Bioinformatics Advances, 2024, **00**, vbae085
https://doi.org/10.1093/bioadv/vbae085
Advance Access Publication Date: 8 June 2024
Original Article

OXFORD

Software

Demultiplexing of single-cell RNA-sequencing data using interindividual variation in gene expression

Isar Nassiri [1,2,3,4,*], Andrew J. Kwok [2,5], Aneesha Bhandari[2], Katherine R. Bull[2], Lucy C. Garner[6], Paul Klenerman [6,7,8], Caleb Webber[9,10], Laura Parkkinen[1,11], Angela W. Lee[2], Yanxia Wu [2], Benjamin Fairfax[12,13], Julian C. Knight [2,14], David Buck[2], Paolo Piazza[1,2,*]

[1]Nuffield Department of Medicine, Centre for Human Genetics, Oxford-GSK Institute of Molecular and Computational Medicine (IMCM), University of Oxford, Oxford, OX3 7BN, United Kingdom
[2]Nuffield Department of Medicine, Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, United Kingdom
[3]Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, OX3 9DU, United Kingdom
[4]Department of Psychiatry, University of Oxford, Oxford, OX3 7JX, United Kingdom
[5]Department of Medicine and Therapeutics, Faculty of Medicine, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, 999077, China
[6]Translational Gastroenterology Unit, Nuffield Department of Medicine, University of Oxford, Oxford, OX3 9DU, United Kingdom
[7]Peter Medawar Building for Pathogen Research, University of Oxford, Oxford, OX1 3SY, United Kingdom
[8]NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, OX3 9DU, United Kingdom
[9]Department of Physiology, Anatomy, Genetics, Oxford Parkinson's Disease Centre, University of Oxford, Oxford, OX1 3PT, United Kingdom
[10]UK Dementia Research Institute, Cardiff University, Cardiff, CF24 4HQ, United Kingdom
[11]Nuffield Department of Clinical Neurosciences, Oxford Parkinson's Disease Centre, University of Oxford, Oxford, OX3 9DU, United Kingdom
[12]MRC–Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, OX3 9DS, United Kingdom
[13]Department of Oncology, University of Oxford & Oxford Cancer Centre, Churchill Hospital, Oxford University Hospitals NHS Foundation Trust, Oxford, OX3 7DQ, United Kingdom
[14]Chinese Academy of Medical Science Oxford Institute, University of Oxford, Oxford, OX3 7BN, United Kingdom

*Corresponding authors. Nuffield Department of Medicine, Centre for Human Genetics, Oxford-GSK Institute of Molecular and Computational Medicine (IMCM), University of Oxford, Oxford, OX3 7BN, United Kingdom. E-mails: isar.nassiri@well.ox.ac.uk (I.N.) and paolo.piazza@well.ox.ac.uk (P.P.)

Associate Editor: Magnus Rattray

Abstract

Motivation: Pooled designs for single-cell RNA sequencing, where many cells from distinct samples are processed jointly, offer increased throughput and reduced batch variation. This study describes expression-aware demultiplexing (EAD), a computational method that employs differential co-expression patterns between individuals to demultiplex pooled samples without any extra experimental steps.

Results: We use synthetic sample pools and show that the top interindividual differentially co-expressed genes provide a distinct cluster of cells per individual, significantly enriching the regulation of metabolism. Our application of EAD to samples of six isogenic inbred mice demonstrated that controlling genetic and environmental effects can solve interindividual variations related to metabolic pathways. We utilized 30 samples from both sepsis and healthy individuals in six batches to assess the performance of classification approaches. The results indicate that combining genetic and EAD results can enhance the accuracy of assignments (Min. 0.94, Mean 0.98, Max. 1). The results were enhanced by an average of 1.4% when EAD and barcoding techniques were combined (Min. 1.25%, Median 1.33%, Max. 1.74%). Furthermore, we demonstrate that interindividual differential co-expression analysis within the same cell type can be used to identify cells from the same donor in different activation states. By analysing single-nuclei transcriptome profiles from the brain, we demonstrate that our method can be applied to nonimmune cells.

Availability and implementation: EAD workflow is available at https://isarnassiri.github.io/scDIV/ as an R package called scDIV (acronym for single-cell RNA-sequencing data demultiplexing using interindividual variations).

1 Introduction

Although single-cell analyses are beginning to unravel the molecular aetiology of diseases, most studies incorporate average gene expression profiles across bulk tissues, which frequently mask variation between individuals visible at single-cell resolution (van der Wijst *et al.* 2018, Yazar *et al.* 2022). The variability between individuals remains an understudied aspect of co-expression relationships between molecules at the single-cell level. Establishing molecular co-dependencies may break down due to genetic variation, uncoupling of transcription and splicing, or it could be caused by pathological insult. The regulation of downstream markers may be disrupted by altered interindividual co-expression relationships, which in turn affect the expression pattern of entire biological pathways (Johansen *et al.* 2023).

Received: May 11, 2024; **Editorial Decision:** May 29, 2024; **Accepted:** June 7, 2024

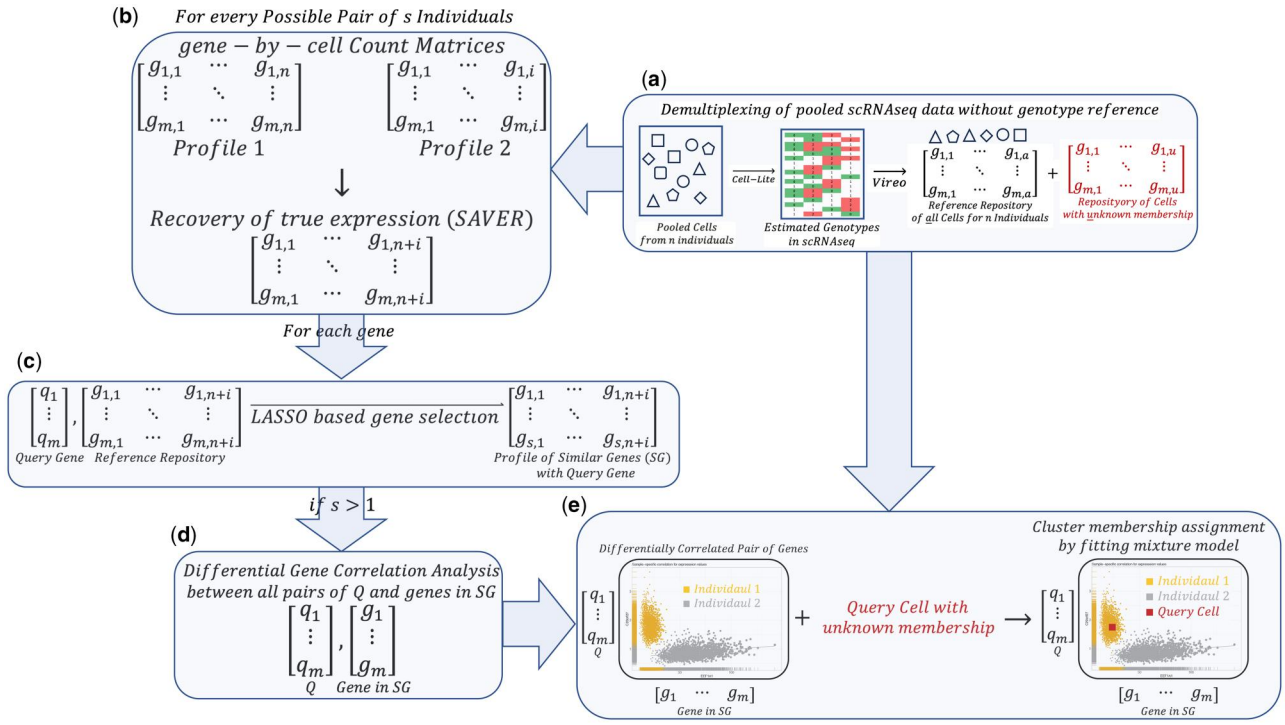


Figure 1. Workflow for computational demultiplexing of unrelated individuals in scRNA-seq. (a) First, we demultiplex pooled samples (vireo) (Huang *et al.* 2019) using genetic differences inferred from scRNA-seq data (cellsnp-lite) (Huang and Huang 2021). (b) Next, for each pair of individuals in the output of genetic demultiplexing, we estimate accurate gene expression values for all genes. (*g*) per cell using the gene expression recovery for single-cell RNA sequencing (SAVER) (Huang *et al.* 2018). (c) We apply the LASSO to obtain the most representative subset of genes (*s*) (Nassiri and McCall 2018) related to query gene (*q*). (d) We apply differential gene correlation analysis between pairs of query gene (*q*) and related genes selected by LASSO (*SG*), to identify the top first DCE genes interindividual (see section 2). (e) The co-expression patterns of the top first DCE genes (*q* and *g*) are used to fit a mixture model and reconstruct the sample identity of each cell.

Previous studies have shown that SNPs can alter co-expression relationships on an individual donor basis by looking for evidence of allele-specific correlation (Fairfax *et al.* 2012, van der Wijst *et al.* 2018, Oelen *et al.* 2022). Here, we develop a computational method, expression-aware demultiplexing (EAD), that harnesses variation of interindividual co-expression signatures at single-cell resolution to stratify donors and improve the demultiplexing of pooled scRNA-seq data (Fig. 1).

The process of separating cells from multiple samples pooled in a single batch is referred to as demultiplexing in this article. Multiplexing can increase the number of donors that can be tested, reduce experiment costs, address batch effects, and make large-scale sample operations feasible. Multiple approaches can be employed to demultiplex pooled single-cell gene expression profiles, including DNA oligonucleotide tagging and natural genetic variation (Huang *et al.* 2019).

Inspired by the differential co-expression (DCE) analysis of two genes between biological groups, we introduce an optimized statistical method for donor stratification guided by variation of co-expression signatures at single-cell resolution (McKenzie *et al.* 2016).

Given cells with known sample identity, we demonstrate that differences in gene–gene relationships exist between individuals by providing multiple examples. The results show top interindividual DCE genes provide a distinct cluster of cells per individual and display the enrichment of cellular macromolecular super-complexes, or organelles related to metabolism (e.g. mitochondria and ribosomes).

We applied this approach to samples from isogenic inbred mice and showed that by controlling genetic and environmental effects, we can resolve interindividual variations related to metabolic pathways.

2 Methods

2.1 Sources of data

Sequencing data for evaluation and application covers 124 10× single-cell RNA-seq samples in 14 batches and 30 bulk RNA samples.

During library preparation for the first and second datasets (see sections 1.1.1 and 1.1.2), every sample (a group of cells that originated from the same donor) was assigned a specific index sequence. Sample indices were incorporated into the sequencing primers on Illumina sequencers. After sequencing, by using the 10× mkfastq pipeline the names of sample index sets were automatically identified and reads per sample were merged into the FASTQ files. Supplementary Tables 4 and 5 included reference tables with the complete set of index codes per donor. Accuracy and reliability of the embedded sample index sequences are hallmarks of the 10× Genomics platform. This provides a solid foundation for evaluating the performance of demultiplexing algorithms without introducing errors from the ground truth itself.

2.1.1 Dataset 1

Eight metastatic melanoma (MM) patients were included in the first scRNA-seq data set, which contains human peripheral blood mononuclear cells (PBMCs) with known donor labels (Fairfax *et al.* 2020). In this dataset, sample indices were used to label cells per sequencing sample without errors. By using sample/donor indices, we were able to map

demultiplexed cells using the expression-aware approach to their original sample donors and assess their performance in singlet assignment (Supplementary Table 1). Two sample pools cover 16 samples pre- and posttreatment at day 21 (eight pre- and eight posttreatment) with immune checkpoint blockade therapies including Nivolumab (NIVO) + ipilimumab (IPI) or Pembrolizumab (Pembro) (Fairfax *et al.* 2020). We chose this dataset to show that interindividual differences and no other confounding factors, such as cell type, underlie differential gene co-expression across individuals. We used the Cell Ranger pipeline (v7.0.1), the GRCh38 reference genome, and 5′ R2-only chemistry to process the data set. The analysis detected up to 36 842 cells per pool, 56 850 mean reads per cell, and 1768 median genes per cell. Raw data for single-cell sequencing datasets have been deposited on the European Genome–phenome Archive (https://ega-archive.org/studies/EGAS00001004081).

2.1.2 Dataset 2

The second scRNAseq data set consists of a pooled sample with known donor labels from six isogenic mice. The samples were exposed to a topical TLR7 agonist, Imiquimod, to induce a systemic lupus erythematosus-like phenotype or vehicle control. In this dataset, cells were labelled per sequencing sample using sample indices (Supplementary Table 3). We applied this dataset to show that by controlling genetic and environmental effects, we can resolve interindividual variations related to metabolic pathways. The Cell Ranger pipeline, mm10-2020-A reference genome, and single-cell 3′ v3 chemistry were employed for data processing. The analysis detected up to 12 685 cells per pool, with an average of 40 101 mean reads per cell and 1553 median genes per cell.

