Center, University of California, San Francisco, California, USA

9. Innovative Genomics Institute, University of California, Berkeley, California, USA

10. Department of Neurology, University of California, San Francisco, San Francisco, California, USA

11. Medical Scientist Training Program (MSTP), University of California, San Francisco, California, USA
12. Developmental and Stem Cell Biology Graduate Program, University of California, San Francisco, California, USA
13. Department of Medicine, University of California, San Francisco
14. Rosalind Russell/Ephraim P Engleman Rheumatology Research Center, University of California, San Francisco, San Francisco, California, USA
15. Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada
16. UCSF Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, CA, USA
17. Chan Zuckerberg Biohub, San Francisco, California, USA
18. Lung Biology Center, University of California, San Francisco, CA, USA
19. Department of Orofacial Sciences, University of California San Francisco, USA

## **Introduction**

Droplet single cell RNA-sequencing (dscRNA-seq) has increased substantially the throughput of single cell capture and library preparation<sup>1, 10</sup>, enabling the simultaneous profiling of thousands of cells. Improvements in biochemistry<sup>11, 12</sup> and microfluidics<sup>13, 14</sup> continue to increase the number of cells and transcripts profiled per experiment. But for differential expression and population genetics studies, sequencing thousands of cells per individual would better capture inter-individual variability than sequencing more cells from a few individuals. However, in standard workflows, dscRNA-seq of many samples in parallel remains challenging to implement.

If the genetic identity of each cell could be determined, pooling cells from different individuals in one microfluidic run would result in lower per-sample library preparation cost and eliminate confounding effects. Furthermore, if droplets containing multiple cells from different individuals could be detected, pooled cells could be loaded at higher concentrations, enabling additional reduction in per-cell library preparation cost.

Here we develop an experimental protocol for multiplexed dscRNA-seq and a computational algorithm, demuxlet, that harnesses genetic variation to determine the genetic identity of each cell (demultiplex) and identify droplets containing two cells from different individuals (**Fig. 2.1**).

While strategies to demultiplex cells from different species<sup>1, 10, 17</sup> or host and graft samples<sup>17</sup> have been reported, simultaneously demultiplexing and detecting doublets from more than two individuals has not been possible. Inspired by models and algorithms developed for detecting contamination in DNA sequencing<sup>18</sup>, demuxlet is fast, accurate, scalable, and compatible with standard input formats<sup>17, 19, 20</sup>.

Demuxlet implements a statistical model for evaluating the likelihood of observing RNA-seq reads overlapping a set of single nucleotide polymorphisms (SNPs) from a single cell. Given a set of best-guess genotypes or genotype probabilities obtained from genotyping, imputation or sequencing, demuxlet uses maximum likelihood to determine the most likely donor for each cell using a mixture model. A small number of reads overlapping common SNPs is sufficient to accurately identify each cell. For a pool of 8 individuals and a set of uncorrelated SNPs each with 50% minor allele frequency (MAF), 4 reads overlapping SNPs are sufficient to uniquely assign a cell to the donor of origin (**Fig. 2.1**) and 20 reads overlapping SNPs can distinguish every sample with >98% probability in simulation (Supplementary Fig. 1). We note that by multiplexing even a small number of individuals, the probability that a doublet contains cells



from different individuals is very high ( $1 - 1/N$ , e.g., 87.5% for  $N=8$  samples) (**Fig. 2.1**). For example, if a 1,000 cell run without multiplexing results in 990 singlets with a 1% undetected doublet rate, multiplexing 1,570 cells each from 63 samples can theoretically achieve the same rate of undetected doublets, producing up to a 37-fold larger number of singlets (36,600) if the sample identity of every droplet can be perfectly demultiplexed (Supplementary Fig. 2, see Methods for details). To minimize the effects of sequencing doublets, profiling 22,000 cells multiplexed from 26 individuals generates 23-fold more singlets at the same effective doublet rate (Supplementary Fig. 3).

## Results

We first assess the performance of multiplexed dscRNA-seq through simulation. The ability to demultiplex cells is a function of the number of individuals multiplexed, the depth of sequencing or number of read-overlapping SNPs, and relatedness of multiplexed individuals. We simulated 6,145 cells (5,837 singlets and 308 doublets) from 2 – 64 individuals from the 1000 Genomes Project<sup>21</sup>. We show that 50 SNPs per cell allows demultiplexing of 97% of singlets and identification of 92% of doublets in pools of up to 64 individuals (Supplementary Figs. 4-5, see Methods for details). Simulating a range of sequencing depths, we determined that 50 SNPs can be obtained with as few as 1,000 unique molecular identifiers (UMIs) per cell (Supplementary Fig. 6), and recommended sequencing depths of standard dscRNA-seq workflows would capture hundreds of SNPs. To assess dependence on the relatedness of multiplexed individuals, we simulated 6,145 cells from a set of 8 related individuals from 1000 Genomes<sup>21</sup>. In this simulation, 50 SNPs per cell would allow demuxlet to correctly assign over 98% of cells (Supplementary Fig. 7). These results suggest optimal multiplexed designs where cells from tens of unrelated individuals should be pooled, loaded at concentrations 2-10x higher than standard workflows, and sequenced to at least 1,000 UMIs per cell.

We evaluate the performance of demuxlet by analyzing a pool of peripheral blood mononuclear cells (PBMCs) from 8 lupus patients. By sequential pairwise pooling, three pools of equimolar concentrations of cells were generated (W1: patients S1-S4, W2: patients S5-S8 and W3: patients S1-S8) and each loaded in a well on a 10X Chromium instrument (**Fig. 2.2**). 3,645 (W1), 4,254 (W2) and 6,205 (W3) cell-containing droplets were sequenced to an average depth of 51,000, 39,000 and 28,000 reads per droplet.

In wells W1, W2 and W3, demuxlet identified 91% (3332/3645), 91% (3864/4254), and 86% (5348/6205) of droplets as singlets (likelihood ratio test,  $L(\text{singlet})/L(\text{doublet}) > 2$ ), of which 25% (+/- 2.6%), 25% (+/- 4.6%) and 12.5% (+/- 1.4%) mapped to each donor, consistent with equal mixing of individuals in each well. From wells W1 and W2, each containing cells from two disjoint sets of 4 individuals, we estimated a demultiplexing error rate (number of cells assigned to individuals not in the pool) of less than 1% of singlets (W1: 2/3332, W2: 0/3864) (**Fig. 2.2**).

We next assess the ability of demuxlet to detect doublets in both simulated and real data. 466/3645 (13%) droplets from W1 were simulated as synthetic doublets by setting the cellular barcodes of 466 cells each from individuals S1 and S2 to be the same. Applied to simulated data, demuxlet identified 91% (426/466) of synthetic doublets as doublets or ambiguous, correctly recovering the sample identity of both cells in 403/426 (95%) doublets (Supplementary Fig. 8). Applied to real data from W1, W2 and W3, demuxlet identified 138/3645, 165/4254, and 384/6205 doublets corresponding to doublet rates of 5.0%, 5.2% and 7.1%, consistent with the expected doublet rates estimated from mixed species experiments (**Fig. 2.2**).

Demultiplexing of pooled samples allows for the statistical and visual comparisons of individual-specific dscRNA-seq profiles. Singlets identified by demuxlet in all three wells cluster into known immune cell types (**Fig. 2.2**) and are correlated with bulk RNA-sequencing of sorted cell populations ( $R=0.76-0.92$ ) (Supplementary Fig. 9). For the same individuals from different wells, t-distributed stochastic neighbor embedding (t-SNE) of dscRNA-seq data are qualitatively consistent, and estimates of cell type proportions are highly correlated ( $R = 0.99$ ) (**Fig. 2.2** and Supplementary Fig. 10). Further, t-SNE projections of the pool and each individual are not confounded by well-to-well effects (Supplementary Fig. 11a). While 6 genes were differentially

expressed between wells W1 and W2 (DESeq2 on pseudobulk counts, FDR < 0.05), only 2 genes were differentially expressed between W1 and W2 individuals in well W3 (FDR < 0.05) (Supplementary Fig. 11b), suggesting multiplexing reduces technical effects due to separate sample processing<sup>22, 23</sup>.

We used multiplexed dscRNA-seq to characterize the cell type specificity and inter-individual variability of response to IFN- $\beta$ , a potent cytokine that induces genome-scale changes in the transcriptional profiles of immune cells<sup>24, 25</sup>. From each of 8 lupus patients, PBMCs were activated with recombinant IFN- $\beta$  or left untreated for 6 hours, a time point we previously found to maximize the expression of interferon-sensitive genes (ISGs) in dendritic cells (DCs) and T cells<sup>26, 27</sup>. Two pools, IFN- $\beta$ -treated and control, were prepared with the same number of cells from each individual and loaded onto the 10X Chromium instrument.

We obtained 14,619 (control) and 14,446 (stimulated) cell-containing droplets, of which demuxlet identified 83% (12,138) and 84% (12,167) as singlets. The estimated doublet rate of 10.9% in each condition is consistent with predicted rates (**Fig. 2.2**) and the observed and expected frequencies of doublets for each pair of individuals are highly correlated (R=0.98) (Supplementary Fig. 12). Detected doublets form distinct clusters near the periphery of other clusters defined by cell type (Supplementary Fig. 13).

Demultiplexing individuals enables the use of the 8 individuals within each pool as biological replicates to quantitatively assess cell type-specific IFN- $\beta$  responses in PBMCs. Consistent with previous reports from bulk RNA-sequencing data, IFN- $\beta$  stimulation induces widespread transcriptomic changes observed as a shift in the t-SNE projections of singlets<sup>24</sup> (**Fig. 2.3**). As expected, IFN- $\beta$  did not affect cell type proportions between control and stimulated cells (Supplementary Fig. 14), and these were consistent with flow cytometry measurements (R=0.88)

(Supplementary Fig. 15). Estimates of abundances for ~2000 homologous genes in each cell type and condition correlated with similar data from mice (Supplementary Fig. 16). We identified 3,055 differentially expressed genes ( $\log_{2}FC > 2$ ,  $FDR < 0.05$ ) in at least one cell type (Supplementary Table 1). For 709 genes, estimates of fold change in response to IFN- $\beta$  stimulation in myeloid and CD4<sup>+</sup> cells are consistent with estimates in monocyte derived dendritic cells<sup>28</sup> and CD4<sup>+</sup> T cells<sup>27</sup>, respectively (Supplementary Fig. 17) and correlated with qPCR results of sorted CD4<sup>+</sup> T cells (Supplementary Fig. 18). Differentially expressed genes cluster into modules of cell type-specific responses enriched for distinct gene regulatory programs (**Fig. 2.3**, Supplementary Table 2). For example, genes upregulated in all leukocytes (Cluster III: 401 genes,  $\log_{2}FC > 2$ ,  $FDR < 0.05$ ) or only in myeloid cells (Cluster I: 767 genes,  $\log_{2}FC > 2$ ,  $FDR < 0.05$ ) are enriched for general antiviral response (e.g. KEGG Influenza A: Cluster III  $P < 1.6 \times 10^{-5}$ ), chemokine signaling (Cluster I  $P < 7.6 \times 10^{-3}$ ) and pathways active in systemic lupus erythematosus (Cluster I  $P < 4.4 \times 10^{-3}$ ). The five clusters of downregulated genes are enriched for antibacterial response (KEGG Legionellosis: Cluster II monocyte down  $P < 5.5 \times 10^{-3}$ ) and natural killer cell mediated toxicity (Cluster IV NK/Th cell down:  $P < 3.6 \times 10^{-2}$ ). The analysis of multiplexed dscRNA-seq data recovers cell type-specific gene regulatory programs affected by interferon stimulation consistent with published IFN- $\beta$  signatures in mouse and humans<sup>29</sup>.

Over all PBMCs, the variance of mean expression across individuals is higher than the variance across synthetic replicates whose cells were randomly sampled (Lin's concordance = 0.022, Pearson correlation = 0.69, **Fig. 2.3**). The variance across synthetic replicates whose cells were sampled matching for cell type proportions is more concordant with the variance across individuals (Lin's concordance = 0.54, Pearson correlation = 0.78, **Fig. 2.3**), suggesting a

contribution of cell type composition on expression variability. However, for each cell type, the variance across individuals<sup>22, 30</sup> is also higher than the variance across synthetic replicates (Lin's concordance = 0.007-0.20) suggesting additional inter-individual variability not due to cell type composition (Supplementary Fig. 19). In CD14<sup>+</sup>CD16<sup>-</sup> monocytes, the correlation of mean expression between pairs of synthetic replicates from the same individual (>99%) is greater than from different individuals (~97%), further indicating inter-individual variation beyond sampling (**Fig. 2.3**). We found between 15 to 827 genes with statistically significant inter-individual variability in control cells and 7 to 613 in stimulated cells (Pearson correlation, FDR < 0.05), with most found in classical monocytes (cM) and CD4<sup>+</sup> helper T (Th) cells. Inter-individual variable genes in stimulated cM and to a lesser extent in Th cells ( $P < 9.3 \times 10^{-4}$  and  $4.5 \times 10^{-2}$ , hypergeometric test, **Fig. 2.3**) are enriched for differentially expressed genes, consistent with our previous discovery of more IFN- $\beta$  response-eQTLs in monocyte-derived dendritic cells than CD4<sup>+</sup> T cells<sup>26, 27</sup>. Comparing to 407 genes previously profiled in bulk monocyte-derived dendritic cells, the proportion of variance explained by inter-individual variability is more correlated in myeloid cells after stimulation ( $R = 0.26 - 0.3$ ) than before ( $R = 0.05 - 0.19$ ). To map genetic variants associated with cell type proportions and cell type-specific expression using multiplexed dscRNA-seq, we sequenced an additional 15,250 (7 donors), 22,619 (8 donors) and 25,918 cells (15 donors; 8 lupus patients, 5 rheumatoid arthritis patients, and 2 healthy controls). Demuxlet identified 71% (10,766/15,250), 73% (16,618/22,619) and 60% (15,596/25,918) of droplets as singlets, correctly assigning 99% of singlets from the first two pools, W1 and W2 (10,740/10,766 and 16,616/16,618). The estimated doublet rates of 18%, 18% and 25% are consistent with the increased concentrations of loaded cells (**Fig. 2.2**). Similar to the IFN- $\beta$  stimulation experiment, we found that expression variability was determined by variability

in cell type proportion (**Fig. 2.4**) and reproducible between batches (Supplementary Fig. 20). Associating >150,000 genetic variants (MAF > 20%) with the proportion of 8 major immune cell populations, we identified a SNP (chr10:3791224) significantly associated ( $P = 1.03 \times 10^{-5}$ , FDR < 0.05) with the proportion of NK cells (**Fig. 2.4**).

Across 23 donors, we conducted an expression quantitative trait loci (eQTL) analysis to map genetic variants associated with expression variability in each major immune cell type. We found a total of 32 local eQTLs (+/- 100kb, FDR < 0.1), 22 of which were detected in only one cell type (**Fig. 2.4**, Supplementary Table 3). Previously reported local eQTLs from bulk CD14<sup>+</sup> monocytes, CD4<sup>+</sup> T cells and lymphoblastoid cell lines are more significantly associated with gene expression in the most similar cell types (cM, Th and B cells, respectively) than other cell types (**Fig. 2.4**). We used an inverse variance weighted meta-analysis to identify genes with pan-cell type eQTLs, including those in the major histocompatibility complex (MHC) class I antigen presentation pathway including *ERAP2* ( $P < 3.57 \times 10^{-32}$ , meta-analysis), encoding an aminopeptidase known to cleave viral peptides<sup>34</sup>, and *HLA-C* ( $P < 1.74 \times 10^{-29}$ , meta-analysis), which encodes the MHC class I heavy chain (**Fig. 2.4**). *HLA-DQA1* has local eQTLs only in some cell types ( $P < 2.11 \times 10^{-15}$ , Cochran's Q) while *HLA-DQA2* has local eQTLs in all antigen presentation cells ( $P < 1.02 \times 10^{-43}$ , Cochran's Q). Among other cell type-specific local eQTLs are *CD52*, a gene ubiquitously expressed in leukocytes that only has eQTLs in monocyte populations, and *DIP2A*, a gene with an eQTL only in NK cells that is associated with immune response to vaccination in peripheral blood<sup>35</sup>. These results demonstrate the ability of multiplexed dscRNA-seq to characterize inter-individual variation in immune response and when integrated with genetic data, reveal cell type-specific genetic control of gene expression, which would be undetectable when bulk tissues are analyzed.

