# A hybrid single cell demultiplexing strategy that increases both cell recovery rate and calling accuracy

Lei Li[1#], Jiayi Sun[1#], Yanbin Fu[1#], Siriruk Changrob[1], Joshua J.C. McGrath[1], Patrick C. Wilson[1*]

**Affiliations:**
[1]Gale and Ira Drukier Institute for Children's Health, Weill Cornell Medicine, New York, NY, 10065, USA
[#] These authors contributed equally
*Correspondence: pcw4001@med.cornell.edu (P.C.W.)

## Abstract

Recent advances in single cell RNA sequencing allow users to pool multiple samples into one run and demultiplex in downstream analysis, greatly increasing the experimental efficiency and cost-effectiveness. However, the expensive reagents for cell labeling, limited pooling capacity, non-ideal cell recovery rate and calling accuracy remain great challenges for this approach. To date, there are two major demultiplexing methods, antibody-based cell hashing and Single Nucleotide Polymorphism (SNP)-based genomic signature profiling, and each method has advantages and limitations. Here, we propose a hybrid demultiplexing strategy that increases calling accuracy and cell recovery at the same time. We first develop a computational algorithm that significantly increases calling accuracy of cell hashing. Next, we cluster all single cells based on their SNP profiles. Finally, we integrate results from both methods to make corrections and retrieve cells that are only identifiable in one method but not the other. By testing on several real-world datasets, we demonstrate that this hybrid strategy combines advantages of both methods, resulting in increased cell recovery and calling accuracy at lower cost.

## Highlights

1. An improved algorithm for cell hashing that distinguishes true positive from background for each individual hashtag at higher accuracy
2. This hybrid strategy increases cell recovery and calling accuracy while lowering experimental cost
3. This hybrid demultiplexing strategy is applicable for single-cell RNA sequencing with different donor species, subjects, and cell populations
4. Doublet rate is a major determinant of the performance of SNP-based demultiplexing method

## Introduction

The technical advances in single-cell sequencing have greatly benefited biological and medical research by enhancing our ability to investigate cellular mechanisms of homeostasis and disease in a more precise, high-resolution, and multi-omic fashion(Tang et al., 2009, Svensson et al., 2018, Stuart and Satija, 2019). In the past decade, more and more single-cell methods have been proposed to improve the quality, magnitude, modality, and economy of single-cell experimental approaches(Tang et al., 2009, Picelli et al., 2013, Klein et al., 2015, Macosko et al., 2015, Cao et al., 2017, Stoeckius et al., 2018, Kang et al., 2018, Heaton et al., 2020). Among them, single cell sample pooling and demultiplexing can greatly reduce the per-cell cost, and therefore have been extensively studied.

To date, there are two major single cell demultiplexing methods: cell hashing and genomic signature profiling (also known as SNP profiling). Cell hashing is one of the key techniques that facilitates super-loading and demultiplexing of single-cell samples (Stoeckius et al., 2018). This method is comprised of an experimental protocol and a computational algorithm. Based on Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-seq) technology, cell hashing involves labelling cells from unique samples with an antibody conjugated to a unique oligonucleotide [also called hashtags, or hashtag oligonucleotides (HTOs)]. In the case of hashing, these antibodies are specific for ubiquitously expressed surface antigens such as β2-microglobulin (Stoeckius et al., 2018). After labelling, cells can be pooled and processed together during single-cell preparations (e.g., 10X Genomics). Subsequently, after the resultant library is sequenced, cells' sample identities can be demultiplexed based on their expression level of all hashtags using computational approaches. The genomic signature profiling utilize unique genetic variations, Single Nucleotide Polymorphisms (SNPs), of each subject to determine sample identity of each single cell. The "demuxlet" method is a computational tool that determines the sample identity by using natural genetic variation of each droplet(Kang et al., 2018). Later, another computational tool called Souporcell was developed to cluster single cells based on their genotypes and detect doublets(Heaton et al., 2020), and has been widely used in the community according to its accuracy.