2.1.3 Dataset 3

The third data include five single-cell multi-omics (RNA-seq + ATAC-seq) batches with unknown donor labels. The dataset consists of circulating haematopoietic progenitor cells samples from seven healthy controls, 15 sepsis patients, and eight convalescent samples (six samples per batch) (Kwok *et al.* 2023). The genetic demultiplexing pool samples were based on the use of an extra 30 bulk RNA-sequencing profiles from the same individuals. This dataset was chosen to evaluate the utility of EAD for challenging biological models involving a mixture of heterogeneous cell types and donors. The data set was processed using the Cell Ranger pipeline, GRCh38 reference genome, and Multiome chemistry. The analysis detected up to 14 806 cells per pool, 49 892 mean reads per cell, and 3250 median genes per cell. Raw data for single-cell sequencing datasets have been deposited on the European Genome–phenome Archive (https://ega-archive.org/studies/EGAS00001006283) (Kwok *et al.* 2023).

2.1.4 Dataset 4

The fourth dataset includes a scRNA-seq of human MAIT cells (Garner *et al.* 2023). We used a subset of the original dataset, comprising five channels of a Chromium Next GEM Chip K. MAIT cells from three donors were either left unstimulated or activated with a TCR, cytokine, or dual TCR + cytokine stimulus. Cells from each donor-condition combination were labelled with TotalSeq-C hashtag antibodies (12 total), pooled, and split across the eight channels of the Chromium Chip (Garner *et al.* 2023). The Cell Ranger (v7.0.1) count pipeline was used to process FASTQ files for gene expression. Next, hashtag oligo demultiplexing (HTO) and the extraction of singlets, doublets, and negative cells from multiplexing experiments were performed using the HTODemux algorithm (Stoeckius *et al.* 2018) implemented in the Seurat tool (Hao *et al.* 2024). We chose to utilize this dataset to demonstrate how interindividual DCE patterns can enhance cell hashing demultiplexing outcomes. The analysis detected up to 16 251 cells per pool, 53 413 mean reads per cell, and 3525 median genes per cell. Raw data for single-cell sequencing datasets have been deposited on the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE194187 and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE194189) (Garner *et al.* 2023).

2.1.5 Dataset 5

The fifth dataset includes ventral and dorsal tiers of the substantia nigra (SN) and the cortex (middle frontal gyrus) from five healthy donors (Agarwal *et al.* 2020). The total number of samples is 12, which includes two replicate samples of SN. The sample libraries contain genes that vary by nuclei from 607 to 3364, and mean Reads that range from 18 377 to 59 513 in both regions (Agarwal *et al.* 2020). To obtain processed single-nuclei RNA-sequencing matrices, use the accession code https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE140231 from the Gene Expression Omnibus. This dataset was used to show that our method can be applied to nonimmune cells.

2.2 Quantification and gene expression analysis

Sequencing data was processed by Cell Ranger (v7.0) pipelines to create a feature barcoding and gene expression library. The filtered gene–cell matrix generated by the cell ranger was converted from an HDF5 gene–cell matrix to a gene–cell count matrix using the cell ranger mat2csv command provided by 10× genomics. We applied the scater package to filter out single-cell profiles that were outliers for any metrics, as they are considered low-quality libraries (McCarthy *et al.* 2017). The SCRAN package was applied to detect and remove doublets using expression profiles as described in Dahlin *et al.* (2018). Cell type annotation on scRNA-seq data was performed using the scQCEA tool (Nassiri *et al.* 2023).

2.3 Gene expression recovery

Single-cell RNA-sequencing studies only sequence a small portion of the transcripts present in each cell. This leads to inaccurate quantification of genes with low or moderate expression levels. We used expression recovery methods to provide accurate expression estimates for all genes in order to address this challenge (Huang *et al.* 2018). The SAVER method was our preference because it recovers the relationship between two marker genes, which we are aware do not correlate (Huang *et al.* 2018).

2.4 Identification of interindividual variation in gene expression

We introduce an optimized statistical method to detect DCE patterns using single-cell data from two individuals (McKenzie *et al.* 2016).

We use cellsnp-lite (v1.2.2) (Huang and Huang 2021) and vireo (vireoSNP/0.3.2) (Huang *et al.* 2019) to infer genetic variants from scRNA-seq data in the first step. Gene–cell

count matrices for all possible pairs of individuals are generated using vireo's best-proposed donor for each cell (or alternative resources, including the use of HTO). The SAVER (Huang *et al.* 2018) tool is used to transform the gene–cell count matrix per pair of individuals (see section 1.2). The glmnet R package (Engebretsen and Bohlin 2019) is used to select representative gene subsets in the expression profile for each pair of individuals. The calculation of correlation coefficients (r) for all possible pairs of selected genes is done using lasso. Interindividual differential gene correlation analysis (IDCA) was conducted for two donors ($D1$ and $D2$) and genes ($G1$ and $G2$) using correlation coefficients as described below.

The Fisher Z-transformation was applied to stabilize the variance of Spearman's rank correlation coefficients (r_{D1} *and* r_{D2}) (McKenzie *et al.* 2016):

$$z = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right) \quad (1)$$

The calculation of the difference in z-scores (dz) between two donors (*e.g.* r_{D1} *and* r_{D2}) was done using (McKenzie *et al.* 2016):

$$dz = \frac{(z_1 - z_2)}{\sqrt{|\text{var}\ (r_{s_{D1}}) - \text{var}\ (r_{s_{D2}})|}} \quad (2)$$

$var(r_{s_x})$ is the variance of z for the group of cells with the identical donor (s_x). The proposed method was implemented as an R package (see section 2.9 for more details).

2.5 Cell assignment to donors using a bivariate mixture model

We use a bivariate mixture model for uncovering correlation classes (donors) for each pair of differentially correlated genes. We reconstruct the identity of each cell by fitting a mixture model. The bivariate Gaussian mixture model is used to measure the joint variability between the expression profile of an indicated cell and two donors (donor label proposed by genetic-based demultiplexing) using maximum posterior probability. We applied flexmix R package(Grun and Leisch 2008) to perform mixtures of regression models using the Expectation Maximization algorithm and model-based clustering.

In order to reconstruct the sample identity of each cell, a mixture model is fitted using the co-expression patterns of the top DCE genes. A mixture model is utilized to predict a cluster of cells that are not labelled and to reconstruct their sample identity. We consider all possible donor pairs for an indicated cell. This procedure is repeated for the top 10 genes that are differentially co-expressed. We verify that the cell is from a donor if we are able to assign it to the donor most of the time (number of assignments to an indicated donor equal or greater than the total number of pairs of donors minus 1). Reconstructing the sample identification of cells can be done using more liberal thresholds, such as the total number of pairs of donors minus 2.

2.6 Calling genotypes from bulk RNA-sequencing data

Bulk RNA-seq profiles of 30 samples were used for genetic variant calling using GATK (Genome Analysis Toolkit) (Deelen *et al.* 2015). Our first step was to align the bulk RNA-seq reads to the hg38 reference genome using the Burrows-Wheeler Alignment tool (Li and Durbin 2009). Next, we used GATK's HaplotypeCaller tool to find genetic variants in the aligned reads (Van der Auwera *et al.* 2013). Using BCFtools, we removed false positives and low-quality variants from the variants after calling them (Li 2011). Bulk RNA-Seq genotypes were inputted into the demuxlet tool with a posterior probability (PRB = 1) of singlet assignment (Kang *et al.* 2018).

2.7 Calling genotypes from scRNA-sequencing data

The first step in assigning genetic donors to samples is to perform SNP genotyping using CellSNP-Lite (v1.2.3) (Huang and Huang 2021) in a given data set. We followed the cellSNP-lite manual's recommended default parameters. Using the cell data from cellSNP-Lite as input, we demultiplexed using vireoSNP (v0.5.8) (Huang *et al.* 2019).

2.8 Data presentation

The pathway enrichment analysis (Wu *et al.* 2021) of interindividual DCE genes (McKenzie *et al.* 2016) was performed using R packages. Box plots and dot plots were generated using ggpubr (v0.2) and customizing ggplot2 (Almeida *et al.* 2018).

2.9 Implementation

EAD workflow is available at https://isarnassiri.github.io/scDIV/ as an R package called scDIV (acronym for Single-Cell RNA-sequencing data Demultiplexing using Interindividual Variations). Our implementation with variable selection and proper data structures has made the EAD computationally efficient and can be run on a laptop with 16 Gb of memory and two 3.5-GHz CPUs. To run the tool on multiple servers simultaneously for large datasets, users should use a shell script. The package website has documentation that includes examples.

3 Results

We developed a generic five-step workflow for demultiplexing scRNA-seq data using interindividual variation in gene expression (Fig. 1). First, we infer genetic variants from scRNA-seq data (Huang and Huang 2021) and demultiplex pooled samples (Fig. 1a). Gene–cell count matrices are generated for all possible pairs of individuals by utilizing the best-proposed donor for each cell in the previous step. The gene–cell count matrix is transformed by an expression recovery method per pair of individuals to provide precise gene expression values for all genes per cell (Fig. 1b) (Huang *et al.* 2018). We apply lasso (least absolute shrinkage and selection operator) to find compact and representative gene subsets in the expression profiles to improve the accuracy and reduce the redundant downstream number of computations steps (Fig. 1c) (Nassiri and McCall 2018, Yang *et al.* 2021). Next, we performed an individual-specific co-expression analysis searching for altered co-expression patterns of gene pairs interindividual (Fig. 1d) (see section 2). A mixture distribution of correlation classes for each pair of differentially correlated genes is used to fit a mixture model. We reconstruct the identity of each cell based on the similarity of their gene expression with donor-specific clusters using the mixture model (Fig. 1e). Cells are assigned if expression-aware and genetic-based demultiplexing propose the same best singlet (Fig. 2 and Supplementary Fig. 1). In this way,

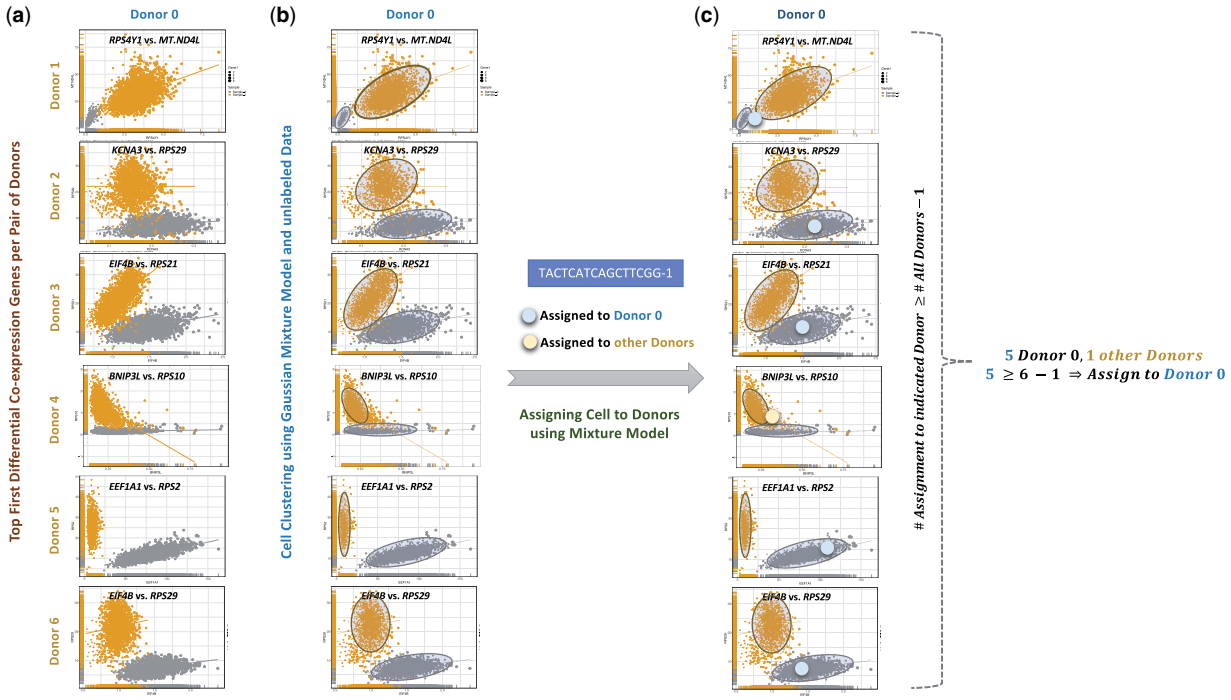


Figure 2. An example of EAD is to assign an indicated cell to one of seven individuals using a mixture model. Genetic-based demultiplexing already suggested that the cell most likely belongs to donor-0 using the partial genotypic data from state of the individuals in a pool of donors. Now, using the gene expression profiles, we want to check the best-guess assignment obtained from genetic-based demultiplexing. (a) We select the top first pair of DCE genes per donor-0 compared to other donors. The expression pattern of a pair of DCE genes is used to create distinct clusters of cells across individuals. (b) We use a mixture model to predict a cluster of unlabelled cells and reconstruct their sample identity. (c) For an indicated cell (TACTCATCAGCTTCGG-1), we consider all possibilities and most of the time the cell is assigned to donor 0. We confirm that the cell belongs to donor 0 if we successfully assign it to donor 0 for an equal or greater number of pairs of donors, minus 1.

the combined method can achieve greater accuracy than the SNP-based or barcode-based methods by selecting the element-wise maximum of the two demultiplexing results.

