There are amendments to this paper

nature biotechnology

Multiplexed droplet single-cell RNA-sequencing using natural genetic variation

Hyun Min Kang[1,20], Meena Subramaniam[2–6,20], Sasha Targ[2–7,20], Michelle Nguyen[8–10], Lenka Maliskova[3,11], Elizabeth McCarthy[7], Eunice Wan[3], Simon Wong[3], Lauren Byrnes[12], Cristina M Lanata[13,14], Rachel E Gate[2–6], Sara Mostafavi[15], Alexander Marson[8–10,13,16,17], Noah Zaitlen[3,13,18], Lindsey A Criswell[3,13,14,19] & Chun Jimmie Ye[3–6]

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 droplet containing a single cell (singlet) and detect droplets containing two cells (doublets). 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 single-nucleotide polymorphisms (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 eight pooled samples, demuxlet correctly recovers the sample identity of >99% of singlets and identifies doublets at rates consistent with previous estimates. We apply demuxlet to assess cell-type-specific changes in gene expression in 8 pooled lupus patient samples treated with interferon (IFN)-β and perform eQTL analysis on 23 pooled samples.

DscRNA-seq has increased substantially the throughput of single-cell capture and library preparation[1,2], enabling the simultaneous transcriptomic profiling of thousands of cells. Improvements in biochemistry[3,4] and microfluidics[5,6] continue to increase the number of cells and transcripts profiled per experiment. But for differential expression and population genetics studies, sequencing thousands of cells each from many individuals 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 droplet containing a single cell (singlet) and identify droplets containing two cells from different individuals (doublets) (**Fig. 1a**). While strategies to demultiplex cells from different species[1,2,7] or host and graft samples[7] 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[8], demuxlet is fast, accurate, scalable, and compatible with standard input formats[7,9,10].

Demuxlet implements a statistical model for evaluating the likelihood of observing RNA-seq reads overlapping a set of SNPs from each cell- containing droplet. 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 donors for each droplet using a mixture model. A small number of reads overlapping common SNPs is sufficient to accurately identify each droplet. For a pool of eight individuals and a set of uncorrelated SNPs, each with 50% minor allele frequency (MAF), four reads overlapping

[1]Department of Biostatistics and Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, Michigan, USA. [2]Biological and Medical Informatics Graduate Program, University of California, San Francisco, San Francisco, California, USA. [3]Institute for Human Genetics (IHG), University of California, San Francisco, San Francisco, California, USA. [4]Institute for Computational Health Sciences, University of California, San Francisco, San Francisco, California, USA. [5]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, California, USA. [6]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, California, USA. [7]Medical Scientist Training Program (MSTP), University of California, San Francisco, San Francisco, California, USA. [8]Department of Microbiology and Immunology, University of California, San Francisco, San Francisco, California, USA. [9]Diabetes Center, University of California, San Francisco, San Francisco, California, USA. [10]Innovative Genomics Institute, University of California, Berkeley, Berkeley, California, USA. [11]Department of Neurology, University of California, San Francisco, San Francisco, California, USA. [12]Developmental and Stem Cell Biology Graduate Program, University of California, San Francisco, San Francisco, California, USA. [13]Department of Medicine, University of California, San Francisco, San Francisco, California, USA. [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, San Francisco, California, USA. [17]Chan Zuckerberg Biohub, San Francisco, California, USA. [18]Lung Biology Center, University of California, San Francisco, San Francisco, California, USA. [19]Department of Orofacial Sciences, University of California, San Francisco, San Francisco, USA. [20]These authors contributed equally to this work. Correspondence should be addressed to H.M.K. (hmkang@umich.edu) or C.J.Y (jimmie.ye@ucsf.edu).

Received 20 March; accepted 16 November; published online 11 December 2017; doi:10.1038/nbt.4042

