Benchmarking single-cell hashtag oligo demultiplexing methods

George Howitt[1, 2], Yuzhou Feng[1], Lucas Tobar[1, 2], Dane Vassiliadis[1,2], Peter Hickey[8,9], Mark A. Dawson[2, 7], Sarath Ranganathan[3, 4, 5], Shivanthan Shanthikumar[3, 4, 5], Melanie Neeland[3,5], Jovana Maksimovic[1, 2]*, Alicia Oshlack[1, 2, 6]*

[1] Computational Biology Program, Peter MacCallum Cancer Centre, Parkville, VIC, Australia
[2] Sir Peter MacCallum Department of Oncology, University of Melbourne, Parkville, VIC, Australia
[3] Respiratory Diseases, Murdoch Children's Research Institute, Parkville, VIC, Australia
[4] Respiratory and Sleep Medicine, Royal Children's Hospital, Parkville, VIC, Australia
[5] Department of Paediatrics, University of Melbourne, Parkville, VIC, Australia
[6] School of Mathematics and Statistics, University of Melbourne, Parkville, VIC, Australia
[7] Centre for Cancer Research, University of Melbourne, Parkville, VIC, Australia
[8] The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, VIC 3052, Australia
[9] The Department of Medical Biology, The University of Melbourne, Parkville, VIC 3010, Australia
*these authors contributed equally and should be considered joint corresponding authors

Abstract

Sample multiplexing is often used to reduce cost and limit batch effects in single-cell RNA (scRNA-seq) sequencing experiments. A commonly used multiplexing technique involves tagging cells prior to pooling with a hashtag oligo (HTO) that can be sequenced along with the cells' RNA to determine their sample of origin. Several tools have been developed to demultiplex HTO sequencing data and assign cells to samples. In this study, we critically assess the performance of six HTO demultiplexing tools: *hashedDrops*, *HTODemux*, *GMM-Demux*, *demuxmix*, *deMULTIplex* and *BFF*. The comparison uses data sets where each sample has also been demultiplexed using genetic variants from the RNA, enabling comparison of HTO demultiplexing techniques against complementary data from the genetic "ground truth". We find that all methods perform similarly where HTO labelling is of high quality, but methods that assume a bimodal counts distribution perform poorly on lower quality data. We also provide heuristic approaches for assessing the quality of HTO counts in a scRNA-seq experiment.

Introduction

Improvements in droplet-based single-cell RNA sequencing (scRNA-seq) technologies have prompted growing interest in exploring variation in gene expression at cellular resolution. While costs continue to decrease, it remains expensive to separately capture and sequence individual samples. Batch effects also confound meaningful differences in gene expression between samples, and robust detection of multiplets (droplets containing two or more cells) solely from

the transcriptome remains an issue (Neavin et al. 2022). One solution to address these problems is to design multiplexed experiments, where samples are pooled prior to droplet capture and sequencing. The cost per sample is reduced by a factor of the number of samples sequenced, while major sample preparation batch effects within the pool are eliminated. Importantly, droplets containing cells from two or more samples can be identified. In addition, the number of cross-sample doublets can be used to estimate the expected number of within-sample doublets and thereby inform the application of other doublet detection algorithms such as *scds* (Bais and Kostka 2020) and *scDblFinder* (Germain et al. 2021).

Despite these advantages, it is important to carefully consider the most appropriate multiplexing protocol for the sample type(s), and if additional information is required to associate the cells with their sample of origin. For genetically distinct samples, demultiplexing can be performed based on genetic variants identified from the transcriptome using a variety of tools such as *vireo* and *demuxlet* (Y. Huang, McCarthy, and Stegle 2019; Kang et al. 2018). However, genetic demultiplexing is not possible where samples from the same individual are sequenced together (e.g. before and after treatment or different tissues from the same individual), or in model organisms, where there is typically little genetic variation between individuals. Additionally, although genetic demultiplexing is able to distinguish cells from genetically distinct individuals, it cannot provide absolute identification of the individual sample within the pool without further information about the samples, such as SNP genotyping.

Cell hashing is an alternative multiplexing technique. Prior to pooling, a barcoded label called a hashtag oligo (HTO) is added, one to each sample. The HTOs attach to either antibodies or lipids on the surface of the cells and the HTOs are captured and sequenced in parallel to the RNA. The antibodies used bind to ubiquitous cell surface proteins (Stoeckius et al. 2018) whilst the lipids incorporate into the plasma cell membrane (McGinnis et al. 2019).

Sequencing of the HTOs produces a HTO counts matrix, an $N_{HTOs} \times N_{droplets}$ matrix consisting of the read counts for each HTO in each droplet. In an ideal scenario, each droplet contains only one cell and each cell contains only counts for the HTO corresponding to its sample of origin. In this ideal case, the demultiplexing algorithm involves simply identifying the non-empty entries in each column of the HTO counts matrix. In practice, the data is noisy; droplets may contain multiple cells, HTO conjugated antibodies/lipid molecules may not bind well to the cells or may dissociate and bind to cells from another sample in the pooling stage, or unbound HTOs may be present in droplets (Stoeckius et al. 2018; McGinnis et al. 2019). Therefore, some sophistication is required for demultiplexing algorithms to distinguish the counts from the "true" HTO against a background of "false" counts.

In this study, we present a comparison of six HTO demultiplexing methods: *hashedDrops*, *HTODemux*, *deMULTIplex*, *GMM-Demux*, *demuxmix* and *BFF*. We discuss the details of each in the Methods section. In all cases, the fundamental goal of each method is the same: to examine the counts of each HTO in a droplet and determine the sample of origin of each cell. Conceptually, this is achieved by separating the signal from the oligo bound to the cell in the sample preparation stage ('positive' HTOs) with ambient counts that arise from contamination

('negative' HTOs). Droplets with no positive HTOs are classified as 'negative' or 'unknown'. Droplets with more than one positive HTO are classified as doublets/multiplets. Those with only one positive HTO are classified as singlets.
Here we use two data sets to assess the performance of each demultiplexing method by comparing the assignments from HTO demultiplexing to assignments from genetic demultiplexing on the same data. Firstly, we suggest some visualisations for assessing the quality of HTO tagging. Next, we compare each method's performance on data whose labelling quality ranges from good to poor. We find that all methods perform similarly when the labelling is of high quality. However, with lower quality labelling, methods that make simplistic, explicit assumptions about the data perform worse than those that take a more flexible approach.