There are multiple methods that can be used to label cells or nuclei with antibodies-based oligonucleotides (Stoeckius *et al.* 2018, Gaublomme *et al.* 2019). By sequencing the cells' RNA molecules with HTOs, a matrix of count for HTOs per droplet is produced and utilized for demultiplexing of the sample pools. Barcode-based multiplexing methods may be used to combine cells from the same donor and different stimulation conditions into sample pools (Garner *et al.* 2023). In these circumstances, the genetic-based demultiplexing cannot differentiate between two samples that came from the same donor but had different stimulation conditions. Our EAD workflow uses alternative methods like HTODemux (Stoeckius *et al.* 2018) to estimate gene expression for each pair of individuals (Fig. 1b) in these cases. In order to evaluate EAD, we utilize various datasets that employ diverse methods to label cells in scRNA-seq data for each sequencing donor. These approaches involve sample indices, genetic variation derived from bulk RNA-sequencing profiles of the same samples, and HTOs.

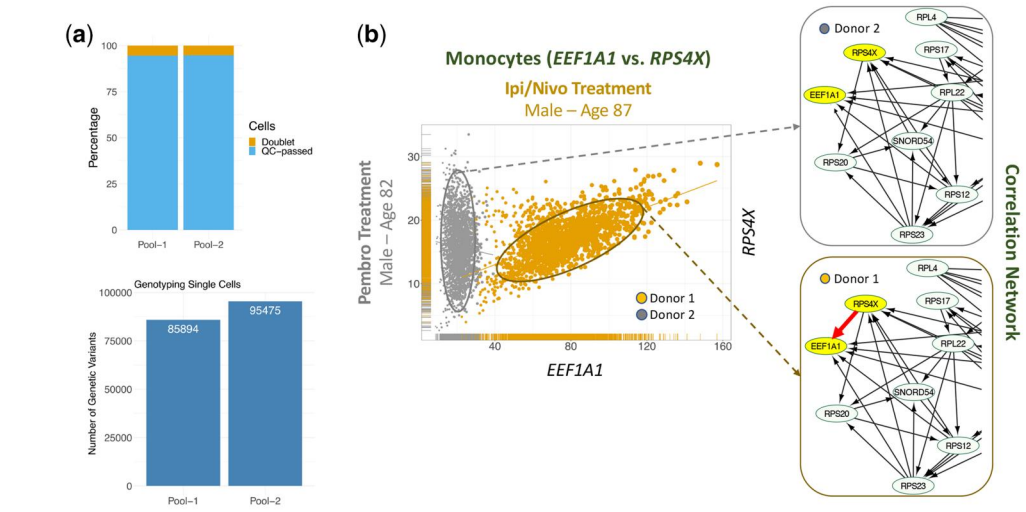
3.1 Model validation using synthetic sample pools

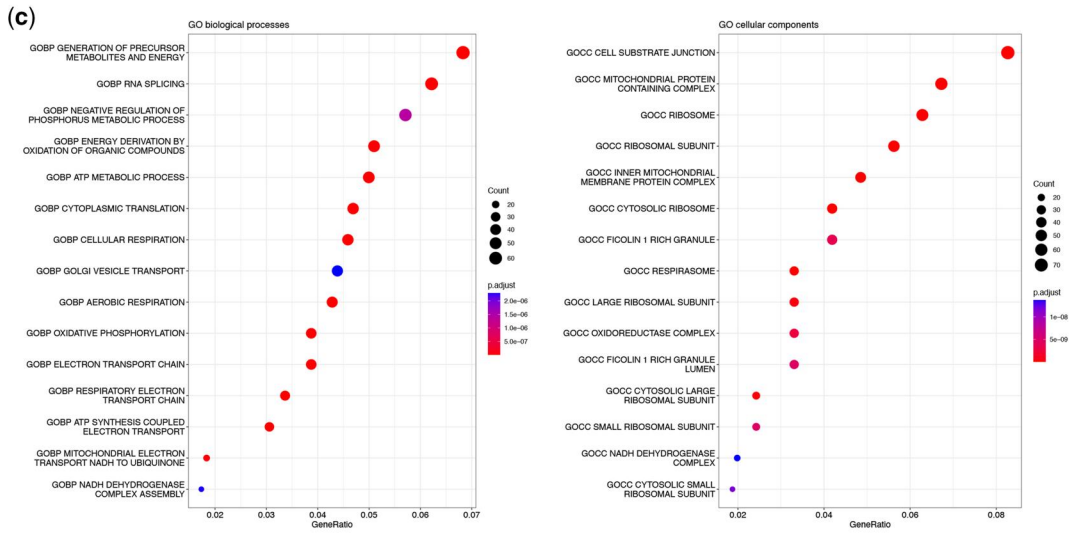
Our initial assessment of EAD was based on synthetic sample pools from eight MM patients (Fairfax *et al.* 2020). To perform a preliminary evaluation of EAD, we mixed intact donor cells, and refer to them as the 'synthetic sample pool'. A unique identification number was given to every cell in the synthetic sample pool to reveal its true origin. By using mixed cells with known donor identities, we were able to evaluate the performance of demultiplexing algorithms without introducing errors from the ground truth itself.

Peripheral blood samples were obtained from patients with MM who were treated with immune checkpoint blockade (Fairfax *et al.* 2020). Samples were collected both before and after the initial treatment cycle. PBMCs were used to isolate monocyte and T cells. Monocyte and T cells were mixed in suspension and the Chromium 10× system was utilized to process the single-cell transcriptome. During library preparation for scRNA sequencing, libraries were tagged using unique indices per donor (Supplementary Table 1). Each sequencing run involved pooling (multiplexing) multiple libraries and sequencing them together (Fairfax *et al.* 2020). The sequencing run concluded with demultiplexing, and the reads produced were separated into different FASTQ files according to donor indices.

Sorting a pile of laundry (sequencing reads) with a tag (donor index) that identifies the owner (donor) is an analogy to demultiplexing in sequencing. This is not in line with the definition of single-cell sample demultiplexing in this article. In this article, demultiplexing is similar to sorting a pile of laundry (sequencing reads) without tags for each item, by incorporating other indicators like size or personal clothing style (genetic variation) (Howitt *et al.* 2023).

We demultiplex monocyte and T cells separately in each synthetic sample pool (Fig. 3a and b). In the original paper, subsetting was performed to select T cells expressing CD8A, CD8B, and CD3D, and monocytes expressing CD14. We applied further subsetting to eliminate heterogeneous cell populations, including monocytes expressing CD3D, CD3E, CD3G, CD8E, or CD19, and T cells expressing CD14 or CD19.





(d)

T Cells

Sample	Cells	Prevalence-EAD	Prevalence-GD	Prevalence-mixedD	Prevalence-Improvement
1A	2303	0.99826	0.99783	1.00000	↑ 0.00217
1B	3096	0.99839	0.99451	0.99935	↑ 0.00484
2A	597	0.93300	0.98995	0.99665	↑ 0.0067
2B	1571	0.97199	0.99554	1.00000	↑ 0.00446
3A	2145	0.99814	0.99347	1.00000	↑ 0.00653
3B	2179	0.98990	0.99771	1.00000	↑ 0.00229
4A	1167	0.98458	0.99486	1.00000	↑ 0.00514
4B	570	0.97193	0.99474	0.99825	↑ 0.00351
5A	1539	0.93502	0.99025	0.99675	↑ 0.0065
5B	1659	0.97227	0.98975	0.99879	↑ 0.00904
6A	64	0.00000	0.96875	0.98438	↑ 0.01562
6B	2713	0.99668	0.99558	1.00000	↑ 0.00442
7A	1426	0.99649	0.99439	1.00000	↑ 0.00561
7B	1182	0.99746	0.99154	0.99915	↑ 0.00761
8A	2241	0.99822	0.99554	1.00000	↑ 0.00446
8B	1764	0.99546	0.99546	1.00000	↑ 0.00454

Monocytes

Sample	Cells	Prevalence-EAD	Prevalence-GD	Prevalence-mixedD	Prevalence-Improvement
1A	1783	0.96971	0.99215	0.99888	↑ 0.00673
1B	1901	0.98527	0.99264	0.99895	↑ 0.00631
2A	467	0.83940	1.00000	1.00000	↓ 0
2B	1155	0.92727	0.98961	0.99654	↑ 0.00693
3A	1190	0.96891	0.99412	1.00000	↑ 0.00588
3B	1548	0.97739	0.99548	0.99935	↑ 0.00388
4A	749	0.97330	0.99466	0.99733	↑ 0.00267
4B	759	0.95916	0.99473	1.00000	↑ 0.00527
5A	1298	0.94145	0.99307	0.99923	↑ 0.00616
5B	943	0.96925	0.99258	0.99788	↑ 0.0053
6A	631	0.99049	0.98891	0.99842	↑ 0.00951
6B	2030	0.98374	0.99113	1.00000	↑ 0.00887
7A	1602	0.98065	0.99313	0.99875	↑ 0.00562
7B	2367	0.99873	0.99916	1.00000	↑ 0.00084
8A	1661	0.97471	0.99699	1.00000	↑ 0.00301
8B	1595	0.99310	0.99687	1.00000	↑ 0.00313

Figure 3. (a) The results of quality control and the number of called variants for two single-cell sample pools made of 16 samples. (b) Example of top DCE genes (EEF1A and RPS4X) in PBMCs, distinguishes clusters of cells per individual. Pairwise correlations can be visualized as a network. The result shows that EEF1A and RPS4X display co-expression only in donor 1 (red arrow), which could not have been detected based on all individual cells or donor 2. (c) The results of ontology gene set enrichment analysis show a significant association of interindividual DCE genes with the regulation of metabolic processes. The enrichment analysis of cellular components in the dot plot shows associations with mitochondria, ribosomes, cellular macromolecular super complexes, or organelles related to metabolism. (d) Comparison between the methods in terms of prevalence (abbreviations: EAD, expression-aware demultiplexing; GD, genetic-based demultiplexing; mixed, a combination of ED and GD results). According to prevalence, a combination (mixed) of GD and EAD leads to better results for Monocytes and T cells. The colour density reflects the continuous data range to compare values. Lower values are shown to be the most profitable. The range of prevalence values is between 0 and 1.0.

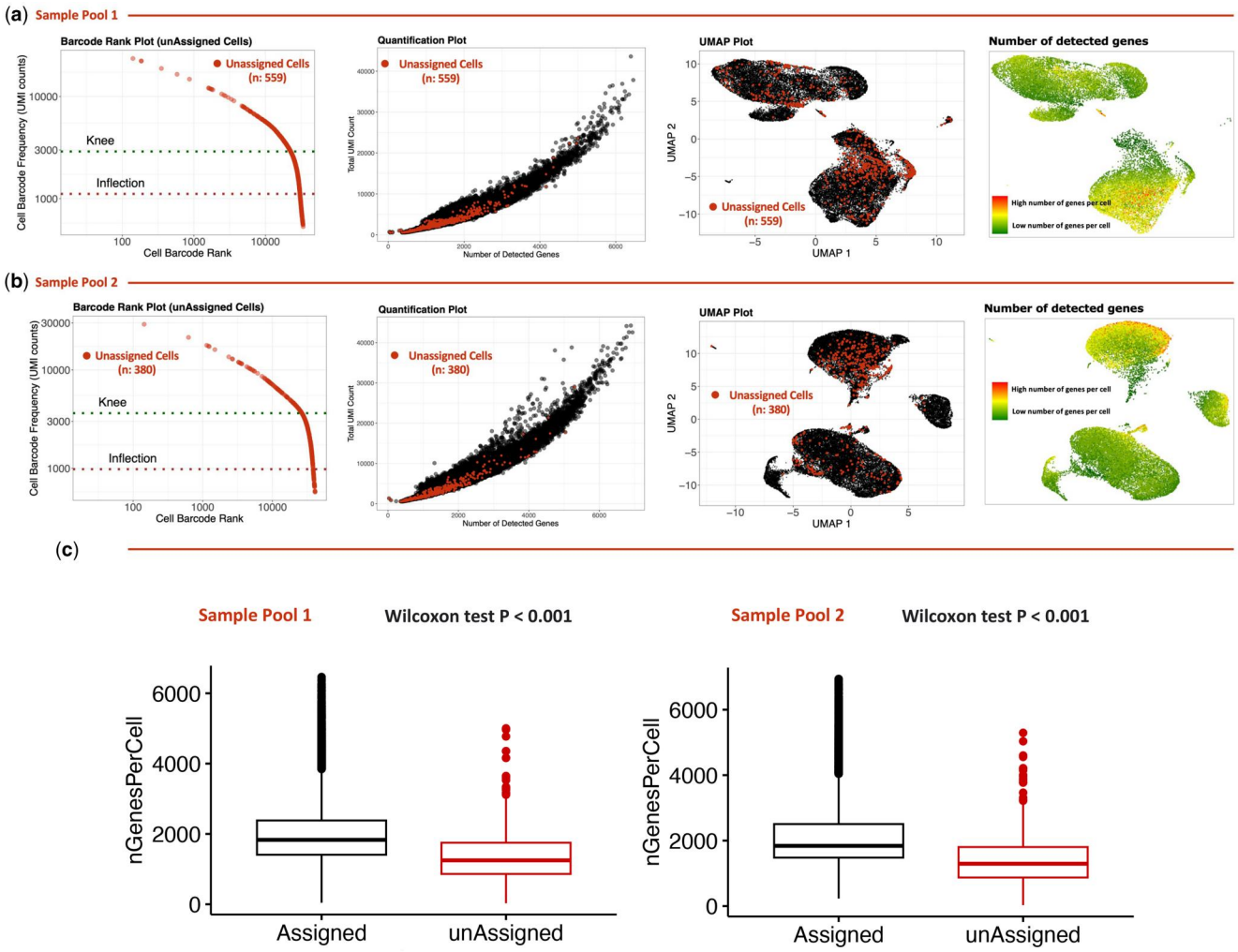


Figure 4. (a, b) The knee, quantification, and UMAP plots show the location of unassigned cells in two sample pools. We ensure that the background red-coloured cells appear on top by dividing the points into different layers and plotting the red points after the black points (Nassiri *et al.* 2023). Number of detected gene plots shows an association between cell assignment and the number of genes per cell. (c) Boxplots show a significant difference in the mean number of genes per cell across the classes of assigned and unassigned cells using EAD.

