Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference

Yuanhua Huang[1,2,*], Davis J. McCarthy[2,3,4,*], Oliver Stegle[2,5,6,*]

[1]Department of Clinical Neurosciences, University of Cambridge, CB2 0QQ, Cambridge, UK; [2]EMBL-European Bioinformatics Institute, Wellcome Genome Campus, CB10 1SD, Hinxton, Cambridge, UK; [3]St Vincent's Institute of Medical Research, Fitzroy, Victoria 3065, Australia; [4]Melbourne Integrative Genomics, University of Melbourne, Parkville, Victoria 3010, Australia; [5]European Molecular Biology Laboratory, Genome Biology Unit, 69117 Heidelberg, Germany; [6]Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), 69120, Heidelberg, Germany.
*Corresponding authors.

Abstract

Common genetic variants have been used as natural barcodes to demultiplex cells from pooled RNA-seq experiments. Existing demultiplexing strategies rely on access to complete genotype data from the set of pooled samples, which greatly limits the applicability of such methods, in particular when genetic variation is not the primary object of study. To address this, we here present Vireo, a computationally efficient Bayesian model to demultiplex data from pooled experimental designs. Our model can be applied to dataset with partial or without any genotype information of the pooled samples. Using simulations and results on real data, we demonstrate the robustness of our model and illustrate the utility of multi-sample experimental designs for common expression analyses.

1 Introduction

Single-cell RNA-seq (scRNA-seq) is a rapidly evolving technology. Robust protocols and reduced costs mean that scRNA-seq is increasingly applicable in biomedicine, for example to identify biomarkers in disease [1, 2], or to characterize the cellular response to treatment and other external stimuli [3, 4].

Across these use cases, pooled experimental designs that combine multiple samples in a single experiment have critical statistical advantages compared to the serial analysis of cells from multiple individuals in independent experiments [5, 6], and in particular allow to differentiate inter individual variation from batch. Pooled designs that jointly process a large number of cells are facilitated by the uptake of droplet sequencing methods, including Drop-seq [7] and the 10x Genomics chromium system [8], which allow for assaying tens of thousands of cells in a single run.

The aforementioned advantages have motivated a series of barcoding strategies to demultiplex samples from pooled experiments. In addition to reduced experimental logistics and reduced batch variation, pooled designs can also facilitate the identification of double cells. Existing barcoding strategies include molecular labelling prior to analysis [9, 10, 11, 12] as well as exploiting natural genetic barcodes of germline variants that segregate between pooled individuals [13]. The latter approach is both elegant and has the advantage that existing experimental workflows can be retained.

However, current methods to demultiplex such genetically barcoded pools, such as Demuxlet [13], require genotype data for all pooled samples. Each cell is then assigned to one of

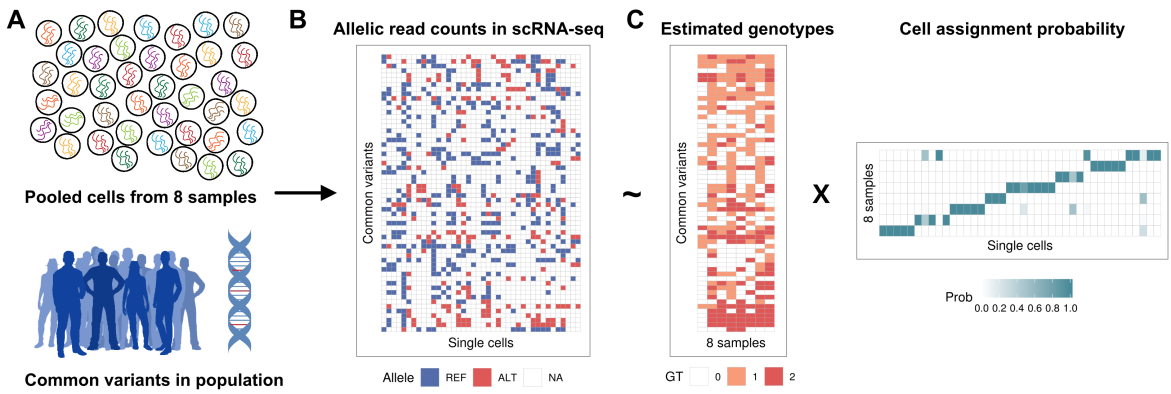

Figure 1: Illustration of Vireo for demultiplexing multi-individual scRNA-seq studies without reference genotyping data. (A, B) The inference is based on genotyped common polymorphic variants in each individual cell, using a reference set of common variants. (B, C) The resulting sparse read count matrices of alternative alleles and reference alleles (displayed as compound matrix for simplicity; NA in white means no observed reads) are then decomposed into a genotype probability matrix for each input sample and a cell assignment probability matrix.

the individuals based on their genetic distance to variants extracted from the scRNA-seq reads. While there is a growing interest in multi-sample analyses that seek to explicitly study the effect of genetic variation between individuals on single-cell states [14, 15], the need of genotype data is prohibitive for studies without a genetic focus per se. Consequently, the potential of pooled experimental designs is currently not fully realized.

To address this, we here present Vireo (Variational Inference for Reconstructing Ensemble Origins), a principled Bayesian method to demultiplex arbitrary pooled designs that combine genetically distinct individuals. Uniquely, Vireo models the genotypes of each individual as latent variables, which are inferred from the observed scRNA-seq reads. The model can also utilize partial genotype information, e.g. when genotype data are available for a subset of individuals, and hence can be applied to a wide range of experimental settings. Vireo is implemented as part of the cardelino R package available under an open-source license at `https://github.com/PMBio/cardelino`.

2 Results and discussion

Vireo jointly assigns each cell to one of K individuals and estimates the genotypic state of these individuals at known polymorphic loci. The model considers a set of common genetic variants (for example derived from 1000 Genomes Project [16]), which are genotyped based on scRNA-seq reads observed in each cell. Even in low coverage single-cell data, this approach typically allows detection of on the order of 100 variants per cell (e.g. using 3' 10x Genomics data; around 50,000 reads per cell, Fig 1 and Methods). This information is sufficient for the partial genotype reconstruction of expressed variants and cell assignment (Fig. 1). Vireo also explicitly accounts for doublets by considering cells that are assigned to a combination of individuals in the original sample pool. Finally, the model can automatically detect the most likely number of individuals in a pool, a feature that is useful if some of the pooled samples drop out for experimental reasons, and the method supports partial genotype data for a subset of the pooled samples.