Results

Evaluation data sets

We perform our comparison of hashtag demultiplexing methods on four tagging experiments across two data sets each using different tagging technologies. The first data set, the BAL data set, contains 24 genetically distinct samples of bronchoalveolar lavage fluid tagged with Totalseq-A antibody-derived tags (ADTs) (Stoeckius et al. 2018). These samples were processed in 3 batches of 8 pooled samples, each with 2 captures per batch. Batch 1 contains 24,823 droplets, batch 2 contains 50,668 droplets and batch 3, 64,842 droplets.
We subsequently perform the same analysis on a second data set, the cell line data set, consisting of three human lung cancer cell lines, which are tagged with MULTI-seq lipid-modified oligos (LMOs) (McGinnis et al. 2019). For each data set *vireo* (Y. Huang, McCarthy, and Stegle 2019) is used to assign cells to individuals with default settings (see Methods) and these are used as the "ground truth" to assess the accuracy of the HTO demultiplexing methods.

QC visualisation

To assess the quality of the HTO labelling and sequencing, we present some qualitative visualisations that can guide more quantitative analysis. Figure 1 shows the probability density function (PDF), approximated using kernel density estimation, of the logarithm of counts per cell of each HTO (labelled by the corresponding genetic donor) across the three batches in the BAL data set (Figure 1a-c). The tSNE dimensional reduction of the PCA of log-normalized HTO counts in each batch are also shown (Figure 1d-f).

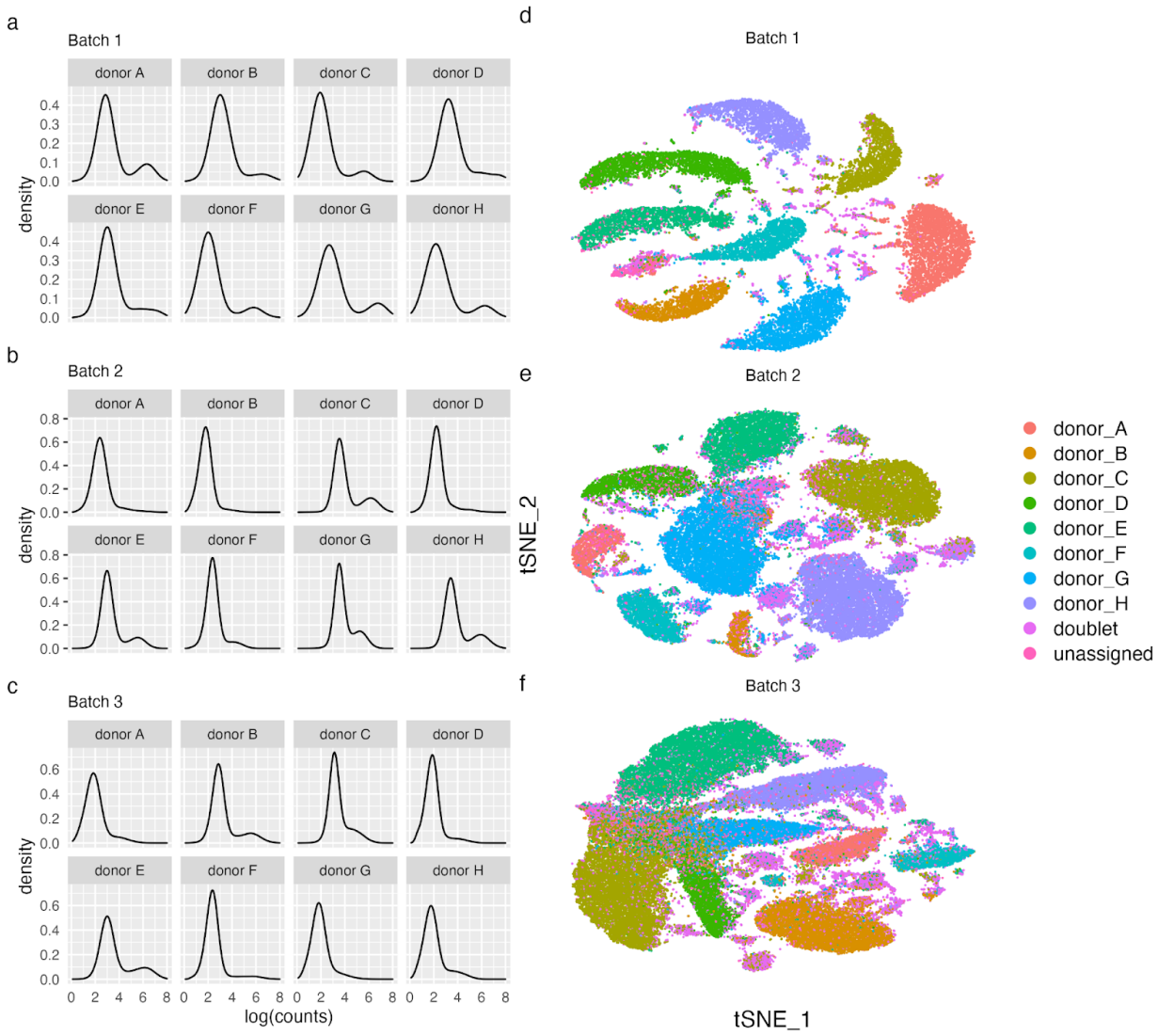

Figure 1: Quality assessment visualisations of the BAL data set. Left column (a-c): Probability density function of the logarithm of HTO counts for each hashtag (labelled by corresponding genetic donor), right column (d-f): tSNE dimensional reduction of HTO counts, coloured by genetic donor in batches 1 (d), 2 (e) and 3 (f). NB: all donors are genetically distinct individuals, generic donor labels are repeated across batches for simplicity.

Each HTO in batch 1 of the BAL data set follows a bimodal distribution (Figure 1a), with a lower peak corresponding to the background counts in the majority of droplets and a higher peak corresponding to the cells from the tagged sample. In batches 2 (Figure 1b) and 3 (Figure 1c), some HTOs (e.g. donor B in batch 2 and donor G in batch 3) appear unimodal, indicating lower

quality labelling. In the right column, the tSNE of batch 1 (Figure 1d) has 8 distinct clusters, corresponding to the 8 individual samples, with a constellation of smaller, interspersed clusters which correspond to doublets and unassigned droplets based on the genetic assignments. Batches 2 (Figure 1e) and 3 (Figure 1f) also show 8 clusters, however, the boundaries of these clusters are closer than in batch 1, and overlap for some samples in batch 3 . While not quantitative, the tSNE plots in Figure 1 indicate that the cells in batch 1 are well-labelled, while those in batches 2 and 3 are labelled more poorly, highlighting that demultiplexing these batches is more challenging. In addition, specific samples within a batch are labelled more poorly than others as indicated by the density plots of the individual HTOs and the overlapping tSNE clusters. Overall, the density and tSNE plots of the HTO counts can be used to evaluate the quality of the HTO labelling. High quality data is indicated by bimodal density plots and tSNE plots with distinct, major clusters corresponding to the number of samples.