We compared the performance of overall demultiplexing matrices generated based on genetic-based (GD), EAD, and mixed methods for demultiplexing results against that from the known sample labels (Fig. 3d and Supplementary Table 2). We accept the best singlet proposed by GD for unassigned cells if it is consistent with the donor proposed by EAD and call it mixed demultiplexing (Huang *et al.* 2019).

To calculate the accuracy and evaluate the results, we define true positive (TP) as a donor assignment result that correctly assigns the donor according to the known label. The term false positive refers to a donor assignment result that incorrectly assigns the donor based on the known label. True negative (TN) is characterized by an unassigned donor where the best guess incorrectly suggests a particular donor based on the known label. A false negative is defined as an unassigned donor result that correctly suggests a particular donor based on the known label.

The results showed that the combination of GD and EAD results improves the prevalence and harmonic mean of precision and sensitivity (F1 score) (Fig. 3d and Supplementary Table 2). The accuracy of the two approaches is comparable, and mixed results improve accuracy (Supplementary Table 2).

Enrichment analysis revealed that genes that are differentially co-expressed between individuals are strongly linked to the regulation of metabolic processes and represent metabolic differences between individuals (Fig. 3c). Therefore, our method to detect interindividual variation in gene expression could be applied to provide insight into challenges presented by interindividual differences in the responses to nutrition and obesity, cardiovascular and endocrine research, and comprehending the alterations that occur with age and the ensuing neurodegenerative conditions (Manach *et al.* 2017, Lotta *et al.* 2021, Johansen *et al.* 2023). Figure 3b shows an example of top DCE genes (*EEF1A* and *RPS4X*) associated with memory decline in normal aging and Alzheimer's disease, which distinguishes clusters of cells per individual (Beckelman *et al.* 2016a,b). Only donor 1 showed co-expression of *EEF1A* and *RPS4X*, which was not present in all cells or donor 2. Changes in interindividual co-expression relationships could have an impact on the regulation of downstream markers whose expression pattern influences entire biological pathways.

We found that 939 cells were not assigned to any donor. The location of detected unassigned cells on the knee plot showed aggregation after the inflection point (Fig. 4a and b). In addition, unassigned cells were mostly found in the bottom-left corner of the quantification plot, as shown in

Fig. 4. It means filtering out nonrelevant cells can improve the accuracy of demultiplexing, especially for samples with a low number of cells. In general, the UMAP projection and number of genes per cell plots showed that unassigned cells in EAD tend to be assigned to regions with low number of genes per cell (Wilcoxon test $P < .001$) (Fig. 4).

We were able to control confounding factors by removing single-cell profiles that did not fit any metric and *in vitro* separating T cells and monocyte cells. This allowed us to show that patterns of differential gene co-expression across individuals are the consequence of interindividual gene expression differences and no other factors such as cell type (Fig. 3b).

Synthetic sample pools are valuable for evaluating scRNA-seq demultiplexing approaches, but they have some shortcomings to consider. The complexity of real pooled single-cell RNA-seq data systems may not be fully captured by synthetic sample pools. Furthermore, a synthetic sample pool may not cover the full range of technical variations that can happen during actual sample pooling. If the synthetic sample pool does not fully replicate real-world scenarios, the observed patterns could reflect the limitations of the synthetic model.

We further evaluate the effectiveness of EAD using several heterogeneous cell populations in the following sections.

In this model, both treatment group and sex might be expected to alter the immune transcriptome. Therefore, we stratified individuals based on the type of treatment and sex as confounding variables to consider their influence on DCE patterns (Bongen *et al.* 2019). We found no consensus interindividual DCE patterns for matched pairs based on sex (e.g. Male and Male) or type of treatment (e.g. Pembro and Pembro) (Fig. 5a).

We evaluated the effectiveness of the expression-aware approach in demultiplexing cells from isogenic individuals. We pooled cells with a known label from six isogenic mice exposed to a topical TLR7 agonist Imiquimod to induce a systemic lupus erythematosus-like phenotype or vehicle control and tried to perform demultiplexing using the expression-aware approach (Fig. 5b and c). Sample indices were used to label cells per sequencing sample/donor in this dataset (Supplementary Table 3).

Interindividual differences were not discernible through DCE patterns across classes of isogenic samples (Fig. 5c). The results indicate that it is challenging to identify variations in metabolic pathways for sample pools with very little genetic diversity among donors. Therefore, demonstrating the efficacy of expression-aware approach for demultiplexing these pooled samples was not possible.

The assumption that genetic and environmental differences lead to DCE patterns between individuals is supported by the absence of individual-specific signatures in isogenic laboratory mice (Figs 3 and 5c). These variables can be controlled to resolve interindividual differences in gene co-expression related to metabolic pathways (Fig. 5c). In conclusion, nonisogenic individual samples are suitable to be pooled for demultiplexing scRNA-seq data using interindividual variations in genetic and gene expression.

3.2 Application to real pooled single-cell RNA-seq

We considered more challenging scenarios involving five pooled sample (batch) and 30 donors (six donors per batch) to test if the method would work well with more heterogeneous cell populations (Fig. 6a and b) (Kwok *et al.* 2023). From every donor, the nuclei of CD34+ circulating hematopoietic stem and progenitor cells (HSPCs) were extracted. Six samples with equal numbers of cells were combined to form a single pool. Each pool was then subjected to cell lysis and nuclear extraction. RNA libraries were sequenced after loading each pool of nuclei across four channels of the 10× genomics lane on the chip (Kwok *et al.* 2023). In this context, a 10× lane would represent a batch loaded onto the chip. To differentiate this dataset from other datasets, we use batch and 10× lane interchangeably.

10× lanes in the scRNA-seq experiment of HSPCs included samples from healthy donors, sepsis donors, and donors with both sepsis symptoms and COVID-19 infection (Kwok *et al.* 2023).

There are four methods that can demultiplex pooled single-cell RNA-seq without relying on reference SNP genotypes: vireo (Huang *et al.* 2019), scSplit (Xu *et al.* 2019), Freemuxlet (Kang *et al.* 2018), and Souporcell (Heaton *et al.* 2020). According to available benchmark studies, vireo outperforms other tools when it comes to demultiplexing pooled single-cell RNA-seq data without genotype reference (Neavin *et al.* 2024, Cardiello *et al.* 2023). Furthermore, vireo outputs have a high level of consistency with the outcomes of other tools (Neavin *et al.* 2024, Cardiello *et al.* 2023). Therefore, we opted for vireo as the genotype-free demultiplexer for pooled single-cell RNA-seq.

We utilized the Demuxlet tool and reference genotypes derived from bulk RNA-sequencing profiles of the same samples to determine the donor identity of every singlet (Kang *et al.* 2018). SNPs in each individual's genome were used in demultiplexing with genotype reference (Demuxlet) to identify individual donor identities per cell with high accuracy. This allows for a precise assessment of how other algorithms compare to this ground truth. The performance of EAD and GD (GD—vireo) (Huang *et al.* 2019) methods for demultiplexing, which do not require a reference SNP genotypes, was evaluated in the next step using Demuxlet results.

The TP rate is determined by the fraction of cells with the corresponding genetic assignment (Demuxlet) for each possible EAD, Mixed-D or GD assignment. The fraction of cells with EAD, Mixed-D or GD assignments that are different from their genetic assignments is called the false positive rate (FP). The false negative rate (FN) is the proportion of cells that have genetic assignment and EAD, Mixed-D or GD un-assignment, and the best guess of EAD, Mixed-D or GD accurately indicates a specific donor. The TN rate is the proportion of cells that have genetic un-assignment and EAD, Mixed-D or GD un-assignment.

We used the sensitivity and F-score to evaluate performance as our main metric. The proportion of positive cases that are correctly identified by the evaluation method is represented by sensitivity. The F-score refers to the harmonic mean of precision and recall, which can range from zero to one. A higher F-score indicates superior performance.

Six samples from genetically distinct donors were represented by cells in each of the five 10× lanes in the data set. The data from each 10× lane was subjected to demultiplexing methods. A cell could be assigned to singlet, a singlet that corresponds to one of the six unique samples or unassigned by each demultiplexing method. Misclassifying true singlets as doublets can be a more significant error source in downstream analysis than misclassifying true doublets (Wolock *et al.* 2019, Howitt *et al.* 2023). Therefore, for vireo, we treat

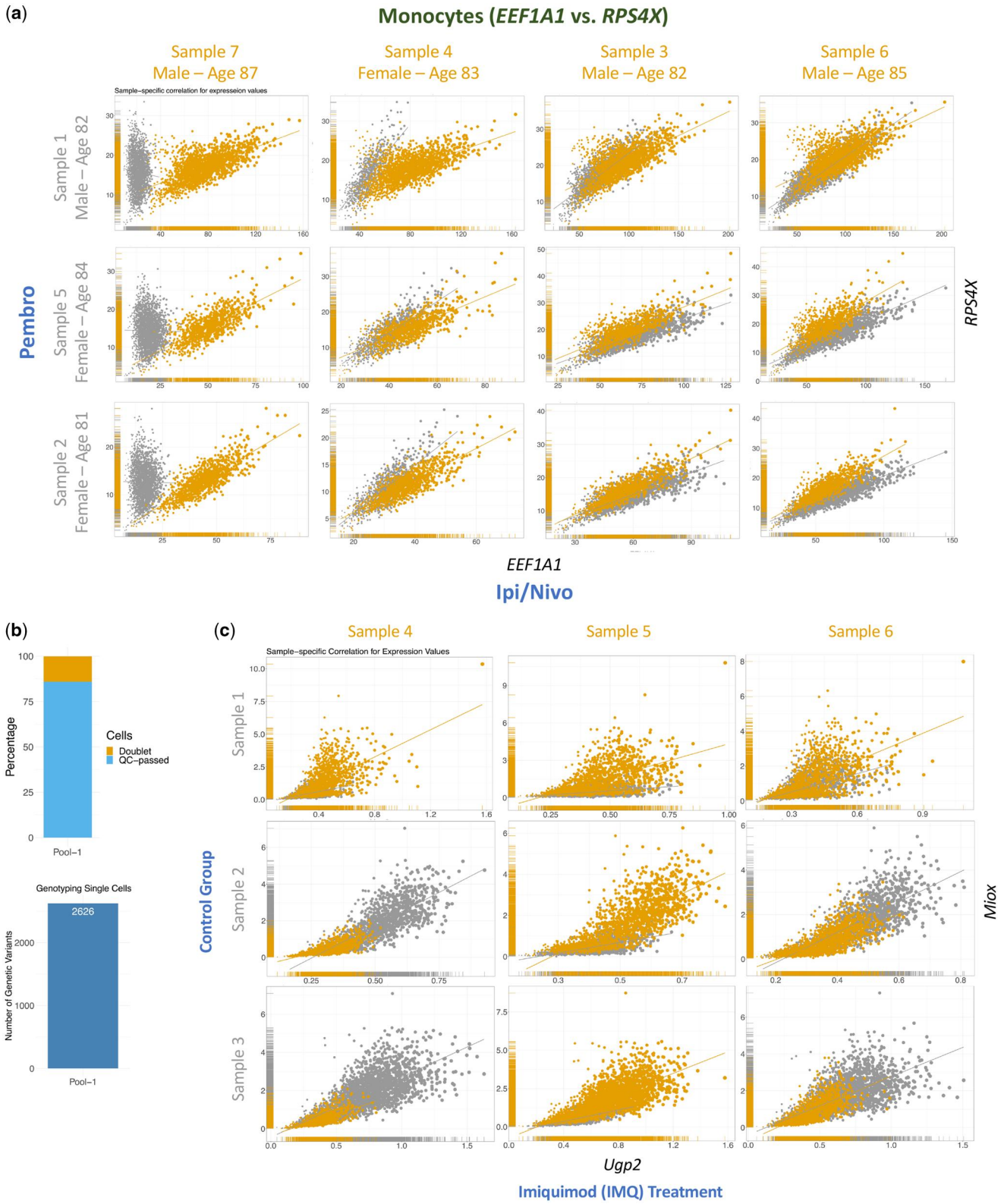


Figure 5. (a) Confounding variables influence the variation in gene expression between individuals. We stratified melanoma patients based on the type of treatment with immune checkpoint blockades (Ipilimumab + Nivolumab (Ipi + Nivo) or Pembrolizumab (Pembro)) and sex (male or female). If these factors cause DCE patterns among individuals, we expect the appearance and disappearance of the top first examples of DCE genes across classes. We found no such an accumulation. For example, the expression of *EEF1A1* and *RPS4X* represent individual 7 and we do not see a similar DCE pattern for matched pairs of individuals based on treatment or sex (e.g. Male and Female). (b) The results of quality control and the number of called variants per cell for a single-cell pool sample made of 6 isogenic mice. We applied the scater package to filter out single-cell profiles that were outliers for any metrics, as they are considered low-quality libraries (McCarthy *et al.* 2017). (c) A pool of six samples with known cell labels from isogenic individuals was used as input for an EAD workflow. DCE patterns across pairs of donors could not distinguish interindividual differences in the gene expression including genes related to the metabolic pathway (e.g. *ugp2* and *Miox* genes), and we only see differences related to the treatment (e.g. sample 4 versus sample 2).

