Genome Biology

METHOD Open Access

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

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

Abstract

Multiplexed single-cell RNA-seq analysis of multiple samples using pooling is a promising experimental design, offering increased throughput while allowing to overcome batch variation. To reconstruct the sample identify of each cell, genetic variants that segregate between the samples in the pool have been proposed as natural barcode for cell demultiplexing. Existing demultiplexing strategies rely on availability of complete genotype data from the pooled samples, which 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 single-cell data from pooled experimental designs. Uniquely, our model can be applied in settings when only partial or no genotype information is available. Using pools based on synthetic mixtures and results on real data, we demonstrate the robustness of Vireo and illustrate the utility of multiplexed experimental designs for common expression analyses.

Keywords: Multiplexing, Single-cell RNA-seq, Genetic variation, Variational Bayes

Background

Single-cell RNA-seq (scRNA-seq) is a rapidly evolving technology. Robust protocols and reduced costs have fostered applications 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, multiplexed experimental designs that combine multiple samples in a single experiment have critical statistical advantages compared to the serial analysis of samples in independent experimental batches [5, 6]. In particular, pooled designs allow disentangling true inter-individual variation from experimental batch variation. Pooled designs whereby a large number of cells from distinct samples are processed in a joint fashion are facilitated by the availability of droplet sequencing methods in particular, including Drop-seq [7] and the 10x Genomics Chromium platform [8], which can assay 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 simplified experimental logistics and reduced batch variation, pooled designs can also facilitate the identification of doublet cells. Existing barcoding strategies include molecular labelling prior to analysis [9–12] as well as exploiting natural genetic barcodes of germline variants that segregate between pooled individuals [13]. While molecular barcoding is in principle applicable to any study design, genetic barcoding is both elegant and can be seamlessly integrated in existing scRNA-seq workflows, without the need to introduce additional processing steps.

Multiplexed designs with genetic barcoding are particularly applicable in biomedical research, where the analysis of larger cohorts of genetically distinct individuals is particularly relevant [14]. However, current methods for demultiplexing genetically barcoded pools, such as Demuxlet [13], require genotype reference data for the pooled samples. Using variant information extracted from the scRNA-seq reads, each cell is then assigned to a sample in the pool based on its genetic distance to genotypic

*Correspondence: yuanhua@ebi.ac.uk; dmccarthy@svi.edu.au; o.stegle@dkfz.de
[1] Department of Clinical Neurosciences, University of Cambridge, CB2 0QQ Cambridge, UK
[2] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, CB10 1SD Cambridge, UK
Full list of author information is available at the end of the article

states in the genotype reference database. While there is a growing interest in multi-sample analyses to study the effect of genetic variation between individuals at single-cell level, e.g., [15–17], the requirement to supply a genotype reference database is prohibitive for studies without a genetic focus per se. Consequently, the potential of pooled experimental designs is currently not 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 leverage 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.

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. Specifically, the model takes a set of common genetic variants as input (for example derived from the 1000 Genomes Project [18]), which are genotyped in each cell based on the scRNA-seq read data. Despite the typically low coverage of single-cell RNA-seq experiments, this approach allows for genotyping on the order of 100 expressed variants per cell (e.g., using 3' 10× Genomics data; approx. 50,000 reads per cell, Fig. 1 and Methods). By aggregating information across cells, these sparse genotype data are sufficient to reconstruct partial genotypic state of the individuals in the pool, which in turn allows for probabilistic demultiplexing whereby each cell is assigned to one of these individuals (Fig. 1). Vireo also accounts for the possibility of doublets (two or more cells processed as a “single cell” in the assay), by considering cells with variants that are most consistent with a genotypic state formed by the combination of two individuals. Finally, the model estimates the most likely number of pooled individuals, a feature that is useful if some of the pooled samples drop out for experimental reasons, and the method can incorporate partial genotype data that are available for a subset of the pooled samples.

Model validation using synthetic data

Initially, we considered synthetic data with a known truth to validate our approach. We considered raw 3' single-cell RNA-seq data from the 10x Genomics platform (v2 kit) for 16 genetically distinct samples from the census of immune cells project that are available from the Human Cell Atlas (Methods) [19]. We then synthetically mixed 8 of these samples (1000 cells per sample and 4000 UMIs per cell on average), and simulated 8% of the cells as doublets, which were included alongside the sampled singlet cells (“singlets"; Methods). Initially, we evaluated Vireo’s ability to estimate the number of input samples, by comparing the marginal likelihood of multiple Vireo runs assuming increasing numbers of samples in the pool, ranging from six to twelve. Notably, models with at least the true number of input samples ($K = 8$) were differentiated from models with too low sample counts based on the elbow plot of the variational lower bound (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 required to explain the data were used, indicating that the model is robust to choosing more samples than necessary during inference (Additional file 1: Figure S1).

