Multiplexing droplet-based single cell RNA-sequencing using natural genetic barcodes

Authors: Hyun Min Kang*$, Meena Subramaniam$, Sasha Targ$, Michelle Nguyen, Lenka Maliskova, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina Lanata, Rachel Gate, Sara Mostafavi, Alexander Marson, Noah Zaitlen, Lindsey A Criswell, Jimmie Ye*

The confluence of microfluidic and sequencing technologies has enabled profiling of the transcriptome[1, 2], epigenome[3], and chromatin conformation of single cells[4] at an unprecedented scale. Initial applications of single cell RNA-sequencing have characterized cellular heterogeneity in tumors[5, 6], tissues[7, 8], and response to stimulation[9]. More recently, droplet-based technologies have significantly increased the throughput of single cell capture and library preparation[1, 10], enabling transcriptome sequencing of thousands of cells from one microfluidic reaction.

While improvements in biochemistry[11, 12] and microfluidics[13, 14] continue to increase the number of cells that can be sequenced per sample, for many applications (e.g. differential expression and genetic studies), sequencing thousands of cells each from many individuals would better capture interindividual variability than sequencing more cells from a few individuals. However, in standard workflows, running a separate microfluidic reaction for each sample remains cost prohibitive[15]. Multiplexing could significantly reduce the per sample cost by allowing cells from several individuals to be processed simultaneously, and reduce the per cell cost by allowing higher flow rates due to the ability to detect and exclude doublets that contain cells from two different individuals. Further, sample multiplexing limits the technical variability associated with sample and library preparation, improving statistical power to accurately estimate true biological effects[16].

We present a simple experimental design and computational algorithm, demuxlet, to multiplex samples in dscRNA-seq without additional experimental modification (**Fig. 1A**). While strategies to demultiplex cells from different species[1, 10, 17] or host and graft samples[17] have been reported, no method is available for simultaneous demultiplexing and doublet detection

of cells from > 2 individuals. Inspired by models and algorithms developed for contamination detection in DNA sequencing data[18], demuxlet is fast, accurate, scalable and works with standard input formats[17, 19, 20].

At the heart of our strategy is a statistical model for predicting the probability of observing a consistent 'genetic barcode', a set of single nucleotide polymorphisms (SNPs), in the RNA-seq reads of a single cell and the genotypes (from SNP genotyping, imputation or DNA sequencing) of donor samples. The model accounts for the base quality score of the RNA-sequencing reads as previously described[18] and genotype uncertainties at unobserved SNPs from imputation to large reference panels[21]. It then uses maximum likelihood to determine the most likely sample identity for each cell using a mixture model. A small number of reads overlapping common SNPs is sufficient to accurately identify the sample of origin. For a pool of 8 samples, 4 SNPs can uniquely assign a cell to the donor of origin (**Fig. 1B**), and 20 SNPs each with minor allele frequency (MAF) of 50% can distinguish every sample with 98% probability.

The mixture model in demuxlet also uses genetic information to identify doublets containing two cells from different individuals, which comprise most droplets containing multiple cells. By multiplexing even a small number of samples, a doublet will have a high probability ($1 - 1/N$, e.g. 87.5% for $N = 8$ samples) of containing cells from two individuals which is detectable by the demuxlet model (**Fig. 1C**). The ability to recover the sample identity of each cell ("demuxing") and identify most doublets enables experimental designs that significantly increase the per sample throughput of current dscRNA-seq workflows.

We first assess the feasibility of our strategy and the performance of demuxlet by analyzing multiplexed peripheral blood mononuclear cells (PBMCs) from 8 patients with

systemic lupus erythematosus (SLE). Using a sequential pooling strategy, 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 Single Cell instrument (**Fig. 2A**). 3,645, 4,254 and 6,205 single cells were obtained from each well and sequenced to an average depth of 23k, 17k and 13k reads per cell.