2.1 Model validation using simulated data

Initially, we used simulated data with known ground truth to validate our model. We simulated a multiplexed dataset consisting of 10,000 cells from 8 pooled samples based on genotypes from individuals of European ancestry that are part of the 1000 Genomes project [16]. We simulated cells with a doublet rate of 10% and a coverage of 54,000 reads per cell on average (Methods). Initially, we evaluated Vireo's ability to estimate the true number of input samples, by comparing the marginal likelihood of multiple Vireo runs assuming increasing numbers of samples in the pool, ranging from four to twelve. Notably, the model with the true sample number ($K = 8$) was identified as the most probable given the data (Fig. 2A). We also observed that models that assume larger pool sizes ($K > 8$) tended to yield sparse solutions, which means that only the relevant subset of latent samples for cell assignment were used, indicating that the model is robust to this parameter choices (Fig. 2B).

Next, we evaluated Vireo's performance for assigning cells to their respective sample of origin. When ranking cells by the assignment confidence (defined as the most likely sample assignment, prob_max), we observed that the majority of cells (i.e., not doublets) were correctly assigned to the sample of origin (Fig. 2C), with an area under the precision-recall curve (AUPRC) of >0.99. Similar to existing methods that depend on genotype reference data [13], Vireo can also be used to identify doublet cells, provided that the doublets are formed of combinations of cells from two genetically distinct samples in the pool. Although the identification of doublets is in general more challenging than the assignment of singlets (AUPRC=0.92), Vireo retained an overall high detection rate (e.g. 87.6% sensitivity, 99.9% specificity, 98.6% precision at prob_doublet > 0.9, Fig. 2C and Supp. Fig. S1). In practice and in the experiments below, we use prob_max >0.9 as a threshold for the singlets assignment, and prob_doublet >0.9 for the detection of doublets (see Methods).

Exploring a wider range of settings, we also evaluated the model when varying the number of multiplexed samples (Fig. 2D) and the number of cells sampled in each experiment (Fig. 2E). As expected, the cell-assignment accuracy decreased with increasing numbers of samples in the pool, but Vireo retained high accuracy for up to 16 multiplexed samples (Fig. 2D and Supp. Fig. S2). Conversely, the cell-assignment accuracy and, in particular the accuracy of doublet detection, increased with the cell count in the experiment (Fig. 2E), as well as for increased sequencing depth per cell (Supp. Fig. S3). Next, we assessed the utility of partial genotype information for a subset of samples in the pool, which as expected increased the model performance (Fig. 2F), in particular in settings with low sequencing coverage (Supp. Fig. S4). We also evaluated the robustness of Vireo when applying the model to biased pools of samples, i.e. settings in which a subset of samples contribute a smaller than expected fraction of cells. Vireo robustly detected and aligned cells to samples with a relative frequency as low as 5% (Supp. Fig. S5), while retaining high accuracy for doublet detection. Finally, we assessed the accuracy of the genotype reconstruction for the pooled samples, finding that Vireo implicitly provides accurate genotypes information for genetic variants that are covered by scRNA-seq, especially for variants covered by more than 10 reads per sample (precision > 0.99; Supp. Fig. S6). Although the genotype estimates are not available genome-wide, these data can be used as a linking key to align the reconstructed samples to other 'omics data or to combine demultiplexed datasets across experiments (Methods).

2.2 Application to pooled real data

Next, we applied Vireo to two real datasets that have previously been considered to benchmark methods for demultiplexing pooled experiments when genotype information is available for all samples [13].