Quantitative comparisons of demultiplexing methods

Each of the three batches in our BAL data set contain cells from eight samples, from genetically distinct donors. Each demultiplexing method (including the genetic demultiplexing) can return one of 10 assignments for a cell: singlet, corresponding to one of the eight unique samples; doublet; or negative. We compare six HTO demultiplexing methods: *BFF* (Boggy et al. 2022); *deMULTIplex* (McGinnis et al. 2019); *demuxmix* (Tuddenham et al. 2022); *GMM-Demux* (Xin et al. 2020); *hashedDrops* (Lun, Riesenfeld, et al. 2019) and *HTODemux* (Stoeckius et al. 2018). *BFF* has two modes, BFF_{raw} and $BFF_{cluster}$, and we present the output of both. All of the methods we consider have some adjustable parameters that affect output, however, in our exploration changing the default options does not significantly change the assignments. We discuss the details of each method and their parameters further in the Methods section below. The exception is hashedDrops, which uses a simple counts threshold to distinguish negatives and singlets. We find that in many cases the default value of this threshold is too high, and performance is improved by lowering its value. To illustrate this we present the hashedDrops classifications with both the default value (`confident.min = 2`) and the value we find gives the best performance (`confident.min = 0.5`). As each batch was processed across two captures, we run the demultiplexing methods on HTO data from each individual capture. However, the results are presented per batch for simplicity as we do not observe significant variation between captures within a batch.

In Figure 2, we show the fraction of assignments in each broad category: singlet, doublet or negative (unassigned) from vireo and each hashtag demultiplexing method for the three batches in the BAL data set with variable labelling quality.

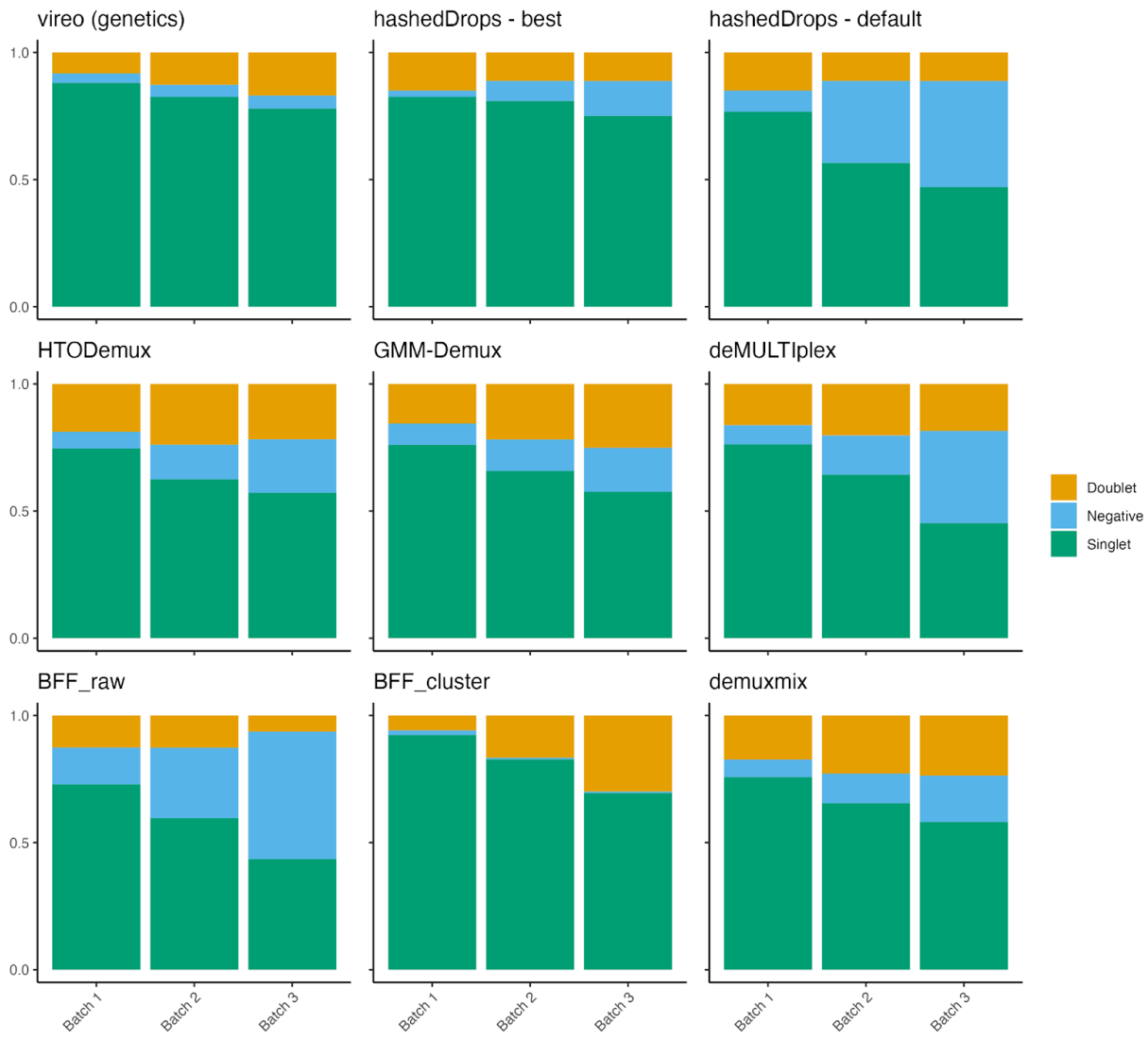

Figure 2: The proportion of cell assignments to singlets, doublets or negative droplets for each demultiplexing method of the BAL data set. Each panel is a method and each bar is a batch.

Two clear trends are apparent in Figure 2. First, *vireo* is able to assign more droplets as singlets than any of the hashtag demultiplexing methods. Second, the hierarchy of HTO tagging quality between the batches suggested by Figure 1 is confirmed in Figure 2. The fraction of negative droplets increases from batch 1 to batch 3 for most methods. The exception to both is $BFF_{cluster}$, which assigns slightly more singlets than *vireo* in batches 1 and 2, and assigns fewer negative droplets in batches 2 and 3 than in batch 1.