Demuxlet identified 91% (3332/3645), 91% (3864/4254), and 86% (5348/6205) of droplets as singlets from wells W1, W2 and W3, respectively. 25% (+/- 2.6%), 25% (+/- 4.6%) and 12.5% (+/- 1.4%) of singlets from wells W1, W2 and W3 mapped to each donor, consistent with equal mixing of 8 individuals. We estimate an error rate (number of cells assigned to individuals not in the mixture) of 2/3332 (W1) and 0/3864 (W2) singlets by analyzing wells W1 and W2, each containing cells from two disjoint sets of 4 individuals (**Fig. 2B**), suggesting > 99% of singlets were assigned to individuals correctly.

We next assess the ability of demuxlet to detect doublets in both simulated and real data. 466/3645 (13%) cells were simulated as synthetic doublets by setting the cellular barcodes of two sets of 466 cells from individuals S1 and S2 to be the same. Applied to the 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 (**fig. S1**). Applied to real data from W1, W2 and W3, demuxlet identified 138/3645, 165/4254, and 384/6205 doublets, corresponding to 5.0%, 5.2% and 7.1%, consistent with the linear relationship between the number of cells sequenced and doublet rates estimated using a mixed species experiment (**Fig. 2C**).

Sample demultiplexing enables individual-specific visualization of single cell data we call 'drop prints'. While both variability in cell type proportion and gene expression have been previously observed in PBMCs, it has not been possible to fully control for batch effects due to separate processing of samples[22, 23]. Singlets identified by demuxlet in all three wells cluster into known PBMC subpopulations (**Fig. 2D**) and are not confounded by well to well effects (**fig. S2A**)**.** While we found 6 differentially expressed genes (FDR < 0.05) between wells W1 and W2, only 2 genes were differentially expressed in well W3 between W1 and W2 individuals (FDR < 0.05) (**fig. S2B**) suggesting sample multiplexing could reduce confounding such as library preparation batch effects. Furthermore**,** for the same individuals, drop prints from two different wells are qualitatively consistent, the estimates of cell type proportions for the same individuals in W1 or W2 and W3 are highly correlated (R = 0.99) (**Fig. 2E** and **fig. S3**), and the inferred cell type-specific expression profiles are correlated with bulk sequencing of sorted cell populations (R=0.76-0.92) (**fig. S4**)**.** These results demonstrate that demuxlet recovers the sample identity of single cells with high accuracy, identifies doublets at the expected rate, and can allow for comparison of individuals within and across wells.

Demuxlet enables multiplexed experimental designs that increase the sample throughput for profiling of interindividual responses across a variety of conditions. We applied such a multiplexing strategy to characterize cell type-specific responses to IFN-β, a potent cytokine that induces genome-scale changes in the transcriptional profiles of immune cells[24, 25]. From 8 lupus patients, 1M PBMCs each were isolated, sequentially pooled, and divided in two aliquots. One sample was activated with recombinant IFN-β 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[26, 27]. A matched control sample was also cultured for 6 hours. From this experiment, we captured and sequenced 14,619 control and 14,446 stimulated cells.

In control and stimulated experiments, demuxlet identified 83% (12138/14619) and 84% (12167/14446) of droplets as singlets, and recovered the sample identity of 99% (12127/12138 and 12155/12167) of singlets. Detected doublets form distinct clusters in t-SNE space at the periphery of other cell types, indicative of the expected enrichment of doublets for mixed cell types in a heterogeneous population (**fig. S5**). The estimated doublet rate of 10.9% is consistent with predicted rates based on the number of cells recovered, and the observed proportion of doublets from each pair of individuals is highly correlated with the expected proportions (R=0.98) (**Fig. 2C** and **fig. S6**).

Demultiplexing individuals enables the use of the 8 samples within a pool as biological replicates to quantitatively assess cell type-specific responses to IFN-β stimulation. Consistent with previous reports from bulk RNA-sequencing data, IFN-β stimulation induces widespread transcriptomic changes observed as a shift in the t-SNE projections of singlets (**Fig. 3A**)[24]. After assigning each singlet to a reference cell population[17], we identified 2,686 differentially expressed genes (logFC > 2, FDR < 0.05) in at least one cell type in response to IFN-β stimulation (**table S1**). These genes cluster into modules of cell type-specific responses enriched for distinct gene regulatory processes (**Fig. 3B, table S2**). For example, the two clusters of upregulated genes, pan-leukocyte (Cluster III: 401 genes, logFC > 2, FDR < 0.05) and $CD14^{+}$ specific (Cluster I: 767 genes, logFC > 2, FDR < 0.05), were enriched for general antiviral response (e.g. KEGG Influenza A: Cluster III $P < 1.6x10^{-5}$), chemokine signaling (Cluster I $P < 7.6x10^{-3}$) and genes implicated in SLE (Cluster I $P < 4.4x10^{-3}$). The five clusters of downregulated genes were