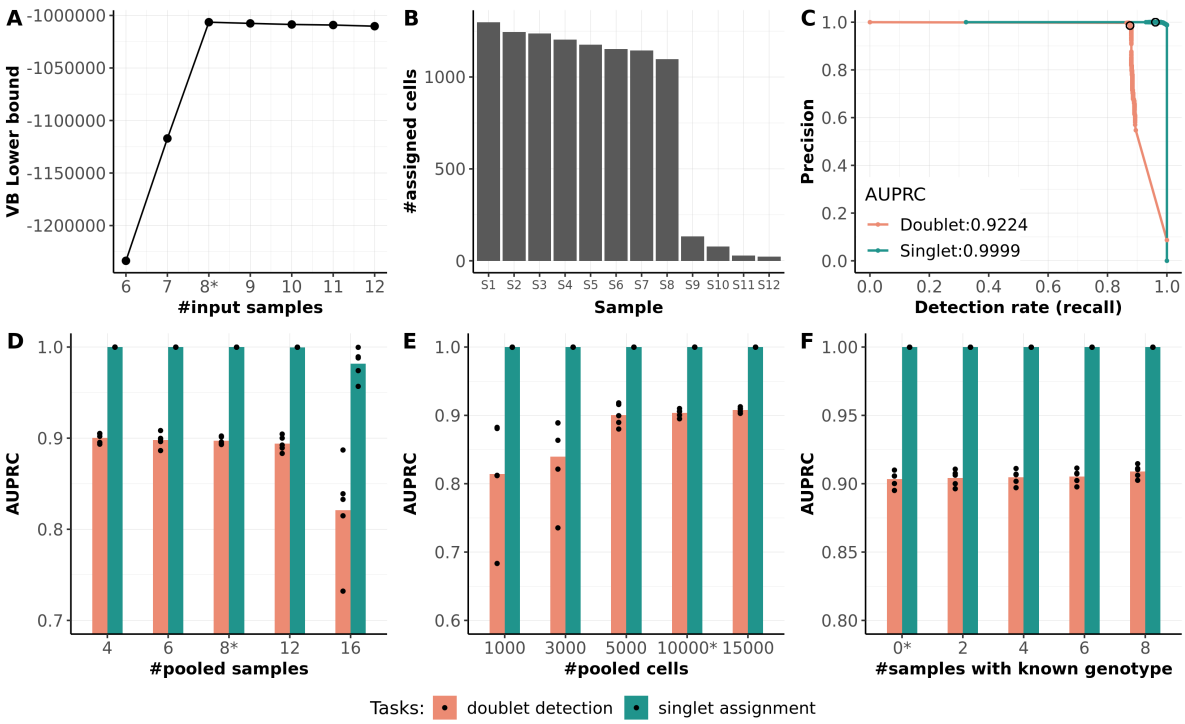

Figure 2: Evaluation of Vireo using simulated data. (A-C) Assessment of Vireo performance on one simulated dataset consisting of 10,000 cells drawn from a pool of 8 samples with 10% doublet rate and 54,000 reads per cell on average. (A) Vireo model evidence (variational lower bound) when varying the pool size assumed in the model. (B) The distribution of number of cells assigned to individual samples when running Vireo assuming a too large pool of size N=12. (C) Precision-recall curve for the assignment of singlet-cells and doublet detection, considering varying assignment thresholds when applying Vireo with the correct pool size of 8 input samples. The 90% thresholds values for singlet (prob_max $>$ 0.9) and doublet (prob_doublet $>$ 0.9) assignment are highlighted with black circles. (D-F) Area under precision-recall curve (AUPRC) for singlet assignment and doublet detection for alternative simulation settings, either varying number of input samples (D), the number of total cells in the dataset (doublet rate varied proportional; 1% to 15%; Methods) (E), as well as the number of samples with known (observed) genotype (F). Parameters not varied in either of these experiments were set to their default values (indicated by the star symbol).

First, we considered a set of three multiplexed experiments (Fig. 3A-C; W1-W3, between 3,639 and 6,145 cells) of peripheral blood mononuclear cells (PBMCs) from eight lupus patients. We applied Vireo without using genotype information to all cells from these three batches (Methods), thereby also creating an implicit link between the samples across all the three experiments. The Vireo cell assignments were markedly consistent with the assignments obtained when using Demuxlet, which however depends on genotype data for all samples (Fig. 3A). Similarly, we observed overall concordant doublet cell assignments, although there were larger differences than for singlet assignments (Fig. 3B). We also applied Vireo separately to each of the three datasets and used the inferred genotype state of the samples to link the samples identify across experiments retrospectively (Fig. 3C). This shows how the inferred genotype data can be used or integrating demultiplexed samples across experiments, or to link them to other (sequencing-based) assays for the same samples (Methods).

As a second use case, we considered two experiments of PBMCs from the same eight patients: one batch with IFN-β stimulation and a matched control experiment without stimulus. Cells