We next compare the specific individuals allocated by the singlet assignments of each HTO demultiplexing method to the "ground truth" of genetic assignments from *vireo*. To quantitatively assess their performance, we calculate the F-score (see Methods), a statistic which is the harmonic mean of precision and recall. The F-score ranges between 0 and 1, with a higher

value indicating better performance. Figure 3 shows the F-score of each method for each possible singlet assignment (as well as the mean F-score) for each of the three batches.

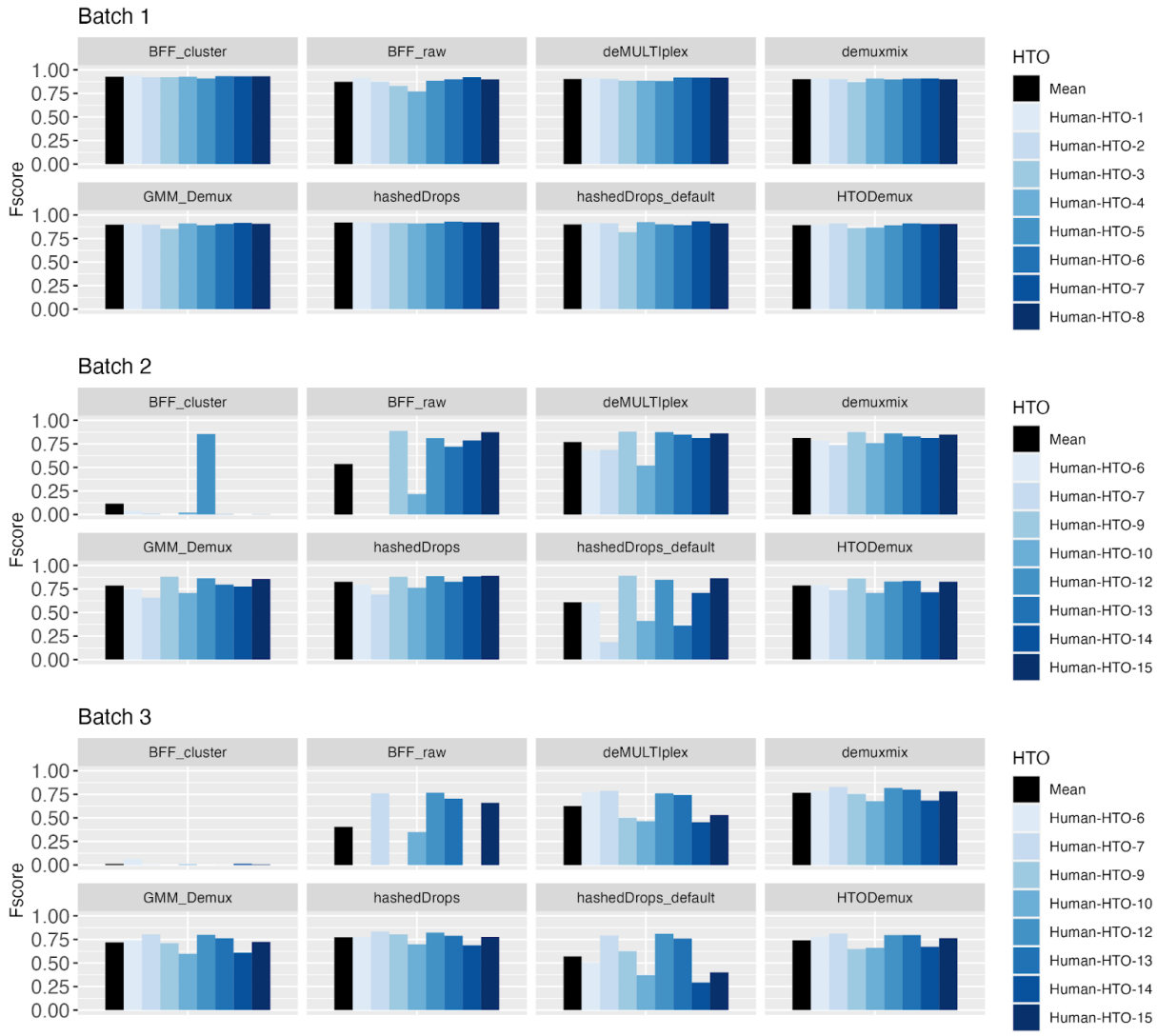

Figure 3: F-scores of each possible singlet assignment with each demultiplexing method for the BAL data set. Batch 1 (top panel), batch 2 (middle panel), batch 3 (bottom panel).

Table 1 shows the mean F-score of each sample for each method, in each of the three batches.

Metric	Batch 1	Batch 2	Batch 3
Genetic singlet fraction	21851/24823	41816/50668	50496/64842
F_{mean} (hashedDrops)	0.919	**0.827**	**0.773**
F_{mean} (hashedDrops - default)	0.901	0.609	0.569
F_{mean} (HTODemux)	0.893	0.788	0.741
F_{mean} (GMM-Demux)	0.899	0.786	0.718
F_{mean} (deMULTIplex)	0.903	0.771	0.627
F_{mean} (BFF_{raw})	0.875	0.537	0.405
F_{mean} ($BFF_{cluster}$)	**0.927**	0.117	0.011
F_{mean} (demuxmix)	0.901	0.814	0.766

Table 1: Singlet fraction and mean F-score of each demultiplexing method for each batch.

Figure 3 and Table 1 show that all methods perform similarly well for batch 1. The overall performance of all methods drops for batches 2 and 3 and some methods begin to show significant performance differences between batches. Notably, $BFF_{cluster}$, which has the highest mean F-score in batch 1, fails for all HTOs except HTO-12 in batch 2, and fails for all HTOs in batch 3. BFF_{raw} is unable to classify any cells from HTO-6 or HTO-7 in batch 2 or HTO-6, HTO-8 and HTO-14 in batch 3. *hashedDrops* has higher scores in all batches with optimised parameters than with the default settings, and has the highest mean F-score of all methods in batches 2 and 3. *Demuxmix*, *GMM-demux* and *HTODemux* show consistent performance across all three batches.

Next, we investigate doublets in more detail. Figure 2 demonstrates that all HTO demultiplexing methods call more doublets and negatives than *vireo*. However, assigning true doublets as singlets is a potentially significant source of error in downstream analysis (Wolock, Lopez, and Klein 2019). Therefore, we reason that it is worse for an algorithm to assign doublets as singlets than to leave them unassigned. In Figure 4 we compare the fraction of genetic doublets assigned by each method as either doublet, singlet or negative.