## Discussion

The capability to demultiplex and identify doublets using natural genetic variation reduces the per-sample and per-cell library preparation cost of single-cell RNA-sequencing, does not require synthetic barcodes or split-pool strategies<sup>36-40</sup>, and captures biological variability among individual samples while limiting unwanted technical variability. We find the optimal number of samples to multiplex is approximately 20, based on sample processing time and empirical doublet rates of current microfluidic devices and anticipate that automated sample handling and lower doublet rates will increase the optimal number of individuals to multiplex.

Compared to sorting known cell types followed by bulk RNA-seq, multiplexed dscRNA-seq is a more efficient and unbiased method for obtaining cell type-specific immune traits<sup>41</sup>. Demuxlet enables reliable estimation of cell type proportion, recovers cell type-specific transcriptional response to stimulation, and could facilitate further genetic and longitudinal analyses in relevant cell types and conditions across a range of sampled individuals, including between healthy controls and disease patients<sup>42-44</sup>. While demuxlet could in principle be applied to sequencing solid tissue, standardizing sample processing and preservation remain major challenges.

Although we developed demuxlet specifically for RNA-sequencing, we anticipate that the computational framework could be easily extended to other single cell assays where synthetic barcodes or natural genetic variation are measured by sequencing.



## Methods

### Identifying the sample identity of each single cell.

We first describe the method to infer the sample identity of each cell in the absence of doublets. Consider RNA-sequence reads from  $C$  barcoded droplets multiplexed across  $S$  different samples, where their genotypes are available across  $V$  exonic variants. Let  $d_{cv}$  be the number of unique reads overlapping with the  $v$ -th variant from the  $c$ -th droplet. Let  $b_{cvi} \in \{R, A, O\}$ ,  $i \in \{1, \dots, d_{cv}\}$  be the variant-overlapping base call from the  $i$ -th read, representing reference (R), alternate (A), and other (O) alleles respectively. Let  $e_{cvi} \in \{0, 1\}$  be a latent variable indicating whether the base call is correct (0) or not (1), then given  $e_{cvi} = 0$ ,  $b_{cvi} \in \{R = 0, A = 1\}$  and  $\sim \text{Binomial}\left(2, \frac{g}{2}\right)$  when  $g \in \{0, 1, 2\}$  is the true genotype of sample corresponding to  $c$ -th droplet at  $v$ -th variant. When  $e_{cvi} = 1$ , we assume that  $\Pr(b_{cvi}|g, e_{cvi})$  follows Supplementary Table 4.  $e_{cvi}$  is assumed to follow Bernoulli  $\left(10^{-\frac{q_{cvi}}{10}}\right)$  where  $q_{cvi}$  is a phred-scale quality score of the observed base call. We use the standard 10X pipeline to process the raw reads which estimates the phred-scale quality score based on the alignment of each read to the reference human transcriptome using the STAR aligner<sup>49</sup>.

We allow uncertainty of observed genotypes at the  $v$ -th variant for the  $s$ -th sample using  $P_{sv}^{(g)} = \Pr(g|\text{Data}_{sv})$ , the posterior probability of a possible genotype  $g$  given external DNA data  $\text{Data}_{sv}$  (e.g. sequence reads, imputed genotypes, or array-based genotypes). If genotype likelihood  $\Pr(\text{Data}_{sv}|g)$  is provided (e.g. unphased sequence reads) instead, it can be converted to a posterior probability scale using  $P_{sv}^{(g)} = \Pr(\text{Data}_{sv}|g)\Pr(g)$  where  $\Pr(g) \sim \text{Binomial}(2, p_v)$  and  $p_v$  is the population allele frequency of the alternate allele. To

allow errors  $\varepsilon$  in the posterior probability, we replace it with  $(1 - \varepsilon)P_{sv}^{(g)} + \varepsilon\Pr(g)$ . The overall likelihood that the  $c$ -th droplet originated from the  $s$ -th sample is

$$L_c(s) = \prod_{v=1}^V \left[ \sum_{g=0}^2 \left\{ \prod_{i=1}^{d_{cv}} \left( \sum_{e=0}^1 \Pr(b_{cvi}|g, e) \right) P_{sv}^{(g)} \right\} \right] \quad (1)$$

In the absence of doublets, we use the maximum likelihood to determine the best-matching sample as  $\operatorname{argmax}_s [L_c(s)]$ .

### Screening for droplets containing multiple samples.

To identify doublets, we implement a mixture model to calculate the likelihood that the sequence reads originated from two individuals, and the likelihoods are compared to determine whether a droplet contains cells from one or two samples. If sequence reads from the  $c$ -th droplet originate from two different samples,  $s_1, s_2$  with mixing proportions  $(1 - \alpha) : \alpha$ , then the likelihood in (1) can be represented as the following mixture distribution<sup>18</sup>,

$$L_c(s_1, s_2, \alpha) = \prod_{v=1}^V \left[ \sum_{g_1, g_2} \left\{ \prod_{i=1}^{d_{cv}} \left( \sum_{e=0}^1 (1 - \alpha) \Pr(b_{cvi}|g_1, e) + \alpha \Pr(b_{cvi}|g_2, e) \right) P_{sv}^{(g_1)} P_{sv}^{(g_2)} \right\} \right]$$

To reduce the computational cost, we consider discrete values of  $\alpha \in \{\alpha_1, \dots, \alpha_M\}$ , (e.g. 5 - 50% by 5%). We determine that it is a doublet between samples  $s_1, s_2$  if and only if

$$\frac{\max_{s_1, s_2, \alpha} L_c(s_1, s_2, \alpha)}{\max_s L_c(s)} \geq t \text{ and the most likely mixing proportion is estimated to be}$$

$\operatorname{argmax}_\alpha L_c(s_1, s_2, \alpha)$ . We determine that the cell contains only a single individual  $s$  if

$$\frac{\max_{s_1, s_2, \alpha} L_c(s_1, s_2, \alpha)}{\max_s L_c(s)} \leq \frac{1}{t}, \text{ and less confident droplets are classified as ambiguous. While we}$$

consider only doublets for estimating doublet rates, we remove all doublets and ambiguous droplets to conservatively estimate singlets. Supplementary Fig. 8 illustrates the distribution of singlet, doublet likelihoods and the decision boundaries when  $t = 2$  was used.

### Theoretical expectation of deconvoluting singlets.

The theoretical distribution of expected singlets with multiplexing (presented in Supplementary Fig. 2) is as follows. Let  $d_0$  (e.g. 0.01) be the proportion of true multiplets when  $x_0$  (1,000) cells are loaded when multiplexing was not used. Then the expected multiplet rates when  $x$  cells are loaded can be modeled exponentially as  $d(x) = 1 - (1 - d_0)^{\frac{x}{x_0}}$ . Let  $\alpha$  be the fraction of true singlets incorrectly classified as non-singlets (i.e. doublet or ambiguous), and  $\beta$  be the fraction of multiplets correctly classified as non-singlets. When multiplexing  $x$  cells equally from  $n$  samples, the expected multiplet rates are  $d(x)$ , and  $\frac{1}{n}d(x)$  are expected to be undetectable doublets mixed between the cells from the same sample. Therefore, the overall effective multiplet rate is  $\left[\frac{n-(n-1)\beta}{n}\right]d(x)$ . Similarly, the expected number of correctly identified singlets becomes  $\frac{(1-\alpha)[1-d(x)]x_0d(x)}{-\log(1-d_0)}$ . Given  $\alpha, \beta$  the expected number of singlets can be calculated by fixing the multiplet rate  $d(x) = d_0$ . We used  $d_0 = 0.01, x_0 = 1000$  for the simulation in Supplementary Fig. 2.

### Dependence of demultiplexing performance on experimental design parameters.

The demuxlet ‘plp’ option was used to generate a pileup format of 6,145 cells from one well of PBMC 10x data. The reads in the pileup were then modified to reflect the genotypes of individuals sampled from the 1000 Genomes Phase 3 cohort. The pileup was downsampled to obtain different numbers of read-overlapping exonic SNPs (ranging from 5,000 to 100,000) for the whole cohort. To create simulated doublets, we randomly sampled and merged pairs of barcodes within a dataset, resulting in a 5% doublet rate in the original data. For simulations with related individuals, we simulated transcriptomes from 8 individuals in 1000 Genomes with

varying degrees of relatedness, ranging from unrelated to parent-child (HG00146, HG00147, HG00500, HG00501, HG00502, HG00512, HG00514, and HG00524).

*Isolation and preparation of PBMC samples.*

Informed consent was obtained from all patients sequenced in this study. Peripheral blood mononuclear cells were isolated from patient donors, Ficoll separated, and cryopreserved by the UCSF Core Immunologic Laboratory (CIL). PBMCs were thawed in a 37°C water bath, and subsequently washed and resuspended in EasySep buffer (STEMCELL Technologies). Cells were treated with DNaseI and incubated for 15 min at RT before filtering through a 40um column. Finally, the cells were washed in EasySep and resuspended in 1x PBMS and 0.04% bovine serum albumin. Cells from 8 donors were then re-concentrated to 1M cells per mL and then serially pooled. At each pooling stage, 1M cells per mL were combined to result in a final sample pool with cells from all donors.

*IFN- $\beta$  stimulation and culture.*

Prior to pooling, samples from 8 individuals were separated into two aliquots each. One aliquot of PBMCs was activated by 100 U/mL of recombinant IFN- $\beta$  (PBL Assay Science) for 6 hrs according to the published protocol<sup>26</sup>. The second aliquot was left untreated. After 6 hrs, the 8 samples for each condition were pooled together in two final pools (stimulated cells and control cells) as described above.

### Fluorescence-activated cell sorting and analysis.

1M PBMCs from each donor were stained using standard procedure (30 min, 4 C) with the following surface antibody panel (CD3-PerCP clone SK7 (BioLegend), CD4-APC clone OKT4 (BioLegend), CD8-BV570 clone RPA-T8 (BioLegend), CD14-FITC clone 63D3 (BioLegend), CD19-BV510 clone SJ25C1 (BD), and Ghost dye A710 viability stain (Tonbo)) (Life Sciences Reporting Summary). Samples were then analyzed and sorted using a BD FACSAria Fusion instrument at the UCSF flow cytometry core. To calculate cell type proportions, the number of events in each of CD3<sup>+</sup> CD4<sup>+</sup> CD8<sup>-</sup> (CD4<sup>+</sup> T cells), CD3<sup>+</sup> CD4<sup>-</sup> CD8<sup>+</sup> (CD8<sup>+</sup> T cells), CD3<sup>-</sup> CD19<sup>+</sup> (B cells), and CD3<sup>-</sup> CD14<sup>+</sup> (monocytes) were divided by the sum of events in these gates (Supplementary Fig. 21).

### Quantitative polymerase chain reaction analysis.

RNA was isolated from sorted CD4<sup>+</sup> T cells following the RNeasy micro kit protocol (QIAGEN), and cDNA was prepared using MultiScribe Reverse Transcriptase (Applied Biosystems cat #4368814). The qPCR primers were chosen from the PrimerBank reference when available<sup>50</sup>. Each sample was run in triplicate with the Luminaris HiGreen qPCR kit (Thermo Scientific #K0992) according to standard protocol using a Roche Light Cycler 96 instrument and fold change was calculated from  $\Delta\Delta CT$  between control and stimulated samples with GAPDH as a reference gene.

### Droplet-based capture and sequencing.

Cellular suspensions were loaded onto the 10x Chromium instrument (10x Genomics) and sequenced as described in Zheng et al<sup>17</sup>. The cDNA libraries were sequenced using a custom

program on 10 lanes of Illumina HiSeq2500 Rapid Mode, yielding 1.8B total reads and 25K reads per cell. At these depths, we recovered >90% of captured transcripts in each sequencing experiment.

#### *Bulk isolation and sequencing.*

PBMCs from lupus patients were isolated and prepared as described above. Once resuspended in EasySep buffer, the EasyEight Magnet was used to sequentially isolate CD14<sup>+</sup> (using the EasySep Human CD14 positive selection kit II, cat #17858), CD19<sup>+</sup> (using the EasySep Human CD19 positive selection kit II, cat #17854), CD8<sup>+</sup> (EasySep Human CD8 positive selection kitII, cat#17853), and CD4<sup>+</sup> cells (EasySep Human CD4 T cell negative isolation kit (cat #17952) according to the kit protocol. RNA was extracted using the RNeasy Mini kit (#74104), and reverse transcription and tagmentation were conducted according to Picelli et al. using the SmartSeq2 protocol<sup>51, 52</sup>. After cDNA synthesis and tagmentation, the library was amplified with the Nextera XT DNA Sample Preparation Kit (#FC-131-1096) according to protocol, starting with 0.2ng of cDNA. Samples were then sequenced on one lane of the Illumina HiSeq4000 with paired end 100bp read length, yielding 350M total reads.

#### *Alignment and initial processing of single cell sequencing data.*

We used the CellRanger v1.1 and v1.2 software with the default settings to process the raw FASTQ files, align the sequencing reads to the hg19 transcriptome, and generate a filtered UMI expression profile for each cell<sup>17</sup>. The raw UMI counts from all cells and genes with nonzero counts across the population of cells were used to generate t-SNE profiles.

### Cell type classification and clustering.

To identify known immune cell populations in PBMCs, we used the Seurat package to perform unbiased clustering on the 2.7k PBMCs from Zheng et al., following the publicly available Guided Clustering Tutorial<sup>17, 53</sup>. The FindAllMarkers function was then used to find the top 20 markers for each of the 8 identified cell types. Cluster averages were calculated by taking the average raw count across all cells of each cell type. For each cell, we calculated the Spearman correlation of the raw counts of the marker genes and the cluster averages, and assigned each cell to the cell type to which it had maximum correlation.

### Differential expression analysis.

Demultiplexed individuals were used as replicates for differential expression analysis. For each gene, raw counts were summed for each individual. We used the DESeq2 package to detect differentially expressed genes between control and stimulated conditions<sup>54</sup>. Genes with  $\text{baseMean} > 1$  were filtered out from the DESeq2 output, and the qvalue package was used to calculate  $\text{FDR} < 0.05$ <sup>55</sup>.

### Estimation of interindividual variability in PBMCs.

For each individual, we found the mean expression of each gene with nonzero counts. The mean was calculated from the  $\log_2$  single cell UMI counts normalized to the median count for each cell. To measure interindividual variability, we then calculated the variance of the mean expression across all individuals. Lin's concordance correlation coefficient was used to compare the agreement of observed data and synthetic replicates. Synthetic replicates were generated by sampling without replacement either from all cells or cells matched for cell type proportion. Cell

type-specific variability estimated as the correlation between synthetic replicates was compared to variability estimates from 23 biological replicates of bulk IFN-stimulated monocyte-derived dendritic cells. Protein coding genes (407/414) originally measured using Nanostring (a hybridization based PCR-free quantification method) were assessed, and variability in the bulk dataset was estimated as repeatability using a linear mixed model<sup>56,26</sup>.

#### Estimation of interindividual variability within cell types.

For each cell type, we generated two bulk equivalent replicates for each individual by summing raw counts of cells sampled without replacement. We used DESeq2 to generate variance-stabilized counts across all replicates. To filter for expressed genes, we performed all subsequent analyses on genes with 5% of samples with > 0 counts. The correlation of replicates was performed on the log<sub>2</sub> normalized counts. Pearson correlation of the two replicates from each of the 8 individuals was used to find genes with significant interindividual variability.