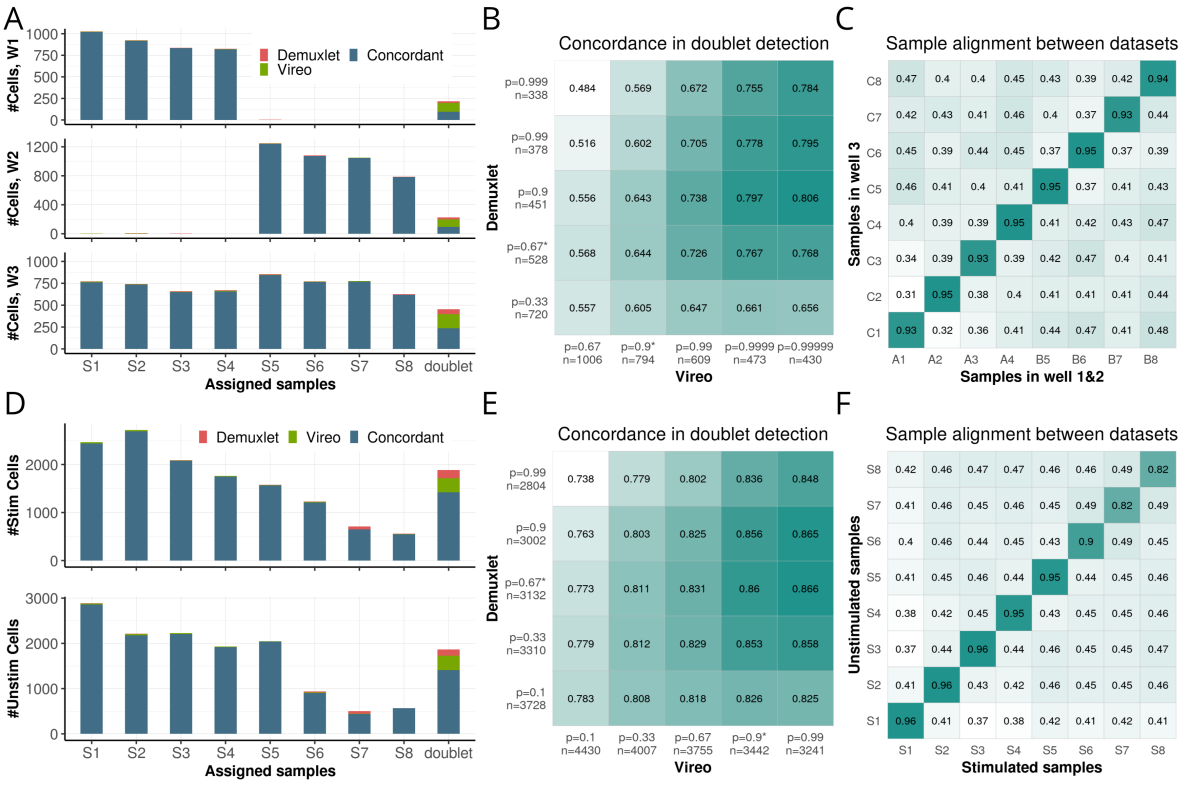

Figure 3: Evaluation of Vireo on data from multiplexed human PBMCs. (A-C) Results obtained on two datasets consisting of two pools of 4 samples each, as well as a third dataset consisting of the union of all 8 samples. (A) Concordance of singlet assignment and doublet detection between Vireo without genotype data and Demuxlet applied with complete genotype reference. Bars denote the number of cells assigned to each cell, either considering cells that were consistently assigned by Vireo and Demuxlet (blue), or assigned exclusively by Vireo (green) or Demuxlet (red). (B) Concordance of doublet detection between Vireo and Demuxlet, when varying the assignment threshold for each method. Note, p denotes the threshold prob_doublet in Vireo (x axis) and Demuxlet (y axis) respectively, and n denotes the number of identified doublets. Assignment of cells in A is based on the most probable sample assignment, considering all cells that were not detected as doublet. Cells with a doublet probability (p_doublet > 0.9 in Vireo; $> 2/3$ in Demuxlet) were labelled as doublet cell and are considered in B. (C) Alignment of samples, when applying Vireo separately to datasets 1&2 and dataset 3. Values in the heatmap denote the fraction of concordant genotype states between pairs of samples from both Vireo runs, considering variants with a read coverage of at least 10 UMIs per sample. (D-F) Results obtained on a second dataset, consisting of two experiments that consider the same 8 pooled samples in two conditions: unstimulated and stimulated. Results shown correspond to the panels in (A-C).

were cultured for six hours after pooling, which in contrast to the first dataset resulted in an imbalanced distribution of cells across samples (Fig. 3D). Despite this distributional bias, Vireo again yielded demultiplexing results that were markedly consistent with the results obtained by methods that require a genotype reference (Fig. 3D-E), and Vireo enabled aligning samples across both experiments (Fig. 3F).

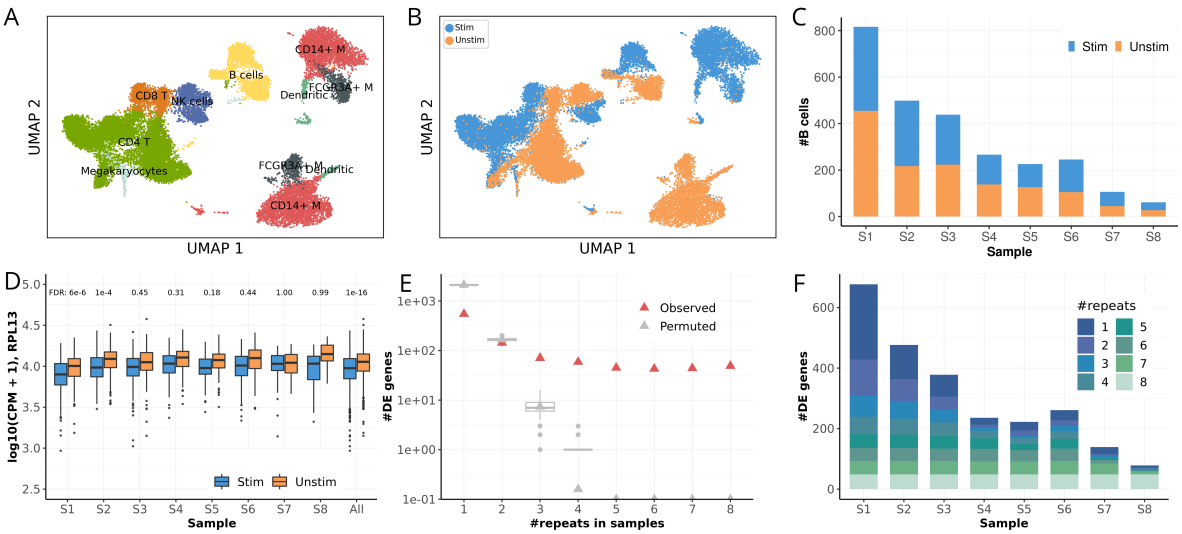

Figure 4: Case study of differential expression analysis of INF-beta stimulated PBMC using two matched pools that consist of 8 samples. (A) UMAP representation of single-cell transcriptome profiles, colored by eight major cell types identified. Note, M denotes Monocytes. (B) Analogous UMAP representation colored by the experimental condition: stimulated versus unstimulated. (C) The number of stimulated and unstimulated B cells sequenced in each of the 8 samples. (D-I) Differential gene expression (DE) analysis between stimulated and unstimulated B cells. (D) The expression level of an example gene RPL13, depicting the distribution of expression levels in both conditions, either considering each sample separately or considering aggregated data. Above each box plot shown is the adjusted P-value (Benjamini-Hochberg) of a DE test between conditions (Likelihood-ratio tests). CPM: count per million. (E) Number of recurrently detected DE genes between conditions (FDR< 0.05), detected in at least one to eight samples. Box plots in grey show recurrence results that reflect the chance expectation (using 200 permutations). (F) The number of DE genes in each of the 8 samples, categorized by the number of recurrent DE discoveries across samples, that is the number of individuals in which a given gene is identified as DE (FDR< 0.05).

2.3 Leveraging multiplexed designs for differential expression analysis

Finally, we considered the de-multiplexed dataset consisting of stimulated and unstimulated cells (Fig. 3D-F) to explore the utility of multi-sample designs for differential gene expression analysis. Graph-based clustering (implemented in Scanpy [17]) applied to the joint dataset consisting of stimulated and unstimulated cells from all eight donors (Fig. 4C) identified eight major clusters, which could be annotated to common cell types (Fig. 4A-B; Supp. Fig. S7). Next, we tested for differential gene expression between the stimulated and unstimulated condition within each cell type (using edgeR, accounting for cells as replicates [18]). Considering B cells as a representative example (see Supp. Fig. S9-12 for full results), this identified between 78 and 477 DE genes in individual donors (FDR<5%; Fig. 4F), with cell count being a major explanatory factor for differences in the number of DE genes (Fig. 4C). Although globally, DE genes tended to be recurrently detected in multiple donors (Fig. 4E), there was a substantial fraction of DE genes that were private to individual donors. For example, the gene RPL13 (Fig. 4D) was differentially expressed in two of eight donors, highlighting the important of inter individual differences. We also explored carrying out joint testing across all donors (using a random effect model implemented in edgeR; Methods), which lead to broadly similar conclusions (Supp. Fig. S8).

3 Conclusion

Here, we have presented Vireo, a Bayesian method for demultiplexing pooled single-cell RNA-seq experiments by exploiting natural genetic barcodes and cell genotyping based on scRNA-seq reads. Uniquely, Vireo does not require the reference genotypes of the input samples to demultiplex an experiment without scarifying accuracy in comparison to methods that require a genotype reference. Vireo is implemented using computationally efficient variational Bayesian inference, which provides a fully Bayesian treatment while retaining scalability to large scRNA-seq datasets.