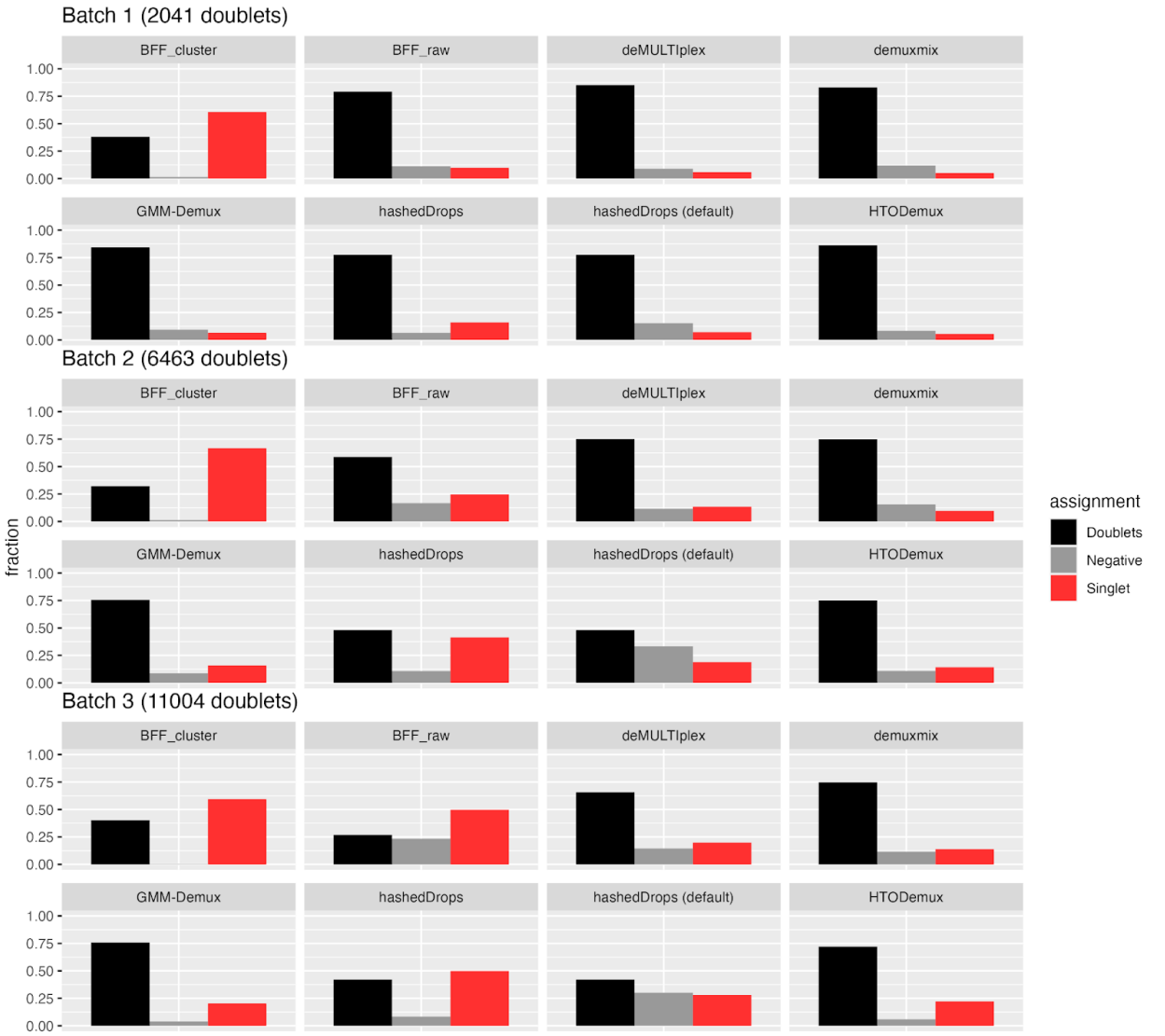

Figure 4: Fraction of genetic doublets assigned by each HTO demultiplexing method to doublets, singlets or negative in the BAL data set.

Figure 4 illustrates several factors not apparent in Figure 3 and Table 1. Firstly, $BFF_{cluster}$, which has the highest F-score for batch 1, unfortunately assigns more than half of the genetic doublets in that batch as singlets. Secondly, while adjusting the parameters of *hashedDrops* from their default values improves the F-score, the number of incorrectly assigned genetic doublets approximately doubles in all batches. Thirdly, the other best-performing methods based on F-score: *demuxmix*, *HTODemux* and *GMM-demux*, perform well on this metric as well, assigning < 20% of genetic doublets as singlets in all batches.

Cell line data

In addition to the BAL data set previously described, we perform the same analysis on a second data set, the cell line data set, consisting of three samples from genetically distinct human lung cancer cell lines. Here, H3122, H358 and H1792 cells were tagged with different MULTI-seq LMOs (McGinnis et al. 2019), a different tagging technology to the ADTs used on the BAL samples. These samples were pooled together and processed in 1 batch across three captures, with 45,977 total droplets.

Since both the ADT and LMO technologies produce an $N_{HTOs} \times N_{droplets}$ counts matrix with similar distributions (see supplementary Figure S6), we expect the demultiplexing methods to perform similarly on the LMO and ADT data. Figure 5 shows the F-score for each method on each of the three samples in this data set, as well as the categorical assignments of genetic doublets.