#### Quantitative trait mapping in major immune cell types.

Genotypes were imputed with EAGLE<sup>57</sup> and filtered for MAF > 0.2, resulting in a total of 189,322 SNPs. Cell type proportions were calculated as number of cells for each cell type divided by the number of total cells for each person. Linear regression was used to test associations between each genetic variant and cell-type proportion with the Matrix eQTL software<sup>58</sup>. Cis-eQTL mapping was conducted in each cell type separately. All genes with at least 50 UMI counts in 20% of the individuals in all PBMCs were tested for each cell type, resulting in a total of 4,555 genes. Variance-stabilized and log-normalized gene expression was calculated using the 'rlog' function of the DESeq2 package<sup>54</sup>. All variants within a window of



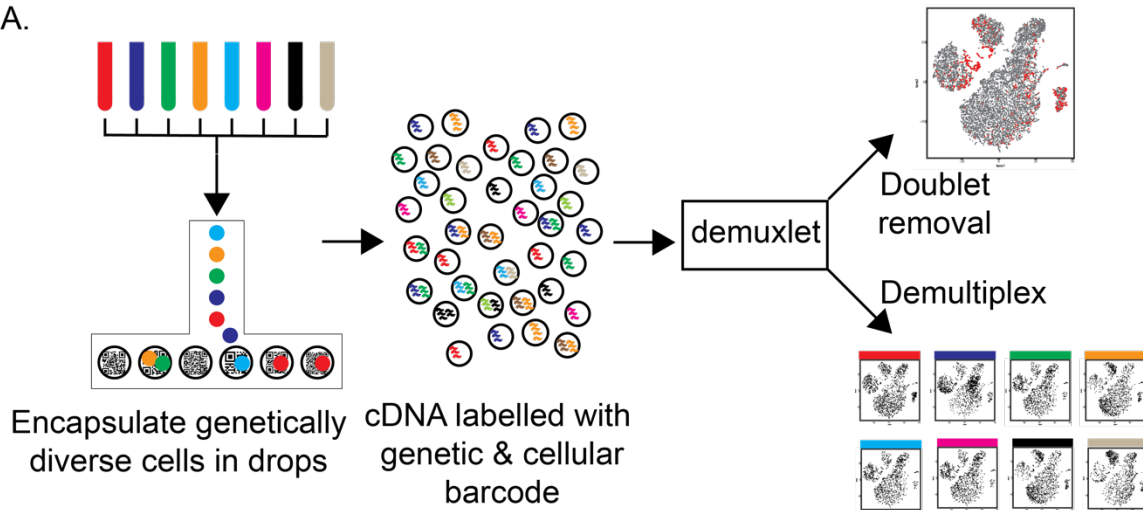
100kbp of each gene were tested with linear regression using Matrix eQTL<sup>58</sup>. Batch information for each sample as well as the first 3 principal components of the expression matrix were used as covariates.

Single cell and bulk RNA-sequencing data has been deposited in the Gene Expression Omnibus under the accession number GSE96583. Demuxlet software is freely available at <https://github.com/statgen/demuxlet>

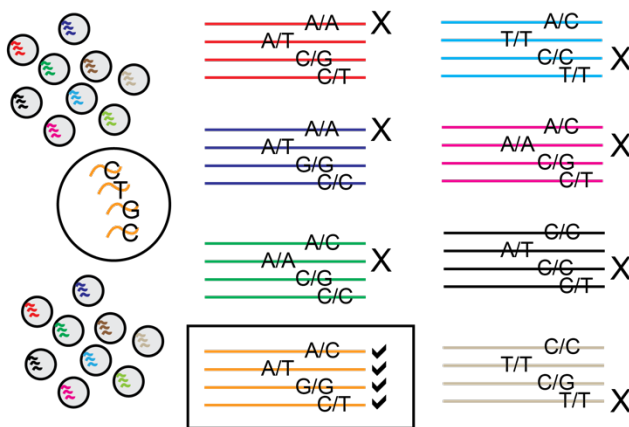
## Figures

Fig. 1

A.



B.



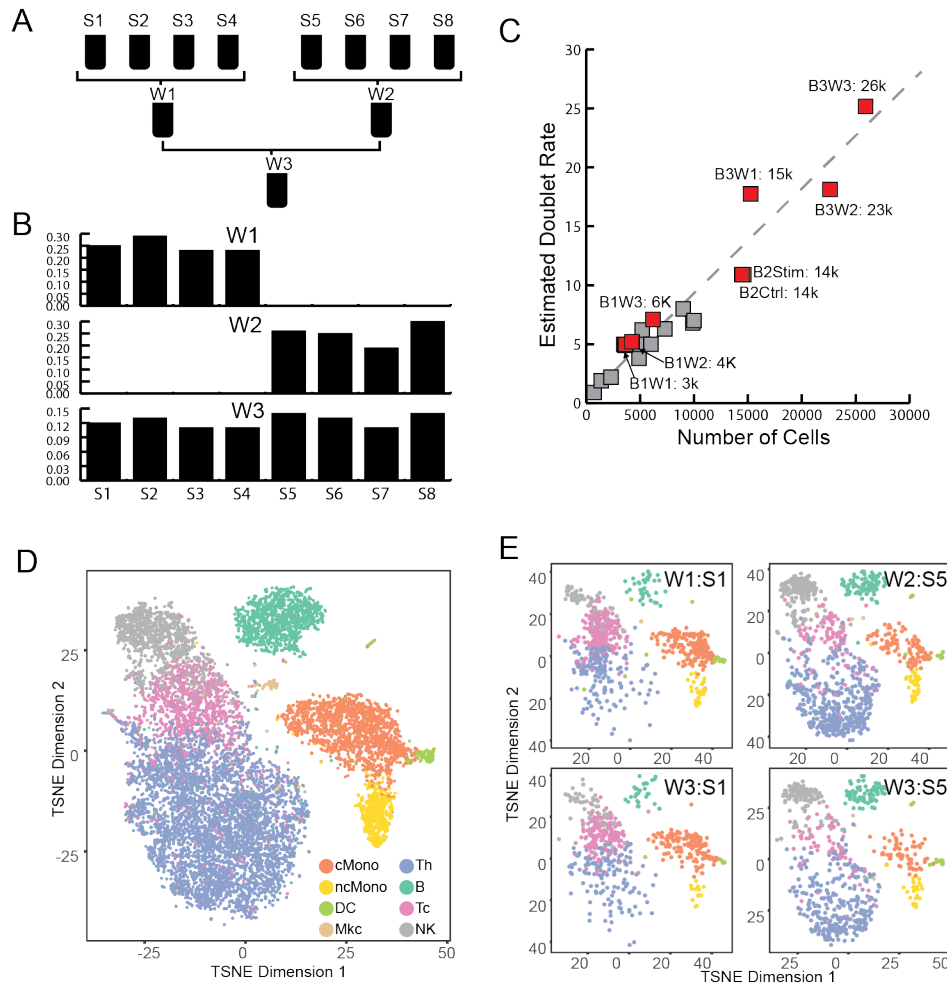
C.



**Figure 2.1: demultiplexing and doublet identification from single cell data.**

a) Pipeline for experimental multiplexing of unrelated individuals, loading onto droplet-based single-cell RNA-sequencing instrument, and computational demultiplexing (demux) and doublet removal using demuxlet. Assuming equal mixing of 8 individuals, b) 4 genetic variants can recover the sample identity of a cell, and c) 87.5% of doublets will contain cells from two different samples.

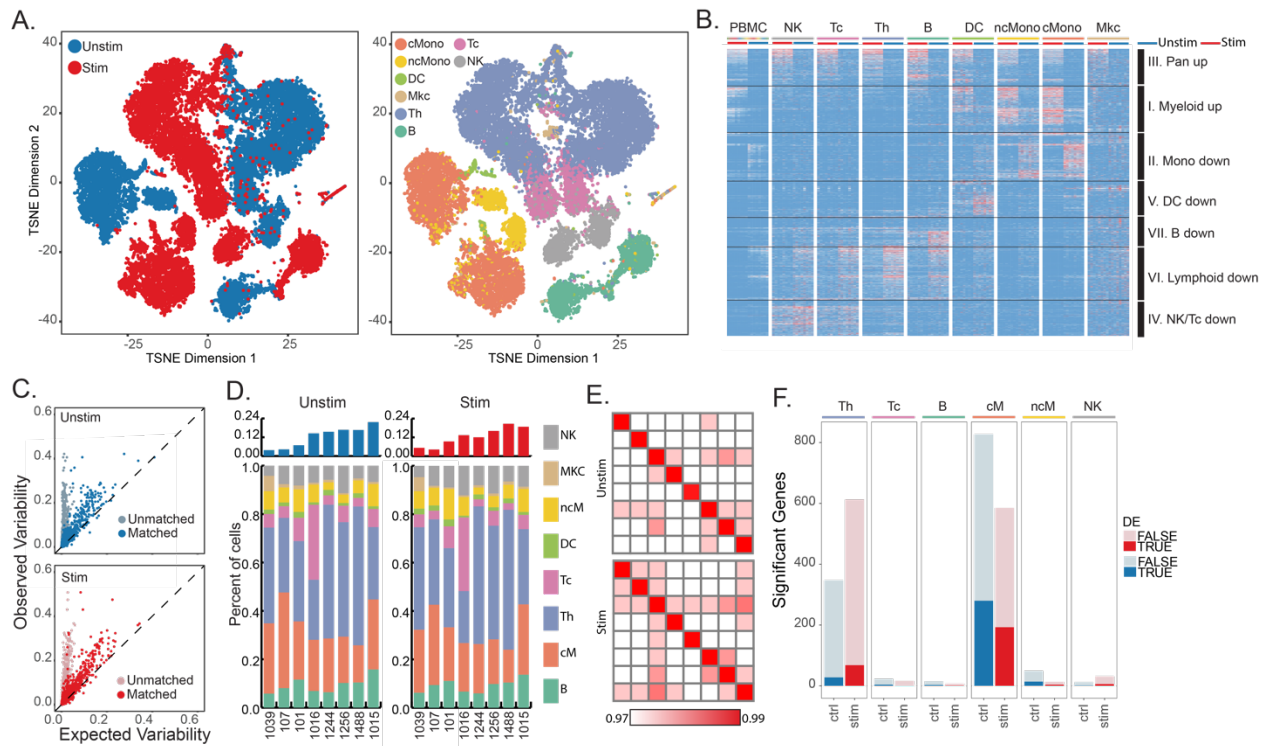
Figure 2



**Figure 2.2: Performance of demuxlet.**

a) Experimental design for equimolar pooling of cells from 8 unrelated samples (S1-S8) into three wells (W1-W3). W1 and W2 contain cells from two disjoint sets of 4 individuals. W3 contains cells from all 8 individuals. b) Demultiplexing single cells in each well recovers the expected individuals. c) Estimates of doublet rates versus previous estimates from mixed species experiments. d) Cell type identity determined by prediction to previously annotated PBMC data. e) t-SNE plot of two individuals (S1 and S5) from different wells are qualitatively concordant.

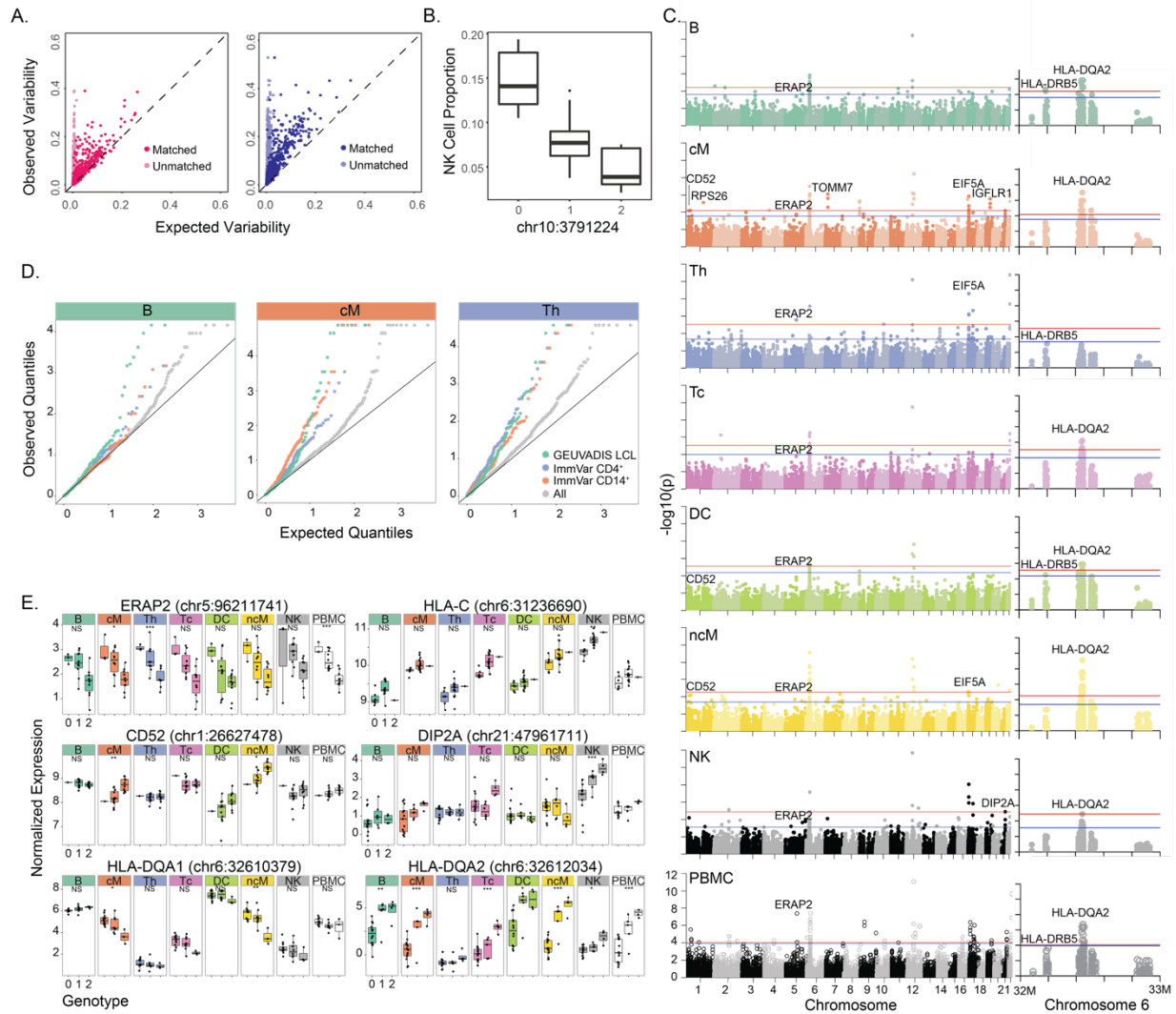
Figure 3



**Figure 2.3: Interindividual variability in IFN- $\beta$  response.**

a) t-SNE plot of unstimulated (blue) and IFN- $\beta$ -stimulated (red) PBMCs and the estimated cell types. b) Cell type-specific expression in stimulated (left) and unstimulated (right) cells. Differentially expressed genes shown (FDR < 0.05,  $|\log(\text{FC})| > 1$ ). Each column represents cell type-specific expression for each individual from demuxlet. c) Observed variance (y-axis) in mean expression over all PBMCs from each of the 8 individuals versus expected variance (x-axis) over synthetic replicates sampled across all cells (light blue, pink) or replicates matched for cell type proportion (blue, red). d) Cell type proportions for each individual in unstimulated and stimulated cells. e) Correlation between sample replicates in control and stimulated cells. f) Number of significantly variable genes in each cell type and condition.

Figure 4



**Figure 2.4 – Genetic control over cell type proportion and gene expression (N=23).**

a) Observed variance (y-axis) in mean expression over all PBMCs from each individual versus expected variance (x-axis) over synthetic replicates sampled across batch 1 (left, N=8) and batch 3 (right, N=15). b) Association of chr10:3791224 with NK cell type proportions. c) Genome-wide and chromosome 6 Manhattan plots across all major cell types. Horizontal lines correspond to FDR < 0.1 (blue) and FDR < 0.05 (red). d) Q-Q plots across all genes and subsets of previously published eQTLs in relevant cell types are shown for B, cM, and Th populations. e) Notable cis-eQTLs across all major immune cell types are marked with \*(FDR < 0.25), \*\* (FDR < 0.1), and \*\*\* (FDR < 0.05). Lack of association is marked with NS (not significant).

## References

1. Macosko, E.Z. et al. in *Cell*, Vol. 161 1202-1214 (2015).
2. Pollen, A.A. et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat Biotech* **32**, 1053-1058 (2014).
3. Buenrostro, J.D. et al. in *Nature*, Vol. 523 486-490 (Nature Research, 2015).
4. Nagano, T. et al. in *Nature*, Vol. 502 59-64 (2013).
5. Patel, A.P. et al. in *Science*, Vol. 344 1396-1401 (American Association for the Advancement of Science, 2014).
6. Tirosh, I. et al. in *Science*, Vol. 352 189-196 (American Association for the Advancement of Science, 2016).
7. Muraro, M.J. et al. A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst* **3**, 385-394 e383 (2016).
8. Baron, M. et al. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst* **3**, 346-360 e344 (2016).
9. Shalek, A.K. et al. in *Nature* (2014).
10. Klein, A.M. et al. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161**, 1187-1201 (2015).
11. Stegle, O., Teichmann, S.A. & Marioni, J.C. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* **16**, 133-145 (2015).
12. Gawad, C., Koh, W. & Quake, S.R. Single-cell genome sequencing: current state of the science. *Nat Rev Genet* **17**, 175-188 (2016).