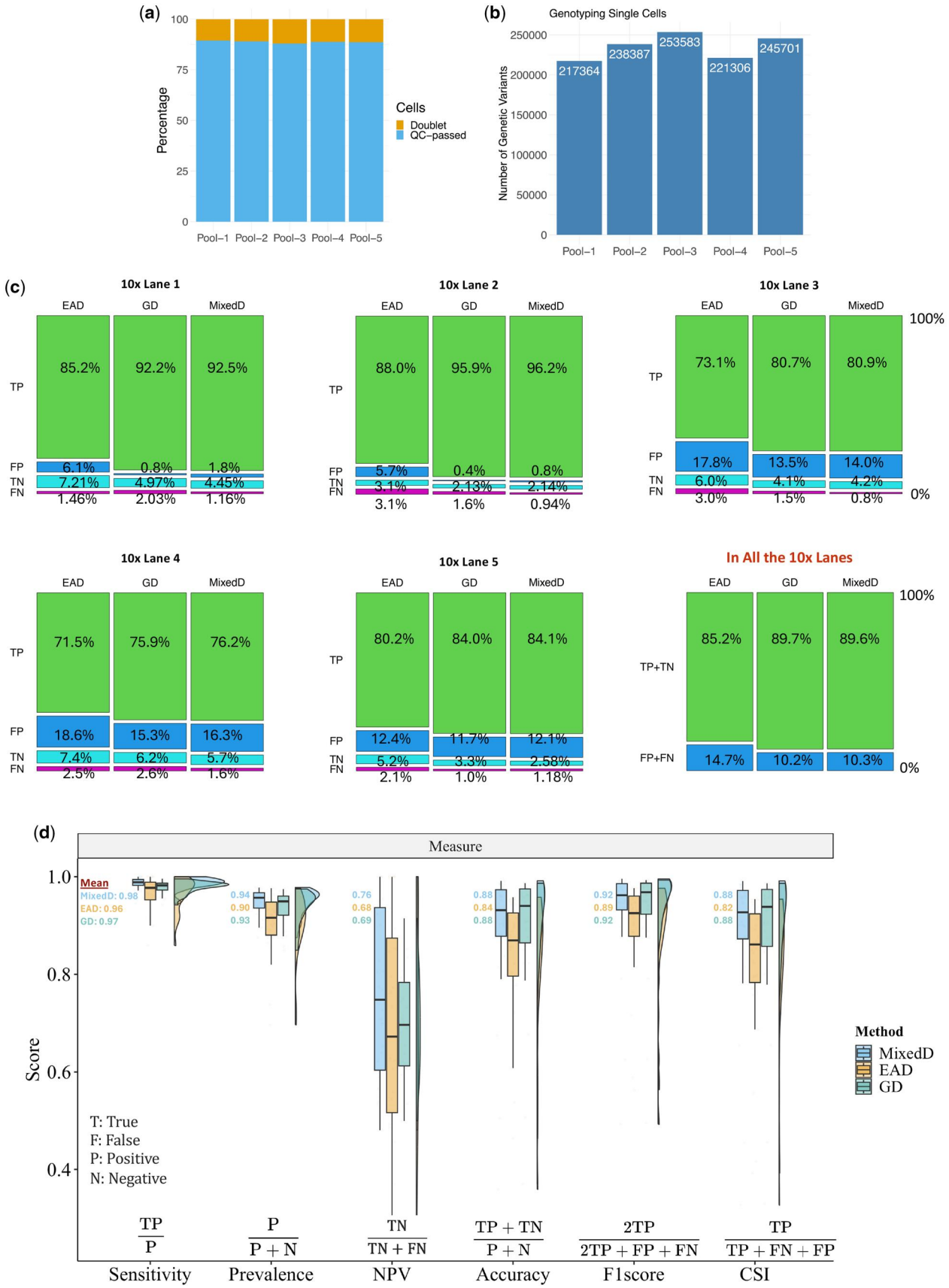


Figure 6. Demultiplexing with genotype reference (Demuxlet) provides a source of ground truth for benchmarking the performance of demultiplexing algorithms. (a and b) The results of quality control including the percentage of filter-out cells as a doublet and the number of called variants per cell for five single-cell pool samples (10× lanes) made of 23 sepsis and seven healthy individuals. (c) Percentage of correct (TP and TN), and incorrect assigned cells (FP and FN) using EAD, GD demultiplexing (vireo), and the combination of genetic-based and expression-aware approaches (Mixed). The outcomes are given for each 10× lane and for all 10× lanes. (d) The confusion matrices were utilized to generate the key metrics and give a comprehensive assessment of demultiplexing methods that do not require a reference genome. Abbreviations: EAD, expression-aware demultiplexing; GD, genetic-based demultiplexing; Mixed-D, mixed demultiplexing; NPV, negative predictive value; CSI, critical success index.

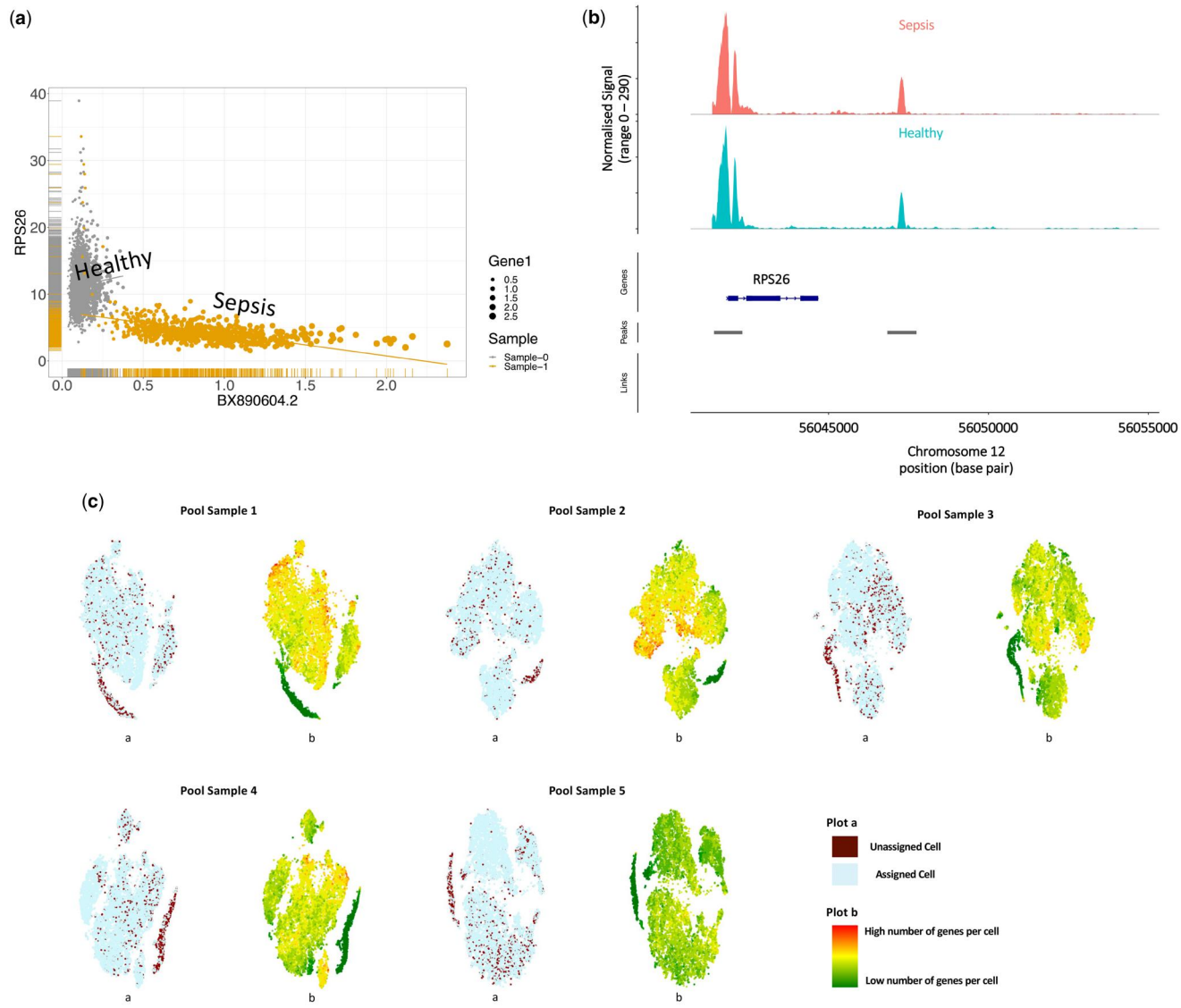


Figure 7. (a) Example of the top interindividual DCE genes provide a distinct cluster of cells per individual. (b) There are no significant differences in chromatin accessibility across states observed in the case of the top DCE gene. (c) Association between cell assignment and the number of genes per cell. Some unassigned cells show a low number of genes per cell, which means filtering out unassigned cells can improve the accuracy of cell calls.

the doublets as unassigned and do not attempt to reassign them through mixed demultiplexing.

We next evaluated the donor assignments made by expression-aware, vireo, and Mixed-D methods against the genetic assignments from Demuxlet that were considered to be 'ground truth'. Figure 6 exhibits mosaic plots of the confusion matrix for demultiplexing methods, divided by lanes. In all 10× lanes, Mixed-D had an increased TP and FP rate, but a decreased FN rate. Mixed-D and vireo demonstrated a consistent correct assignment (TP+TN) in all batches (Fig. 6c).

To quantitatively evaluate the performance of demultiplexing methods, we calculated the key metrics derived from the confusion matrix (Fig. 6d). Mixed-D had the most significant mean sensitivity, prevalence, and negative predictive value among all methods for every 10× lane. All 10× lanes displayed consistent F-scores and accuracy for Mixed-D and vireo (Fig. 6d).

While expression-based methods could not significantly improve the percentage of correct assignment (TP+TN) in 10× lanes of this heterogeneous dataset, Mixed-D could provide more confirmation of the vireo results and improve the accuracy of assignments (Fig. 6c).

Despite the diversity of cells, the top interindividual DCE genes produced a distinct cluster of cells per individual (Fig. 7a). Consistent with previous results, ontology gene set enrichment analysis for DCE genes showed a significant association with the regulation of metabolic processes (Supplementary Fig. 2). While the top interindividual DCE genes are related to metabolic functions or organelles, the breakdown of co-expression patterns can represent context-specific pathognomonic signatures. An illustrative example occurs in the differential co-expression of genes specific to sepsis in the context of COVID-19 infection such as *NEDD9* (Rizzo and Yuan 2022), *BACE2* (Tang *et al.* 2022), and *DHX30* (Apostolidou *et al.* 2021). The application of interindividual DCE genes allows for the differentiation of subtle pathology-induced patterns in a context-specific manner.