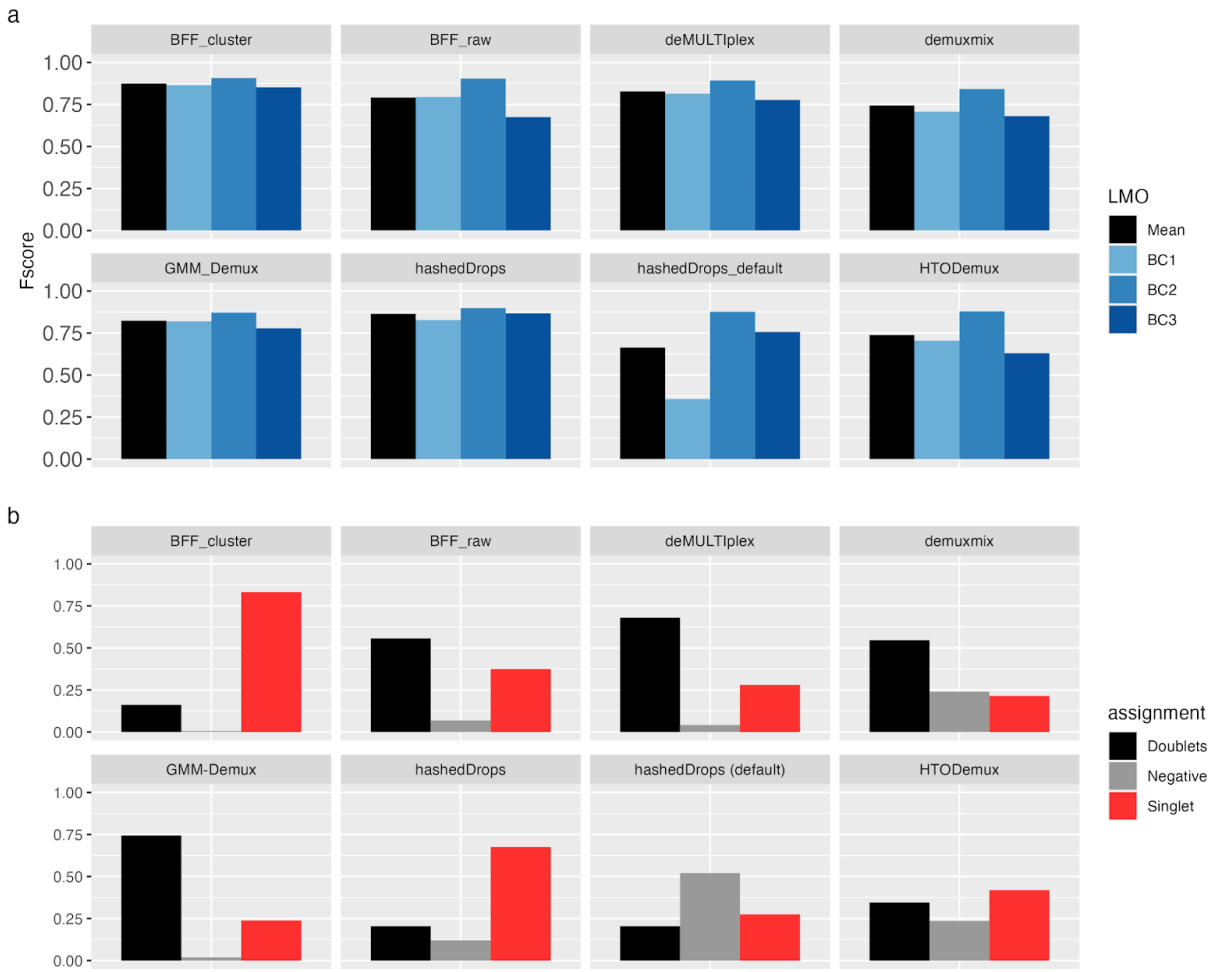

Figure 5: (a) F-score for each demultiplexing method on each sample from the cell line data set. (b) Fraction of the 4945 genetic doublets assigned to different categories by each method.

Figure 5 and Figure S6 show that the LMO data is of comparable quality to batch 1 of the ADT data, and the performance of each of the methods is similar. Based on F-score alone (Figure 5a), $BFF_{cluster}$ performs best, however, looking at Figure 5b we see that more than 75% of genetic doublets are assigned as singlets. Based on the two metrics, we find that *deMULTIplex*, *GMM-demux* and *demuxmix* perform well, *hashedDrops* with default parameters and *HTODemux* perform relatively poorly, and *hashedDrops* with lowered thresholds performs well on the F-score, but misidentifies ≈ 60% of genetic doublets as singlets, more than twice as many as with default values.

Discussion

As sample multiplexing becomes more common in scRNA-seq experiments, reliable demultiplexing of cells becomes paramount. We benchmark six methods for cell demultiplexing based on hashtag oligo data. Of the methods we consider, *demuxmix* shows the best overall performance. However, the difference between *demuxmix*, *GMM-Demux* and *HTODemux* is small, and all should perform similarly well on most data sets. Our results are consistent for hashtagging using ADTs and LMOs, indicating that the performance of the demultiplexing methods is agnostic to the choice of tagging protocol.

Although most of the tools are straightforward to run, and interact well with popular single-cell analysis packages, there are some important usability differences. *Demuxmix* is part of the *Bioconductor* ecosystem and can easily be run in R. As it only requires a HTO counts matrix to return assignments it can be incorporated as part of a *Bioconductor* or *Seurat*-based single-cell analysis pipeline. *HTODemux* is part of the *Seurat* package and requires a *Seurat* object as input, therefore runs most easily alongside other *Seurat* tools for single-cell analysis. *GMM-Demux* is a command-line tool, which may provide a barrier to entry for some users, although wrappers such as *cellhashR* (Boggy et al. 2022) can be used to run it from R.

We demonstrate two simple visualisation methods to assess the quality of hashtag data, and show that if the probability density of counts follows a bimodal distribution, and the counts separate into well-defined clusters on a dimensional reduction plot, then all demultiplexing methods perform well. However, if these conditions are not met, demultiplexing algorithms that explicitly assume bimodal distributions (such as *deMULTIplex* and *BFF*) fail to correctly assign some droplets to their samples of origin. Threshold-based methods, such as *hashedDrops*, can perform well but make a trade-off between greater recovery of singlets and false positives. More sophisticated methods, such as the clustering-based *HTODemux* and *demuxmix*, and the Bayesian estimation-based method *GMM-Demux* perform best and most consistently on both high and low-quality hashtag data.

Low-quality hashtag data does not imply low-quality RNA expression data; importantly, the two are largely uncorrelated (see Figure S1 in Supplementary Materials). We show that the difference between demultiplexing methods becomes more pronounced as the quality of the hashtag data reduces. Therefore, maximising performance of demultiplexing methods on lower-quality hashtag data is particularly important to prevent otherwise good quality cells being excluded in a single-cell analysis.

Methods

Single-cell data generation

The BAL data set is derived from CITE-seq experiments of 24 samples of paediatric BAL. Samples were collected, cryopreserved, and thawed as previously described (Shanthikumar et al. 2020). Live, single cells were sorted using a BD FACS Aria fusion and resuspended in 25µL of cell staining buffer (BioLegend). Human TruStain FcX FC blocking reagent (BioLegend) was added according to manufacturers' instructions for 10min on ice. Each tube was made up to 100µL with cell staining buffer and TotalSeq Hashtag reagents (BioLegend) were added to each sample for 20min on ice. Cells were washed with 3mL cell staining buffer and centrifuged at 400xg for 5min at 4°C. Supernatant was discarded and each sample resuspended at 62,500 cells/100µL following which 100µL of each sample were pooled into one tube. Pooled cells were centrifuged at 400xg for 5min at 4°C, supernatant discarded, and resuspended in 25µL cell staining buffer and 25ul of TotalSeqA Human Universal Cocktail v1.0 (BioLegend) for 30min on ice. This cocktail contains 154 immune related surface proteins. Cells were washed in 3mL cell staining buffer and centrifuged at 400xg for 5min at 4°C. Following two more washes, cells were resuspended in PBS + 0.04% BSA for Chromium captures. Single-cell captures, library preparation, and sequencing was performed as we have described previously (Maksimovic et al. 2022).