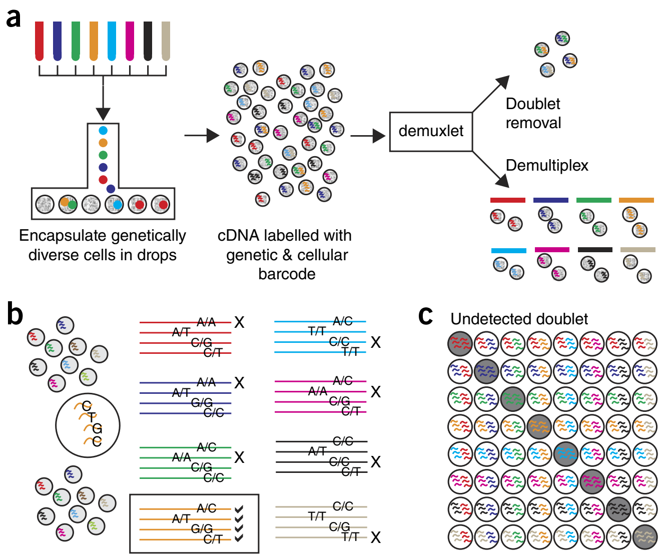
a
Encapsulate genetically diverse cells in drops
cDNA labelled with genetic & cellular barcode
demuxlet
Doublet removal
Demultiplex
b
A/A
A/T
G/G
C/T
A/C
T/T
C/G
A/A
G/G
C/C
A/T
X
c
Undetected doublet

Figure 1 Demuxlet: 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. (**b**,**c**) Assuming equal mixing of eight individuals, four genetic variants can recover the sample identity of a cell (**b**), and 87.5% of doublets will contain cells from two different samples (**c**).

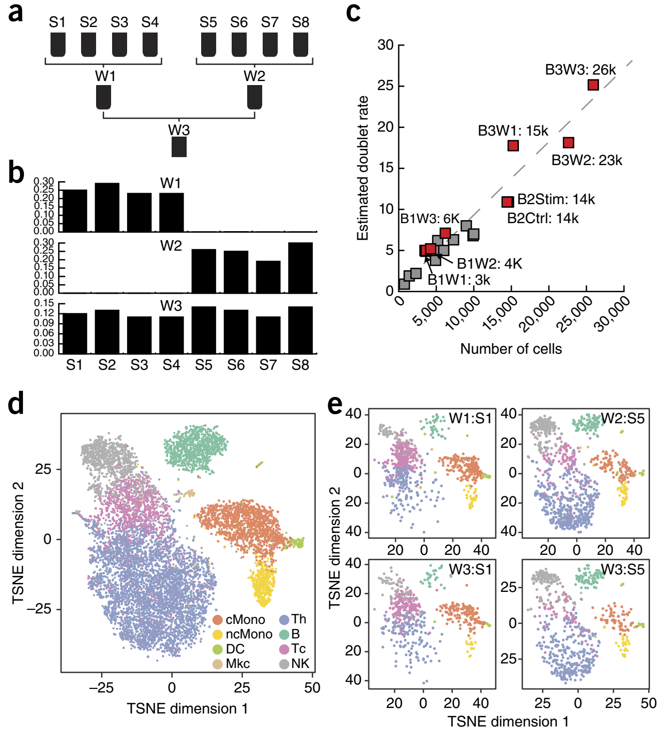

Figure 2 Performance of demuxlet. (**a**) Experimental design for equimolar pooling of cells from eight unrelated samples (S1–S8) into three wells (W1–W3). W1 and W2 contain cells from two disjoint sets of four individuals. W3 contains cells from all eight individuals. (**b**) Demultiplexing singlets 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.

SNPs are sufficient to uniquely assign a singlet to the donor of origin (**Fig. 1b**) 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, for example, 87.5% for N = 8 samples) (**Fig. 1c**). 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 37-fold more singlets (36,600) if the sample identity of every droplet can be perfectly demultiplexed (**Supplementary Fig. 2**). Profiling 30,000 cells multiplexed from 26 individuals can theoretically generate 23-fold more singlets at the same effective doublet rate, minimizing the effects of sequencing doublets (**Supplementary Fig. 3**).

We first assessed the performance of multiplexed dscRNA-seq through simulation. The ability to demultiplex droplets is a function of the number of individuals multiplexed, the depth of sequencing or number of read-overlapping SNPs, and the relatedness of multiplexed individuals. We simulated 6,145 droplets (5,837 singlets and 308 doublets) from 2–64 individuals from the 1000 Genomes Project[11]. We show that 50 SNPs per droplet allows demultiplexing of 97% of singlets and identification of 92% of doublets in pools of up to 64 individuals (**Supplementary Figs. 4–5**). 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 droplet (**Supplementary Fig. 6**), and that the 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 singlets from a set of eight related individuals from 1000 Genomes[11]. In this simulation, 50 SNPs per singlet would allow demuxlet to correctly assign over 98% of singlet (**Supplementary Fig. 7**). These results suggest optimal multiplexed designs where cells from tens of unrelated individuals should be pooled, loaded at concentrations 2–10× higher than standard workflows, and sequenced to at least 1,000 UMIs per droplet.