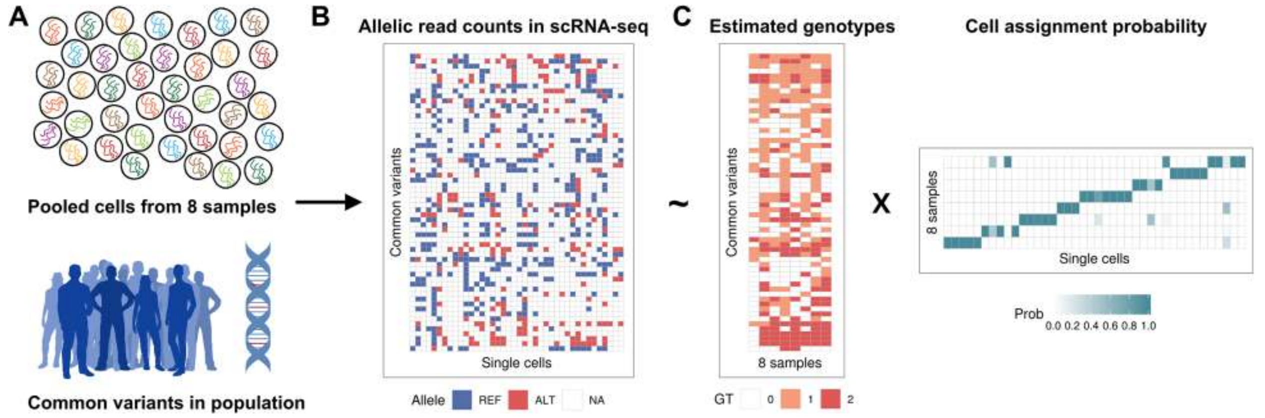

Fig. 1 Illustration of Vireo for demultiplexing multi-sample scRNA-seq studies without reference genotype data. **a**, **b** The inference is based on genotyped common polymorphic variants in each cell, defined based on a standard reference of common human variants. **b**, **c** The resulting sparse read count matrices of alternative and reference alleles (displayed as compound matrix for simplicity; NA in white denotes no observed reads) are then decomposed into a matrix of estimated genotypes for each input sample and a probabilistic cell assignment matrix