13. Streets, A.M. et al. Microfluidic single-cell whole-transcriptome sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 7048-7053 (2014).
14. Zilionis, R. et al. Single-cell barcoding and sequencing using droplet microfluidics. *Nat. Protocols* **12**, 44-73 (2017).
15. Ziegenhain, C. et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell* **65**, 631-643.e634 (2017).
16. Hicks, S.C., Teng, M. & Irizarry, R.A. in bioRxiv 025528 (Cold Spring Harbor Labs Journals, 2015).
17. Zheng, G.X.Y. et al. in Nature Communications | doi:10.1038/ncomms9687, Vol. 8 14049 (Nature Publishing Group, 2017).
18. Jun, G. et al. in The American Journal of Human Genetics, Vol. 91 839-848 (2012).
19. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).
20. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
21. The Genomes Project, C. A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
22. Aguirre-Gamboa, R. et al. Differential Effects of Environmental and Genetic Factors on T and B Cell Immune Traits. *Cell Reports* **17**, 2474-2487.
23. Li, Y. et al. A Functional Genomics Approach to Understand Variation in Cytokine Production in Humans. *Cell* **167**, 1099-1110.e1014 (2016).

24. Mostafavi, S. et al. Parsing the Interferon Transcriptional Network and Its Disease Associations. *Cell* **164**, 564-578.
25. Stark, G.R., Kerr, I.M., Williams, B.R.G., Silverman, R.H. & Schreiber, R.D. in <http://dx.doi.org/10.1146/annurev.biochem.67.1.227>, Vol. 67 227-264 ( Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA, 2003).
26. Lee, M.N. et al. in *Science*, Vol. 343 1246980-1246980 (2014).
27. Ye, C.J. et al. in *Science*, Vol. 345 1254665-1254665 (2014).
28. Andrés, A.M. et al. Balancing Selection Maintains a Form of ERAP2 that Undergoes Nonsense-Mediated Decay and Affects Antigen Presentation. *PLOS Genetics* **6**, e1001157 (2010).
29. Mostafavi, S. et al. Parsing the Interferon Transcriptional Network and Its Disease Associations. *Cell* **164**, 564-578 (2016).
30. Palmer, C., Diehn, M., Alizadeh, A.A. & Brown, P.O. Cell-type specific gene expression profiles of leukocytes in human peripheral blood. *BMC Genomics* **7**, 115 (2006).
31. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506-511 (2013).
32. Orrù, V. et al. Genetic Variants Regulating Immune Cell Levels in Health and Disease. *Cell* **155**, 242-256 (2013).
33. Brodin, P. et al. Variation in the Human Immune System Is Largely Driven by Non-Heritable Influences. *Cell* **160**, 37-47 (2015).
34. Saveanu, L. et al. Concerted peptide trimming by human ERAP1 and ERAP2 aminopeptidase complexes in the endoplasmic reticulum. *Nat Immunol* **6**, 689-697 (2005).



35. Franco, L.M. et al. Integrative genomic analysis of the human immune response to influenza vaccination. *eLife* **2**, e00299 (2013).
36. Cao, J. et al. Comprehensive single cell transcriptional profiling of a multicellular organism by combinatorial indexing. *bioRxiv* (2017).
37. Dixit, A. et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853-1866.e1817 (2016).
38. Adamson, B. et al. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **167**, 1867-1882.e1821 (2016).
39. Jaitin, D.A. et al. Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell* **167**, 1883-1896.e1815 (2016).
40. Datlinger, P. et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nat Meth* **14**, 297-301 (2017).
41. Farh, K.K.-H. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337-343 (2015).
42. Buettner, F. et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotech* **33**, 155-160 (2015).
43. Tung, P.-Y. et al. Batch effects and the effective design of single-cell gene expression studies. *Scientific Reports* **7**, 39921 (2017).
44. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331-338 (2017).
45. Habib, N. et al. Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science* **353**, 925 (2016).

46. Lake, B.B. et al. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science (New York, N.Y.)* **352**, 1586-1590 (2016).
47. Habib, N. et al. DroNc-Seq: Deciphering cell types in human archived brain tissues by massively-parallel single nucleus RNA-seq. *bioRxiv* (2017).
48. Wills, Q.F. et al. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat Biotech* **31**, 748-752 (2013).
49. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
50. Wang, X., Spandidos, A., Wang, H. & Seed, B. PrimerBank: a PCR primer database for quantitative gene expression analysis, 2012 update. *Nucleic Acids Research* **40**, D1144-D1149 (2012).
51. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Meth* **10**, 1096-1098 (2013).
52. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protocols* **9**, 171-181 (2014).
53. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotech* **33**, 495-502 (2015).
54. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biology* **11**, R106 (2010).
55. Dabney, A., Storey, J.D. & Warnes, G.R. qvalue: Q-value estimation for false discovery rate control. *R package version 1* (2010).
56. Falconer, D.S., Mackay, T.F. & Frankham, R. Introduction to quantitative genetics (4th edn). *Trends in Genetics* **12**, 280 (1996).

57. Loh, P.R., Palamara, P.F. & Price, A.L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat Genet* **48**, 811-816 (2016).
58. Shabalin, A.A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353-1358 (2012).

## **Chapter 3: Multiplexed RNA-sequencing of 1M immune cells reveals the cellular, molecular, and genetic correlates of systemic lupus erythematosus.**

### **Introduction**

The approach described in Chapter 2 enables the large-scale profiling of disease cohorts in efficient and cost-effective manner. Chapter 3 highlights the biological insights that can be gained from applying this method to a disease cohort. In this chapter, we study Systemic Lupus Erythematosus using our multiplexed single cell sequencing workflow and demuxlet, profiling 120 lupus cases and 46 healthy controls.

Systemic lupus erythematosus (SLE) is a systemic autoimmune disease that disproportionately affects women<sup>1</sup> and is characterized by a broad range of manifestations across multiple organs<sup>2</sup>. Molecular analyses have implicated the production of autoantibodies, dysregulation of antigen presentation and lymphocyte signaling<sup>3,4</sup>, activation of the interferon signaling pathway<sup>3,4</sup>, and failure of apoptotic clearance as hallmarks of the disease<sup>5,6</sup>. Many genes that participate in these immunological processes are proximal to the ~100 genetic variants thus far associated with SLE<sup>7</sup>. However, while these results implicate multiple immune pathways in SLE<sup>3,8</sup>, mapping the cell types and states underlying the pathogenesis of the disease remains incomplete and annotating the molecular function of disease-associated variants remains challenging.

Historically, separate approaches have been used to characterize changes in cell composition and state in SLE. Flow cytometry analysis that estimates composition based on known cell surface markers has reported frequency changes of circulating immune populations<sup>9,10</sup>. Bulk peripheral

blood gene expression analyses with or without sorting for specific subsets have found elevated levels of interferon signaling with pleiotropic effects across immune cell types<sup>3,4,11</sup>. However, flow cytometry relies on the initial set of markers (and thus biased by prior knowledge) and bulk expression averages across diverse cells between and within types. Moreover, neither can simultaneously measure the frequencies and activation states of cell types or capture heterogeneity within sorted populations. Additionally, it is challenging to apply both methods at scale across the large cohorts necessary to detect subtle shifts in cell composition and gene expression caused by disease or disease-associated variants.

Massively parallel single-cell RNA-sequencing (scRNA-seq) holds enormous potential as a comprehensive approach to simultaneously estimate the composition and characterize the state of circulating immune cells. When integrated with dense genotyping data, there is a further opportunity to ascribe molecular functionality to disease-associated variants across a number of cellular contexts. However, profiling large population cohorts using scRNA-seq has been limited by sample throughput and susceptibility to technical and biological variability. To overcome these limitations, we recently described a sample multiplexing approach that leverages single nucleotide polymorphisms (SNPs) to enable systematic and cost-effective profiling of  $10^4$  cells from 10-100 genetically distinct samples in one microfluidic reaction<sup>12</sup>.

Here, we used multiplexed single cell sequencing to profile ~1 million peripheral blood mononuclear cells (PBMCs) isolated from 58 healthy controls and 136 SLE patients of Asian and European ancestry, including patients experiencing active disease flares. We analyzed this rich dataset to define changes in cellular composition, cell-type-specific gene expression, and

immune repertoire in cases. To place our findings in the context of the known molecular signatures of SLE, we estimated the contribution of cellular composition and cell-type-specific expression to inter-individual expression variability calculated over aggregated PBMCs, hereinafter pseudobulk. We further explored the contribution of genetics to interindividual variability by mapping common genetic variants associated with gene expression (eQTLs) across 8 immune cell types, identifying eQTLs whose effects are modified by cell-type identity and interferon activation. Finally, we leveraged published genome-wide association studies (GWAS) summary statistics to annotate cell types that may mediate genetic associations in SLE and other autoimmune diseases. Our work demonstrates the power of single-cell RNA-seq as a compelling tool for quantitative high-dimensional immunological phenotyping of immunological disease.

## Results

### Changes in helper T cell composition causally associated with SLE

We generated a cross-sectional dataset of 834,096 cell profiles across 169 donors (119 cases from the California Lupus Epidemiology Study<sup>13</sup> and 50 controls from the ImmVar Consortium<sup>14-17</sup>). PBMCs were profiled using multiplexed single cell sequencing in 13 pools each containing 16 donors<sup>12</sup> (**Fig. 3.1**). A total of 1,134,700 cell-containing droplets were sequenced to an average depth of 18,201 reads per droplet. 834,096 cells remained after quality control filtering and removal of droplets containing two cells using demuxlet<sup>12</sup> (doublet rate 26.5%, expected 22-25%) resulting in 4,590 singlets (+/- 1,572) assigned to each donor (**Supp. Fig. 1**).

From the multiplexed single cell sequencing profiles, we estimated the composition of circulating immune cells per sample and assessed the robustness of the estimates. Following batch correction, normalization, principal component analysis, *k*-nearest neighbor graph construction, and Leiden clustering (see **Methods**), we assigned each of 32 resulting clusters to 11 immune cell types based on known gene signatures including: classical (cM) and non-classical monocytes (ncM), conventional (cDCs) and plasmacytoid dendritic cells (pDCs), CD4<sup>+</sup> (T4) and CD8<sup>+</sup> T cells (T8), natural killer cells (NK), B cells (B), proliferating lymphocytes (Prolif), megakaryocytes (MK), and progenitor cells (Progen). Uniform Manifold Approximation and Projection (UMAP)<sup>18</sup> revealed distinct regions of the embedding occupied by cells of different types (**Fig. 3.1**) and to a lesser extent by cells from cases versus controls (**Fig. 3.1**). For each sample, we constructed a personalized projection and obtained highly reproducible

estimates of cell composition between biological replicates (Mean  $R^2 = 0.81$ ) (**Fig. 3.1, Supp. Fig. 2-3**). Notably for 107 cases, estimates of monocyte (ncM+cM) and lymphocyte (T4+T8+NK+B) abundances are extremely well correlated with those measured by automated white blood cell counts with differential reported in the UCSF Electronic Health Records (EHR) ( $R_{\text{mono}} = 0.88$ ,  $P < 9.30 \times 10^{-36}$ ,  $R_{\text{lympho}} = 0.97$ ,  $P < 1.40 \times 10^{-63}$ , **Fig. 3.1**).

We used least squares regression weighted by the total number of cells per donor to quantify composition differences between cases and controls. Cases were marked by higher percentages of monocytes (cM: +10.7%,  $P < 1.68 \times 10^{-8}$ ; ncM: +1.7%,  $P < 5.32 \times 10^{-4}$ , Linear Regression) and a corresponding lower T4 percentage because composition estimates are relative (-13.3%,  $P < 1.78 \times 10^{-13}$ ;  $R_{\text{cM.vs.T4}} = -0.41$ ) (**Fig. 3.1, Supp. Fig. 4-5**). Additionally, SLE patients have higher percentages of ProlifT (+0.34%,  $4.29 \times 10^{-4}$ ), and a small but significant percentage of pDCs (-0.55%,  $P < 5.16 \times 10^{-24}$ ) consistent with most reports<sup>19</sup>. No significant effects of treatment on composition were detected in patients currently receiving mycophenolate mofetil, hydroxychloroquine, or oral steroids (**Supp. Fig. 6**), consistent with reports that suggest mycophenolate mofetil has no effect on white blood cells<sup>20</sup> and prednisone has only transient effects on CD4<sup>+</sup> T cells<sup>21,22</sup>.

A higher ratio of monocytes to T4 cells could be due to mutually antagonistic regulation of myeloid versus lymphoid lineages during hematopoiesis or the enrichment or depletion of one lineage. Analysis of lymphocyte and monocyte abundances reported in the EHR of an additional 117 cases and 1,688 matched controls found no difference in the abundance of monocytes but depletion of lymphocytes in cases (Caucasians:  $P < 8.00 \times 10^{-9}$ , African Americans:  $P < 1.81 \times 10^{-5}$ ,



Asians:  $P < 1.66 \times 10^{-14}$ , **Fig. 3.1**). To assess if lymphocyte depletion is causative for disease, we performed Mendelian randomization using Generalised Summary-data-based Mendelian Randomisation (GSMR)<sup>23</sup> using summary statistics on SLE and blood composition traits from the UK Biobank (Bycroft et al. 2018). The causal effect size of SLE-associated variants on disease status is negatively correlated with their effect sizes on lymphocyte ( $\beta_{\text{SLE.lymph}} = -0.11$ ,  $P < 0.03$ ; **Fig. 3.1**) but not monocyte abundances (**Supp. Fig. 7**). These results suggest that single cell sequencing can reliably detect broad changes in peripheral blood cell types and is concordant with other approaches to estimate cellular composition.

### **Composition accounts for more inter-individual expression variability in SLE**

Bulk profiling of circulating immune cells have identified transcriptomic signatures linked to interferon signaling, lymphocyte activation, and cytolytic function in SLE<sup>4</sup>. However, pinpointing the pathogenic cells underlying bulk transcriptional signatures may be confounded by the functional overlap and frequency variation of immune cell types. To identify expression changes across cell types in SLE, we computed pseudobulk PBMC or cell-type-specific profiles for each sample and identified 141 differentially expressed (DE) genes in PBMCs and an additional 57 in at least one of eight cell types between cases and controls (cM, ncM, cDC, pDC, T4, T8, NK, or B) ( $\text{FDR} < 0.01$ ,  $\text{abs}(\log\text{FC}) > 1$ ; **Supp. Table 1, Fig. 3.2**, see **Methods**). The 198 DE genes clustered into 6 up-regulated and two down-regulated modules in SLE (**Fig. 3.2, Supp. Fig. 8**). Down-regulated modules  $M_{\text{pDC}}$  and  $M_{\text{T4}}$  are comprised of lineage-specific genes reflective of the decrease in the frequencies of T4 (i.e. *CCR7*) and pDC (i.e. *LILRA4*). The up-regulated modules include  $M_{\text{Pan}}$ , enriched for interferon-stimulated genes (ISG) across all cell

types (**Supp. Table 1**), and two modules ( $M_{\text{Mono}}$ ,  $M_{\text{ncM}}$ ) expressing genes specific to the myeloid lineage.  $M_{\text{Pan}}$  and  $M_{\text{Mono}}$  capture 21/30 previously described ISG genes in SLE<sup>5</sup> while  $M_{\text{ncM}}$  is composed of components of the complement system. A pseudobulk ISG signature score calculated over all PBMCs is positively correlated with myeloid cell percentage ( $R = 0.58$ ) and negatively correlated with lymphoid cell percentage ( $R = -0.22$ ) (**Fig. 3.2**). Additional up-regulated modules  $M_{\text{Lymph}}$ ,  $M_{\text{T8}}$  and  $M_{\text{B}}$  consist of genes expressed in T4, T8 and B cells including cytotoxicity ( $M_{\text{T8}}$ : *GZMB*, *GZMH*), activation and checkpoint ( $M_{\text{Lymph}}$ : *TIGIT*, *KLRB1*), and major histocompatibility complex molecules and cytokines ( $M_{\text{B}}$ : *HLA-DRB5*, *IL6*). Genes in these modules were largely not differentially expressed in PBMCs likely due to the low frequency of cells in circulation (e.g. T8 and Bs) and the opposing actions of cell depletion and increased expression (e.g. T4), highlighting an important advantage of single-cell analysis.