We investigated the potential for identifying interindividual single-cell expression variability using multiome ATAC. The results showed that chromatin accessibility within the same

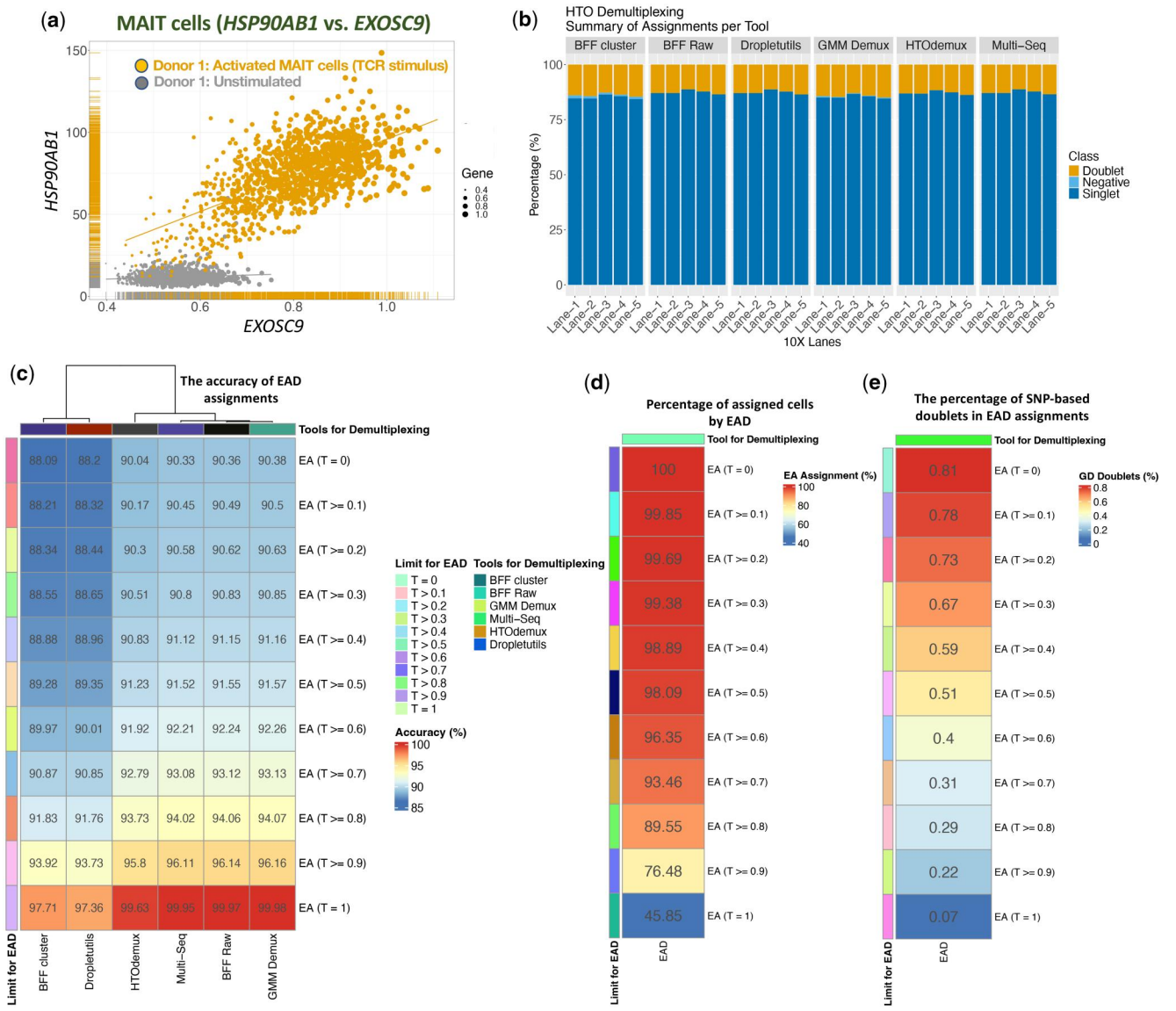


Figure 8. (a) A distinct cluster of cells from the same donor but a different stimulation condition is present in an example of the top DCE genes. (b) Summaries of cell hashing demultiplexing results showing the number of singlets called per 10× lane along with the percentage of doublets and negative cells (both filtered out). (c–e) Evaluating the impact of different thresholds for EA sample demultiplexing on accuracy, percentage of assigned cells, and percentage of genetic-based demultiplexing doublets in EAD assignments by specifying various thresholds (*7*).

cell does not provide a similar explicit model for co-expression relationships between molecules at single-cell level (Fig. 7b).

The tSNE projection plot revealed that unassigned cells were accumulating in regions with low numbers of genes per cell (Fig. 7c). Therefore, the accuracy of cell identification can be improved by removing cells which have very few genes per cell. Functional enrichment can be combined with unassigned cells to maintain statistical power for pooled scRNA-seq data analysis as an alternative solution (Fairfax *et al.* 2020).

The authors in the original paper (Kwok *et al.* 2023) used the combined gene expression profiles to group cells according to their gene expression and chromatin data (seven clusters labelled as C1–7). The differences in gene expression between the clusters of the dataset are likely driven by their skew towards different lineages (e.g. C4: lymphocyte progenitors, C5: emergency granulopoiesis, and C7: normal granulopoiesis). We did not observe strong correlation between the proportion of cells in each cluster and number of unassigned cells including singlets and doublets.

These results demonstrate the potential utility of demultiplexing pooled single-cell RNA-sequencing samples using interindividual variation in gene expression in various biological models.

3.3 Integrate the demultiplexing results from expression-based and barcode-based methods

Multiple methods are available for labelling cells or nuclei with oligonucleotide-barcoded antibodies (Stoeckius *et al.* 2018, Gaublomme *et al.* 2019). A matrix of counts per HTOs per droplet is created by sequencing the cells' RNA molecules with HTOs.

Using a scRNA-seq dataset of stimulated human MAIT cells (Garner *et al.* 2023), we present a demonstration of how expression-aware and barcode-based demultiplexing methods can complement each other (Fig. 8). We used a subset of the original dataset, comprising five channels of a Chromium

Next GEM Chip K. MAIT cells from three donors were either left unstimulated or activated with a TCR, cytokine, or dual TCR+cytokine stimulus. Cells from each donor-condition combination were labelled with TotalSeq-C hashtag antibodies (12 total), pooled, and split across the eight channels of the Chromium Chip (Garner *et al.* 2023).

The dataset contained three donors and four conditions per donor per 10× lane. Genetic-based demultiplexing (vireo) was unable to differentiate two samples from the same donor but different stimulation conditions (Howitt *et al.* 2023). Therefore, we were unable to produce results for mixed demultiplexing generated using genetic-based and expression-aware methods on this dataset. To estimate gene expression for each pair of samples (Fig. 1a) in our EAD workflow, we utilized the output of Seurat HTODemux (Stoeckius *et al.* 2018) instead of vireo. We chose HTODemux as our preferred tool because it is one of the top three most effective methods and performs fairly well on different types of datasets with minor differences between its output and other tools (Howitt *et al.* 2023). The EAD technique was capable of separating all 12 samples per 10× lane, including those that were derived from the same donor but different stimulation conditions (Fig. 8a). Several methods were used for demultiplexing of cell hashing data: BFFcluster (Boggy *et al.* 2022), BFFraw (Boggy *et al.* 2022), GMM-Demux (Xin *et al.* 2020), MULTI-seq (McGinnis *et al.* 2019), HTODemux (Stoeckius *et al.* 2018), demuxEM (Gaublomme *et al.* 2019), and DropletUtils hashedDrops (Griffiths *et al.* 2018, Boggy *et al.* 2022). Each algorithm assigns cells as singlets (one hashtag antibody detected), doublets (two or more hashtag antibodies detected), or negative (no hashtag antibody detected) (Fig. 8b). We treat the doublets and negatives as unassigned and do not attempt to reassign them through EA demultiplexing. To assess the performance of the EAD method, we compared its assignments to singlet calls from the six HTO demultiplexing tools. To calculate accuracy, we compared EAD classifications to HTO algorithm classifications, and then divided the number of concordant classifications by the total number of classifications.

By using 70 000 singlets from five 10× lanes, EAD was able to assign the 90% (62 682/70 000) of cells with an average of 93.2% accuracy (Fig. 8c). EAD's results were similar across all HTO tools, but there was a slight decline for DropletUtils and BFFcluster (Fig. 8c). BFFcluster and DropletUtils determine doublets by thresholding barcode counts, resulting in the algorithm's performance being highly dependent on the correct selection of this parameter (Boggy *et al.* 2022). The default parameters recommended by cellhashR were used in this study (Boggy *et al.* 2022). Adjusting the parameters could potentially yield more consistent results. It is probable that there are problematic singlets for the expression-aware method as it did not consistently assign 6.8% of singlets. Our next step was to delve deeper into the cases where demultiplexing with hashtag oligos and EAD is not in accordance.

Figures 1 and 2c explain that EAD takes into account the assignments for an indicated cell across all possible donor pairs. A cell that is successfully assigned to a donor for an equal or greater number of donor pairs minus 1 (0.9 of all the donor pairs that are possible), is confirmed as belonging to the donor by EAD (Fig. 2c). It is possible to utilize either more liberal or rigid thresholds (T) for EA sample demultiplexing to evaluate its impact on accuracy (Fig. 8c–e). According to the results, EAD produces balanced output for accuracy and the percentage of assigned cells when T is ≥ 0.8. Increasing the threshold to 1 could lead to an increase in accuracy to 99.9%, but at the expense of decreasing the percentage of assigned cells to 46% on average (Fig. 8d). Since there are three genetically distinct donors per 10× lane, we can use SNP-based multiplexing to estimate cells with mismatched SNP profiles (doublets). According to the findings, elevating the threshold for EA sample demultiplexing consistently decreases the number of detected doublets by SNP-based demultiplexing (vireo) (Fig. 8e). This trend remains the same even with different thresholds that yield similar percentages of assigned cells (e.g. 0.1–0.6) (Fig. 8d and e). By using the EAD with a high threshold for sample demultiplexing (e.g. $T = 1$), it is possible to obtain reliable results for cells with sufficient barcode counts for accurate classification, as evidenced by this result (Fig. 8c and d).

The 1968 discordant assignments occurred totally among singlet calls from six HTO demultiplexing tools implemented in cellhashR (Boggy *et al.* 2022). cellhashR defines discordant cells when there is at least one hashing algorithm that produces a different result. The majority of discordant results are caused by the algorithms having slightly different thresholds, as demonstrated in the previous study (Boggy *et al.* 2022). To offer a demonstration of an EAD application that can be utilized in conjunction with existing barcode-based and SNP-based multiplexing methods, we attempted to assign cells using Mixed-D for each singlet that was discordant across HTO demultiplexing tools. When studying a biological phenomenon, it is typical to use several complementary techniques (such as Western blotting and qPCR) to enhance the reliability of findings. If the results of different techniques are similar, there is a boost in confidence in the overall conclusion. By using complementary techniques, we can analyse cell assignments from various angles, such as barcode-based and expression-based, and the convergence of results indicates the reliability of the assignments. Mixed-D only approved singlets proposed by EAD ($T = 1$) that were identical to the proposed donor by at least three tools in cellhashR output. The mixed-D was able to save 50.4% (993/1968) of discordant singlets (Min. 40.95%, Median 50.00%, Max. 59.05%) and it improved the HTO demultiplexing results by an average of 1.4% (993/70 000) (Min. 1.25%, Median 1.33%, Max. 1.74%). Since three HTO demultiplexing tools have previously verified the Mixed-D assignments and we employ a threshold equal to 1 for EAD, we can anticipate secure assignments for the discordant cells.

The cost of single-cell sequencing is influenced by various factors, such as the specific technology employed, the number of cells sequenced, the sequencing depth, and the provider. In most cases, the cost of scRNA-seq is ~£1–5 per cell. Considering the dataset of blood and liver MAIT cells as an example, mixed demultiplexing can reduce the number of discordant cells (from 1968 to 975) and save around £3000. Expression-based demultiplexing is a promising complementary approach to identifying individual samples in a pooled sequencing experiment. By combining the results of both methods, we can obtain a more precise picture of the actual demultiplexing results.

Interindividual DCE analysis within the same cell type can be used to identify cells from the same donor in different activation states, as indicated by the results from the MAIT cell dataset (Garner *et al.* 2023). The use of interindividual DCE analysis for scRNA-seq is not limited to EAD. It can potentially be applied to discover small changes in gene expression