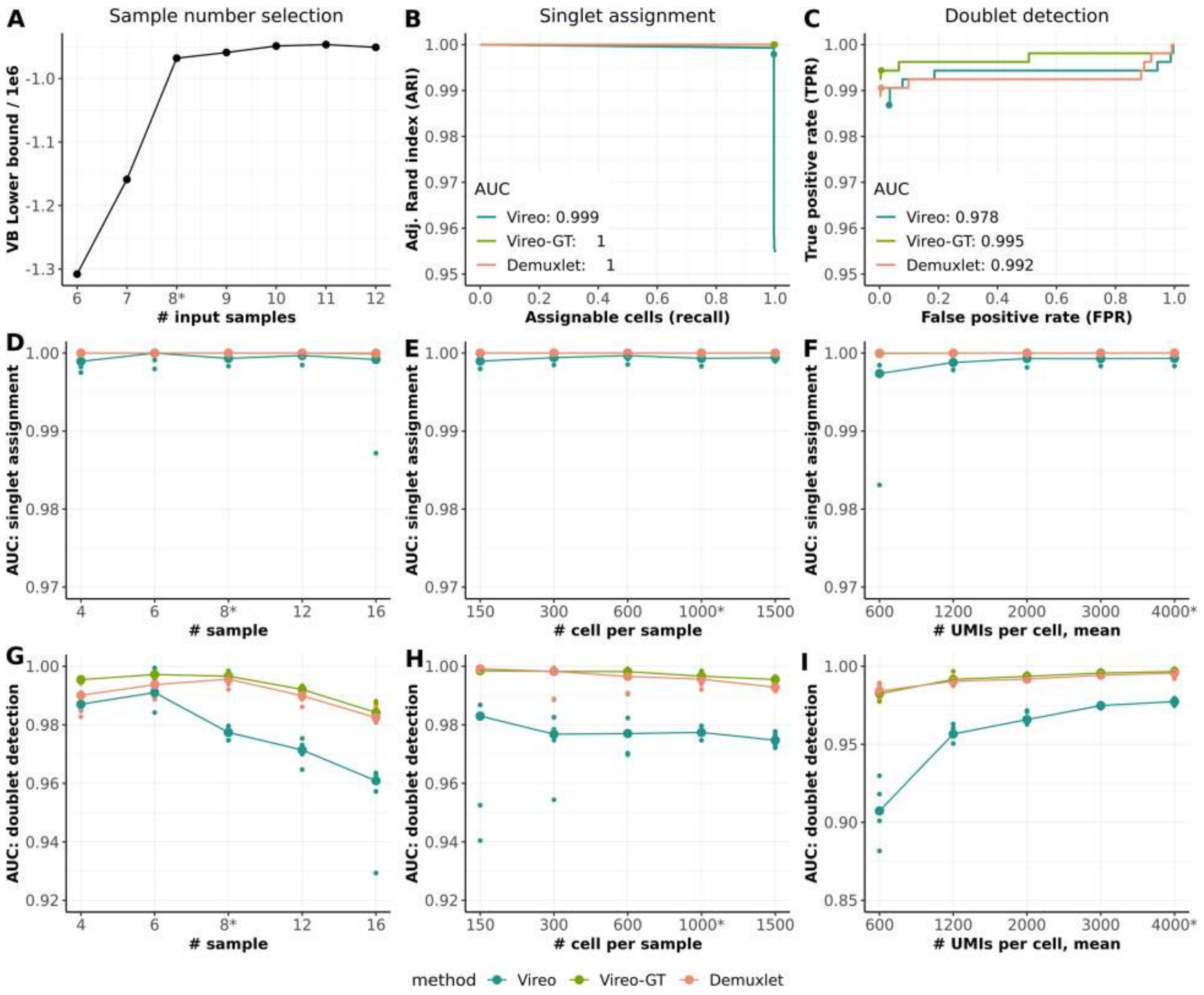

Fig. 2 Evaluation of Vireo using synthetic mixtures of up to 16 scRNA-seq datasets, including a comparison to models that require genotype data of the pooled samples (Vireo-GT and Demuxlet). **a**–**c** Assessment of Vireo performance on one representative simulated dataset consisting of 8 pooled samples with 1000 cells per sample, 8% doublet rate, and 4000 UMIs per cell on average. **a** Vireo model evidence (variational lower bound) when varying the pool size assumed in the model. **b** Adjusted Rand index (ARI) between the most likely inferred and the true singlet assignment, when varying the assignment confidence. The recommended cutoff (prob_max >0.9 for Vireo and Vireo-GT, and PRB.SNG1 >0.67 for Demuxlet) are highlighted as dot. **c** Receiver operating characteristic (ROC) curve for detecting doublets, when varying the assignment confidence. The recommended cutoff (prob_doublet >0.9 for Vireo and Vireo-GT, and 0.67 for Demuxlet) are highlighted with dots. **d**–**i** Systematic assessment of Vireo and alternative methods on simulated data using a range of parameter choices, using five replicate runs. **d**–**f** Area under ARI for singlet assignment considering alternative simulation settings, either varying the number of input samples **d**, the number of total cells in the dataset (doublet rate varied proportional; 1.2 to 12%; Methods) **e**, as well as the number of UMIs per cell **f**. **g**-**i** Area under the ROC for doublet detection, considering the same simulation parameters as in panels **d**–**f**. Parameters not varied in either of these experiments were set to their default values (indicated by the star symbol). Small dots in each experiment denote the five replicate simulation experiments, and the big dot denotes the median performance across replicates

Next, we evaluated the performance of Vireo for singlet assignment and doublet detection, where for comparison we also considered alternative models that require full genotype data of the pooled samples (Demuxlet [13] and Vireo-GT, i.e., Vireo with full genotype data; Methods). By measuring the adjusted Rand index (ARI) of the most likely assignment of singlet cells to samples with regard to the true assignments, we found that Vireo achieved markedly accurate results, yielding comparable performance as Vireo-GT and Demuxlet (Fig. 2b). We also varied the assignment confidence (Methods), finding that all three methods achieve near-perfect assignments of the full set of singletons (recall = 1). In the following, we consider the area under the ARI-recall curve (AUC) as a measure to systematically assess the performance of singlet assignment across a wider range of settings.

Similar to Demuxlet, 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. Vireo without genotype achieves doublet detection with an overall AUC = 0.978 (98.7% sensitivity and 96.7% specificity at prob_doublet >0.9, Fig. 2c), which is only marginally lower than the performance achieved when using genotype data (Vireo-GT or Demuxlet, both AUC $\approx$ 0.995). In practice, and in the experiments reported below, we recommend prob_max >0.9 as the threshold for the singlet 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, g), the number of cells assayed in each experiment (Fig. 2e, h), and the number of UMIs per cell (Fig. 2f, i). As expected, the cell-assignment accuracy decreased with increasing numbers of samples in the pool, but Vireo retained high accuracy for up to 12 multiplexed samples (Fig. 2d, g). Beyond 12 samples, there is a risk that the Vireo solution represents a local optimum of the variational lower bound, omitting one or multiple samples present in the pool (Fig. 2d). Using current experimental technologies, however, such high multiplexes are not commonly considered, as the necessary cell counts are associated with greatly increased doublet rates (e.g., on the 10x Chromium platform). Conversely, the accuracy of cell-assignment was consistently high across a large range of cell counts per sample (Fig. 2e), where larger numbers of cells tended to result in increased accuracy. Similarly, increasing the sequencing coverage resulted in improved accuracy for doublet detection (Fig. 2i), whereas accurate singleton assignments were achieved even with extremely low UMI counts per cell (Fig. 2f).

Next, we assessed the utility of partial genotype data for a subset of samples in the pool, which as expected increased the model performance, particularly in settings with low sequencing coverage (1200 UMIs per cell, Additional file 1: Figure S2). We also evaluated the robustness of Vireo when applying the model to biased pools of samples, i.e., settings in which some samples contribute a smaller than expected fraction of cells. Vireo robustly detected and aligned cells to samples with a relative frequency as low as 10% (Additional file 1: Figure S3), while retaining high accuracy for doublet detection. However, rare samples that were represented by fewer than 100 cells could be be missed in some settings.

Finally, we assessed the accuracy of the genotype reconstruction of the pooled samples, finding that Vireo implicitly provides accurate genotype information for expressed variants (10 or more UMIs) detected in the scRNA-seq data (overall precision > 0.96, with heterozygous sites of lowest precision = 0.91; Additional file 1: Figure S4). Although such estimated genotypic states are intrinsically not available genome-wide, these partial genotype profiles can be used as a linking key to align the reconstructed samples to other omics data or to combine demultiplexed datasets across experiments (Methods).

Application to real pooled data

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

First, we considered a set of three multiplexed experiments (Fig. 3a-c; W1-W3, between 3639 and 6145 cells) of peripheral blood mononuclear cells (PBMCs) from eight lupus patients. We applied Vireo without using genotype information to all cells across these three batches (Methods), thereby also creating an implicit link across all the three experiments. The Vireo cell assignments were markedly consistent with the assignments obtained when using Demuxlet, which, however, relies crucially 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 sample identity across experiments retrospectively (Fig. 3c). This result demonstrates the utility of the inferred genotype data for integrating demultiplexed samples across experiments, and shows how the inferred genotypes can also be used to link demultiplexed scRNA-seq samples to other (sequencing-based) assay data available from 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 were cultured for 6 h 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).