We evaluated the performance of demuxlet by analyzing a pool of peripheral blood mononuclear cells (PBMCs) from eight 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 were loaded in a well on a 10× Chromium instrument (**Fig. 2a**). 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% (3,332/3,645), 91% (3,864/4,254), and 86% (5,348/6,205) of droplets as singlets (likelihood ratio test, L(singlet)/L(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 four individuals, we estimated a demultiplexing error rate (number of cells assigned to individuals not in the pool) of less than 1% of singlets (W1: 2/3,332, W2: 0/3,864) (**Fig. 2b**).

We next assessed the ability of demuxlet to detect doublets in both simulated and real data. 466/3,645 (13%) droplets from W1 were simulated as synthetic doublets by setting the cellular barcodes of 466 droplets, each from individuals S1 and S2, to be the same. Applied to simulated data, demuxlet identified 91% (426/466)

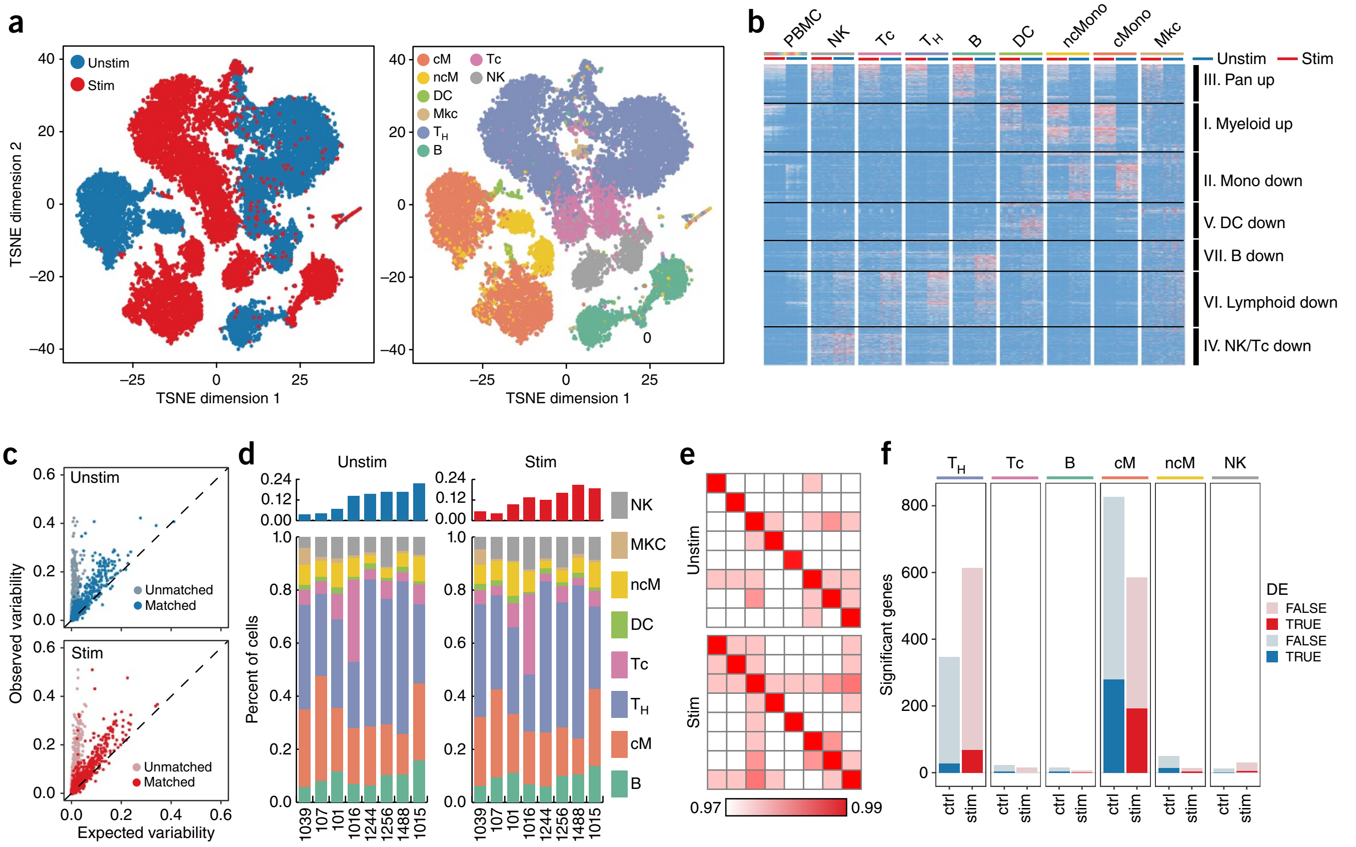

Figure 3 Inter-individual variability in IFN-β response. (**a**) t-SNE plot of unstimulated (blue) and IFN-β-stimulated (red) PBMCs and the presumed cell types. cM, CD14$^+$CD16$^-$ monocytes; ncM, CD14$^+$CD16$^+$ monocytes; DC, dendritic cells; Mkc, megakaryocytes; Th, CD4$^+$ T cells; B, B cells; Tc, CD8$^+$ T cells; NK, natural killer cells. (**b**) Cell-type-specific expression in stimulated (left) and unstimulated (right) cells. Differentially expressed genes shown (FDR < 0.05, |log(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 eight 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.

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 138/3,645, 165/4,254, and 384/6,205 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. 2c**).

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. 2d**) 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. 2e** 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 six genes were differentially expressed between wells W1 and W2 (DESeq2 on pseudobulk counts, FDR < 0.05), only two genes were differentially expressed between W1 and W2 individuals in well W3 (FDR < 0.05) (**Supplementary Fig. 11b**), suggesting multiplexing reduces technical effects owing to separate sample processing[12,13].

We used multiplexed dscRNA-seq to characterize the cell-type specificity and inter-individual variability of response to IFN-β, a potent cytokine that induces genome-scale changes in the transcriptional profiles of immune cells[14,15]. From each of eight lupus patients, PBMCs were activated with recombinant IFN-β or left untreated for 6 h, a time point we previously found to maximize the expression of interferon-sensitive genes in dendritic cells and T cells[16,17]. Two pools, IFN-β-treated and control, were prepared with the same number of cells from each individual and loaded onto the 10× 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), respectively, as singlets. The estimated doublet rate of 10.9% in each condition is consistent with predicted rates (**Fig. 2c**), and the observed and expected frequencies of doublets for each pair of individuals were 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 eight individuals within each pool as biological replicates to quantitatively assess cell-type-specific IFN-β responses in PBMCs. Consistent with previous reports from bulk RNA-sequencing data, IFN-β stimulation induces widespread transcriptomic changes observed as a shift in