Using simulated data, we have demonstrated the accuracy of Vireo for demultiplexing pooled samples, and its flexibility in handling partial genotype data for a subset of samples, should these data be available. Unsurprisingly, we observed that the accuracy of the genotype estimation step per sample is primarily linked to the sequencing coverage, which also substantially affects the ability to detect doublet cells. As the exact requirements on the sequence coverage depend on the cell count and the number of pooled samples, we provide a simulation framework that enables the user to explore parameters thereby aiding the experimental design of pooled studies (see tutorials in Vireo's manual: `https://huangyh09.github.io/vireo-manual/`). If cells from the same individuals are assayed in multiple batches, Vireo can also demultiplex them jointly, which boosts the assignment accuracy, especially in experiments with lower read coverage. Furthermore, the estimated genotypes for individual samples enables aligning samples from the scRNA-seq data with from other 'omics data for the same samples (Fig. 3C), which provides a flexible approach to link samples across experiments, including multi-omics treatment-control designs.

We noticed that demultiplexing without a genotype reference starts to deteriorate for large sample pools ($>$ 16 samples). Increased sequencing coverage will allow for demultiplexing even larger pools, but there will be practical limitations. In practice, if only a small number of sample replicates are required, we recommend pooling 6 to 10 samples per experiment. For extremely large pools with more than twelve samples, genotype data is still required to enable accurate demultiplexing.

4 Methods

4.1 Vireo model

Given a list of N common variants, we extract allelic expression of these variants in each of M cells with RNA-seq data (see below for details on the read pileup approach for variant genotyping). Let A and D respectively denote the read or UMI count matrices for the alternative allele (i.e., ALT) and the total read depth (i.e., sum of ALT and REF) for N variants across M cells. Vireo models variation in these counts matrices by employing a clustering model with clusters corresponding to K individuals in the pool, with (unknown) genotype states G. The values of G take on values of 0, 1, or 2, corresponding to homozygous REF, heterozygous and homozygous ALT alleles.

The observed alternative allele counts A are modelled as binomial distributed given the read depths

$$p(a_{i,j}|d_{i,j}, \theta_t) = \texttt{Binom}(a_{i,j}|d_{i,j}, \theta_t), t \in T = \{0, 1, 2\}, \tag{1}$$

where t is the true genotype of variant i in cell j, and θ_t is the binomial rate parameter that encodes the corresponding allele dosage of the alternative allele for genotype t. Theoretically, the allele dosage is $\theta_t = t/2$, whereas in practice we allow for deviations to account for sequencing errors, genotype estimation errors, and allelic imbalance.

The genotype in a given cell is defined by a clustering model where the latent genotype t for variant i in cell j is coded by two indicator variables: the cell assignment vector Z_j, which assigns cell j to a latent sample in the pool, and the genotype identity $G_{i,k}$, which defines the allelic state of variant i in sample k. Specifically, the indicator variable $Z_{j,k} = 1$ if cell j is assigned to sample k and 0 otherwise; we also impose the constraint $\sum_k Z_{j,k} = 1$, which means that in expectation each cell originates from exactly one sample. Analogously, the indicator variable $G_{i,k,t} = 1$ if the genotype of variant i in sample k is t, and 0 otherwise, and we again require $\sum_t G_{i,k,t} = 1$. The cell assignment matrix Z is strictly unknown and needs to be estimated from the observed data. In general, the genotype matrix G is also unknown and is estimated jointly with Z. If genotype information are available for one or multiple samples in the pool, this information can be encoded as informative prior on G; see below.

The likelihood of the full datasets, spanning all N variants that were genotyped in each of M cells given the cell assignment matrix Z, the genotype matrix G and binomial parameter $\boldsymbol{\theta}$ follows as:

$$p(A, D|Z, G, \boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{j=1}^{M} \prod_{k=1}^{K} \prod_{t \in T} p(a_{i,j}|d_{i,j}, \theta_t)^{(Z_{j,k}==1)\times(G_{i,k,t}==1)}. \tag{2}$$

To complete the definition of the model, we introduce prior distributions on the latent variables, which results in the following joint distribution over both observed and latent variables

$$p(A, D, Z, G, \boldsymbol{\theta}) = p(A, D|Z, G, \boldsymbol{\theta})p(Z|\boldsymbol{\pi})p(G|U)p(\boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{\beta}). \tag{3}$$

For computational convenience, we use conjugate prior distributions, namely beta distribution for $\boldsymbol{\theta}$ and multinomial distributions for both Z and G.

$$\begin{aligned} p(Z_{j,k} = 1|\boldsymbol{\pi}) &= \texttt{Multinom}(\boldsymbol{\pi}) = \pi_k \\ p(G_{i,k,t} = 1|U) &= \texttt{Multinom}(\boldsymbol{u}_{i,k}) = u_{i,k,t} \\ p(\theta_t|\alpha_t^{(0)}, \beta_t^{(0)}) &= \texttt{beta}(\theta_t|\alpha_t^{(0)}, \beta_t^{(0)}) \end{aligned} \tag{4}$$

The hyper parameters are constant and set as follows. We use an uninformative prior for Z: $\pi_k = 1/K$, which corresponds to a uniform assignment probability of cells to samples. The user can define other multinomial probabilities, for example to encode known bias in the sample representation. Similarly, we employ a uniform prior on genotype G, i.e., $u_{i,j,t} = 1/3$ if no genotype data are available. If the genotypes are partially known for a subset of samples and/or variants, a corresponding informative prior is encoded. Specifically, $u_{i,j,t}$ takes the known genotype value with a relax rate ξ, i.e., $u_{i,j,t} = 1 - \xi$ if the known genotype is t, otherwise $u_{i,j,t} = \xi$. The error rate parameter is set to $\xi = 0.05$ by default.