Leveraging multiplexed designs for differential expression analysis

Finally, we considered the demultiplexed 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 [20]) applied to the joint dataset consisting of stimulated and unstimulated cells from all eight samples (Fig. 4c) identified eight major clusters, which could be annotated by common cell types (Fig. 4a-b; Additional file 1: Figure S5). Next, we tested for differential

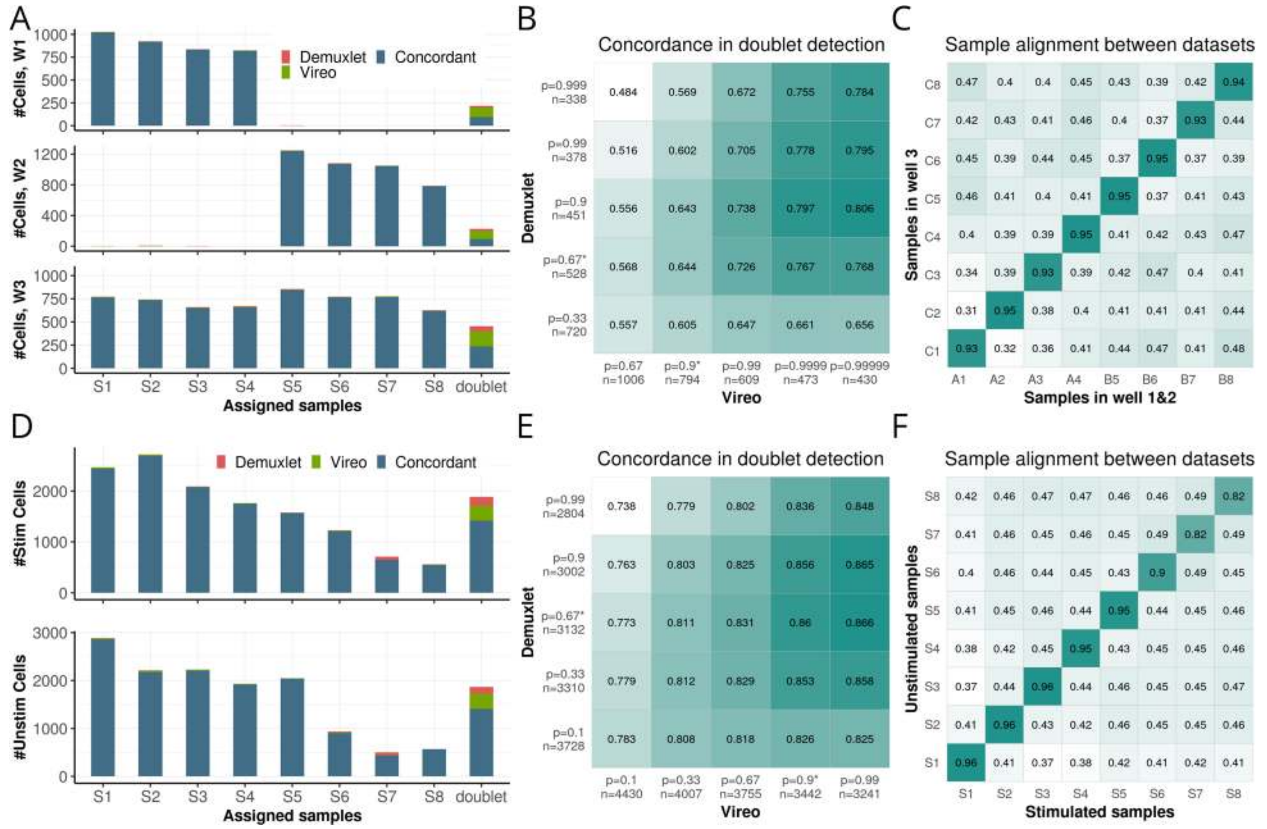

Fig. 3 Evaluation of Vireo on data from multiplexed human PBMCs. **a**–**c** Results obtained on two datasets consisting of two pools of 4 samples each, as well as a third dataset consisting of the union of all 8 samples. **a** Concordance of singlet assignment and doublet detection between Vireo without genotype data and Demuxlet applied with complete genotype reference. Bars denote the number of cells assigned to each sample, either considering cells that were consistently assigned by both 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 detected 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 cells and are considered in **b**. **c** Alignment of samples, when applying Vireo separately to the three datasets considered in **a**. 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 from a second experiment, consisting of two datasets with the same 8 samples pooled in two different conditions: unstimulated and stimulated. Results shown correspond to the panels in **a**–**c**

gene expression between the stimulated and unstimulated condition within each cell type (using edgeR, considering cells as replicates [21]). Considering B cells as a representative example (see Additional file 1: Figure S8–S11 for full results), this analysis identified between 78 and 477 DE genes in individual samples (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 samples (Fig. 4e), there was a substantial fraction of DE genes that were private to individual samples. For example, the gene OAZ1 (Fig. 4d) was differentially expressed in four of eight samples, highlighting the relevance of inter-individual differences (more examples in Additional file 1: Figure S6). We also explored carrying out joint testing across all samples (using samples as an explanatory factor in the model in edgeR; Methods), which led to broadly similar conclusions (Additional file 1: Figure S7).

Conclusion

Here, we have presented Vireo, a Bayesian method for demultiplexing pooled single-cell RNA-seq datasets by exploiting natural genetic barcodes and cell genotyping based on scRNA-seq reads. Uniquely, Vireo does not require any reference genotype data of the specific samples that are pooled in the experiment, while achieving demultiplexing accuracies that are comparable to methods that require a genotype reference. Vireo is