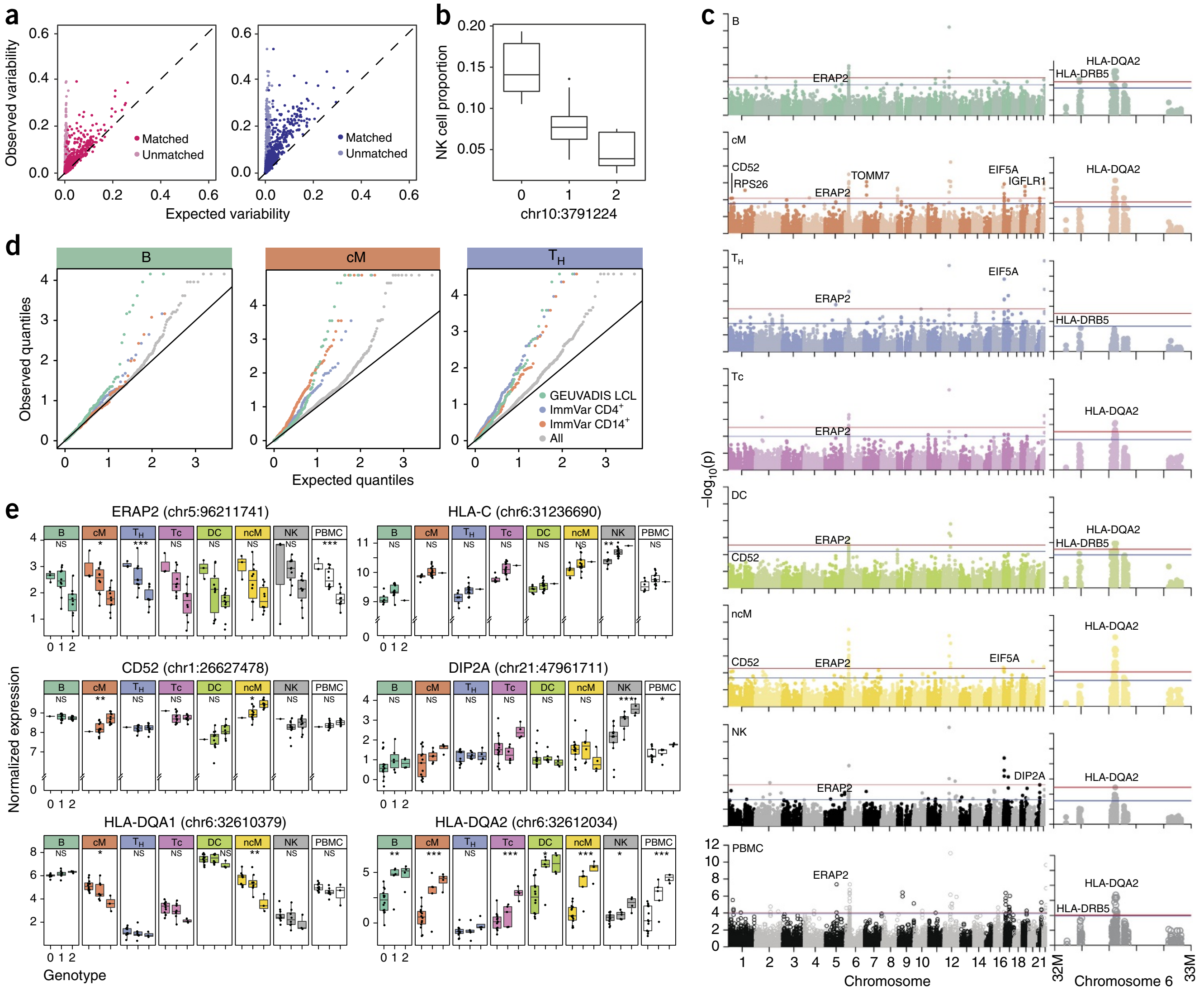

Figure 4 Genetic control over cell type proportion and gene expression (N = 23). (**a**) Observed variance (y axis) in mean expression over all PBMCs from each individual versus expected variance (x axis) over synthetic replicates sampled across batch 1 (left, N = 8) and batch 3 (right, N = 15). (**b**) Association of chr10:3791224 with NK cell type proportions. (**c**) Genome-wide and chromosome six 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 T_H 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).

the t-SNE projections of singlets[14] (**Fig. 3a**). As expected, IFN-β 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 ~2,000 homologous genes in each cell type and condition correlated with similar data from mice (**Supplementary Fig. 16**). We identified 3,055 differentially expressed genes abs(logFC) > 1, FDR < 0.05) in at least one cell type (**Supplementary Table 1**). For 709 genes, estimates of fold change in response to IFN-β stimulation in myeloid and CD4$^+$ cells were consistent with estimates in monocyte-derived dendritic cells[18] and CD4$^+$ T cells[17], respectively (**Supplementary Fig. 17**) and correlated with qPCR results of sorted CD4$^+$ T cells (**Supplementary Fig. 18**). Differentially expressed genes clustered into modules of cell-type-specific responses enriched for distinct gene regulatory programs (**Fig. 3b** and **Supplementary Table 2**). For example, genes upregulated in all leukocytes (Cluster III: 401 genes, abs(logFC) > 1, FDR < 0.05) or only in myeloid cells (Cluster I: 767 genes, abs(logFC) > 1, FDR < 0.05) were 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 were 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/T_H 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-β signatures in mouse and humans[14].