Leveraging the ability to simultaneously estimate the frequency and expression profile of each cell type, we used variance component analysis to quantify the contribution of cellular composition ( $V_{\text{comp}}$ ) and cell-type-specific expression ( $V_{\text{exp}}$ ) to inter-individual expression variability across PBMCs (see **Methods**). Composition explains more PBMC variance for differentially expressed genes ( $V_{\text{comp}} = 48\%$ ) than all genes ( $V_{\text{comp}} = 25\%$ ) (**Fig. 3.2**). Partitioning of  $V_{\text{comp}}$  and  $V_{\text{exp}}$  implicates specific cell types responsible for the inter-individual variability of each module (see **Methods**). For  $M_{\text{Pan}}$ , PBMC variance is mostly determined by the percentage (31%) and expression (25%) of cMs with composition contributing substantially more in cases than controls ( $V_{\text{cM,comp}} = 25\%$  vs 1.2%) (**Fig. 3.2**). For  $M_{\text{T8}}$ , highlighted by *IFNG*, T8 percentage contributes most to PBMC variance and was higher in cases than controls ( $V_{\text{T8,comp}} = 32\%$  vs 13%) (**Fig. 3.2**). Beyond modules, an intriguing example is the proinflammatory cytokine *IL6*,

one of the only genes whose PBMC variance is determined by B cells in cases but not controls ( $V_{B,comp} = +16.8\%$ ,  $V_{B,exp} = +19.2\%$ ) (**Fig. 3.2**). While *IL6* is known to induce B cell hyperactivity in SLE<sup>25,26</sup>, its expression by the cognate cells it activates suggests autoregulatory mechanisms orthogonal to canonical sources, possibly through the spontaneous formation of germinal centers<sup>27</sup>, to promote the production of autoantibodies and systemic autoimmunity.

We next assessed whether models using estimates of cellular composition and cell-type-specific expression features could better predict disease status and activity than known bulk gene expression. Here we compared models that used monocyte/lymphocyte composition estimates, refined cell type label estimates, cell type specific gene expression, and bulk gene expression as features. Using elastic nets (see **Methods**), all models except one that only used monocyte/lymphocyte composition were highly predictive of disease status (10-fold cross-validation  $R^2 > 0.93$ , **Fig. 3.2**). Within cases, although no model predicted the *Systemic Lupus Erythematosus Disease Activity Index* (SLEDAI) particularly well ( $R^2 \sim 0.09-0.23$ ), a model that used only 11 composition features better predicted individual SLEDAI components than one that used the pseudobulk expression of the 30 published ISG genes<sup>5</sup> (low complement  $R^2$ : 0.71 vs 0.66, anti-dsDNA  $R^2$ : 0.60 vs 0.59, rash  $R^2$ : 0.87 vs 0.80, **Fig. 3.2**). Lupus nephritis is a major complication of SLE and a model that included both cell-type-specific expression and composition components was able to predict past kidney complications significantly better than one that used the pseudobulk expression of ISG genes (kidney  $R^2$ : 0.60 vs 0.53, **Fig. 3.2**).

## Myeloid effects in SLE are positively correlated with interferon activity

The significant inter-individual variability explained by their intrinsic expression in myeloid populations suggests additional heterogeneity within the myeloid compartment that underlies the bulk ISG signature in SLE. To test this, we re-clustered cM, ncM, pDC, cDC into 10 clusters including two differentiating the monocyte lineage (cM: *CD14*<sup>+</sup> classical, ncM: *CD16*<sup>+</sup> non-classical) and three differentiating the dendritic cell lineage (cDC1: *CLEC9A*<sup>+</sup> conventional, cDC2: *FCERIA*<sup>+</sup> conventional, pDC: *IRF7*<sup>+</sup> plasmacytoid) (**Fig. 3.3, Supp. Fig. 9**). Importantly, several functionally distinct clusters were also detected including *IL1B*<sup>+</sup> pro-inflammatory monocytes (cM<sub>inf</sub>), activated monocytes expressing ISGs (cM<sub>act</sub>), complement-expressing monocytes (ncM<sub>comp</sub>), and two populations of macrophages (Mac1 and Mac2, both expressing *CSF3R* and distinguished by the expression of *ISG15*) (**Fig. 3.3**). cM<sub>act</sub>, cM<sub>inf</sub>, and the macrophage clusters all express *CD14* indicative of their origin from classical monocytes while ncM<sub>comp</sub> expresses *FCGR3A* (*CD16*) indicative of their origin as non-classical monocytes (**Fig3.3**).

Monocytes defined by function and dendritic cells defined by lineage occur at different frequencies between cases and controls. As a percentage of all PBMCs, pDCs remain reduced in cases while the two cDC populations do not change in frequency (**Fig. 3.3**). Two monocyte populations, cM<sub>act</sub> (+5.66%,  $P < 2.47 \times 10^{-5}$ ) and ncM<sub>comp</sub> (+0.27%,  $P < 3.72 \times 10^{-5}$ ) and Mac2 (+0.19%,  $P < 5.29 \times 10^{-5}$ ) are notably increased in frequency (**Fig. 3.3**). The percentages of these cell types are positively correlated with the pseudobulk ISG signature score across all donors (cM<sub>act</sub>:  $R = 0.60$ ,  $P < 6.95 \times 10^{-14}$ ; ncM<sub>comp</sub>:  $R = 0.45$ ,  $P < 1.01 \times 10^{-7}$ ; Mac2:  $R = 0.52$ ,  $2.87 \times 10^{-7}$ )

and in cases (cM<sub>act</sub>: R = 0.57, P < 2.52x10<sup>-11</sup>; ncM<sub>comp</sub>: R = 0.41, P < 4.27x10<sup>-6</sup>; Mac2: R = 0.46, 3.00x10<sup>-7</sup>) suggesting that these myeloid populations are the main producers of the ISG signature (**Fig. 3.3**). This is confirmed by the elevated expression of the ISG signature score component modules (M<sub>Pan</sub>, M<sub>Mono</sub>) in these clusters in cases but not controls (**Fig. 3.3**). Ordering myeloid cells along a diffusion pseudotime (DPT) based on the degree of IFN activation revealed a shift toward higher activation in cases as a function of the SLEDAI (**Fig. 3.3, see Methods**). This shift was also observed when ordering cells based on comparison to an independent *in vitro* stimulation dataset<sup>12</sup> (**Supp. Fig. 10**) but not observed when ordering cells by the expression of lineage markers *CD14* and *FCGR3A* (**Fig. 3.3**). Compared to controls, even cells from cases with 0 SLEDAI are shifted toward higher IFN activation indicative of subclinical disease. These results suggest that IFN production specific to monocytes is a potential indicator of disease status as well as severity.

### **SLE marked by naive helper T cell depletion and cytotoxic T cell expansion**

While lymphopenia is near-universal in pediatric and adult SLE, precisely which lymphocyte subpopulations are depleted during disease remains unknown. Our initial analysis provided evidence for the depletion of CD4<sup>+</sup> T cells in cases while the abundances of CD8<sup>+</sup>, natural killer, and B cells remain unchanged. To further characterize the changes in the composition and state of the lymphoid compartment, we re-clustered T4, T8, NK, B and Prolif cells into 19 clusters (**Fig. 3.4**).

In the T cell compartment, we identified canonical subpopulations of naive ( $T4_{naive}$ : annotated by *CCR7* expression) and central memory  $CD4^+$  cells ( $T4_{cm}$ : annotated by *ANXA1* and *IL7R* expression), and the corresponding  $CD8^+$  cells ( $T8_{naive}$ : *CCR7* and *CD8B*,  $T8_{cm}$ : *SBF2*) (**Fig. 3.4**). Additional populations detected include regulatory ( $T4_{reg}$ : *RTKN2*, *TIGIT* and *FOXP3*) and interferon-activated cells ( $T_{IFN}$ : tagged ISGs such as *ISG15*) within the  $CD4$  lineage; mucosal-associated invariant cells ( $T8_{em,MAIT}$ : *KLRB1*) and two effector memory populations ( $T8_{em,cyto1}$  and  $T8_{em,cyto2}$ ) within the  $CD8$  lineage (**Fig. 3.4**). The effector memory T8 populations both express the chemokine *CCL5*, effector molecules *PRF1* and *GZMA*, and exhaustion markers *LAG3* and *PDCD1*, and can be distinguished by the expression of granzymes ( $T8_{em,cyto1}$ : *GZMB* and *GZMH*,  $T8_{em,cyto2}$ : *GZMK*; **Fig. 3.4**, **Supp. Fig. 11**).

The distribution of T cells, especially  $CD8$ s, was shifted toward effector phenotypes in cases versus controls (**Fig. 3.4**). While both  $T4_{naive}$  and  $T8_{naive}$  percentages were reduced ( $T4$ : -12.7%,  $P < 4.03 \times 10^{-23}$ ,  $T8$ : -3.49%,  $P < 9.99 \times 10^{-8}$ , **Fig. 3.4**),  $T4_{naive}$  but not  $T8_{naive}$  percentage is negatively correlated with the pseudobulk PBMC ISG signature score ( $R = -0.62$ ,  $P < 4.63 \times 10^{-15}$  vs  $R = -0.08$ ,  $P < 0.39$ ) (**Fig. 3.4**). Strikingly, both  $T8_{em,cyto1}$  and  $T8_{em,cyto2}$  percentages were significantly increased (+4.57%,  $P < 2.34 \times 10^{-5}$ ; +1.16%,  $P < 8.45 \times 10^{-3}$ ) while  $T8_{em,MAIT}$  percentages were decreased (-2.12%,  $P < 1.75 \times 10^{-18}$ ) (**Fig. 3.4**). Previous studies have implicated  $GZMB^+/PRF1^+$   $CD8^+$ s in SLE pathogenesis possibly by generating nontolerogenic granzyme-B mediated autoantigen fragments that may overwhelm physiologic clearance pathways and contribute to antigenic feeding of dendritic cells<sup>30</sup>.

To investigate the potential causal role for changes in the T cell compartment, we amplified and sequenced the CDR3 region of the T cell receptor (TCR), recovering productive paired *TCRA* and *TCRB* sequences from 10.2% of T4s and 8.7% of T8s with no differences in recovery between 119 cases and 22 controls (see **Methods**). Intriguingly, T8 cells from 48 of 119 cases (compared to 8 of 22 controls) and T4 cells from 1 case (compared to no controls) expressed at least one TCR sequence in at least two cells suggestive of clonal expansion of T8 and not T4 cells in SLE. This was confirmed by a higher Gini coefficient (a measure of repertoire restriction) in cases for T8 ( $P < 0.006$ , t-test) but not T4 cells ( $P < 0.62$ ; **Fig. 3.4**). Expanded T8 clones (defined as those detected in more than 1 cell) were enriched within the  $T8_{em,cyto1}$  ( $P < 0.004$ ) and  $T8_{em,cyto2}$  ( $P < 0.03$ ) clusters (**Fig. 3.4**). As a positive control, clones expressing the invariant *TRAV1-2* and *TRAJ33* chain aggregated almost exclusively within the  $T8_{em,MAIT}$  cluster (**Supp Fig. 12**). The lack of correlation between the percentages of the T8 subsets and the ISG signature score within cases suggests that the expansion of effector memory T8 cells is independent of type 1 interferon activation. Although relatively few TCRs were sampled in this experiment from each individual, these results suggest further TCR sequencing efforts that can be used to establish more robust criteria for clonal expansion in SLE patients as compared to healthy controls.

Within other lymphocyte compartments, we observed more subtle changes in cases. We identified three NK cell subpopulations distinguished by *XCL1/2* ( $NK_{bright}$ ), *PRF1* ( $NK_{dim}$  and  $NK3$ ) and *HBA1* ( $NK3$ ) and four B cell subpopulations distinguished by *TCL1A* ( $B_{naive}$ ), *HLA-DRA* (all B cells), *MZB1* ( $B_{plasma}$ ) and cytotoxic markers including *GZMB* and *PRF1* ( $B_{doublets}$ ) (**Fig. 3.4**). In cases, the NK clusters did not change in frequency while  $B_{mem}$  percentages were

decreased (-2.73%,  $P < 1.96 \times 10^{-7}$ ). The percentage of B<sub>doublets</sub> cells, expressing high levels of cytolytic and B cell markers, was also increased in cases. These cells resemble the recently described interacting pairs of B cells and either NK or T8 cells<sup>31</sup>.

### Context specific genetic effects on gene expression

The integration of multiplexed dscRNA-seq and dense genotyping provides an opportunity to examine the prevalence and magnitude of genetic effects associated with composition, cell-type-specific expression, and cellular response to prolonged stimulation in disease states. No genetic variants were associated at genome-wide significance with either lymphocyte or monocyte percentage, likely reflective of the small effect sizes of common variants<sup>32</sup> and the effect of disease and treatment on these traits (**Supp Fig. 13**). On the other hand, hundreds to thousands of *cis* expression quantitative loci (*cis*-eQTLs) were detected in each cell type (1,118 in T4, 1,180 in T8, 403 in NK cells, 538 in B cells, 1,686 in cM, 889 in ncM, 337 in cDCs, and 39 in Mkc; FDR < 0.1; **Supp. Table 2**) using the pseudobulk gene expression in each of 118-119 individuals. Out of the 3,092 *cis*-eQTLs detected in at least one cell type, 2,132 were not detected in pseudobulk PBMCs, suggesting that the majority of *cis*-eQTLs have heterogeneous effects across cell types. While the genetic correlations of *cis*-eQTLs between pairs of cell types are generally high ( $r_G = 0.25-0.61$ ), clustering based on either the genetic correlations and the number of overlapping *cis*-eQTLs reflect the known lineage relationships between circulating immune cell types (**Fig 3.5**). Further, compared to *cis*-eQTLs detected in PBMCs, *cis*-eQTLs detected in each cell type were more enriched for accessible regions of the genome measured in the same cell type by ATAC-seq<sup>33</sup> (Mann-Whitney test, **Fig 3.5**).



We integrated published GWAS data to assess the enrichment of cell-type-specific *cis*-eQTLs for autoimmune disease loci. T4-specific and T8-specific *cis*-eQTLs were most enriched for SLE-associated variants suggesting T lymphocytes as potential mediators for genetic variants causal for disease (**Fig 3.5**). One example is the SLE-associated variant rs6671847, which has a significant effect on the expression of *HSPA6* in T8 cells (**Fig 3.5**). *HSPA6* is also within a risk locus for Ulcerative Colitis, and part of a family of heat shock proteins known to influence autoimmunity and tumor immunity<sup>34,35</sup>. Another example is the SLE-associated variant rs7258015<sup>36</sup>, which is associated with the expression of *ICAM3* only in ncM cells (**Fig. 3.5**). Circulating ICAM3 is upregulated in patients with autoimmune diseases<sup>37</sup> and serum levels of ICAM3 can be used as an indicator of lymphocyte stimulation in PBMCs<sup>38</sup>. Beyond SLE, we also found enrichment of type-1 diabetes variants within T8 and NK *cis*-eQTLs, and multiple sclerosis variants within B cell *cis*-eQTLs (**Supp Fig. 14**) consistent with the known pathogenesis of each disease.

We and others have previously shown that *in vitro* stimulation with recombinant IFNB can modify the effects of genetic variants on the expression of myeloid cells<sup>15,39</sup>. We assessed if the ISG signature reflective of type-1 interferon activation *in vivo* can similarly modify genetic effects (*cis*-IFN-eQTLs) on gene expression in SLE. Using a model that explicitly tests for interactions between genetic variants and the ISG signature score (**Methods**), we detected 1 *cis*-IFN-eQTL in cMs and 4 *cis*-IFN-eQTLs in PBMCs (FDR < 0.1). Despite the limited power, previous interferon response eQTLs<sup>40</sup> featured a more prominent deviation from null in the quantile-quantile plot compared to all variants (**Fig. 3.5**). The paucity of signals could be due to

imperfect estimates of *in vivo* interferon activation, small interacting effect sizes, and population heterogeneity of the samples. Results from population-specific analyses mirrored those from the full cohort: of the top 100 *cis*-IFN-eQTLs in all cases, 53 were nominally significant ( $P < 0.05$ ) in the European cases ( $P < 8.81 \times 10^{-42}$ , binomial test) and 42 in the Asian cases ( $P < 3.53 \times 10^{-28}$ , binomial test), suggesting minimal effects due to population heterogeneity (**Supp Fig. 15**).