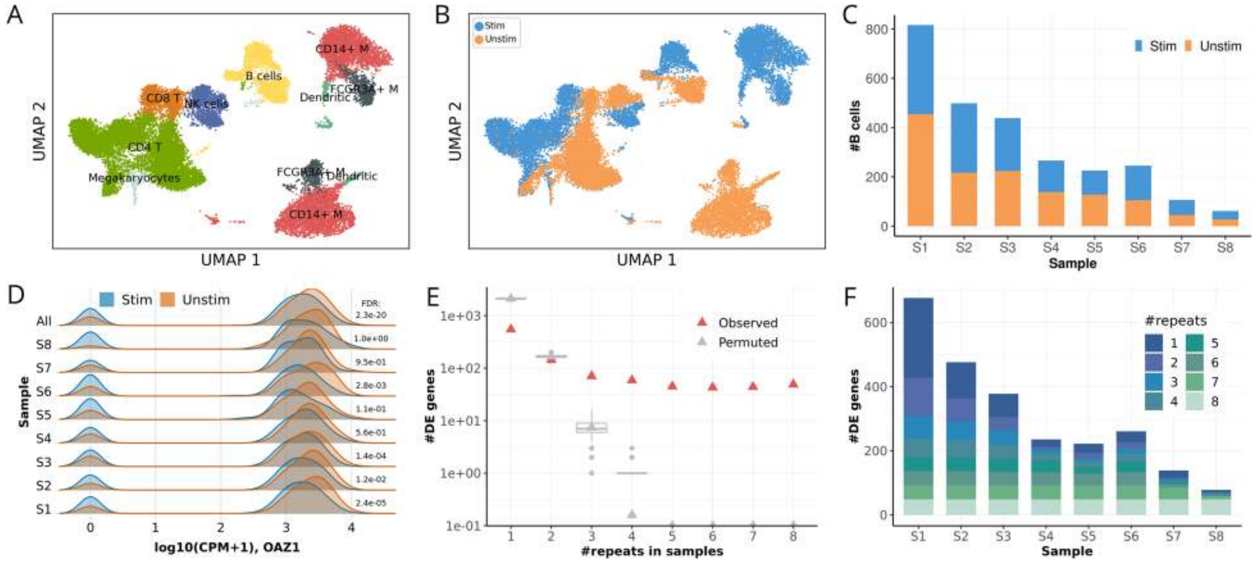

Fig. 4 Case study of differential expression analysis of INF-β stimulated PBMCs using two matched pools that consist of 8 samples. **a** UMAP representation of single-cell transcriptome profiles, colored by eight major cell types identified. Note, M denotes Monocytes. **b** Analogous UMAP representation 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**–**f** Differential gene expression (DE) analysis between stimulated and unstimulated B cells. **d** The expression level of an example gene OAZ1, depicting the distribution of expression levels in both conditions, either considering each sample separately (S1..S8) or considering aggregated data pooled across all samples (All). FDR: adjusted P value (Benjamini-Hochberg) of each DE test between conditions with likelihood-ratio test. 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 gray show the recurrence expected by chance (based on 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)

implemented using computationally efficient variational Bayesian inference, which provides a fully Bayesian treatment while retaining scalability to large datasets.

Using synthetic mixtures of cells, we have evaluated the accuracy of Vireo for demultiplexing pooled samples, and found it robust to a variety of settings. We also demonstrated the model's flexibility for handling partial genotype data for some of the 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 for the optimal 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. 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 enable aligning samples from the scRNA-seq data with other 'omics data for the same samples (Fig. 3c), which provides a flexible approach for linking samples across experiments, including multi-omics treatment-control designs.

We noticed that the accuracy of demultiplexing without a genotype reference starts to deteriorate for pools with more than 12 samples. Increased sequencing coverage may allow for demultiplexing even larger pools, but there remain general experimental limitations for such designs. In particular, as long as the doublet rates scales with increased cell count such designs remain of limited interest.

As future technologies that motivate even larger pool sizes become available, extensions of Vireo that can handle such settings may be warranted. Notably, the demultiplexing accuracy is also linked to read coverage per cell as well as total cell count, two characteristic quantities that are likely to improve as single-cell technologies continue to mature.

As a reference-free method, Vireo is particularly useful in settings where samples are treated as biological replicates and the primary object is the variation between samples or groups, which does not require the explicit identification of individual samples in the pool (Fig. 3 and 4). Beyond that, Vireo has the intrinsic limitation that the inferred samples cannot be directly identified or linked to metadata. However, when the necessity for sample identity arises, the estimated genotype states are readily available for linking the samples to other 'omics data, e.g., other scRNA-seq batches (Fig. 3c, f) or bulk RNA-seq (Additional file 1: Figure S12). These principles can be applied to any read-based assay, which provides genotypes. Finally, it is straightforward to generate targeted qPCR-based genotypes for a minimal set of discriminatory variants (Additional file 1: Figure S13). The Vireo software provides helper functions for designing such experiments, which directly leverages the reconstructed genotypes in the pool to define a small set of discriminatory variants (Methods). Vireo may also prove suitable for demultiplexing pooled samples for other read-based single-cell assays such as single-cell ATAC-seq, but further benchmarking on appropriate datasets would be needed and was not explored in this work.

Molecular barcoding strategies, e.g., [9–12], have recently emerged as an alternatives to genetic barcoding in many respects courtesy of their more universal applicability. For example, molecular barcoding enables pooling multiple treatment conditions or tissues from the same individual or from individuals with the same genetic background (e.g., inbred model organisms). Nevertheless, natural genetic variants as barcode, which thanks to Vireo now can be applied even when no genotype data are available, have the advantage of avoiding additional laboratory work, thus reducing the logistical complexity, which can impact cost, processing efficiency and data quality.

Methods

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

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

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 is available for one or multiple samples in the pool, this information can be encoded as an 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:

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

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 a beta distribution for $\boldsymbol{\theta}$ and multinomial distributions for both Z and G.

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

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: $\theta_0 \sim \mathtt{beta}(0.3, 29.7)$, $\theta_1 \sim \mathtt{beta}(3, 3)$, and $\theta_2 \sim \mathtt{beta}(29.7, 0.3)$, with which the posterior of $\boldsymbol{\theta}$ will be obtained by fitting to the dataset.