enriched for antibacterial response (KEGG Legionellosis: Cluster II monocyte down $P < 5.5x10^{-3}$) and natural killer cell mediated toxicity (Cluster IV NK/Th cell down: $P < 3.6x10^{-2}$). The differential expression using cell type-specific estimates from single cell data recovers known gene regulatory programs affected by interferon stimulation.

We next characterize interindividual variability in PBMC expression at baseline and in response to IFN-β stimulation. In both control and stimulated cells, the variance of mean expression among individuals is substantially higher than expected from synthetic replicates (**Fig. 3C**). As previously reported[22, 28], cell type proportion varied significantly among individuals and contributes to variability in gene expression **(fig. S7)**. The variance estimated from synthetic replicates with matched cell type proportions is more concordant with the observed variance (Lin's concordance = 0.54 versus 0.022, Pearson correlation = 0.78 versus 0.69, **Fig. 3C**-**D**). However, comparing mean expression from synthetic replicates within cell types (Lin's concordance = 0.007 - 0.20, Pearson correlation = 0.27 – 0.68) shows that there is interindividual variability not explained simply by cell type proportion (**fig. S8**).

We then explored interindividual variability in expression within one cell type, $CD14^{+}CD16^{-}$ monocytes. The correlation of mean expression between pairs of synthetic replicates from the same individual (>99%) was greater than between different individuals (~97%), indicating variation beyond sampling (**Fig. 3E**). We found 585 genes that have significant interindividual variability in stimulated $CD14^{+}$ $CD16^{-}$ monocytes and 827 in control by correlating the synthetic replicates across individuals (Pearson correlation, FDR < 0.05). The variable genes in stimulated $CD14^{+}CD16^{-}$ monocytes and to a lesser extent in $CD4^{+}$ T cells ($P < 9.3x10^{-4}$ and $4.5x10^{-2}$, hypergeometric test, **Fig. 3F**) are enriched for differentially expressed

genes, consistent with our previous discovery of more IFN-β response-eQTLs in monocyte-derived dendritic cells than $CD4^+$ T cells[26, 27]. We hypothesize that natural genetic variation could explain interindividual variability in gene expression in our multiplexed data. For example, schlafen family member 5 (SLFN5) and guanylate binding protein 3 (GBP3) expression are highly correlated between replicates after IFN-β stimulation (R=0.92, $P < 0.0011$ and 0.80, $P < 0.017$). The average expression of the two synthetic replicates are associated with known eQTLs in $CD14^+$ monocytes and lymphoblastoid cell lines, respectively (SLFN5: rs11080327 $P < 3.1 \times 10^{-4}$, GBP3: rs10493821 $P < 2.1 \times 10^{-2}$, **Fig. 3G**)[26, 29]. These results suggest that single cell sequencing recovers repeatable interindividual variation in gene expression and in two genes, is associated with known genetic determinants.

We introduce demuxlet, a new computational method that enables simple and efficient sample multiplexing for dscRNA-seq, validate its performance in simulated and real data, and characterize single cell expression of PBMCs from SLE patients in several different conditions. Our results demonstrate demuxlet provides reliable estimation of cell type proportion across individuals, recovers cell type-specific transcriptional programs from mixed populations consistent with previous reports, and identifies genes with interindividual variability[24]. The capability to demultiplex and identify doublets using natural genetic variation significantly reduces the per-sample and per-cell cost of single-cell RNA-sequencing, does not require synthetic barcodes or split-pool strategies[30-34], and captures biological variability among individual samples while limiting the effects of unwanted technical variability.