For the cell line data set, three human lung cancer cell lines: H1792, H3122 and H358 were labelled with a different 3' lipid modified oligo (LMO) as in (McGinnis et al. 2019). Cell lines were pooled in a 1:1:1 ratio and the pool was used for three separate captures with the 10x Chromium system using the 10X Genomics NextGEM 3′ Single-cell Gene Expression Solution (10x Genomics). Post single-cell capture, scRNA libraries were generated according to the manufacturer's recommendations and LMO library preparation was performed as described previously (McGinnis et al. 2019). LMO count matrices were generated from fastq files using CITE-seq-count v 1.4.3

Genetic demultiplexing

For both data sets, genetic donors were assigned to the samples by first performing SNP genotyping using cellSNP-lite (v1.2.0 for the BAL data; v1.2.1 for the cell line data) (X. Huang and Huang 2021). We used a list of common variants from the 1000 genome project (1000 Genomes Project Consortium et al. 2015) and filtered SNPs with < 20 UMIs or < 10% minor alleles, as recommended in the cellSNP-lite manual. We then used vireoSNP 0.5.6 (Y. Huang, McCarthy, and Stegle 2019) for demultiplexing using the output of cellSNP-lite as the cell data and no additional donor information. More detail is provided in (Maksimovic et al. 2022).

Calculating the F-score

For each possible HTO assignment we calculate the true positive rate TP, which is the fraction of cells with that HTO assignment that have the corresponding vireo assignment; the false positive rate FP, which is the fraction of cells with that HTO assignment and a different genetic assignment; and the false negative rate FN, which is the fraction of cells with the corresponding genetic assignment but a different HTO assignment. Our key metric is the F-score, which is defined as:

$$F = \frac{TP}{TP + 1/2\,(FP + FN)}\,.$$

F is the harmonic mean of the precision and recall, and can vary between 0 and 1, with a higher F-score implying better performance.

Overview of demultiplexing methods in this comparison

hashedDrops

hashedDrops, part of the *DropletUtils* package (Lun, participants in the 1st Human Cell Atlas Jamboree, et al. 2019), is a simple threshold-based classifier. First, the HTO counts matrix is corrected for the ambient counts of each HTO in the data (either before or after filtering out empty droplets). It then ranks the HTO counts in each droplet. Assignments are determined solely by the log-fold change (LFC) between the highest and second highest counts in a droplet, relative to the median counts for that HTO. Firstly, doublets are called where the LFC of the second highest HTO is greater than a user-defined number of median absolute deviations (MAD) above the median and also greater than another user-defined threshold. If a droplet is not assigned as a doublet, singlet assignments are determined by checking that the LFC of the HTO with the highest count in each droplet is greater than a user-defined threshold and is also not less than a user-defined number of MADs below the median. While less sophisticated than other demultiplexing methods, *hashedDrops* has the advantage of making very few assumptions about the data, and is easily configurable by the user. However, as the results are very sensitive to the

choice of the hard thresholds their values should be carefully considered. We explore the effect of varying the singlet threshold parameter in Figure S3 in the supplementary materials.

HTODemux

HTODemux (Stoeckius et al. 2018), included in the *Seurat* package, uses a clustering-based approach. The HTO counts are normalized using the centred log ratio (CLR) transformation. Then an unsupervised *k*-medoids clustering is performed, with $k = N_{HTOs} + 1$. For each HTO, cells are identified as positive or negative in a two step procedure. Firstly, the cluster with the lowest expression count for each HTO is defined as the "negative" cluster, and a negative binomial distribution is fitted to the counts in that cluster. For the droplets outside that cluster, droplets with HTO counts above a user-defined quantile (0.99 by default) are assigned as positive for the HTO. After performing this procedure on all HTOs, droplets that have been assigned positive for more than one HTO are classified as multiplets, droplets with no positive assignments are classified as negative, and the droplets assigned positive for only one HTO are classified as singlets. We explore the effect of varying the quantile threshold in Figure S4 in the supplementary materials.

GMM-Demux

Like *HTODemux*, *GMM-Demux* (Xin et al. 2020) uses the CLR-transformed HTO counts. In well-behaved data, the distribution of the CLR-transformed counts of each HTO is bimodal, with the lower peak corresponding to the 'negative' background and the higher peak corresponding to the true 'positive' counts. *GMM-Demux* fits a two-component Gaussian mixture model to the distribution of each HTO, and uses Bayesian estimation to assign each droplet to the higher-or lower-peaked distribution for each HTO. Droplets with only one positive HTO assignment are classified as singlets, droplets with no positive assignments are classified as negative, while droplets with multiple positive assignments are classified as multiplets, with the identity of the most-probable HTOs in each multiplet included in the output. Every positive assignment is given a confidence score between 0 and 1, and a user-defined confidence threshold (0.8 by default) can be adjusted to be more or less strict with the output classifications. We explore the effect of varying the confidence threshold set in Figure S5 in the supplementary materials.

demuxmix

demuxmix (Tuddenham et al. 2022) is similar to *GMM-Demux* but uses a negative binomial mixture model on the untransformed HTO counts, rather than a mixed Gaussian on the CLR-transformed counts. For each HTO, all cells are clustered into positive and negative clusters using k-means clustering. Cells with very high counts are marked as outliers, and the non-outliers are fitted to a two-component negative binomial distribution using an expectation-maximisation algorithm. *demuxmix* can also leverage the RNA counts to improve performance, using the number of detected genes in the RNA library as a covariate in the mixture model. Using this additional RNA information with the BAL data set showed no

significant improvement on either data set in thi paper, so the results presented are based on the HTO counts only.

deMULTIplex

deMULTIplex (McGinnis et al. 2019) uses an iterative approach. First, a kernel density estimator is used to smooth the log-normalized HTO counts. For each HTO, an initial threshold for positive classification is defined as the highest maximum (assuming a bimodal normalized counts distribution), while the initial threshold for negative classification is the mode. Then, the algorithm sweeps through the quantile range between these two thresholds to find the value that classifies the largest proportion of the data as singlets. Each droplet is then compared against each HTO-specific threshold, being classified as negative, singlet or multiplet based on the number of HTOs for which it passes. All negatively-classified droplets are removed from the counts matrix, and the process is repeated until successive iterations identify no additional negative droplets. While the thresholds for singlets and doublets can be adjusted manually, the default option searches for the value which maximises the fraction of singlet assignments, and our results use this automatic threshold-determining mode.