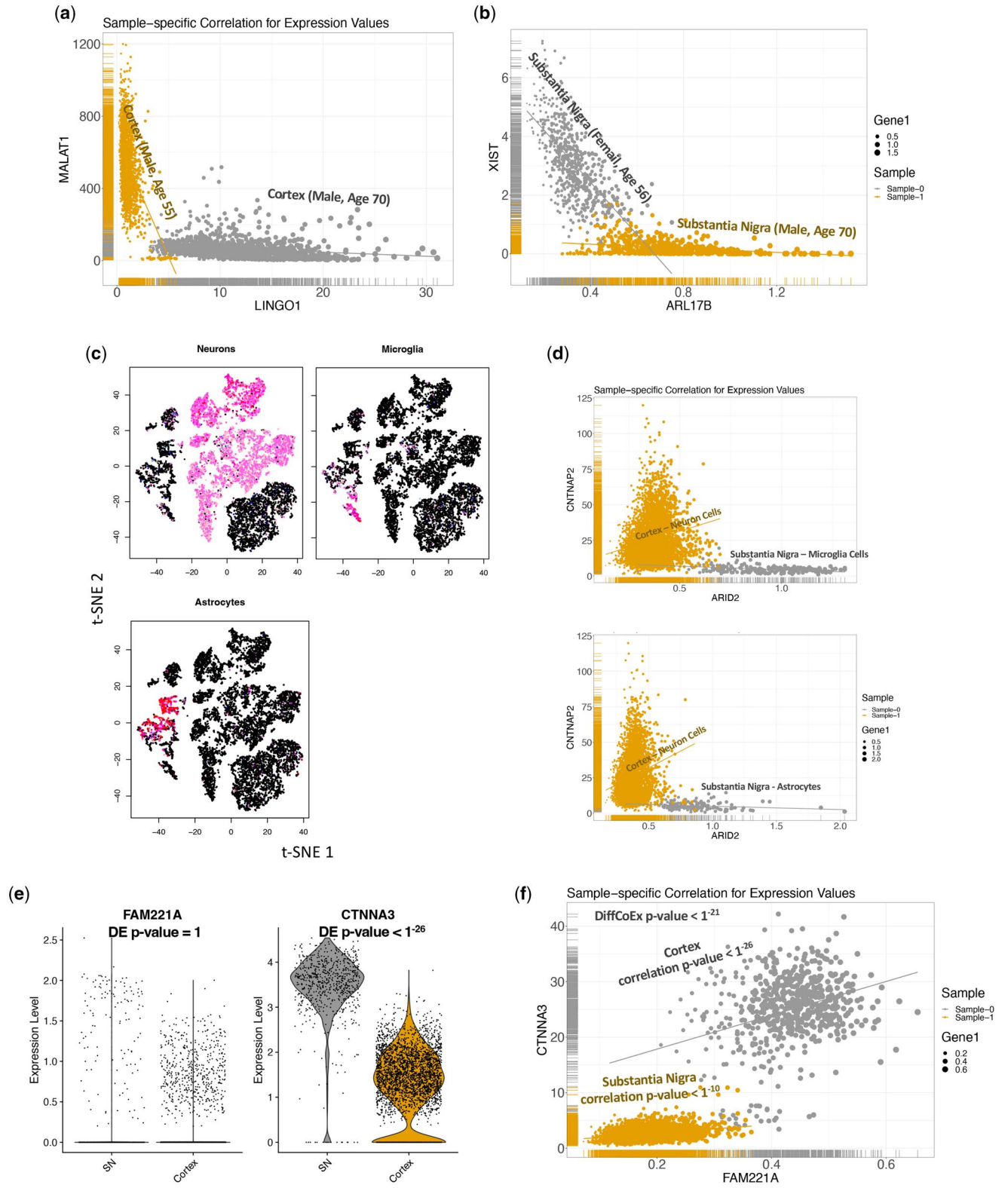


Figure 9. A few examples of the many interindividual DCE genes that have been identified in the substantia nigra and cortex. These genes play important roles in a variety of neurological processes, and their dysregulation can contribute to the development of neurological disorders. More investigation is required to fully comprehend the roles of these genes in the brain and their potential as therapeutic targets for neurological disorders. (a, b) Examples of interindividual DCE pattern in the substantia nigra and cortex. (c) The t-SNE project of transcriptionally and functionally distinct clusters, highlighting microglia and neuron cell type groups, is presented. Pink/Red cells have passed the threshold of cell type enrichment (Nassiri *et al.* 2023). (d) Examples of DCE patterns between cell types in the substantia nigra and cortex. (e) A pair of genes that exhibit differential co-expression but not differential expression. The analyses were carried out by employing single-cell expression profiles across the SN and cortex regions of a donor. In each region, the single-cell expression distributions of *CTNNA3* and *FAM221A* genes are visualized by a violin plot. The expression levels of both genes decrease from SN to cortex, but only *CTNNA3* has a significant differential expression. (f) The *CTNNA3* and *FAM221A* gene pairs exhibit DCE patterns in the SN and cortex.

that are linked to various conditions or cell states. Researchers who want to study the transcriptional heterogeneity of cell states within the same cell type can benefit from this technique (Garner *et al.* 2023). These applications are covered in more detail in the following section.

3.4 Leveraging interindividual variation in gene expression for precision therapeutic strategies

In this section, we show that it is possible to identify genes that are differentially co-expressed for each pair of donors for nonimmune cells. Using DCE pairs between individuals allows for EAD. Furthermore, we examine examples of variations in co-expressed gene patterns among various individuals that can shed light on the molecular mechanisms underlying differences in therapeutic response.

To demonstrate that our method can be utilized for nonimmune cells, we used single-nuclei transcriptome profiles of the SN and cortex (middle frontal gyrus) (in total 12 samples, including two SN replicates) (Agarwal *et al.* 2020). The analysis of cell-type enrichment showed that there were five different cell populations in each sample of the SN and six different cell populations in the cortex (e.g. astrocytes) (Agarwal *et al.* 2020). We were able to identify distinct clusters of cells across individuals based on the expression pattern of DCE genes despite cell heterogeneity (Fig. 9). In at least one pair of samples, we found 3207 genes with DCE (FDR $< 1^{-10}$), with 568 of them being detected in only one pair of samples for the indicated region. We did not observe DCE patterns between replicated samples (FDR $< 1^{-4}$). Our achievement of finding representative DCE genes for each pair of donors allows us to use EAD for this data type.

One example is the differential co-expression of *MALAT1* and *LINGO1* in the cortex of different individuals (Fig. 9a). *MALAT1* is a long noncoding RNA (lncRNA) that is highly expressed in the brain, particularly in the cortex. *MALAT1* has the potential to be a promising therapeutic target for a variety of neurological disorders (Wang *et al.* 2022). Neuronal survival and axonal growth are regulated by *LINGO1*, which is a leucine-rich single transmembrane protein, in the cortex. It has been observed that it promotes the death of neurons and hinders the growth of axons. It is suggested that *LINGO1* may be involved in both the normal pruning of neurons during development and the pathological loss of neurons that occurs in neurodegenerative diseases (Inoue *et al.* 2007). *LINGO1* has the potential to be a promising therapeutic target for several neurological disorders. *MALAT1* and *LINGO1* have been demonstrated to interact and have opposite effects on neuronal survival (Fan *et al.* 2018). Neuronal survival is promoted by *MALAT1*, while neuronal death is promoted by *LINGO1* (Inoue *et al.* 2007, Wang *et al.* 2022). The balance between *MALAT1* and *LINGO1* expression may play a significant role in regulating neuronal survival in the cortex. Further research is needed to fully understand the differential co-expression relationship between upregulation of *MALAT1* and downregulation of *LINGO1*, as well as its roles in neurological disorders.

Other examples of interindividual DCE pairs include *XIST* and *ARL17B* in SN (Fig. 9b). The SN expresses *XIST*, which has been shown to regulate the expression of genes crucial for dopaminergic neuron function (Wang *et al.* 2021). *ARL17B* is present in the SN and has been discovered to play a significant role in the survival of dopaminergic neurons (Reus *et al.* 2021). Parkinson's disease, which is characterized by the loss of dopaminergic neurons in the SN, has been linked to mutations in *XIST* and *ARL17B* (Wang *et al.* 2021). Our result indicates that there is a settled co-expression relationship between *XIST* and *ARL17B*. Further exploration is required to fully comprehend the functional consequence of losing association between *XIST* and *ARL17B* expression in SN and their possible contribution to Parkinson's disease.

The expression patterns of many genes are not directly co-regulated within cell types, but they are differentially co-expressed across cell types, as shown by our results. For instance, the interaction between *ARID2* and *CNTNAP2* resolves in glial cells of the SN (Fig. 9c and d). This demonstrates differential co-expression between cortical neurons and SN glial cell types. The regulation of gene expression and DNA replication by chromatin remodelling is facilitated by ARID2 (AT-rich interaction domain 2) protein (Kang *et al.* 2021). It is particularly important for glial cells to maintain homeostasis and protect against infection and neurodegeneration. In glial cells, *CNTNAP2* (contactin-associated protein 2) is a key component of glial-neuronal communication, neuroprotection, and cognitive function (Gandhi *et al.* 2023, St George-Hyslop *et al.* 2023). The interaction between CNTNAP2 and ARID2 leads to the stabilization of the ARID2 protein (Moffat *et al.* 2022). ARID2 needs this stabilization to properly regulate gene expression (Moffat *et al.* 2022). Our results suggest that the co-expression of *ARID2* and *CNTNAP2* is resolved in the microglia and astrocyte cells in SN (Fig. 9d). Further research is needed to fully elucidate the effects of losing the interaction between *ARID2* and *CNTNAP2* on microglial and astrocyte cell functions in SN.

Our observation revealed a distinct set of genes that represent each brain region and cell-type. In addition, we found examples of interindividual variations in co-expressed gene patterns, which provide insight into the challenges presented by variation in therapeutic response (Crowell *et al.* 2020, Guo *et al.* 2023).

Typical differential gene expression analysis (DiffEx) does not provide a similar explicit model for co-expression relations between molecules at the single-cell level. The focus of DiffEx is on the expression levels of individual genes (Gaublomme *et al.* 2019). Differential co-expression analysis (DiffCoEx) is focused on discovering genes whose expression levels change significantly under various conditions. DiffCoEx offers insight into the coordinated response of genes and assists us in understanding the underlying biological networks (McKenzie *et al.* 2016, Crowell *et al.* 2020). As a demonstration, we compare DiffEx and DiffCoEx results across SN and cortex regions in a 56-year-old female donor (Fig. 9e and f). Differential expression of genes across the SN and cortex was found by the Seurat tool, and 64.8% (5483/8456) of them were also differentially co-expressed (FDR$<1^{-3}$) (Hao *et al.* 2024). The example of *CTNNA3* and *FAM221A* genes illustrates the distinction between differential expression and differential co-expression. In both SN and cortex contexts, Fig. 9e illustrates where the *CTNNA3* and *FAM221A* genes are expressed. The expression values of both genes decrease from SN to cortex, but *FAM221A* is not a significant differentially expressed gene ($P = 1$), as shown in Fig. 9e. These genes have significant differential co-expression in the SN and cortex, which cannot be predicted solely based on differential expression relationships (Fig. 9f).

4 Discussion

The capability of multiplex scRNA-seq samples has attracted research attention to lowering experiment costs and addressing batch effects. Several multiplexing methods and bioinformatics tools have been developed for demultiplexing pooled datasets (Kang *et al.* 2018, Stoeckius *et al.* 2018, Gaublomme *et al.* 2019, Guo *et al.* 2019). The main approaches include barcode-based (Boggy *et al.* 2022) and SNP-based sample pool demultiplexing (Kang *et al.* 2018, Huang *et al.* 2019). Compared to other approaches, EAD facilitates feature selection (marker gene selection) for clustering single-cell data by obtaining individual-specific variability (Ranjan *et al.* 2021). EAD has added benefits to previous methods by improving the accuracy of cell assignments into individual samples, without the need for additional experimental steps (Nassiri *et al.* 2023).

The accurate assigning of cells to their respective donors can provide valuable insights for researchers into the biological processes and disease mechanisms that vary across individuals. The findings indicate that combining expression-based demultiplexing with SNP-based or barcode-based methods is the most accurate approach for demultiplexing single-cell and single-nuclei sequencing data. The combined demultiplexing results are more accurate due to the fact that they take into account the strengths of methods. For example, cell-multiplexing oligos using cells or nuclei samples can be technically challenging and have limited performance (Stoeckius *et al.* 2018). On the other hand, the identity of donors (donor-specific information) is not specified by EAD. When cells are assigned to donor groups using EAD, the identity of donors can be specified by incorporating demultiplexing results provided by barcode-based multiplexing assays (Kim *et al.* 2022).

Each person has a distinct set of SNPs. In the event that two cells from different individuals are combined to form a doublet, the combined SNP profile will appear unbalanced. SNP demultiplexing algorithms have the ability to identify these inconsistencies and label them as potential doublets (Kang *et al.* 2018, Huang *et al.* 2019). SNP demultiplexing for doublet detection is not perfect and can lead to false positives. Factors like sequencing depth and quality, and completeness of reference panels used for SNP calling determine the accuracy of doublet detection (Kang *et al.* 2018). False positives can be caused by sequencing errors, allelic dropout, batch effects, somatic mutations, and natural variability in gene expression (Huang *et al.* 2019). While expression-based demultiplexing cannot detect doublets, the combination of SNP and expression demultiplexing can offer more confirmation or reassign doublet predictions. Filtering out cells that are classified as true doublets using tools like SCAN is recommended before reassigning a predicted doublet using Mixed-D (Dahlin *et al.* 2018).

The validation approach for characterization of co-expression variations is limited by the use of datasets from the 10× Genomics Chromium platform. In general, it is plausible that targeted scRNA-seq is less likely to obtain enough SNPs for SNP-based demultiplexing and has a limited number of genes for DCE analysis. By incorporating other single-cell RNA and protein expression technologies that have been developed, our approach can be easily enhanced.

The application of interindividual DCE analysis of scRNA-seq goes beyond EAD. It can potentially be applied to reveal biological activity that is useful for patient stratification, identifying biomarkers of target engagement, and connecting genomic programs with biological functions (McKenzie *et al.* 2016, Badia-i-Mompel *et al.* 2023). The study of interindividual variation in single-cell RNA-sequencing data is a constantly evolving field that involves continuous methodological advancement and refinement (Kumasaka *et al.* 2023). Most single-cell studies employ average gene expression profiles across cell types or states of interest, which often obscures differences among individuals that are apparent at single-cell resolution (Murdock and Tsai 2023). Variability between individuals remains an understudied aspect of relationships between molecules at the single-cell level (Crowell *et al.* 2020). Cell-level mixed models are currently employed for differential state analysis across multiple samples (e.g. donors) and experimental conditions using cell-level measurements (Crowell *et al.* 2020). The cell-level mixed models generate log-fold changes and the proportion of cells that have expressed a particular gene in each sample or group. These methods greatly underestimate the differences in gene expression between different cell subpopulations, such as those with low expression (Crowell *et al.* 2020, Auerbach *et al.* 2021). The cell-level solution to improve this issue is provided by DCE analysis, which directly models gene expression. In addition, it provides the possibility of modelling related gene regulatory networks and following each pair of DCE genes along the axis of samples. While the interindividual differential co-expression approach identifies how disease signatures vary between clinical subtypes, these changes ultimately might converge on shared signalling pathways that present biomarkers for new precision therapeutic strategies (Johansen *et al.* 2023).

We anticipate that EAD will be applied to improve the results of SNP-based and barcode-based multiplexing assays, provide insight into challenges presented by variation in therapeutic response, and consider the effect of these stratification strategies on proposed candidate biomarker genes (Stoeckius *et al.* 2018, Li *et al.* 2020).