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 [22] to obtain an approximate solution, thereby retaining the benefits of a Bayesian treatment while achieving 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 minimize the Kullback-Leibler ($\mathcal{KL}$) divergence between the approximate distribution $q(\mathbf{Y})$ and the actual posterior distribution $p(\mathbf{Y}|\mathbf{X})$

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

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,

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

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

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

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

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

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 Additional file 1: Supplementary Methods Equation (S1-6)). Following Eq. 8, it is possible to derive iterative update equations for the Q distribution of the latent variables (see Additional file 1: Supplementary Methods Equation (S7-12) for full details).

$$q^*(Z) = \prod_{j=1}^{M} \prod_{k=1}^{K} r_{j,k}^{Z_{j,k}};$$

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

$$q^*(G) = \prod_{i=1}^{N} \prod_{k=1}^{K} \prod_{t \in T} g_{i,k,t}^{G_{i,k,t}};$$

$$g_{i,k,t} = \frac{u_{i,k,t} \exp \sum_{j=1}^{M} \left\{\tilde{r}_{j,k}\left[a_{i,j}\varphi(\tilde{\alpha}_t) + b_{i,j}\varphi(\tilde{\beta}_t)\right]\right\}}{\sum_{h \in T} u_{i,k,h} \exp \sum_{j=1}^{M} \left\{\tilde{r}_{j,k}\left[a_{i,j}\varphi(\tilde{\alpha}_h) + b_{i,j}\varphi(\tilde{\beta}_h)\right]\right\}} \quad (10)$$

$$q^*(\boldsymbol{\theta}) = \prod_{t \in T} \mathtt{beta}(\theta_t | \alpha_t, \beta_t);$$

$$\alpha_t = \alpha_t^{(0)} + \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{K} \tilde{r}_{j,k} \tilde{g}_{i,k,t} a_{i,j}, \quad (11)$$

$$\beta_t = \beta_t^{(0)} + \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{K} \tilde{r}_{j,k} \tilde{g}_{i,k,t} b_{i,j}$$

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: 50 restarts) and the solution that maximizes the variational lower bound is selected. Thanks to our implementation with sparse matrix data structures and the support of multiple threads, the Vireo model is computationally efficient. On a laptop with 16 G memory and two 3.5-GHz CPUs, Vireo finishes a two-run estimation (see next section) in 6.7 min for 14,619 cells in an eight sample pool and 58.1 s for 6145 cells in another eight sample pool (results in Fig. 3).

Vireo with known genotype or partial genotype

Besides demultiplexing pooled scRNA-seq without any genotype information, Vireo is also able to leverage any available genotype information. In the case that the genotype is available for all pooled samples, we only use the variants with known genotype and set the genotype probability variable G as known and fixed, which can be derived from the GT tag (for categorical genotype), GP or GL tag (genotype probability or likelihood) in the VCF file. By default, we use GT as it is the most commonly available tag.

Alternatively, Vireo also supports the use of any partial genotypes via a two-step run approach. In the first run, Vireo does not use any genotype information but infers the genotype for each sample. Then, we align the samples with known genotype to these identified samples in this run and replace the estimated genotype probability with the input known values. Therefore, we obtain a genotype probability matrix with mixed known and inferred samples, which we then use as a prior of G, instead of the default uniform prior in the second run. Finally, we report the results of the second run as the result of Vireo.

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 with the elbow plot (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 post hoc, as the variational Bayes model is self-regularizing and hence avoids over-fitting (see Additional file 1: Figure S1). 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.

Multiple random initializations

When genotype is not given, Vireo uses a pre-step with multiple random initializations to avoid local optima. By default, Vireo runs for 50 random initializations, each with a short iterations (15 by default). Then the initialization with highest log likelihood will be continued.

As discussed in the above subsection, another strategy to find all K pooled samples is searching from a larger number of clusters. By default, we search $K+\sqrt{K}$ clusters in this pre-step, and only keep the K clusters with largest number of assigned cells to continue, and discarded the $\sqrt{K}$ smaller clusters.

Doublet detection

To detect doublets, we 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

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

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. 9.

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

$$p(\pi_h) = \begin{cases} (1-\eta)/K, & 1 \le h \le K; \text{(singlet)} \\ \eta/K_2, & K < h \le K+K_2; \text{(doublet)} \end{cases} \tag{14}$$

where $K_2 = (K-1) \times K/2$ as number of the combined sample pairs. 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., $\eta = M/100,000$ following [13], which by default is used in our experiments.

Therefore, we can obtain the posterior of each cell's sample identity, i.e., the probability of cell j coming from any of the K input samples or K_2 combined sample pairs (i.e., doublet). We use the highest assignment probability of the K input samples, prob_max, as the confidence score for singlet assignment and use the summarized probability of all K_2 sample pairs as the confidence score of a doublet, namely prob_doublet.

Alignment of samples between multiple data sets

Vireo implicitly estimates the genotypes for the subset of variants with sufficient coverage (good accuracy for variants with >10 reads per sample; see Additional file 1: Figure S4). 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 data sets that consist of the same samples, the estimated genotype also allows for aligning samples across data sets (e.g., Fig. 3c, f for multiple pools, and Additional file 1: Figure S12 for multiple 'omics). The software implementation of Vireo provides support functions for this step by calculating the fraction of variants with matched genotype between two or multiple experiments.

Identification of discriminatory variants

Given a set of variants for which estimated genotypes are available, the Vireo software implements a heuristic to define a minimal and informative set of discriminatory variants. This set of variants can be used to perform qPCR-based genotyping or for other targeted genoytping methods. Briefly, the algorithm implemented in Vireo prioritizes variants with largest information gain in splitting samples, as follows.

1 Remove variants with <20 UMIs per sample.
2 Initialize the variant set $S = \{\}$, and the split T among K samples, and calculate the initial entropy $H(T) = 0$
3 Rank variants by the information gain $IG(T, v) = H(T) - H(T|v)$
4 Select the variant with highest information gain and update S, T, and $H(T)$
5 If $H(T) = \log_2(K)$, return S and T, otherwise go to step 3.