BFF

Bimodal Flexible Fitting (BFF) (Boggy et al. 2022) also assumes a bimodal counts distribution. It operates in two modes, BFF_{raw} and $BFF_{cluster}$. The first, BFF_{raw} smooths the counts distribution using a kernel density estimator, much like *deMULTIplex*. The threshold between positive and negative classification in this case is the local minimum between the two peaks. The second mode, $BFF_{cluster}$, is similar, but includes an additional layer of normalization, called bimodal quantile normalization, before finalizing classifications. The level of smoothing on the counts can be selected by the user, however our results are based on the default, which searches for an optimal value.

Data availability

Code to reproduce the analysis in this paper, as well as all necessary data is available online at https://github.com/Oshlack/hashtag-demux-paper

Acknowledgements

We thank WEHI Advanced Genomics Facility (Casey Anttila, Ling Ling and Daniela Zalcenstein for cell capture and library preparation, Sarah MacRaild and Stephen Wilcox for sequencing). We also thank Anna Trigos for helpful discussion. This work was supported by CZI Inflammation grant DAF2020-217531, AO NHMRC Ideas Grant GNT1187748 and Investigator Grant GNT1196256.

Bibliography

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P.

Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74.
Bais, Abha S., and Dennis Kostka. 2020. “Scds: Computational Annotation of Doublets in Single-Cell RNA Sequencing Data.” *Bioinformatics* 36 (4): 1150–58.
Boggy, Gregory J., G. W. McElfresh, Eisa Mahyari, Abigail B. Ventura, Scott G. Hansen, Louis J. Picker, and Benjamin N. Bimber. 2022. “BFF and cellhashR: Analysis Tools for Accurate Demultiplexing of Cell Hashing Data.” *Bioinformatics* 38 (10): 2791–2801.
Germain, Pierre-Luc, Aaron Lun, Carlos Garcia Meixide, Will Macnair, and Mark D. Robinson. 2021. “Doublet Identification in Single-Cell Sequencing Data Using *scDblFinder*.” *F1000Research* 10 (September): 979.
Huang, Xianjie, and Yuanhua Huang. 2021. “Cellsnp-Lite: An Efficient Tool for Genotyping Single Cells.” *Bioinformatics* , May. https://doi.org/10.1093/bioinformatics/btab358.
Huang, Yuanhua, Davis J. McCarthy, and Oliver Stegle. 2019. “Vireo: Bayesian Demultiplexing of Pooled Single-Cell RNA-Seq Data without Genotype Reference.” *Genome Biology* 20 (1): 273.
Kang, Hyun Min, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, et al. 2018. “Multiplexed Droplet Single-Cell RNA-Sequencing Using Natural Genetic Variation.” *Nature Biotechnology* 36 (1): 89–94.
Lun, Aaron T. L., participants in the 1st Human Cell Atlas Jamboree, Samantha Riesenfeld, Tallulah Andrews, The Phuong Dao, Tomas Gomes, and John C. Marioni. 2019. “EmptyDrops: Distinguishing Cells from Empty Droplets in Droplet-Based Single-Cell RNA Sequencing Data.” *Genome Biology*. https://doi.org/10.1186/s13059-019-1662-y.
Lun, Aaron T. L., Samantha Riesenfeld, Tallulah Andrews, The Phuong Dao, Tomas Gomes, participants in the 1st Human Cell Atlas Jamboree, and John C. Marioni. 2019. “EmptyDrops: Distinguishing Cells from Empty Droplets in Droplet-Based Single-Cell RNA Sequencing Data.” *Genome Biology* 20 (1): 63.
Maksimovic, Jovana, Shivanthan Shanthikumar, George Howitt, Peter F. Hickey, William Ho, Casey Anttila, Daniel V. Brown, et al. 2022. “Multimodal Single Cell Analysis of the Paediatric Lower Airway Reveals Novel Immune Cell Phenotypes in Early Life Health and Disease.” *bioRxiv*. https://doi.org/10.1101/2022.06.17.496207.
McGinnis, Christopher S., David M. Patterson, Juliane Winkler, Daniel N. Conrad, Marco Y. Hein, Vasudha Srivastava, Jennifer L. Hu, et al. 2019. “MULTI-Seq: Sample Multiplexing for Single-Cell RNA Sequencing Using Lipid-Tagged Indices.” *Nature Methods* 16 (7): 619–26.
Neavin, Drew, Anne Senabouth, Jimmy Tsz Hang Lee, Aida Ripoll, sc-eQTLGen Consortium, Lude Franke, Shyam Prabhakar, et al. 2022. “Demuxafy: Improvement in Droplet Assignment by Integrating Multiple Single-Cell Demultiplexing and Doublet Detection Methods.” *bioRxiv*. https://doi.org/10.1101/2022.03.07.483367.
Shanthikumar, Shivanthan, Matthew Burton, Richard Saffery, Sarath C. Ranganathan, and Melanie R. Neeland. 2020. “Single-Cell Flow Cytometry Profiling of BAL in Children.” *American Journal of Respiratory Cell and Molecular Biology* 63 (2): 152–59.
Stoeckius, Marlon, Shiwei Zheng, Brian Houck-Loomis, Stephanie Hao, Bertrand Z. Yeung, William M. Mauck 3rd, Peter Smibert, and Rahul Satija. 2018. “Cell Hashing with Barcoded Antibodies Enables Multiplexing and Doublet Detection for Single Cell Genomics.” *Genome Biology* 19 (1): 224.
Tuddenham, John F., Mariko Taga, Verena Haage, Tina Roostaei, Charles White, Annie Lee,

Masashi Fujita, et al. 2022. “A Cross-Disease Human Microglial Framework Identifies Disease-Enriched Subsets and Tool Compounds for Microglial Polarization.” *bioRxiv*. https://doi.org/10.1101/2022.06.04.494709.

Wolock, Samuel L., Romain Lopez, and Allon M. Klein. 2019. “Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data.” *Cell Systems* 8 (4): 281–91.e9.

Xin, Hongyi, Qiuyu Lian, Yale Jiang, Jiadi Luo, Xinjun Wang, Carla Erb, Zhongli Xu, et al. 2020. “GMM-Demux: Sample Demultiplexing, Multiplet Detection, Experiment Planning, and Novel Cell-Type Verification in Single Cell Sequencing.” *Genome Biology* 21 (1): 188.