The application of single cell sequencing methods such as dscRNA-seq to larger numbers of individuals is a promising approach to characterizing cellular heterogeneity among

154 individuals at baseline and in different environmental conditions, a crucial area for further
155 understanding of health and disease[35-37]. Experimental and computational methods for reliable
156 and efficient sample multiplexing could enable broad adoption of droplet-based RNA-seq for
157 population-scale studies, facilitating genetic and longitudinal analyses in relevant cell types and
158 conditions across a range of sampled individuals[38].

Fig. 1

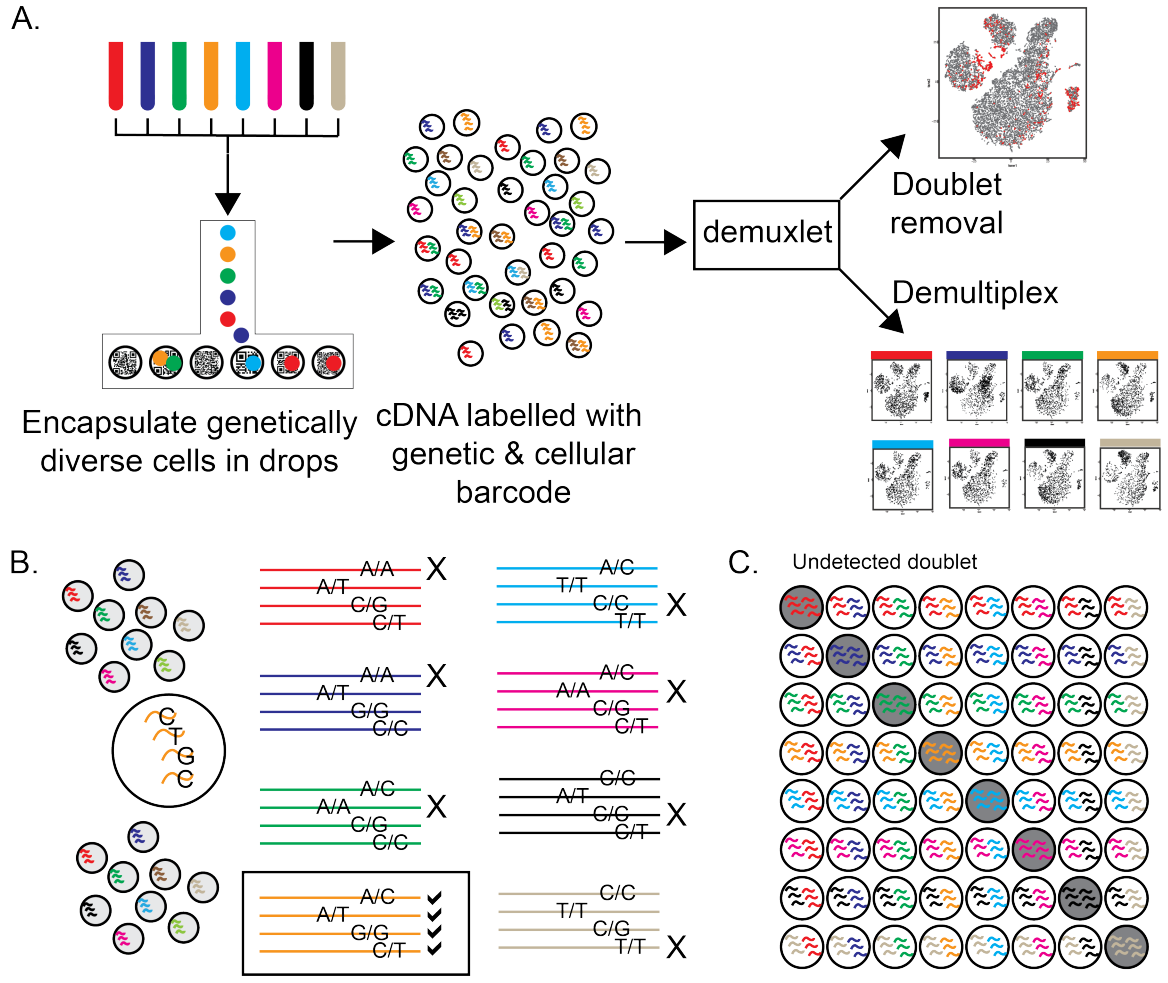

Figure 1 – Demuxlet: demultiplexing and doublet identification from single cell data.