Finally, the hyper parameter for the beta prior on the allelic rate $\boldsymbol{\theta}$ is determined using known germline variants with high coverage: $\theta_0 \sim \texttt{beta}(0.3, 29.7)$, $\theta_1 \sim \texttt{beta}(3, 3)$, and $\theta_2 \sim \texttt{beta}(29.7, 0.3)$, with which the posterior of $\boldsymbol{\theta}$ will be obtained by fitting to the dataset.

4.2 Variational Bayesian inference

Analytical calculation of the posterior distribution of all latent variables given the observed data $p(Z, G, \boldsymbol{\theta}|A, D)$ is not tractable. Thus, we consider variational Bayesian inference [19] to obtain an approximate solution, thereby retaining the benefits of a Bayesian treatment while retaining computational scalability to larger scRNA-seq datasets. Briefly, the objective of variational inference is to approximate the exact (intractable) posterior distribution of the latent variables $p(\mathbf{Y}|\mathbf{X})$ by a factorized distribution $q(\mathbf{Y}) = \prod_i q_i(Y_i)$, where $\mathbf{Y}$ denotes a set of latent variables

and $\mathbf{X}$ denotes the observed variables. The parameters of the variational distribution $q(\mathbf{Y})$ are determined with the objective to minimise the Kullback-Leibler ($\mathcal{KL}$) divergence between the approximate distribution $q(\mathbf{Y})$ and the actual posterior distribution $p(\mathbf{Y}|\mathbf{X})$

$$\mathcal{KL}(q(\mathbf{Y})||p(\mathbf{Y}|\mathbf{X})) = -\int q(\mathbf{Y}) \log \frac{q(\mathbf{Y})}{p(\mathbf{Y}|\mathbf{X})} \mathrm{d}\mathbf{Y}. \tag{5}$$

This objective is equivalent to maximizing the lower bound of the full distribution $\mathtt{L}(q)$, as the log marginal probability of the observed variables is a constant, as follows,

$$p(\mathbf{X}) = \mathtt{L}(q) + \mathcal{KL}(q(\mathbf{Y})||p(\mathbf{Y}|\mathbf{X})), \tag{6}$$

where the lower bound $\mathtt{L}(q)$ is defined as follows,

$$\mathtt{L}(q) = \int q(\mathbf{Y}) \log \frac{p(\mathbf{Y}, \mathbf{X})}{q(\mathbf{Y})} \mathrm{d}\mathbf{Y}. \tag{7}$$

A set of iterative update equations can be derived, which are guaranteed to increase the lower bound

$$q_j(Y_j) = \frac{\exp\{\mathbb{E}_{i \neq j} \log(p(\mathbf{Y}, \mathbf{X}))\}}{\int \exp\{\mathbb{E}_{i \neq j} \log(p(\mathbf{Y}, \mathbf{X})\} \mathrm{d}\mathbf{Y}_j}. \tag{8}$$

Here, $\mathbb{E}_{i \neq j}$ denotes an expectation with respect to the distributions $q_i(\mathbf{Y}_i)$ for all $i \neq j$.

For inference in Vireo, we assume a fully factorized distribution $q(Z, G, \boldsymbol{\theta}) = q(Z)q(G)q(\boldsymbol{\theta})$ to approximate the true posterior distribution $p(Z, G, \boldsymbol{\theta}|A, D)$, and we assume that Z and G follow categorical distributions, and $\boldsymbol{\theta}$ follows beta distributions. Based on this assumption, the lower bound can be computed as in Eq(7) (See Supplementary Methods Eq(S1-6)). Following Eq(8), it is possible to derive iterative update equations for the Q distribution of the latent variables (see Supplementary Methods Eq (S7-12) for full details).

$$q^*(Z) = \prod_{j=1}^{M} \prod_{k=1}^{K} r_{j,k}^{Z_{j,k}}; \qquad r_{j,k} = \frac{\pi_k \exp \sum_{i=1}^{N} \sum_{t \in T} \left\{ \tilde{g}_{i,k,t} [a_{i,j} \varphi(\tilde{\alpha}_t) + b_{i,j} \varphi(\tilde{\beta}_t)] \right\}}{\sum_{h=1}^{K} \pi_h \exp \sum_{i=1}^{N} \sum_{t \in T} \left\{ \tilde{g}_{i,h,t} [a_{i,j} \varphi(\tilde{\alpha}_t) + b_{i,j} \varphi(\tilde{\beta}_t)] \right\}} \tag{9}$$

$$q^*(G) = \prod_{i=1}^{N} \prod_{k=1}^{K} \prod_{t \in T} g_{i,k,t}^{G_{i,k,t}}; \qquad g_{i,k,t} = \frac{u_{i,k,t} \exp \sum_{j=1}^{M} \left\{ \tilde{r}_{j,k} [a_{i,j} \varphi(\tilde{\alpha}_t) + b_{i,j} \varphi(\tilde{\beta}_t)] \right\}}{\sum_{h \in T} u_{i,k,h} \exp \sum_{j=1}^{M} \left\{ \tilde{r}_{j,k} [a_{i,j} \varphi(\tilde{\alpha}_h) + b_{i,j} \varphi(\tilde{\beta}_h)] \right\}} \tag{10}$$

$$q^*(\boldsymbol{\theta}) = \prod_{t \in T} \mathtt{beta}(\theta_t | \alpha_t, \beta_t); \alpha_t = \alpha_t^{(0)} + \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{K} \tilde{r}_{j,k} \tilde{g}_{i,k,t} a_{i,j}, \beta_t = \beta_t^{(0)} + \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{K} \tilde{r}_{j,k} \tilde{g}_{i,k,t} b_{i,j} \tag{11}$$

Here, we introduce $b_{i,j} = d_{i,j} - a_{i,j}$ to simplify the notation and $\varphi(\cdot)$ denotes the digamma function. To mitigate potential local optima, multiple random restarts are considered (default: 4 restarts) and the solution that maximizes the variational lower bound is selected. Thanks to the implementation with sparse matrix and support of multiple threads, Vireo model is computationally efficient. By using four threads on a laptop with 16G memory and two 3.5GHz CPUs, Vireo finishes a two-run estimation (see next section) in 2.7 minutes for 14,619 cells in an eight sample pool and 30.3 seconds for 6,145 cells in another eight sample pool (results in Fig. 3).