Over all PBMCs, the variance of mean expression across individuals was higher than the variance across synthetic replicates whose cells were randomly sampled (Lin's concordance = 0.022, Pearson correlation = 0.69, **Fig. 3c**). The variance across synthetic replicates whose cells were randomly sampled; matching for cell type proportions was more concordant with the variance across individuals (Lin's concordance = 0.54, Pearson correlation = 0.78, **Fig. 3c**,**d**), suggesting a contribution of cell type composition on expression variability. However, for each cell type, the variance across individuals[12,14,19] was 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$^+$CD16$^-$ 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. 3e**). We found 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$^+$ helper T (T_H) cells. Inter-individual variable genes in stimulated cM and to a lesser extent in T_H cells ($P < 9.3 \times 10^{-4}$ and 4.5×10^{-2}, hypergeometric test, **Fig. 3f**) were enriched for differentially expressed genes, consistent with our previous discovery of more IFN-β response–expression quantitative trait loci (eQTLs) in monocyte-derived dendritic cells than in CD4$^+$ T cells[16,17]. Comparing these genes to 407 genes previously profiled in bulk monocyte-derived dendritic cells, we found the proportion of variance explained by inter-individual variability was correlated more strongly 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 droplets (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, respectively, as singlets, correctly assigning 99% of singlets from the first two pools, W1 (10,740/10,766) and W2 (16,616/16,618). The estimated doublet rates of 18%, 18%, and 25% are consistent with the increased concentrations of loaded cells (**Fig. 2c**). Similar to the IFN-β stimulation experiment, we found that expression variability was determined by variability in cell type proportion (**Fig. 4a**) and reproducible between batches (**Supplementary Fig. 20**). Associating >150,000 genetic variants (MAF > 20%) with the proportion of eight 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 natural killer (NK) cells (**Fig. 4b**).