The most striking example of a *cis*-IFN-eQTL is associated with *APOBEC3B* ( $P < 9.55 \times 10^{-7}$ ) (**Fig 3.5F**) where IFN activation as captured by the ISG signature score significantly modifies the effect of genotype of rs12628403 on *APOBEC3B* expression in monocytes (positive for major homozygotes and negative for heterozygotes). This results in low variability in *APOBEC3B* expression in cases with low ISG signature scores and high variability in cases with high ISG signature scores. *APOBEC3B* is a cytidine deaminase implicated in RNA-editing and autoimmunity, and it has been shown to be upregulated in SLE patients with managed disease<sup>41</sup> and further upregulated during flares<sup>42</sup>. Our results suggest that polymorphisms at the *APOBEC3B* locus could contribute to increased variability of its expression resulting in heterogeneous clinical presentation of SLE related to the function of the gene.

### **Periods of heightened disease marked by the presence of macrophages**

One of the clinical complications of SLE is the development of flares that require a change of therapeutic strategy. To characterize the molecular features of SLE during periods of heightened disease, we recruited an additional 8 healthy controls and 17 SLE flare patients (Flare), 8 of whom provided an additional sample three months after change in treatment (Treated) (**Fig. 3.6**).

To facilitate comparisons, 10 healthy controls and 5 non-flaring SLE patients (Managed) from the original cross-sectional cohort (**Fig. 3.1**) were sampled again. Using a panel of 20 oligo-tagged antibodies, we performed sample multiplexed Ab-seq of four pools (ranging from 8 to 17 individuals per pool) (**Fig. 3.6**). In this longitudinal cohort, a total of 218,030 cell-containing droplets were sequenced to an average depth of 42,268 reads per droplet, 153,955 were retained after quality control filtering and removal of droplets containing more than one cell using demuxlet (29.38%, expected 22-25%).

We identified 37 Leiden clusters and assigned them to 11 immune cell types (**Fig. 3.6**). Changes in composition between flare cases and controls were highly correlated with those observed between cross-sectional cases and controls ( $R = 0.81$ ,  $P < 0.005$ ) (**Fig. 3.6**, **Supp. Fig. 16**). To further validate and refine the observed differences in cellular composition, each of 37 clusters was assigned to one of 24 subpopulations that transcriptionally overlaps well with the same subpopulations identified in the cross-sectional cohort (**Fig. 3.6**). Analysis of the subpopulations between flare cases and controls generally confirmed the findings from the cross-sectional cohort ( $R = 0.73$ ,  $P < 1.22 \times 10^{-4}$ ) including the following: decreased percentage of  $T4_{naive}$  (-10.18%,  $P < 8.54 \times 10^{-5}$ ),  $T8_{MAIT}$  (-2.59%,  $P < 4.23 \times 10^{-6}$ ), pDC (-0.53%,  $P < 1.14 \times 10^{-4}$ ); increased percentages of  $T8_{em,cyto1}$  (+9.40%,  $P < 1.97 \times 10^{-4}$ ),  $cM_{act}$  (+5.21%,  $P < 5.90 \times 10^{-4}$ ), and  $ncM_{comp}$  (+0.56%,  $P < 3.44 \times 10^{-4}$ ) cells (**Fig. 3.6**). Surprisingly, we observed a significant increase in macrophages in cases that were not observed in the cross-sectional (+4.00%,  $P < 6.38 \times 10^{-5}$ , Wilcoxon rank-sum, **Fig. 3.6**). Although no significant differences were found in response to treatment overall, all three patients receiving rituximab were depleted of B cells (**Supp. Fig. 17**).

We used protein abundance estimates for 20 cell-surface markers to validate the major findings and identify differences between protein and mRNA features. Comparing protein and mRNA abundances revealed a range of correlations for cell surface markers from 0.03(FAS) to 0.68 (CD14) (pearson r; **Supp. Fig. 18**). Leiden clustering and UMAP projection using protein features revealed cluster assignments that broadly recapitulated those obtained from using mRNA features (**Supp Fig. 19**). The notable exception is T8<sub>em,cyto1</sub> identified from the mRNA analysis which projected onto both T4 and T8 regions of the protein UMAP (**Fig. 3.6**) and express both CD4 and CD8 proteins (**Fig. 3.6**). Only the percentages of cytotoxic T8 and not T4 cells increase in abundance in flare cases (T8: +6.27%,  $P < 2.7 \times 10^{-4}$ ; **Fig. 3.6**) further supporting CD8<sup>+</sup> cytotoxic T cells as an important mediator of SLE. Their presence and the production of IFNG could recruit macrophages detected in flare patients to initiate the recurrence of disease.

## Methods

### **PBMC processing:**

PBMCs were thawed, suspended and multiplexed according to the protocol in Kang et al. and loaded onto the 10x Chromium instrument. Following library prep according to the standard 10x protocol, libraries were sequenced on the HiSeq4000 at a depth of 6306-29862 reads/cell.

### **Single cell preprocessing:**

We sequenced a total of 1,352,730 droplets from cells in the healthy Immvar, Flare, and California Lupus Epidemiological Study cohorts. Demuxlet was used with an error probability of 0.1 to assign each cell to a donor of origin, preserving a total of 991,016 singlets. Using Scanpy version 1.4, we preprocessed the cross-sectional and flare cohort separately by first adjusting for pool using COMBAT, then regressing total nUMIs, percentage mitochondrial UMIs, gender, and principal components capturing a platelet signature. Regression for the platelet signature was performed because of the detection of platelet markers across cell types that likely reflect low levels of contamination due to imperfect ficoll in the CLUES cohort. This claim is supported by the detection of platelet markers only in controls samples pooled with case samples but not with each other. After an initial round of regressing for total nUMIs, percentage mitochondrial UMIs and gender, principal components were computed and those correlated with the expression of *PF4* ( $R > 0.4$ ) were identified as platelet specific. Cell filtering and expression normalization followed default settings, Subsequently, we performed k-nearest neighbor (knn) graph construction, leiden clustering, and plotted UMAP projections. Diffusion Pseudotime analysis (DPT) was performed through the scanpy function ``api.tl.dpt`` with default parameters.

### **Cell Type Annotation and Proportion Calculations:**

Scanpy version 1.4 was used to cluster singlets into leiden communities with parameter settings of resolution of 3 and controlling for random state. For the Flare cohort, a resolution of 2 was used. We found differentially expressed genes between communities in addition to most abundantly expressed genes for each community. We used gene expression profiles of known cell type populations identified in previous literature to identify our communities. The proportion of cells for each cell type was calculated as the number of cells belonging to the cell type divided by the total number of cells assigned to the sample. 2 samples with less than 100 cells total were excluded from the analyses. Cell type counts were calculated by multiplying the proportions by the total white blood cell count for each patient.

### **Electronic health record query:**

SLE cases with available monocyte and lymphocyte counts were selected according to the following criteria: 4532 healthy female controls were selected according to {Rappoport et. al, JALM 2018}. In short, outpatients without abnormal findings of adult patients aged 20-90 was extracted from the EHR system at the University of California, San Francisco (UCSF) Medical Center Data was extracted at February 2018, covering about 6 years of medical service coverage. In case there were multiple healthy encounters were found for a subject, a single random one was chosen. 403 Cases were defined as patients in the same age range who have a diagnosis of an ICD-10 code M32.\* appearing at least twice 30 days apart. Lab tests results for cases were taken from encounters for which MS was assigned as a primary diagnosis or principal problem

diagnosis. Patients with a monocyte count less than 5 and a lymphocyte count less than 6 were excluded.

### **Mendelian randomization:**

To test putative causal associations between risk factors and diseases performed Mendelian randomization using the GSMR (Generalized Summary-data-based Mendelian Randomization) package in gcta\_1.91.5beta. We searched for causal associations between blood count quantitative trait loci (qtls) in UK biobank (lymphocytes, monocytes, red blood cells, white blood cells, and platelets) and lupus qtls. We used 1000 genomes phase3 for our reference, a gwas significance threshold of  $5e-18$ , heidi outlier threshold of 0.15, and a linkage disequilibrium threshold of 0.01.

### **Cell Type Specific Differential Expression:**

For each cell type, we calculated a bulk profile summing all of the counts for each individual. The DESeq2 R package was used to estimate the log<sub>2</sub> fold change and the p-value of gene expression differences between the SLE and the healthy cohorts, and batch was included as a covariate. For visualization and variance decomposition, COMBAT (R package sva) was used to adjust batch effects across all genes and all batches. With the batch adjusted matrix, the differential expression signature was calculated as the first principal component of gene expression corresponding to the 190 differentially expressed genes from PBMCs.

### **Variance Decomposition of PBMC expression:**

Variance decomposition into composition and gene expression components was performed according to the following model:

Raw counts for each cell type were normalized to the total number of PBMCs per donor. PBMC variance was decomposed first into cell composition components using linear regression and the following model:  $y = bp_1 * cp_1 + bp_2 * cp_2 + \dots - 1$ . The residual from this fit was then regressed with the expression of each cell type:  $y_{res} = be_1 * ce_1 + be_2 * ce_2 + \dots - 1$ . The contribution from each cell type (for both proportion and cell-type-expression) was computed using the following:  $ci = \text{sum}(\text{cov\_mat}[i,] * bi * [b_1, b_2, \dots,]) / \text{var}(y)$ . This model accounts for both the variance contribution from each cell type but also allocates the covariance between any pair of cell types equally to each cell type.

### **Sample Genotyping:**

CLUES SLE patients were genotyped on the Affymetrix World LAT Array and the Immvar and flaring SLE patients were genotyped on the OmniExpressExome54 chip. A total of 21,412,068 SNPs were imputed from the Haplotype Reference Consortium version 1.1 with a MAF < 0.01.

### **eQTL discovery:**

1,220,450 SNPs with a MAF < 0.1 were used to map cis-eQTLs within a cis window +/- 100kbp of each gene, and a total of 8,905 genes were tested. Gene expression for each cell type was normalized using the rlog function in the DESeq2 package, and eQTLs were called using the MatrixEQTL package. The first 10 principal components of gene expression and 7 genotype PCs were included as covariates in all of the eQTL linear models. For the IFN interaction model, an



additional interaction term with the IFN signature and genotype, and the effect size and significance of the interaction term were calculated. The IFN signature was calculated as the first principal component of gene expression of the 25 type I interferon genes as listed in Crow et al.

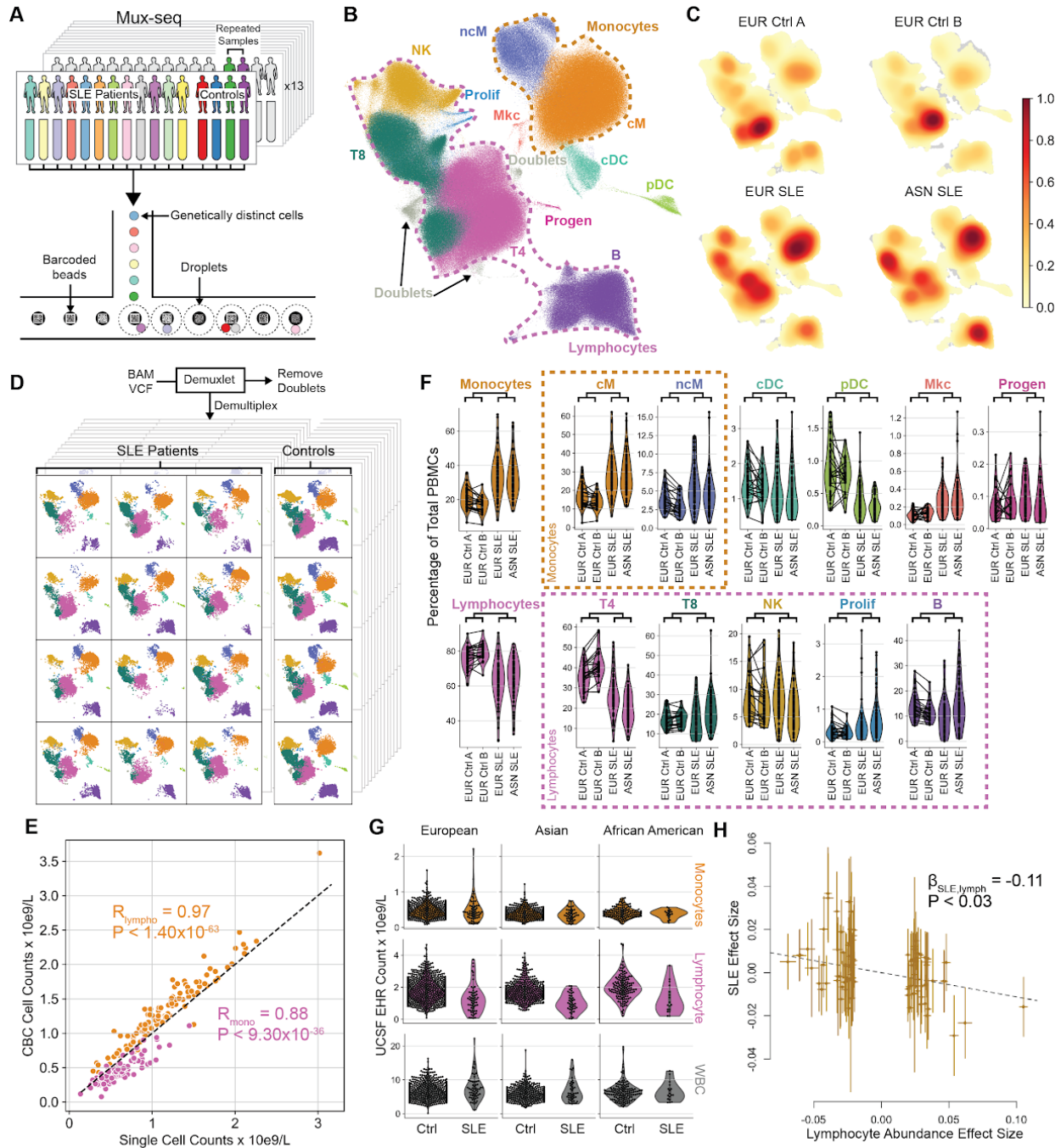
### **ATAC-seq and GWAS enrichment:**

Cell type specific ATAC-seq peaks were downloaded from Calderon et al ([https://web.stanford.edu/group/pritchardlab/dataArchive/immune\\_atlas\\_web/index.html](https://web.stanford.edu/group/pritchardlab/dataArchive/immune_atlas_web/index.html)). For each set of eQTLs and peaks, we applied a Mann-Whitney test to determine the enrichment for significant SNPs residing within each set of cell type specific peaks. GWAS enrichment was calculated using the GREGOR package, and the set of significant SNPs for each disease were downloaded from the UCSC Genome Browser.

### **GEMMA**

GEMMA 0.98.1 was run using the genotypes from SLE patients in PLINK binary format. Both a standardized kinship matrix and gender were adjusted for in the results. Lymphocyte and monocyte counts from the EHR of our SLE patients were used.