4.3 Estimation of the number of pooled samples

Access to the variational lower bound (Eq. 7) allows for estimating the number of samples in a given pool. Briefly, by comparing alternative Vireo runs with increasing numbers of samples

it is possible to identify the most probable value (e.g., Fig. 2A), which provides an objective means to define this parameter.

A second strategy is to set a large number of samples and prune some of the samples posthoc, as variational Bayes model is self-regularizing and hence avoids over-fitting (see Fig. 2B). In practice this approach can also increase robustness as the effective number of samples in a pool can be larger than anticipated due to doublets; see Section below.

4.4 Doublet detection

Vireo use a two-step approach to detect doublet cells. Initially, the model is fit using a large number of samples than present in the pool, where by default we assume 1.5 the number of true samples in the pooled (or estimated number; see above). This over-parametrisation provides additional degrees of freedom to fit both true sample genotype states and combinations that reflect doublets. This fitted solution is pruned after training, thereby resulting in genotype inferences for K samples. To detect doublets we then construct the genotype of each pair of samples and expand the K biological samples by introducing in additional $(K-1)*K/2$ doublet competitors. For simplicity, we assume that the genotype of a doublet sample can be described as the average between two combined samples. Specifically, for a given variant and genotype probability vectors for two samples are $\boldsymbol{x} = [x_0, x_1, x_2]$ and $\boldsymbol{y} = [y_0, y_1, y_2]$, we define the expected genotype for the doublet sample as follows,

$$\begin{aligned} p(t=0) &= x_0 y_0 \\ p(t=1) &= x_1 y_1 + x_0 y_2 + x_2 y_0 \\ p(t=2) &= x_2 y_2 \\ p(t=0.5) &= x_0 y_1 + x_1 y_0 \\ p(t=1.5) &= x_1 y_2 + x_2 y_1 \end{aligned} \tag{12}$$

where we introduce two pseudo-genotypes $t = 0.5$ and $t = 1.5$ respectively for combinations of genotype 0 & 1 and 1 & 2 in the doublet sample. For convenience, we consider the binomial parameters for the alternative allelic reads and assume that the binomial parameters $\theta_{0.5}$ and $\theta_{1.5}$ also follow beta distributions. We approximate the hyper-parameters of the beta distribution empirically by respectively taking the ratio and shapes with the arithmetic and geometric means from the two ordinary genotypes. The resulting distribution of $\theta_{0.5}$ can be expressed as follows

$$\begin{aligned} p(\theta_{0.5}) &= \texttt{beta}(\alpha_{0.5}, \beta_{0.5}) \\ \frac{\alpha_{0.5}}{\alpha_{0.5}+\beta_{0.5}} &= \frac{1}{2}\left(\frac{\alpha_0}{\alpha_0+\beta_0} + \frac{\alpha_1}{\alpha_1+\beta_1}\right) \\ (\alpha_{0.5}+\beta_{0.5})^2 &= (\alpha_0+\beta_0) \times (\alpha_1+\beta_1). \end{aligned} \tag{13}$$

Similarly, we define the distribution of $\theta_{1.5}$.

In this augmented model, we have the full distribution for the extended genotype reference G and $\boldsymbol{\theta}$, consisting of K biological and $(K-1)*K/2$ doublet samples. In this model we can calculate the probability that a cell originates from one of the doublet samples using Eq(8).

As an additional refinement, we specify a non-uniform prior on η to define the *a priori* believe of observing a doublet. Specifically, the prior binomial distribution $\boldsymbol{\pi}$ is constructed as follows

$$p(\pi_h) = \begin{cases} (1-\eta)/K, & 1 \le h \le K \quad \text{(ordinary sample)} \\ \eta/((K-1)*K/2), & K < h \le K + (K-1)*K/2 \quad \text{(doublet sample)} \end{cases} \tag{14}$$

The prior probability for doublet cells is low in most assays, e.g., $\eta = 0.05$. In case of the 10X chromium platform, the prior value can be estimated as a function of the number of loaded cells M, e.g. $\eta = M/100000$ following [13]. In the experiments we use a uniform prior for doublets, which corresponds to $p(\pi_h) = 1/(K + (K-1) * K/2)$.

4.5 Alignment of samples between multiple data sets

Vireo implicitly estimates the genotypes for subset of variants with sufficient coverage (good accuracy for variants with >10 reads per sample; see Supp. Fig. S6). Among other use cases, these estimated genotypes allow for aligning scRNA-seq profiles from samples in a pool to other 'omics data by matching genotype profiles.

When Vireo is applied to multiple pools that consist of the same samples, this also allows for aligning samples across Vireo runs (e.g. Fic. 3C and 3F). The software implementation of Vireo provides support functions for this step, which also matches the set of variants that are accessible in two or multiple experiments.

4.6 Data processing and scRNA-seq genotyping

In this study, we considered two existing multiplexed scRNA-seq datasets that consist of a total of five batches [13]. Raw .bam files were obtained from the Gene Expression Omnibus (GEO; accession number GSE96583). The processed results from Demuxlet for these five batches were directly downloaded from `https://github.com/yelabucsf/demuxlet_paper_code`. Approximately 37 million common variants (allele frequency > 0.0005) extracted from the 1000 Genome Project, phase 3 [16] were used as candidate variants for scRNA-seq genotyping. We provide an companion Python package `cellSNP` (`https://github.com/huangyh09/cellSNPl`) for this task, which enables generating selected pile-ups from scRNA-seq data. We discarded non-bi-allelic variants as well as variants with fewer than 20 total UMIs across all cells or minor (i.e., second) allele has less than 10% of total UMIs. The final output of `cellSNP` are two variants-by-cells matrices, A and D, for UMI counts of alternative allele and the total counts respectively, which are used as input for the Vireo model.

4.7 Simulation

Allele counts were A were drawn from the generative model or Vireo given the total read depth D as well as the genotypes G, for a fixed of samples and binomial sampling parameter $\boldsymbol{\theta}$. To mimic a realistic sequencing coverage distribution, we considered the expression profile of a 10x genomics experiment (10k PBMCs from a Healthy Donor (v3 chemistry), `https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/pbmc_10k_v3`) as a seed D_{seed}, which covers 15,427 variants across 11,769 cells with average of 54,000 reads per cell (Supp. Fig. S13). Cells were simulated as independently identically distributed, randomly selecting a seed cell to define the variant coverage. This process is repeated with replacement of seed cells until the entire matrix D for a specified number of M cells is generated.