Across 23 donors, we conducted an eQTL analysis to map genetic variants associated with expression variability in each major immune cell type. We found a total of 32 local eQTLs (±100 kb, FDR < 0.1), 22 of which were detected in only one cell type (**Fig. 4c** and **Supplementary Table 3**). Previously reported local eQTLs from bulk CD14$^+$ monocytes, CD4$^+$ T cells, and lymphoblastoid cell lines were more significantly associated with gene expression in the most similar cell types (cM, T_H, and B cells, respectively) than other cell types (**Fig. 4d**). 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, such as, *ERAP2* ($P < 3.57 \times 10^{-32}$, meta-analysis), which encodes an aminopeptidase known to cleave viral peptides[20], and *HLA-C* ($P < 1.74\text{x}10^{-29}$, meta-analysis), which encodes the MHC class I heavy chain (**Fig. 4e**). *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, which only has eQTLs in monocyte populations, and *DIP2A*, a gene with an eQTL only in NK cells, which is associated with immune response to vaccination in peripheral blood[21]. 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.

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, removes the need for synthetic barcodes or split-pool strategies[22–26], and allows the capture of biological variability among individual samples while limiting unwanted technical variability. We find the optimal number of samples to multiplex is ~20, based on sample processing time and empirical doublet rates of current microfluidic devices, and anticipate that number to increase with automated sample handling and lower doublet rates.

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[27]. 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 patients[28–30]. 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

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

M.S. and C.J.Y. are supported by NIH R01AR071522 and R21AI133337. S.T. is supported by NIH F30DK115167. H.M.K. is supported by U01HL137182. N.Z. is supported by NIH K25HL121295, R03DE025665, and Department of Defense W81WH-16-2-0018.

AUTHOR CONTRIBUTIONS

H.M.K. and C.J.Y. conceived the project. M.S., S.T., L.M., R.G., L.B., E.W., S.W., and M.N. performed all experiments. H.M.K., M.S., S.T., E.M., S.M., and C.J.Y. analyzed the data. C.L. and L.A.C. provided the patient samples. N.Z. and A.M. provided helpful comments and discussion. H.M.K., M.S., S.T., and C.J.Y. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Macosko, E.Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
2. Klein, A.M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
3. Stegle, O., Teichmann, S.A. & Marioni, J.C. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **16**, 133–145 (2015).
4. Gawad, C., Koh, W. & Quake, S.R. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* **17**, 175–188 (2016).

5. Streets, A.M. *et al.* Microfluidic single-cell whole-transcriptome sequencing. *Proc. Natl. Acad. Sci. USA* **111**, 7048–7053 (2014).
6. Zilionis, R. *et al.* Single-cell barcoding and sequencing using droplet microfluidics. *Nat. Protoc.* **12**, 44–73 (2017).
7. Zheng, G.X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
8. Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91**, 839–848 (2012).
9. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
10. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
11. Auton, A. *et al.* The Genomes Project. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
12. Aguirre-Gamboa, R. *et al.* Differential effects of environmental and genetic factors on T and B cell immune traits. *Cell Rep.* **17**, 2474–2487 (2016).
13. Li, Y. *et al.* A functional genomics approach to understand variation in cytokine production in humans. *Cell* **167**, 1099–1110.e14 (2016).
14. Mostafavi, S. *et al.* Parsing the interferon transcriptional network and its disease associations. *Cell* **164**, 564–578 (2016).
15. Stark, G.R., Kerr, I.M., Williams, B.R.G., Silverman, R.H. & Schreiber, R.D. How cells respond to interferons. *Annu. Rev. Biochem.* **67**, 227–264 (1998).
16. Lee, M.N. *et al.* Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* **343**, 1246980 (2014).
17. Ye, C.J. *et al.* Intersection of population variation and autoimmunity genetics in human T cell activation. *Science* **345**, 1254665 (2014).
18. Andrés, A.M. *et al.* Balancing selection maintains a form of ERAP2 that undergoes nonsense-mediated decay and affects antigen presentation. *PLoS Genet.* **6**, e1001157 (2010).
19. 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).
20. Saveanu, L. *et al.* Concerted peptide trimming by human ERAP1 and ERAP2 aminopeptidase complexes in the endoplasmic reticulum. *Nat. Immunol.* **6**, 689–697 (2005).
21. Franco, L.M. *et al.* Integrative genomic analysis of the human immune response to influenza vaccination. *eLife* **2**, e00299 (2013).
22. Cao, J. *et al.* Comprehensive single cell transcriptional profiling of a multicellular organism by combinatorial indexing. Preprint at bioRxiv https://doi.org/10.1101/104844 (2017).
23. 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).
24. Adamson, B. *et al.* A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867–1882.e21 (2016).
25. Jaitin, D.A. *et al.* Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-Seq. *Cell* **167**, 1883–1896.e15 (2016).
26. Datlinger, P. *et al.* Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 (2017).
27. Farh, K.K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
28. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33**, 155–160 (2015).
29. Tung, P.-Y. *et al.* Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* **7**, 39921 (2017).
30. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331–338 (2017).