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. 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

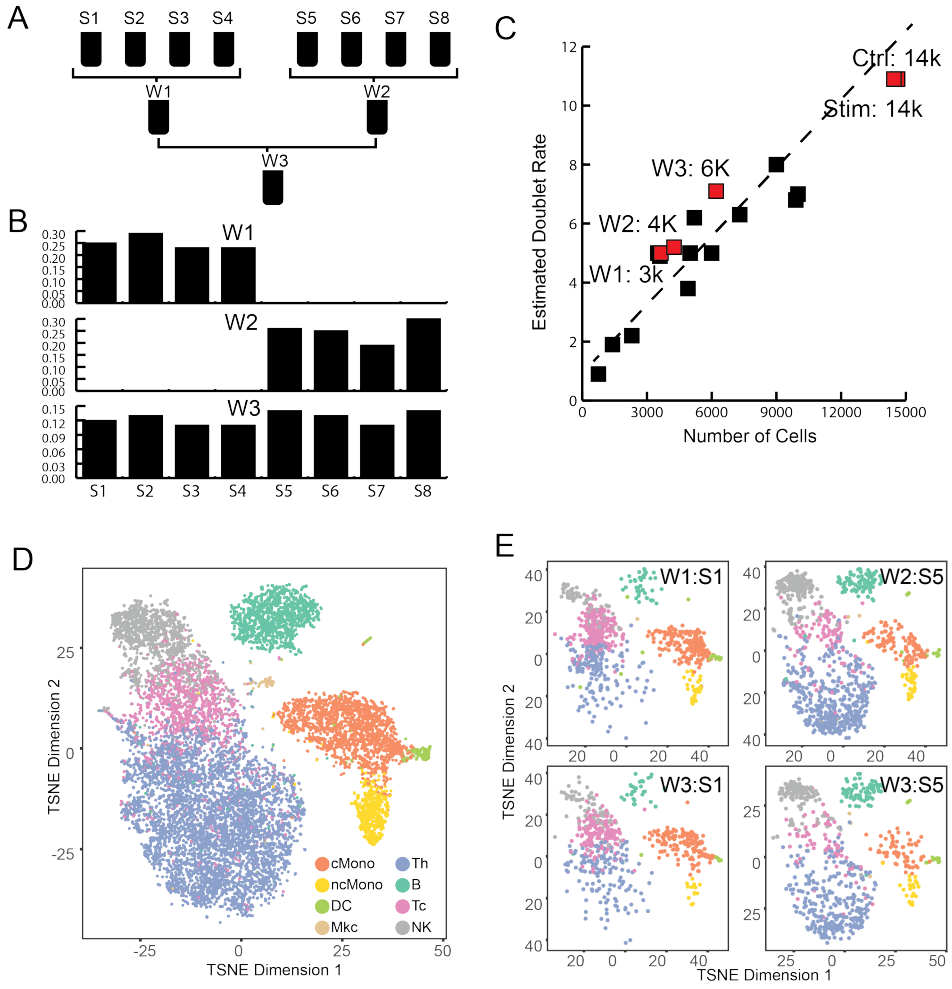

Figure 2 – Performance of demuxlet. A) Experimental design for equimolar pooling of cells from 8 unrelated samples (S1-S8) into three wells (W1-W3). W1 and W2 contain cells from two disjoint sets of 4 individuals. W3 contains cells from all 8 individuals. B) Demultiplexing single cells in each well recovers 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