The genotype reference G of K samples was created by randomly selecting K unrelated individuals (British in England and Scotland) from the 1000 Genomes Project, selecting $N = 18,005$ distinguishing variants. Furthermore, following the input donor fraction $\boldsymbol{\pi}$, we proportionally sample the K genotype profiles to the M simulated cells.

Third, we can simulate the count for alternative allele $a_{i,j}$ of any variant i in any cell j by following a beta-binomial distribution with its depth $d_{i,j}$ and the parameter (α_t, β_t) given the genotype $g_{i,j} = t$. The parameters for each genotype are: (0.2, 99.8) for genotype $t = 0$, (0.1,

0.1) for $t = 1$, and (99.8, 0.2) for $t = 2$, which are fit from standard scRNA-seq data set (not multiplexed) across wide range of coverage.

For simulating doublets with rate η, we generate ηM additional cells following the same sample fraction. Then we randomly choose ηM out of M initial cells to one-to-one merge with the ηM additional cells, each of which consequently contains the aggregated reads of two cells and are defined as doublets. Note, not all these ηM doublets contain variants from two individuals but some are from the same individual.

When simulating a data set, we approximated the total sequenced reads proportional to the number of UMIs covering the input SNPs. For diluting sequencing coverage, a more lowly (or highly) covered seed coverage matrix $\widetilde{D}_{\text{seed}}$ can be generated by a multinomial distribution with the amplification ratio ρ and the default seed coverage matrix D_{seed}, i.e., $\widetilde{D}_{\text{seed}} \sim \text{Multinomial}(\rho S, D_{\text{seed}}/S)$ where S is the summary of D_{seed}.

Author contributions

O.S. and D.J.M conceived and guided the study. Y.H. developed and implemented the model. Y.H., D.J.M and O.S. carried out the experiments. Y.H., O.S. and D.J.M interpreted the results and wrote the paper.

Acknowledgements

We would like to thank Raghd Rostom, Maria Ban, Marc Jan Bonder and Stephen Sawcer for fruitful discussions.

References

[1] Michael JT Stubbington, Orit Rozenblatt-Rosen, Aviv Regev, and Sarah A Teichmann. Single-cell transcriptomics to explore the immune system in health and disease. *Science*, 358(6359):58–63, 2017.

[2] Jellert T Gaublomme, Nir Yosef, Youjin Lee, Rona S Gertner, Li V Yang, Chuan Wu, Pier Paolo Pandolfi, Tak Mak, Rahul Satija, Alex K Shalek, et al. Single-cell genomics unveils critical regulators of Th17 cell pathogenicity. *Cell*, 163(6):1400–1412, 2015.

[3] Detu Zhu, Zuxianglan Zhao, Guimei Cui, Shiehong Chang, Lingling Hu, Yi Xiang See, Michelle Gek Liang Lim, Dajiang Guo, Xin Chen, Paul Robson, et al. Single-cell transcriptome analysis reveals estrogen signaling coordinately augments one-carbon, polyamine, and purine synthesis in breast cancer. *Cell reports*, 25(8):2285–2298, 2018.

[4] Monica Golumbeanu, Sara Cristinelli, Sylvie Rato, Miguel Munoz, Matthias Cavassini, Niko Beerenwinkel, and Angela Ciuffi. Single-cell RNA-Seq reveals transcriptional heterogeneity in latent and reactivated HIV-infected cells. *Cell reports*, 23(4):942–950, 2018.

[5] Po-Yuan Tung, John D Blischak, Chiaowen Joyce Hsiao, David A Knowles, Jonathan E Burnett, Jonathan K Pritchard, and Yoav Gilad. Batch effects and the effective design of single-cell gene expression studies. *Scientific reports*, 7:39921, 2017.

[6] Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, 2017.

[7] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[8] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8:14049, 2017.

[9] Marlon Stoeckius, Shiwei Zheng, Brian Houck-Loomis, Stephanie Hao, Bertrand Z Yeung, William M Mauck, Peter Smibert, and Rahul Satija. Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome biology*, 19(1):224, 2018.

[10] Jase Gehring, Jong Hwee Park, Sisi Chen, Matthew Thomson, and Lior Pachter. Highly Multiplexed Single-Cell RNA-seq for Defining Cell Population and Transcriptional Spaces. *bioRxiv*, page 315333, 2018.

[11] Christopher S McGinnis, David M Patterson, Juliane Winkler, Marco Y Hein, Vasudha Srivastava, Daniel N Conrad, Lyndsay M Murrow, Jonathan S Weissman, Zena Werb, Eric D Chow, et al. MULTI-seq: Scalable sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *bioRxiv*, page 387241, 2018.

[12] Dongju Shin, Wookjae Lee, Ji Hyun Lee, and Duhee Bang. Multiplexed single-cell RNA-seq via transient barcoding for drug screening. *bioRxiv*, page 359851, 2018.

[13] Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M Lanata, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature biotechnology*, 36(1):89, 2018.

[14] Alice Giustacchini, Supat Thongjuea, Nikolaos Barkas, Petter S Woll, Benjamin J Povinelli, Christopher AG Booth, Paul Sopp, Ruggiero Norfo, Alba Rodriguez-Meira, Neil Ashley, et al. Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. *Nature medicine*, 23(6):692, 2017.

[15] Chunhong Zheng, Liangtao Zheng, Jae-Kwang Yoo, Huahu Guo, Yuanyuan Zhang, Xinyi Guo, Boxi Kang, Ruozhen Hu, Julie Y Huang, Qiming Zhang, et al. Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. *Cell*, 169(7):1342–1356, 2017.

[16] 1000 Genomes Project Consortium and others. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[17] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):15, 2018.

[18] Davis J McCarthy, Yunshun Chen, and Gordon K Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic acids research*, 40(10):4288–4297, 2012.

[19] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.