Author contributions

Isar Nassiri (Conceptualization [lead], Formal analysis [lead], Methodology [lead], Project administration [lead], Resources [lead], Software [lead], Validation [lead], Visualization [lead], Writing—original draft [lead], Writing—review & editing [lead]), Andrew J. Kwok (Data curation [supporting], Resources [supporting], Validation [supporting], Writing—original draft [supporting]), Aneesha Bhandari (Data curation [supporting], Resources [supporting], Validation [supporting], Writing—original draft [supporting]), Katherine R. Bull (Data curation [equal], Resources [equal], Validation [equal], Writing—original draft [equal]), Lucy C. Garner (Data curation [equal], Formal analysis [equal], Resources [equal], Validation [equal], Writing—original draft [equal], Writing—review & editing [equal]), Paul Klenerman (Data curation [equal], Resources [equal], Writing—review & editing [equal]), Caleb Webber (Data curation [equal], Resources [equal], Writing—review & editing [equal]), Laura Parkkinen (Conceptualization [equal], Data curation [equal], Resources [equal], Writing—review & editing [equal]), Angela W. Lee (Data curation [equal], Resources [equal], Writing—original draft [equal]), Yanxia Wu (Conceptualization [equal], Data curation [equal], Investigation [equal], Writing—original draft [supporting], Writing—review & editing [supporting]), Benjamin Fairfax (Data curation

[equal], Methodology [supporting], Resources [equal], Writing—original draft [equal], Writing—review & editing [supporting]), Julian C. Knight (Investigation [equal], Resources [equal], Writing—original draft [equal], Writing—review & editing [equal]), David Buck (Conceptualization [equal], Resources [equal], Writing—original draft [equal]), and Paolo Piazza (Funding acquisition [lead], Methodology [equal], Project administration [supporting], Resources [equal], Validation [equal], Writing—original draft [equal], Writing—review & editing [equal])

Supplementary data

Supplementary data are available at *Bioinformatics Advances* online.

Conflict of interest

None declared.

Funding

The research was supported by the Wellcome Trust Core Award [203141/Z/16/Z]. I.N. was supported by the National Institute for Health Research (NIHR) Oxford Health Biomedical Research Centre (BRC). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References

Agarwal D, Sandor C, Volpato V *et al.* A single-cell atlas of the human substantia nigra reveals cell-specific pathways associated with neurological disorders. *Nat Commun* 2020;**11**:4183.

Almeida A, Loy A, Hofmann H. ggplot2 compatible quantile-quantile plots in R. *R J* 2018;**10**:248–61.

Apostolidou S, Harbauer T, Lasch P *et al.* Fatal COVID-19 in a child with persistence of SARS-CoV-2 despite extensive multidisciplinary treatment: a case report. *Children (Basel)* 2021;**8**:564.

Auerbach BJ, Hu J, Reilly MP *et al.* Applications of single-cell genomics and computational strategies to study common disease and population-level variation. *Genome Res* 2021;**31**:1728–41.

Badia-I-Mompel P, Wessels L, Müller-Dott S *et al.* Gene regulatory network inference in the era of single-cell multi-omics. *Nat Rev Genet* 2023;**24**:739–54.

Beckelman BC, Day S, Zhou X *et al.* Dysregulation of elongation factor 1A expression is correlated with synaptic plasticity impairments in Alzheimer's disease. *J Alzheimers Dis* 2016a;**54**:669–78.

Beckelman BC, Zhou X, Keene CD *et al.* Impaired eukaryotic elongation factor 1A expression in Alzheimer's disease. *Neurodegener Dis* 2016b;**16**:39–43.

Boggy GJ, McElfresh GW, Mahyari E *et al.* BFF and cellhashR: analysis tools for accurate demultiplexing of cell hashing data. *Bioinformatics* 2022;**38**:2791–801.

Bongen E, Lucian H, Khatri A *et al.* Sex differences in the blood transcriptome identify robust changes in immune cell proportions with aging and influenza infection. *Cell Rep* 2019;**29**:1961–73.e4.

Cardiello JF, Joven Araus A, Giatrellis S *et al.* Evaluation of genetic demultiplexing of single-cell sequencing data from model species. *Life Sci Alliance* 2023;**6**:e202301979.

Crowell HL, Soneson C, Germain P-L *et al.* Muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nat Commun* 2020;**11**:6077.

Dahlin JS, Hamey FK, Pijuan-Sala B *et al.* A single-cell hematopoietic landscape resolves 8 lineage trajectories and defects in kit mutant mice. *Blood* 2018;**131**:e1–11.

Deelen P, Zhernakova DV, de Haan M *et al.* Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels. *Genome Med* 2015;**7**:30.

Engebretsen S, Bohlin J. Statistical predictions with glmnet. *Clin Epigenetics* 2019;**11**:123.

Fairfax BP, Makino S, Radhakrishnan J *et al.* Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat Genet* 2012;**44**:502–10.

Fairfax BP, Taylor CA, Watson RA *et al.* Peripheral CD8+ T cell characteristics associated with durable responses to immune checkpoint blockade in patients with metastatic melanoma. *Nat Med* 2020;**26**:193–9.

Fan B, Wei Z, Yao X *et al.* Microenvironment imbalance of spinal cord injury. *Cell Transplant* 2018;**27**:853–66.

Gandhi T, Canepa CR, Adeyelu TT *et al.* Neuroanatomical alterations in the CNTNAP2 mouse model of autism spectrum disorder. *Brain Sci* 2023;**13**:891.

Garner LC, Amini A, FitzPatrick MEB *et al.* Single-cell analysis of human MAIT cell transcriptional, functional and clonal diversity. *Nat Immunol* 2023;**24**:1565–78.

Gaublomme JT, Li B, McCabe C *et al.* Nuclei multiplexing with barcoded antibodies for single-nucleus genomics. *Nat Commun* 2019;**10**:2907.

Griffiths JA, Richard AC, Bach K *et al.* Detection and removal of barcode swapping in single-cell RNA-seq data. *Nat Commun* 2018;**9**:2667.

Grun B, Leisch F. FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. *J Stat Soft* 2008;**28**:1–35.

Guo C, Kong W, Kamimoto K *et al.* CellTag indexing: genetic barcode-based sample multiplexing for single-cell genomics. *Genome Biol* 2019;**20**:90.

Guo MG, Reynolds DL, Ang CE *et al.* Integrative analyses highlight functional regulatory variants associated with neuropsychiatric diseases. *Nat Genet* 2023;**55**:1876–91.

Hao Y, Stuart T, Kowalski MH *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* 2024;**42**:293–304.

Heaton H, Talman AM, Knights A *et al.* Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat Methods* 2020;**17**:615–20.

Howitt G, Feng Y, Tobar L *et al.* Benchmarking single-cell hashtag oligo demultiplexing methods. *NAR Genom Bioinform* 2023;**5**:lqad086.

Huang M, Wang J, Torre E *et al.* SAVER: gene expression recovery for single-cell RNA sequencing. *Nat Methods* 2018;**15**:539–42.

Huang X, Huang Y. Cellsnp-lite: an efficient tool for genotyping single cells. *Bioinformatics* 2021;**37**:4569–71.

Huang Y, McCarthy DJ, Stegle O. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome Biol* 2019;**20**:273.

Inoue H, Lin L, Lee X *et al.* Inhibition of the leucine-rich repeat protein LINGO-1 enhances survival, structure, and function of dopaminergic neurons in Parkinson's disease models. *Proc Natl Acad Sci USA* 2007;**104**:14430–5.

Johansen N, Somasundaram S, Travaglini KJ *et al.* Interindividual variation in human cortical cell type abundance and expression. *Science* 2023;**382**:eadf2359.

Kang E, Kang M, Ju Y *et al.* Association between ARID2 and RAS-MAPK pathway in intellectual disability and short stature. *J Med Genet* 2021;**58**:767–77.

Kang HM, Subramaniam M, Targ S *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol* 2018;**36**:89–94.

Kim H-J, Booth G, Saunders L *et al.* Nuclear oligo hashing improves differential analysis of single-cell RNA-seq. *Nat Commun* 2022;**13**:2666.

Kumasaka N, Rostom R, Huang N *et al.* Mapping interindividual dynamics of innate immune response at single-cell resolution. *Nat Genet* 2023;**55**:1066–75.

Kwok AJ, Allcock A, Ferreira RC *et al.*; Emergency Medicine Research Oxford (EMROx). Neutrophils and emergency granulopoiesis drive

immune suppression and an extreme response endotype during sepsis. *Nat Immunol* 2023;**24**:767–79.
Li B, Gould J, Yang Y *et al.* Cumulus provides cloud-based data analysis for large-scale single-cell and single-nucleus RNA-seq. *Nat Methods* 2020;**17**:793–8.
Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 2011;**27**:2987–93.
Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;**25**:1754–60.
Lotta LA, Pietzner M, Stewart ID *et al.*; MacTel Consortium. A cross-platform approach identifies genetic regulators of human metabolism and health. *Nat Genet* 2021;**53**:54–64.
Manach C, Milenkovic D, Van de Wiele T *et al.* Addressing the inter-individual variation in response to consumption of plant food bioactives: towards a better understanding of their role in healthy aging and cardiometabolic risk reduction. *Mol Nutr Food Res* 2017; **61**:1600557.
McCarthy DJ, Campbell KR, Lun ATL *et al.* Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* 2017;**33**:1179–86.
McGinnis CS, Patterson DM, Winkler J *et al.* MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat Methods* 2019;**16**:619–26.
McKenzie AT, Katsyv I, Song W-M *et al.* DGCA: a comprehensive R package for differential gene correlation analysis. *BMC Syst Biol* 2016;**10**:106.
Moffat JJ, Smith AL, Jung E-M *et al.* Neurobiology of ARID1B haploinsufficiency related to neurodevelopmental and psychiatric disorders. *Mol Psychiatry* 2022;**27**:476–89.
Murdock MH, Tsai LH. Insights into Alzheimer's disease from single-cell genomic approaches. *Nat Neurosci* 2023;**26**:181–95.
Nassiri I, Fairfax B, Lee A *et al.* scQCEA: a framework for annotation and quality control report of single-cell RNA-sequencing data. *BMC Genomics* 2023;**24**:381.
Nassiri I, McCall MN. Systematic exploration of cell morphological phenotypes associated with a transcriptomic query. *Nucleic Acids Res* 2018;**46**:e116.
Neavin D, Senabouth A, Arora H *et al.* Demuxafy: improvement in droplet assignment by integrating multiple single-cell demultiplexing and doublet detection methods. *Genome Biol* 2024;**25**:94.
Oelen R, de Vries DH, Brugge H *et al.*; BIOS Consortium. Single-cell RNA-sequencing of peripheral blood mononuclear cells reveals widespread, context-specific gene expression regulation upon pathogenic exposure. *Nat Commun* 2022;**13**:3267.
Ranjan B, Sun W, Park J *et al.* DUBStepR is a scalable correlation-based feature selection method for accurately clustering single-cell data. *Nat Commun* 2021;**12**:5849.
Reus LM, Pasaniuc B, Posthuma D *et al.*; International FTD-Genomics Consortium. Gene expression imputation across multiple tissue types provides insight into the genetic architecture of frontotemporal dementia and its clinical subtypes. *Biol Psychiatry* 2021;**89**:825–35.
Rizzo AN, Yuan JXJ. NEDD9 provides mechanistic insight into the coagulopathy of COVID-19. *Pulm Circ* 2022;**12**:e12087.
St George-Hyslop F, Haneklaus M, Kivisild T *et al.* Loss of CNTNAP2 alters human cortical excitatory neuron differentiation and neural network development. *Biol Psychiatry* 2023;**94**:780–91.
Stoeckius M, Zheng S, Houck-Loomis B *et al.* Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol* 2018;**19**:224.
Tang L, Zhang D, Han P *et al.* Structural basis of SARS-CoV-2 and its variants binding to intermediate horseshoe bat ACE2. *Int J Biol Sci* 2022;**18**:4658–68.
Van der Auwera GA, Carneiro MO, Hartl C *et al.* From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics* 2013;**43**:11.10.11–.10.33.
van der Wijst MGP, Brugge H, de Vries DH *et al.*; BIOS Consortium. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat Genet* 2018;**50**:493–7.
Wang L, Li S, Stone SS *et al.* The role of the lncRNA MALAT1 in neuroprotection against hypoxic/ischemic injury. *Biomolecules* 2022; **12**:146.
Wang W, Min L, Qiu X *et al.* Biological function of long non-coding RNA (LncRNA) xist. *Front Cell Dev Biol* 2021;**9**:645647.
Wolock SL, Lopez R, Klein AM. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst* 2019;**8**: 281–91.e9.
Wu TZ *et al.* clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation (Camb)* 2021;**2**:100141.
Xin H, Lian Q, Jiang Y *et al.* GMM-Demux: sample demultiplexing, multiplet detection, experiment planning, and novel cell-type verification in single cell sequencing. *Genome Biol* 2020; **21**:188.
Xu J, Falconer C, Nguyen Q *et al.* Genotype-free demultiplexing of pooled single-cell RNA-seq. *Genome Biol* 2019;**20**:290.
Yang PY, Huang H, Liu CL. Feature selection revisited in the single-cell era. *Genome Biol* 2021;**22**:321.
Yazar S, Alquicira-Hernandez J, Wing K *et al.* Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* 2022;**376**:eabf3041.

Bioinformatics Advances, 2024, 00, 1–18
https://doi.org/10.1093/bioadv/vbae085
Original Article