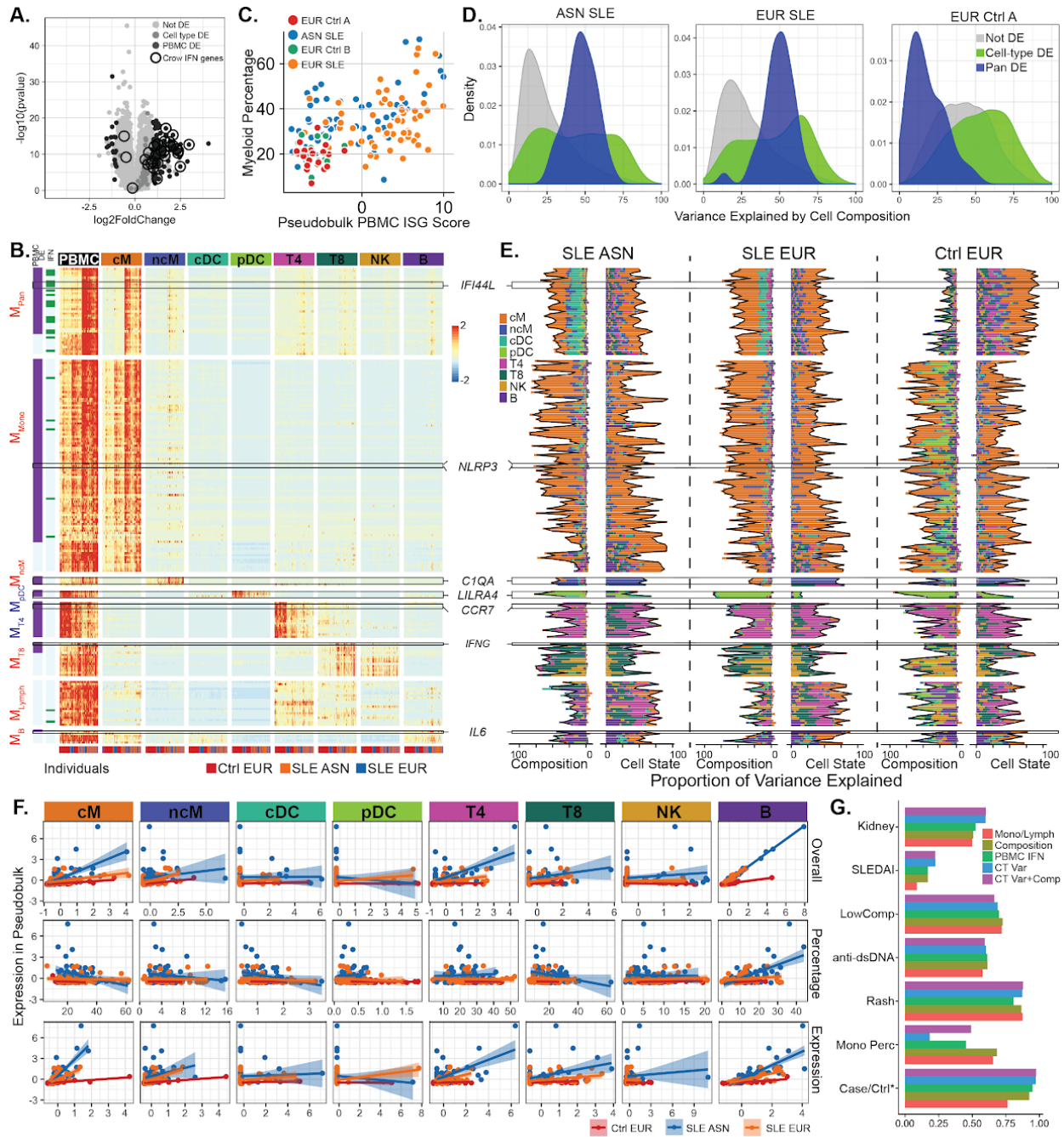
## Figures



**Figure 3.1: Overview and compositional changes in SLE.**

A.) Multiplexed single cell sequencing applied to 119 cases and 50 healthy controls. B.) Assignment of each cell to 11 major cell types. C.) UMAP projection depicting density of cells assigned to cases and controls. D.) UMAP projection for each sample in a pool after demultiplexing using demuxlet. E.) Correlation of single cell counts normalized to the number of cells multiplied by  $10^9$  that are expected per liter of blood (x-axis) and CBC estimates (y-axis)

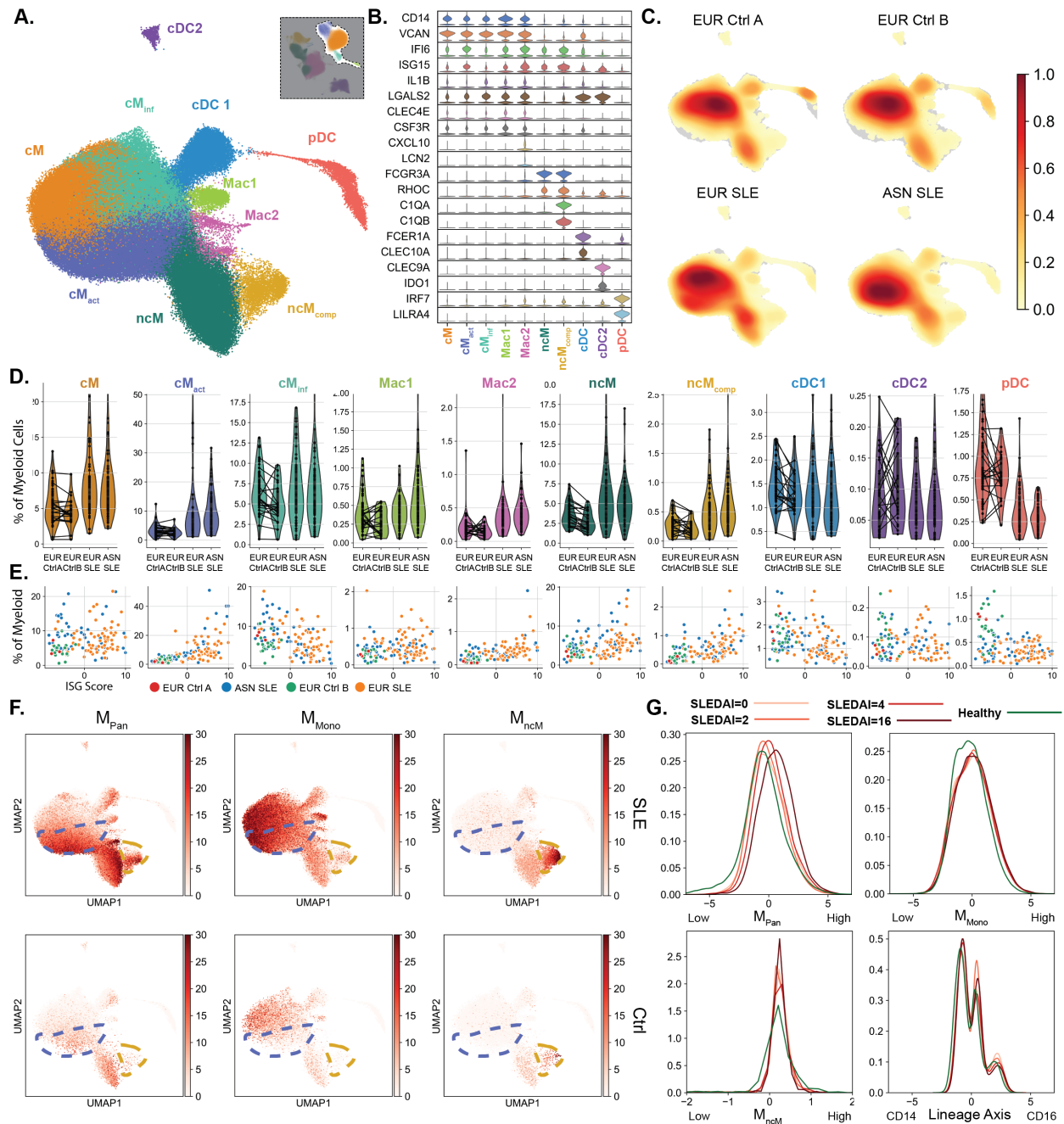
of monocyte and lymphocyte abundances. F.) Cell type percentage differences between cases and controls. Repeated controls are connected by a line. G.) Monocyte and lymphocyte abundances across populations in the UCSF EHR database. H.) Causal effect size correlation between Lymphocyte Count and SLE disease status reported in the UKBK.



**Figure 3.2: Bulk expression differences and variance decomposition.**

A.) Volcano plot of effect size (x-axis) and  $\log_{10}(\text{p-value})$  (y-axis) for differential expression in PBMCs. Differentially expressed genes in PBMCs (black) or only in specific cell types (gray) are colored. Previously identified IFN signature genes are circled. B.) Correlation between ISG score (x-axis) and myeloid percentage (y-axis). C.) Expression heatmap and cluster assignment of 209 differentially expressed genes between cases and controls. D.) Distribution of the contribution of cell type composition to gene expression variability in PBMCs. E.) Contribution of percentage (left) or cell-type specific expression (right) to the expression variability of each DE gene in PBMCs. F.) Correlation of *IL6* and *IFNG* expression in PBMCs with the percentage

of each cell type. G.) Correlation of observed and predicted SLEDAI scores based on gene expression and cell type composition features, including broad composition estimates (Mono/Lymph), refined cell types (Composition), bulk gene expression for IFN genes (PBMC), variable genes in cell type specific expression (CT Var) and the combination of features from cell type specific expression and composition (CT Var+Comp).



**Figure 3.3: Myeloid changes in SLE**

A.) UMAP projection and cluster annotation of myeloid cells. B.) Violin plot of marker genes differentiating annotated clusters. C.) Cell density plots for cases and controls split by ethnicity and sequencing site. D.) Cluster percentage differences between cases and controls. Lines connect control samples replicated across pools. E.) Correlation of cluster percentage with ISG score. red: EUR Ctrl, blue: ASN SLE, green: EUR Ctrl, orange: EUR SLE. F.) UMAP projection of single cells colored by the average expression of clusters of DE genes. G.) The density of cells along DPT ordering based on lineage markers, differentially expressed clusters or 25 known interferon sensitive genes.

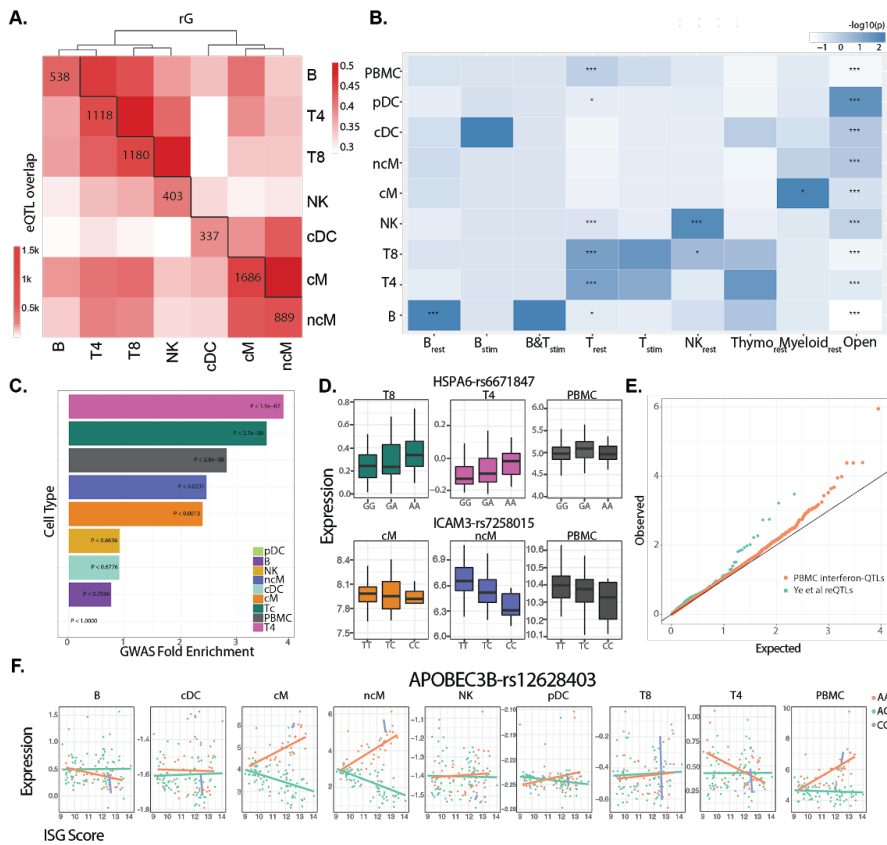


**Figure 3.4: Lymphoid changes in SLE.**

A.) Classification of each lymphocyte into 19 cell types. B.) Violin plot of key gene markers across different cell types. Lines connect control samples replicated across pools. C.) Cell

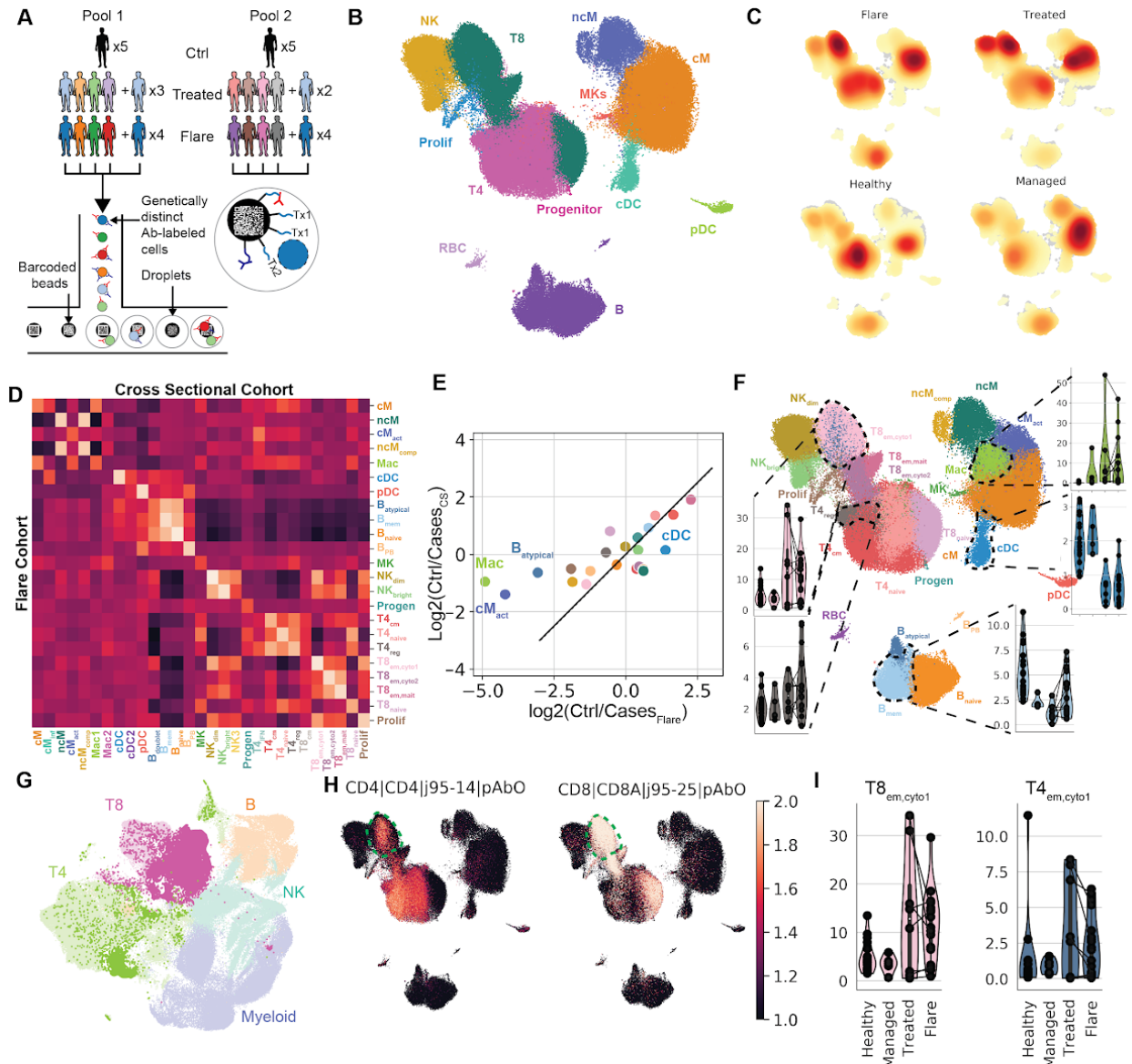
density plots for cases and controls split by ethnicity and sequencing site. D.) Cluster percentage differences between cases and controls E.) Percentage of clusters (y-axis) versus ISG score (x-axis). red: EUR Ctrl, blue: ASN SLE, green: EUR Ctrl, orange: EUR SLE. F.) Box-whisker plot of Gini coefficients marking clonal expansion for T4 and T8 cells in cases and controls. G.) UMAP projection with clonally expanded cells highlighted.





**Figure 3.5: cis-eQTL mapping demonstrates cell type specificity and environmental specificity in genetic effects.**

A.) eQTL overlap (lower triangle) and average genetic correlation (upper triangle) for each pair of cell types. B.) Enrichment measured by Mann Whitney p-value of cell type specific ATAC-seq peaks in each cell type. C.) Enrichment of SLE GWAS variants in each cell type. D.) Genetic variants with cell type specific genetic effects on *HSPA6* and *ICAM3*. E.) Quantile-quantile plot of *cis*-IFN-QTLs (orange) subsetted for previously published *cis*-IFN-QTLs (green). F.) IFN-specific genetic effects for each mutational status on *APOBEC3B* gene expression in each cell type.