ONLINE METHODS

Identifying the sample identity of each droplet. We first describe the method to infer the sample identity of each droplet 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 $b_{cvi} \in \{R, A, O\}$, $i \in \{1, \ldots, d_{cv}\}$ 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 the 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

$$\text{Bernoulli}\left(10^{-\frac{q_{cvi}}{10}}\right)$$

where q_{cvi} is a phred-scale quality score of the observed base call. We use the standard 10× 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[31].

We allow uncertainty of observed genotypes at the v-th variant for the s-th sample using $P_{sv}^{(g)} = \Pr(g \mid \text{Data}_{sv})$, the posterior probability of a possible genotype g given external DNA data 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 $P_{sv}^{(g)} = \Pr(\text{Data}_{sv} \mid g)\Pr(g)$ where $\Pr(g) \sim \text{Binomial}(2, p_v)$ and p_v is the population allele frequency of the alternate allele. To allow errors in the posterior probability, we replace it with $(1 - \in)P_{sv}^{(g)} + \varepsilon \Pr(g)$. The overall likelihood that the c-th droplet originated from the s-th sample is

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

In the absence of doublets, we use the maximum likelihood to determine the best-matching sample as $\text{argmax}_s\,[L_c(s)]$. See **Supplementary Code** for implementation details.

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[8],

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

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

$$\frac{\max_{s_1, s_2, \alpha} L_c(s_1, s_2, \alpha)}{\alpha \max_s L_c(s)} \geq t$$

and the most likely mixing proportion is estimated to be $\text{argmax}_\alpha L_c(s_1, s_2, \alpha)$. We determine that the cell contains only a single individual s if

$$\frac{\max_{s_1, s_2, \alpha} L_c(s_1, s_2, \alpha)}{\max_s L_c(s)} \leq \frac{1}{t},$$

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 Figure 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 (**Supplementary Fig. 2**) is as follows. Let d_o (e.g., 0.01) be the proportion of true multiplets when x_o (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

$$d(x) = 1 - (1 - d_0)^{\frac{x}{x_o}}$$

Let α be the fraction of true singlets incorrectly classified as non-singlets (i.e., doublet or ambiguous), and α 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

$$\frac{(1-\beta)[1 - d(x)]x_0 d(x)}{-\log(1 - d_0)}$$

Given α, β 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 Figure 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 10× 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 data set, resulting in a 5% doublet rate in the original data. For simulations with related individuals, we simulated transcriptomes from eight 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 40-µum column. Finally, the cells were washed in EasySep and resuspended in 1× PBMS and 0.04% bovine serum albumin. Cells from eight donors were then re-concentrated to 1 M cells per mL and then serially pooled. At each pooling stage, 1 M cells per mL were combined to result in a final sample pool with cells from all donors.

IFN-β stimulation and culture. Prior to pooling, samples from eight individuals were separated into two aliquots each. One aliquot of PBMCs was activated by 100 U/mL of recombinant IFN-β (PBL Assay Science) for 6 h according to the published protocol[16]. The second aliquot was left untreated. After 6 h, the eight 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. 1 M 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$^+$CD4$^+$CD8$^-$ (CD4$^+$ T cells), CD3$^+$CD4$^-$ CD8$^+$

(CD8$^+$ T cells), CD3$^-$ CD19$^+$ (B cells), and CD3$^-$ CD14$^+$ (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$^+$ 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[32]. 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 ΔΔCT between control and stimulated samples with GAPDH as a reference gene.