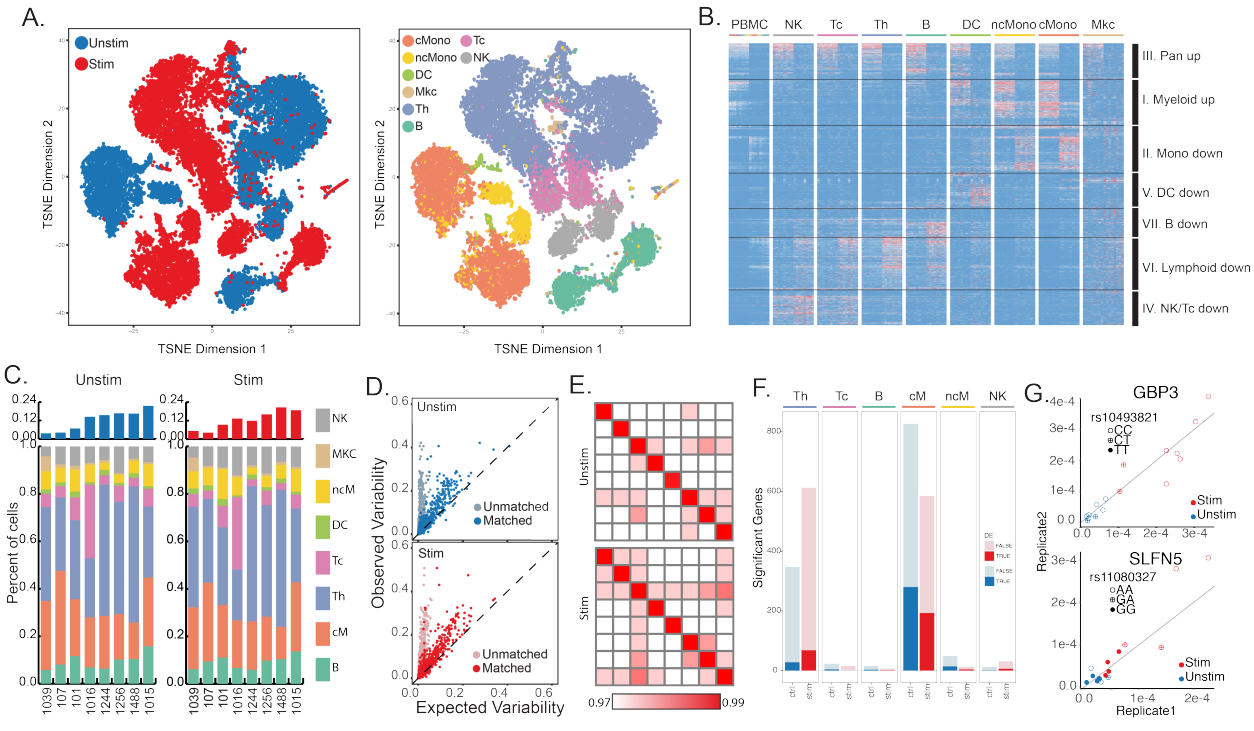

Figure 3 – Interindividual variability in IFN-β response. A) t-SNE plot of unstimulated (blue) and IFN-β-stimulated (red) PBMCs and the estimated cell types. B) Cell type-specific expression in stimulated (left) and unstimulated (right) cells. Differentially expressed genes shown (FDR < 0.05, |log(FC)| > 1). Each column represents cell type-specific expression for each individual from demuxlet. C) Cell type proportions for each individual in unstimulated and stimulated cells. D) Observed variance (y-axis) in mean expression over all PBMCs from each individual versus expected variance (x-axis) over synthetic replicates sampled across all cells (light blue, pink) or replicates matched for cell type proportion (blue, red). E) Correlation between sample replicates in control and stimulated cells. F) Number of significantly variable genes in each cell type and condition. G) Mean expression of SLFN5 and GPB3 in two sample replicates labeled by genotype.

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 $b_{cvi} \in \{R, A, O\},\ i \in \{1, \cdots, d_{cv}\}$ 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 $e_{cvi} = 0, b_{cvi} \in \{R, A\}$ 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 table S3. 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 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 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 ε in the posterior probability, we replace it to $(1-\varepsilon)P_{sv}^{(g)} + \varepsilon\Pr(g)$. 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 $\mathrm{argmax}_s[L_c(s)]$.