Additionally, variants with homozygous alternative alleles in the pooled samples can also be filtered out before hand if needed. Examples of discriminatory variant sets for the six-sample pool from HipSci project are shown in Additional file 1: Figure S13.

Differential expression analysis

Differential expression analysis was performed with edgeR [21] between stimulated and control samples (Fig. 4 and Additional file 1: Figure S7–S11). A generalized linear model (negative binomial regression) is applied in edgeR to test whether the stimulation contributes to the expression variation on a certain gene by using a likelihood-ratio test. Using the raw UMI counts as input to edgeR, we performed cell type specific DE analysis with the following three different strategies for all cells jointly (Additional file 1: Figure S7).

Method 1: $y \sim \text{cdr} + \text{condition} + \text{sample}$, where y is the expression count for a specific gene, which is regressed on three covariates: cdr, the cell detection rate (i.e., the fraction of expressed gene in each cell), stimulation condition and the sample identity.

Methods 2 and 3: $y \sim \text{cdr} + \text{condition}$, where we ignore the sample identity of each cell in the pool (Method 2). This same model can also be used in a pseudo-bulk manner where we aggregate the counts for all cells of the same type in a sample (Method 3). Alternatively, we can always apply this model at single-cell level for each sample separately (Fig. 4d-f).

ScRNA-seq data from Demuxlet paper

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 [18] were used as candidate variants for scRNA-seq genotyping. We provide an companion Python package `cellSNP` [23] 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 outputs 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.

Bulk RNA-seq and scRNA-seq from HipSci project

In order to link the inferred samples to other 'omics data, we used one scRNA-seq pool for iPSC differentiation in the HipSci project (10x Genomics platform, experiment 44, day 0) with six samples: *pipw*, *jejf*, *qehq*, *juuy*, *uilk*, and *toco* [15], and their corresponding bulk RNA-seq data for each sample [24] (http://www.hipsci.org). Both scRNA-seq and bulk RNA-seq data sets were downloaded in .bam files and genotyped on 7.4 millions common bi-allelic variants (minor allele frequency >5%) extracted from the 1000 Genome Project with the `cellSNP` package. For single-cell data, we only keep variants with minor allele frequency $\geq$ 0.1 and $\geq$ 20 UMIs. For each bulk RNA-seq sample, we also only keep variants with minor allele frequency $\geq$ 0.1 but require $\geq$ 100 read counts. Then the genotypes of each bulk RNA-seq sample can be used to align to the samples that are demultiplexed from scRNA-seq data.

Synthetic data

We obtained raw 3' scRNA-seq data based on 10x Genomics platform (v2 kit) for 16 genetically distinct samples from the Human Cell Atlas (Census of Immune Cells) [19]. These data set are not pooled and each sample has its own sequencing run. We only used data from the first channel (each sample with around 100 million reads), which is in the range of a standard 10x sequencing run. We first mapped the raw fastq files to the human genome hg38 by CellRanger v2.1 provided by 10x Genomics (cellranger count command line). Then we used `cellSNP` to genotype 7.4 million common variants (minor allele frequency >5%) extracted from the 1000 Genome Project for these 16 samples in a pseudo-bulk manner. We only keep variants with: 1) >100 UMIs summarized across 16 samples, 2) >10% UMIs from the minor allele, and 3) <5 UMIs for other alleles (i.e., not annotated reference and alternative alleles). Therefore, we obtained the genotypes of 62,193 variants for these 16 samples, which are fed into Demuxlet and Vireo-GT.

By only keeping cells with >500 genes and >1000 UMIs, we had in total 66,410 cells across 16 samples, with each sample having 2495 to 4909 cells. On average, there are 4000 UMIs per cell (median 2700 UMIs). In the synthetic mixture, we pooled reads for a subset of cells from each sample (in .bam format, aligned reads) and generated multiplexed scRNA-seq data (also in .bam format). The script to generate these synthetic data is provided in Vireo's GitHub repository. Doublets were added into the pooled data by adding proportional extra cells and combining them with another cells randomly. The doublet rate is $N/100,000$ where N is the total number of cells in the pool.

By default, we pooled 1000 cells from each of 8 samples with doublet rate of 8%. This simulator also allows setting different size of input samples, for example by setting one sample with fewer cells ranging from 50 to 500 (Additional file 1: Figure S3). With the synthetic data in .bam format, we can even further subsample reads by using `samtools` with -s argument, e.g., 15–75% in Fig. 3f.

All these simulations were randomly repeated for five times to account for the variability in the simulation.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-019-1865-2.

Additional file 1: Supplementary methods and supplementary figures S1–S13.

Additional file 2: Review history.

Acknowledgements

We would like to thank Raghd Rostom, Maria Ban, Marc Jan Bonder, Stephen Sawcer, and Sarah Teichmann for fruitful discussions. We thank the authors from the Census of Immune Cells project for making their data available prior to publication.

Review history

The review history is available as Additional file 2.

Authors' contributions

O.S. and D.J.M conceived and guided the study. Y.H. developed and implemented the model. YH, DJM and OS carried out the experiments. YH, OS, and DJM interpreted the results and wrote the paper. All authors read and approved the final manuscript.

Funding

Y.H. is supported by the University of Cambridge and EMBL-EBI through an EBPOD postdoctoral fellowship. D.J.M. is supported by National Health and Medical Research Council of Australia through an Early Career Fellowship (GNT1112681) and a Project Grant (GNT1162829). O.S. is supported by core funding from EMBL and DKFZ and the EU (ERC project DECODE 810296). We acknowledge funding from the Wellcome Trust and Medical Research Council (WT098503).