**Figure 3.6: SLE flare cohort analysis demonstrates reproducibility of our findings and disease flare specific alterations**

A.) Multiplexed single cell sequencing and CITE-seq applied over 13 cases and 10 healthy controls. B.) Classification of each PBMC into 11 cell types. C.) Cell type proportion differences between cases and control. Lines connect matched flare and treated cases. D.) Cell density plots for controls, treated cases and flare cases. E.) Assignment of each PBMC to 26 cell types and changes in key clusters between flare and treated cases depicted by log<sub>2</sub> fold changes with respect to controls. F.) Pseudobulk gene expression correlation heatmap between the cross sectional and flare cohorts. G.) UMAP of all PBMCs colored by CD4 (left) and CD8 (right) antibody normalized abundance. T8<sub>em,cyto1</sub> is circled.

## Discussion

SLE remains one of the most challenging autoimmune diseases to diagnose and treat. The lack of effective targeted therapies<sup>47–53</sup>, the heterogeneous symptoms, and the treatment response variability highlight the significant need for improved molecular characterization of SLE.

Analyses of almost 1 million cells from nearly 200 individuals revealed key cell types whose frequency, state or both change in SLE. Compositionally, the depletion of naive CD4<sup>+</sup> T cells, especially in patients of Asian ancestry, refines the known observation of lymphopenia<sup>54</sup> and the depletion of antigen-presenting cells (APCs) is consistent with their localization in tissues<sup>55</sup>.

Beyond composition, we observed elevated expression of interferon sensitive genes (ISGs) attributable to the activation of classical monocytes and complement expressing non-classical monocytes. The negative correlation between the frequency of naive CD4<sup>+</sup> T cells and the expression of ISGs in monocytes suggests interferon activation as a cause for T4 lymphopenia. A model consistent with these observations is that APCs localize to sites of inflamed tissue and produce high levels of type-1 interferons with distinct effects on myeloid and lymphoid lineages. In monocytes, CD14<sup>+</sup> cells are polarized into macrophages and CD16<sup>+</sup> cells increase the production of complements. Interferon activation of T lymphocytes results in their sequestration in sites of inflammation through the regulation of CD69 and S1PR1<sup>29</sup>.

The most striking observation is the detection of expanded cytotoxic CD8<sup>+</sup> T cells in SLE patients with managed disease, which is even more abundant during periods of heightened disease. Clonal expansion and proliferation of cytotoxic lymphocytes have previously been observed in independent works<sup>3056</sup> and are consistent with a model of prolonged adaptive immune response in SLE, potentially initiated by foreign and autoantigens. In response to

antigen, activated and expanded cytotoxic T cells lyse antigen-presenting cells through the release of cytotoxic granules<sup>56</sup>. *In vitro*, the granzyme-B-dependent-cytotoxic pathway efficiently cleaves autoantigens observed in human systemic autoimmune diseases, generating unique fragments not observed in any other form of apoptosis and could drive antigenic feeding of APCs<sup>34,57</sup>. Prolonged contact between APCs and CD8<sup>+</sup> T cells leads to the formation of a mature stimulatory synapse and the secretion of interferon-gamma<sup>58</sup>, which would in turn activate macrophages consistent with increased macrophage frequency during flare. As several therapeutic strategies have been developed to target CD8<sup>+</sup> T cells including autologous regulatory T cell transfer with or without low dose IL-2 treatment, our results suggest that similar strategies could be considered as a novel therapeutic avenue to treat SLE.

Integrating measurements of cellular composition and cell-type-specific expression with dense genotyping provides a unique opportunity to partition inter-individual expression variability in PBMCs, assess its genetic determinants, and ascribe functionality to disease-associated variants. This was demonstrated by the detection of thousands of cell-type-specific *cis*-eQTLs enriched for *cis*-regulatory elements active in each respective cell type. T lymphocyte *cis*-eQTLs, in particular, are enriched for SLE-associated variants. In addition, we mapped genetic variants whose effects are modified by elevated interferon levels, a critical disease environment in SLE, suggesting that simultaneous genetic and single-cell transcriptomic profiling could be used to molecularly phenotype patients with systemic autoimmunity.

Looking forward, single-cell analysis of larger and more diverse cross-sectional cohorts is likely important for understanding the differences in SLE risk between genetic ancestries and the

involvement of environmental triggers. Longitudinal profiling of patients with or at risk for SLE could reveal new insights into the initiation of disease, escalation of symptoms, and response to treatment. More efficient transcript capture, higher sequencing depth, and larger sample sizes will undoubtedly improve the definition of molecular signatures to subphenotype SLE, the power to detect *cis*-eQTLs, and the resolution for annotating disease associations. Profiling of matched tissue and blood will provide a more complete picture of how immune cells are dynamically trafficked during disease, especially in cases with organ involvement and inform the development of future treatments for SLE.

## References

1. Carter, E. E., Barr, S. G. & Clarke, A. E. The global burden of SLE: prevalence, health disparities and socioeconomic impact. *Nat. Rev. Rheumatol.* **12**, 605–620 (2016).
2. Kaul, A. *et al.* Systemic lupus erythematosus. *Nat Rev Dis Primers* **2**, 16039 (2016).
3. Sharma, S. *et al.* Widely divergent transcriptional patterns between SLE patients of different ancestral backgrounds in sorted immune cell populations. *J. Autoimmun.* **60**, 51–58 (2015).
4. Banchereau, R. *et al.* Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients. *Cell* **165**, 1548–1550 (2016).
5. Crow, M. K. Type I Interferon in the Pathogenesis of Lupus. *The Journal of Immunology* **192**, 5459–5468 (2014).
6. Lauwerys, B. R., Ducreux, J. & Houssiau, F. A. Type I interferon blockade in systemic lupus erythematosus: where do we stand? *Rheumatology* **53**, 1369–1376 (2014).
7. Guerra, S. G., Vyse, T. J. & Cunninghame Graham, D. S. The genetics of lupus: a functional perspective. *Arthritis Res. Ther.* **14**, 211 (2012).
8. Dozmorov, M. G. *et al.* B-Cell and Monocyte Contribution to Systemic Lupus Erythematosus Identified by Cell-Type-Specific Differential Expression Analysis in RNA-Seq Data. *Bioinform. Biol. Insights* **9**, 11–19 (2015).
9. Kaminski, D. A., Wei, C., Rosenberg, A. F., Lee, F. E.-H. & Sanz, I. Multiparameter flow cytometry and bioanalytics for B cell profiling in systemic lupus erythematosus. *Methods Mol. Biol.* **900**, 109–134 (2012).

10. Liu, M.-F. & Wang, C.-R. Increased Th17 cells in flow cytometer-sorted CD45RO-positive memory CD4 T cells from patients with systemic lupus erythematosus. *Lupus Sci Med* **1**, e000062 (2014).
11. Mostafavi, S. *et al.* Parsing the Interferon Transcriptional Network and Its Disease Associations. *Cell* **164**, 564–578 (2016).
12. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
13. Lanata, C. M. *et al.* Genetic contributions to lupus nephritis in a multi-ethnic cohort of systemic lupus erythematosus patients. *PLoS One* **13**, e0199003 (2018).
14. Ye, C. J. *et al.* Intersection of population variation and autoimmunity genetics in human T cell activation. *Science* **345**, 1254665 (2014).
15. Lee, M. N. *et al.* Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* **343**, 1246980 (2014).
16. Raj, T. *et al.* Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science* **344**, 519–523 (2014).
17. De Jager, P. L. *et al.* ImmVar project: Insights and design considerations for future studies of ‘healthy’ immune variation. *Semin. Immunol.* **27**, 51–57 (2015).
18. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018).
19. Klarquist, J., Zhou, Z., Shen, N. & Janssen, E. M. Dendritic Cells in Systemic Lupus Erythematosus: From Pathogenic Players to Therapeutic Tools. *Mediators Inflamm.* **2016**, 5045248 (2016).

20. Subedi, A., Magder, L. S. & Petri, M. Effect of mycophenolate mofetil on the white blood cell count and the frequency of infection in systemic lupus erythematosus. *Rheumatol. Int.* **35**, 1687–1692 (2015).
21. Yu, D. T. *et al.* Human lymphocyte subpopulations. Effect of corticosteroids. *J. Clin. Invest.* **53**, 565–571 (1974).
22. Slade, J. D. & Hepburn, B. Prednisone-induced alterations of circulating human lymphocyte subsets. *J. Lab. Clin. Med.* **101**, 479–487 (1983).
23. Zhu, Z. *et al.* Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* **9**, 224 (2018).
24. Dhir, V., Singh, A. P., Aggarwal, A., Naik, S. & Misra, R. Increased T-lymphocyte apoptosis in lupus correlates with disease activity and may be responsible for reduced T-cell frequency: a cross-sectional and longitudinal study. *Lupus* **18**, 785–791 (2009).
25. Jourdan, M. *et al.* IL-6 supports the generation of human long-lived plasma cells in combination with either APRIL or stromal cell-soluble factors. *Leukemia* **28**, 1647–1656 (2014).
26. Stoycheva, D. *et al.* IFN- $\gamma$  regulates CD8<sup>+</sup> memory T cell differentiation and survival in response to weak, but not strong, TCR signals. *J. Immunol.* **194**, 553–559 (2015).
27. Arkatkar, T. *et al.* B cell-derived IL-6 initiates spontaneous germinal center formation during systemic autoimmunity. *J. Exp. Med.* **214**, 3207–3217 (2017).
28. Kamphuis, E., Junt, T., Waibler, Z., Forster, R. & Kalinke, U. Type I interferons directly regulate lymphocyte recirculation and cause transient blood lymphopenia. *Blood* **108**, 3253–3261 (2006).
29. Shiow, L. R. *et al.* CD69 acts downstream of interferon- $\alpha/\beta$  to inhibit S1P1 and lymphocyte egress from lymphoid organs. *Nature* **440**, 540–544 (2006).



30. Blanco, P. *et al.* Increase in activated CD8<sup>+</sup> T lymphocytes expressing perforin and granzyme B correlates with disease activity in patients with systemic lupus erythematosus. *Arthritis Rheum.* **52**, 201–211 (2005).
31. Jelcic, I. *et al.* Memory B Cells Activate Brain-Homing, Autoreactive CD4<sup>+</sup> T Cells in Multiple Sclerosis. *Cell* **175**, 85–100.e23 (2018).
32. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415–1429.e19 (2016).
33. Calderon, D. *et al.* Landscape of stimulation-responsive chromatin across diverse human immune cells. *bioRxiv* 409722 (2018). doi:10.1101/409722
34. Anderson, C. A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.* **43**, 246–252 (2011).
35. Calderwood, S. K., Stevenson, M. A. & Murshid, A. Heat shock proteins, autoimmunity, and cancer treatment. *Autoimmune Dis.* **2012**, 486069 (2012).
36. Márquez, A. *et al.* A combined large-scale meta-analysis identifies COG6 as a novel shared risk locus for rheumatoid arthritis and systemic lupus erythematosus. *Ann. Rheum. Dis.* **76**, 286–294 (2017).
37. Martin, S. *et al.* Circulating forms of ICAM-3 (cICAM-3). Elevated levels in autoimmune diseases and lack of association with cICAM-1. *J. Immunol.* **154**, 1951–1955 (1995).
38. Pino-Otín, M. R. *et al.* Existence of a soluble form of CD50 (intercellular adhesion molecule-3) produced upon human lymphocyte activation. Present in normal human serum and levels are increased in the serum of systemic lupus erythematosus patients. *J. Immunol.* **154**, 3015–3024 (1995).

39. Gamazon, E. R. *et al.* Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet.* **50**, 956–967 (2018).
40. Ye, C. J. *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).
41. Roth, S. H. *et al.* Increased RNA Editing May Provide a Source for Autoantigens in Systemic Lupus Erythematosus. *Cell Rep.* **23**, 50–57 (2018).
42. Crow, M. K., Olfieriev, M. & Kirou, K. A. Identification of Candidate Predictors of Lupus Flare. *Trans. Am. Clin. Climatol. Assoc.* **126**, 184–196 (2015).
43. Nakayama, W. *et al.* CD163 expression is increased in the involved skin and sera of patients with systemic lupus erythematosus. *Eur. J. Dermatol.* **22**, 512–517 (2012).
44. Endo, N. *et al.* Urinary soluble CD163 level reflects glomerular inflammation in human lupus nephritis. *Nephrol. Dial. Transplant* **31**, 2023–2033 (2016).
45. Olmes, G. *et al.* CD163+ M2c-like macrophages predominate in renal biopsies from patients with lupus nephritis. *Arthritis Res. Ther.* **18**, 90 (2016).
46. Blanco, P., Viallard, J.-F., Pellegrin, J.-L. & Moreau, J.-F. Cytotoxic T lymphocytes and autoimmunity. *Curr. Opin. Rheumatol.* **17**, 731–734 (2005).
47. Amisshah-Arthur, M. B. & Gordon, C. Contemporary treatment of systemic lupus erythematosus: an update for clinicians. *Ther. Adv. Chronic Dis.* **1**, 163–175 (2010).
48. Chatham, W. W. & Kimberly, R. P. Treatment of lupus with corticosteroids. *Lupus* **10**, 140–147 (2001).
49. Dubey, A. K. *et al.* Belimumab: First targeted biological treatment for systemic lupus erythematosus. *J. Pharmacol. Pharmacother.* **2**, 317–319 (2011).

50. Guerreiro Castro, S. & Isenberg, D. A. Belimumab in systemic lupus erythematosus (SLE): evidence-to-date and clinical usefulness. *Ther. Adv. Musculoskelet. Dis.* **9**, 75–85 (2017).
51. Petri, M. *et al.* Derivation and validation of the Systemic Lupus International Collaborating Clinics classification criteria for systemic lupus erythematosus. *Arthritis Rheum.* **64**, 2677–2686 (2012).
52. Kalunian, K. C. *et al.* A Phase II study of the efficacy and safety of rontalizumab (rhuMAb interferon- $\alpha$ ) in patients with systemic lupus erythematosus (ROSE). *Ann. Rheum. Dis.* **75**, 196–202 (2016).
53. Alert, N. P. R. & T-Cells, C. A. R. Update on TULIP 1 Phase III trial for anifrolumab in systemic lupus erythematosus.
54. Baechler, E. C. *et al.* Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 2610–2615 (2003).
55. Chan, V. S.-F. *et al.* Distinct roles of myeloid and plasmacytoid dendritic cells in systemic lupus erythematosus. *Autoimmun. Rev.* **11**, 890–897 (2012).
56. Mato, T. *et al.* Correlation of clonal T cell expansion with disease activity in systemic lupus erythematosus. *Int. Immunol.* **9**, 547–554 (1997).
57. Walport, M. J., Davies, K. A. & Botto, M. C1q and systemic lupus erythematosus. *Immunobiology* **199**, 265–285 (1998).
58. Siegert, C., Daha, M., Westedt, M. L., van der Voort, E. & Breedveld, F. IgG autoantibodies against C1q are correlated with nephritis, hypocomplementemia, and dsDNA antibodies in systemic lupus erythematosus. *J. Rheumatol.* **18**, 230–234 (1991).
59. Coremans, I. E. *et al.* Changes in antibodies to C1q predict renal relapses in systemic lupus erythematosus. *Am. J. Kidney Dis.* **26**, 595–601 (1995).

60. Frémeaux-Bacchi, V., Weiss, L., Demouchy, C., Blouin, J. & Kazatchkine, M. D. Autoantibodies to the collagen-like region of C1q are strongly associated with classical pathway-mediated hypocomplementemia in systemic lupus erythematosus. *Lupus* **5**, 216–220 (1996).
56. Bots, Michael, and Jan Paul Medema. "Granzymes at a glance." *Journal of cell science* 119.24 (2006): 5011-5014.
57. Casciola-Rosen, Livia, et al. "Cleavage by granzyme B is strongly predictive of autoantigen status: implications for initiation of autoimmunity." *Journal of Experimental Medicine* 190.6 (1999): 815-826.
58. Faroudi M, *et al.* Lytic versus stimulatory synapse in cytotoxic T lymphocyte/target cell interaction: manifestation of a dual activation threshold. *Proc Natl Acad Sci USA* 2003;100: 14145–14150.

**Publishing Agreement**

*It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.*

***Please sign the following statement:***

*I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.*



---

Author Signature

September 4, 2019

---

Date