Droplet-based capture and sequencing. Cellular suspensions were loaded onto the 10× Chromium instrument (10× Genomics) and sequenced as described in Zheng *et al.*[7]. The cDNA libraries were sequenced using a custom program on ten lanes of Illumina HiSeq2500 Rapid Mode, yielding 1.8 B total reads and 25 K reads per droplet. 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 EasyEights Magnet was used to sequentially isolate CD14$^+$ (using the EasySep Human CD14 positive selection kit II, cat #17858), CD19$^+$ (using the EasySep Human CD19 positive selection kit II, cat #17854), CD8$^+$ (EasySep Human CD8 positive selection kitII, cat#17853), and CD4$^+$ 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[33,34]. 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 0.2 ng of cDNA. Samples were then sequenced on one lane of the Illumina Hiseq4000 with paired end 100-bp read length, yielding 350 M 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 droplet[7]. 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 2.7 k PBMCs from Zheng *et al.*[7], following the publicly available Guided Clustering Tutorial[7,35]. The FindAllMarkers function was then used to find the top 20 markers for each of the eight identified cell types. Cluster averages were calculated by taking the average raw count across all cells of each cell type. For each singlet, we calculated the Spearman correlation of the raw counts of the marker genes and the cluster averages, and assigned each singlet 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[36]. Genes with baseMean > 1 were filtered out from the DESeq2 output, and the qvalue package was used to calculate FDR < 0.05 (ref. 37).

Estimation of inter-individual 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 inter-individual 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 data set was estimated as repeatability using a linear mixed model[16,38].

Estimation of inter-individual variability within cell types. For each cell type, we generated two bulk equivalent replicates for each individual by summing raw counts of singlets 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 $\log_2$-normalized counts. Pearson correlation of the two replicates from each of the eight individuals was used to find genes with significant inter-individual variability.

Quantitative trait mapping in major immune cell types. Genotypes were imputed with EAGLE[39] 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[40]. *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[36]. All variants within a window of 100 kbp of each gene were tested with linear regression using Matrix eQTL[40]. Batch information for each sample as well as the first three principal components of the expression matrix were used as covariates.

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

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

31. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
32. Wang, X., Spandidos, A., Wang, H. & Seed, B. PrimerBank: a PCR primer database for quantitative gene expression analysis, 2012 update. *Nucleic Acids Res.* **40**, D1144–D1149 (2012).
33. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
34. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
35. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
36. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
37. Dabney, A., Storey, J.D. & Warnes, G.R. qvalue: Q-value estimation for false discovery rate control. *R package version* **1** (2010).
38. Falconer, D.S. & Mackay, T.F. *Introduction to Quantitative Genetics*, 4th edn. (Pearson, 1996).
39. 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).
40. Shabalin, A.A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).

doi:10.1038/nbt.4042

nature research

Corresponding author(s): Jimmie Ye

☐ Initial submission ☒ Revised version ☐ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

At N=23 we expect to detect > 55% of eQTLs with MAF >0.2 given an average effect size of 10%.

2. Data exclusions

Describe any data exclusions.

Samples were included based pre-established criteria on age, gender, ethnicity and medication status.

3. Replication

Describe whether the experimental findings were reliably reproduced.

Our findings were reproduced based on experimental validation (IFN response) and comparison to previously published genetic data (eQTL analysis)

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

No randomization was explicitly performed. Though across batches, individuals were randomly processed.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Yes. The statistical analyses comparing groups were chosen and implemented before the results of the analysis.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

▸ Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

We designed and implemented demuxlet available at: https://github.com/statgen/demuxlet

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

▸ Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

There are no restrictions to the availability of materials.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

Kits for cell type-specific isolation are described in the Methods section on page 20 in the section titled "Bulk isolation and sequencing"
and the section titled "Fluorescence-Activated cell sorting and analysis"

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

No cell lines were used in this study

b. Describe the method of cell line authentication used.

No cell lines were used in this study

c. Report whether the cell lines were tested for mycoplasma contamination.

No cell lines were used in this study

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No cell lines were used in this study

▸ Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

No animal models were used in this study.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

Samples used in this study are patients of Caucasian descent with lupus, rheumatoid arthritis or are age and gender matched controls.