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[18],

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

To reduce the computational cost, we consider discrete values of $\alpha \in \{\alpha_1, \cdots, \alpha_M\}$, (e.g. 5 - 50% by 5%). We determine that it is a doublet between samples s_1, s_2 if and only if $\frac{\max_{s_1,s_2,\alpha} L_c(s_1,s_2,\alpha)}{\max_s L_c(s)} \geq t$ and the most likely mixing proportion is estimated to be $\mathrm{argmax}_\alpha L_c(s_1, s_2, \alpha)$. We determine that the cell contains only a single individual *s* if $\frac{\max_{s_1,s_2,\alpha} L_c(s_1,s_2,\alpha)}{\max_s L_c(s)} \leq \frac{1}{t}$. The less confident droplets, we classify cells as ambiguous. While we consider only doublets for estimating doublet rates, we remove all doublets and ambiguous droplets to conservatively estimate singlets. Figure S1 illustrates the distribution of singlet, doublet likelihoods and the decision boundaries when t = 2 was used.

Isolation and preparation of PBMC samples.

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. 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-β 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-β (PBL Assay Science) for 6 hours according to the published protocol[26]. The second aliquot was left untreated. After 6 hours, the 8 samples for each condition were pooled together in two final pools (stimulated cells and control cells) as described above.

Droplet-based capture and sequencing.

Cellular suspensions were loaded onto the 10x Chromium instrument (10x Genomics) and sequenced as described in Zheng et al[17]. The cDNA libraries were sequenced using a custom program on 10 lanes of Illumina HiSeq 2500 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 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[39, 40]. 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 HiSeq 4000 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[17]. 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[17, 41]. 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[42]. Genes with baseMean > 1 were filtered out from the DESeq2 output, and the qvalue package was used to calculate FDR < 0.05 [43].

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 log2 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.

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 and QTL detection was performed on the log2 normalized counts. Pearson correlation of the two replicates from each of the 8 individuals was used to find genes with significant interindividual variability.

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/hyunminkang/apigenome.

1. Macosko, E.Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **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. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, **523** 486-490 (2015).
4. Nagano, T. et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, **502** 59-64 (2013).
5. Patel, A.P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, **344** 1396-1401 (2014).
6. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, **352** 189-196 (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. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510,** 363–369 (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. *Proc Natl Acad Sci* **111**, 7048-7053 (2014).
14. Zilionis, R. et al. Single-cell barcoding and sequencing using droplet microfluidics. *Nat Protoc* **12**, 44-73 (2017).
15. Ziegenhain, C. et al. Comparative analysis of single-cell RNA sequencing methods. *Mol Cell* **65**, 631-643.e634 (2017).
16. Hicks, S.C., Teng, M. & Irizarry, R.A. On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. bioRxiv 025528 (2015).
17. Zheng, G.X.Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8,** 14049 (2017).
18. 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).
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. 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).
22. Aguirre-Gamboa, R. et al. Differential effects of environmental and genetic factors on T and B cell immune traits. *Cell Rep* **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. How cells respond to interferon. *Annu Rev Biochem* **67,** 227-264 (2003).
26. Lee, M.N. et al. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* **343,** 1246980-1246980 (2014).
27. Ye, C.J. et al. Intersection of population variation and autoimmunity genetics in human T cell activation. *Science* **345,** 1254665-1254665 (2014).
28. 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).
29. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506-511 (2013).
30. Cao, J. et al. Comprehensive single cell transcriptional profiling of a multicellular organism by combinatorial indexing. bioRxiv 104844 (2017).

31. 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).
32. 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).
33. Jaitin, D.A. et al. Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell* **167**, 1883-1896.e1815 (2016).
34. Datlinger, P. et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nat Meth* **14**, 297-301 (2017).
35. 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).
36. Tung, P.-Y. et al. Batch effects and the effective design of single-cell gene expression studies. *Sci Rep* **7**, 39921 (2017).
37. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331-338 (2017).
38. Wills, Q.F. et al. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat Biotech* **31**, 748-752 (2013).
39. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Meth* **10**, 1096-1098 (2013).
40. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* **9**, 171-181 (2014).
41. 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).
42. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106 (2010).
43. Dabney, A., Storey, J.D. & Warnes, G.R. qvalue: Q-value estimation for false discovery rate control. *R package version* **1** (2010).