Availability of data and materials

Vireo model has been implemented as a standard Python package, which is freely available at https://github.com/single-cell-genetics/vireo with Apache License 2.0. The version 0.1.5 used in for analysis in paper is deposited in Zenodo [25]. All scripts to replicate the simulations in this paper are also included in linked GitHub repository. Vireo's manual with examples is available at https://vireosnp.readthedocs.io. A companion Python package cellSNP for genotyping cells in scRNA-seq data is freely available at https://github.com/single-cell-genetics/cellSNP with Apache License 2.0. The version 0.1.6 used in paper is deposited in Zenodo [23]. The processed common SNPs from 1000 Genome Project are available at https://sourceforge.net/projects/cellsnp/files/SNPlist.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Department of Clinical Neurosciences, University of Cambridge, CB2 0QQ Cambridge, UK. [2] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, CB10 1SD Cambridge, UK. [3] St Vincent's Institute of Medical Research, Fitzroy, 3065 Victoria, Australia. [4] Melbourne Integrative Genomics, University of Melbourne, Parkville, 3010 Victoria, 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.

Received: 3 April 2019 Accepted: 25 October 2019
Published online: 13 December 2019

References

1. Stubbington MJ, Rozenblatt-Rosen O, Regev A, Teichmann SA. Single-cell transcriptomics to explore the immune system in health and disease. Science. 2017;358(6359):58–63.
2. Gaublomme JT, Yosef N, Lee Y, Gertner RS, Yang LV, Wu C, Pandolfi PP, Mak T, Satija R, Shalek AK, et al. Single-cell genomics unveils critical regulators of Th17 cell pathogenicity. Cell. 2015;163(6):1400–12.
3. Zhu D, Zhao Z, Cui G, Chang S, Hu L, See YX, Lim MGL, Guo D, Chen X, Robson P, et al. Single-cell transcriptome analysis reveals estrogen signaling coordinately augments one-carbon, polyamine, and purine synthesis in breast cancer. Cell Rep. 2018;25(8):2285–98.
4. Golumbeanu M, Cristinelli S, Rato S, Munoz M, Cavassini M, Beerenwinkel N, Ciuffi A. Single-cell RNA-Seq reveals transcriptional

heterogeneity in latent and reactivated HIV-infected cells. Cell Rep. 2018;23(4):942–50.

5. Tung P-Y, Blischak JD, Hsiao CJ, Knowles DA, Burnett JE, Pritchard JK, Gilad Y. Batch effects and the effective design of single-cell gene expression studies. Sci Rep. 2017;7:39921.
6. Hicks SC, Townes FW, Teng M, Irizarry RA. Missing data and technical variability in single-cell rna-sequencing experiments. Biostatistics. 2017;19(4):562–78.
7. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161(5):1202–14.
8. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, et al. Massively parallel digital transcriptional profiling of single cells. Nature Commun. 2017;8:14049.
9. Stoeckius M, Zheng S, Houck-Loomis B, Hao S, Yeung BZ, Mauck WM, Smibert P, Satija R. Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol. 2018;19(1):224.
10. Gehring J, Park JH, Chen S, Thomson M, Pachter L. Highly Multiplexed Single-Cell RNA-seq for Defining Cell Population and Transcriptional Spaces. BioRxiv. 2018;315333.
11. McGinnis CS, Patterson DM, Winkler J, Conrad DN, Hein MY, Srivastava V, Hu JL, Murrow LM, Weissman JS, Werb Z, et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nature Meth. 2018;16(7):619–26.
12. Shin D, Lee W, Lee JH, Bang D. Multiplexed single-cell RNA-seq via transient barcoding for simultaneous expression profiling of various drug perturbations. Sci Adv. 2019;5(5):2249.
13. Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, Wan E, Wong S, Byrnes L, Lanata CM, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nature Biotechnol. 2018;36(1):89.
14. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell rna-sequencing for biomedical research and clinical applications. Genome Med. 2017;9(1):75.
15. Cuomo AS, Seaton DD, McCarthy DJ, Martinez I, Bonder MJ, Garcia-Bernardo J, Amatya S, Madrigal P, Isaacson A, Buettner F, et al. Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. BioRxiv. 2019;630996.
16. McCarthy DJ, Rostom R, Huang Y, Kunz DJ, Danecek P, Bonder MJ, Hagai T, Wang W, Gaffney DJ, Simons BD, et al. Cardelino: Integrating whole exomes and single-cell transcriptomes to reveal phenotypic impact of somatic variants. BioRxiv. 2018;413047.
17. van der Wijst MG, Brugge H, de Vries DH, Deelen P, Swertz MA, Franke L. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. Nat Genet. 2018;50(4):493.
18. 1000 Genomes Project Consortium and others. A global reference for human genetic variation. Nature. 2015;526(7571):68.
19. Li B, Kowalczyk MS, Dionne D, Ashenberg O, Tabaka M, Tickle T, Lee J, Shekhar K, Slyper M, Waldman J, Rozenblatt-Rosen O, Regev A, Census of Immune Cells. 2018. https://data.humancellatlas.org. Accessed 12 April 2019.
20. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19(1):15.
21. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012;40(10):4288–97.
22. Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: A review for statisticians. J Am Stat Assoc. 2017;112(518):859–77.
23. Huang Y. CellSNP version 0.1.6. 2019. https://doi.org/10.5281/zenodo.3516640. https://zenodo.org/record/3516640.
24. Kilpinen H, Goncalves A, Leha A, Afzal V, Alasoo K, Ashford S, Bala S, Bensaddek D, Casale FP, Culley OJ, et al. Common genetic variation drives molecular heterogeneity in human iPSCs. Nature. 2017;546(7658):370.
25. Huang Y, McCarthy DJ, Stegle O. Vireo version 0.1.5. 2019. https://doi.org/10.5281/zenodo.3516639. https://zenodo.org/record/3516639.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.