Both demultiplexing methods have their advantages and disadvantages. First, in terms of scope, cell hashing has a much wider application range than genomic signature profiling. Cell hashing can be applied to most datasets regardless of cell type and donor species, whereas genomic signature profiling only works with individuals that have distinct genetic variations (e.g., different human donor), therefore cannot be used in most mice studies or on samples from longitudinal studies of the same donor simultaneously. Second, in terms of economy, cell hashing is more cost-effective despite both methods have extra reagent cost and experimental steps. Specifically, cell hashing adds one extra step to the single-cell sequencing workflow before pooling all samples together: staining cells with HTOs. Computational tool of genomic signature profiling, such as souporcell, can divide all cells into multiple genotype groups based on single cell transcriptional data without any extra experimental steps. However, to link the sample identities to these genotype groups, souporcell requires genetic variation references of all samples generated by bulk DNA- or RNA-seq which results in extensive downstream experimental costs and efforts, including DNA/RNA extraction, library preparation, deep sequencing, and computational analysis. Of note, since cell hashing uses barcoded antibody (HTO), the reagent cost is proportional to total cell numbers of all samples combined whereas the cost of genomic signature profiling is only related to sample numbers. In addition, genomic signatures of the same sample are consistent and therefore can be shared across different datasets (e.g., different cell populations from the same donor), whereas all cell populations must be stained with HTOs when using cell hashing. Third, in terms of performance, genomic signature profiling has much higher cell recovery and accuracy compared with cell hashing. In general, genomic signature profiling has an average of 90% cell recovery rate with decent accuracy and reproducibility, whereas the performance of cell hashing can vary hugely between experiments. Bad staining increases percentages of negative cells and doublet rate also increases with the number of hashtags being used, both of which inevitably produce unidentifiable cells and drop cell recovery rate to 80% or less by sample pair that compounds as additional samples are combined. In conclusion, cell hashing has a wider range of application and higher cost-effectiveness, whereas genomic signature profiling has better performance.

Here, we propose a hybrid single cell demultiplexing strategy that combines both methods and improves computational algorithms, aiming to increase both cell recovery rate and calling accuracy while decreasing the experimental cost. By applying the hybrid strategy on multiple single-cell datasets that have cell hashing profiles and single cell transcriptomes, our results show that the cell recovery rate can be increased to 90% just by running the bioinformatic pipeline, which is significantly higher than the original cell hashing approach. Next, we apply this strategy to multiple integrated datasets that share the same group of donors and demonstrate that it greatly decreases the reagent cost while maintaining the same level of performance. For samples that don't have genetic variations, this strategy can still increase the cell recovery rate to a great extent with our improved demultiplexing algorithm on cell hashing. Therefore, this strategy can be applied to a large variety of single-cell experimental setups and consistently generates high quality results. Together, our hybrid single cell demultiplexing strategy is generalizable, highly cost-effective, and excellent in performance.

# Results

### The existing demultiplexing methods have limitations in performance, efficiency, and economy

To date, two methods for single-cell demultiplexing have been developed and widely used by the community, cell hashing and SNP profiling. However, each method has certain limitations that prevent it from being an universal demultiplexing approach.

Demultiplexing with cell hashing has suffered from low cell recovery rate and reproducibility. Using the algorithm in Seurat package, the overall cell recovery rates of a benchmark dataset (dataset 1, Figure 1-A), two unpublished datasets (dataset 2 and 3, Figure 1-B,C), and two published dataset (dataset 4 and 5, Figure 1-D,E) are 81.67%, 79.57%, 80.31%, 62.29%, and 30.56%, respectively. In dataset 2, 3, and 4, a significant number of cells with low level of a single hashtag were falsely classified as "negative", partially explaining the low cell recovery rates (Figure 1-B,C,D). In dataset 5, 68.98% of cells were not detected for any hashtag, indicating that cell hashing is highly sensitive to antibody staining quality and insufficient staining results in extremely low cell recovery (Figure 1-E). Notably, even in dataset 1, the benchmark dataset with perfect staining, 16% of cells were detected positive for multiple hashtags and classified as "doublets" (Figure 1-A). These were either multiple individually-hashtagged cells sticking together or single cells labeled with other hashtags after sample pooling. As a result, the design of cell hashing inevitably leads to ~20% cells loss, regardless of experimental quality.

SNP profiling has a much narrower range of application and requires extra resources for downstream analysis. Since it exploits natural genetic variations as a reference to distinguish individuals, it can't be used in most mice studies or to separate samples from the same individual. To establish SNP reference, bulk RNA-seq or genomic DNA-seq data of each individual donor must be obtained, making this approach labor-intensive, time-consuming, and less cost-effective. Furthermore, we found that SNP-based demultiplexing could generate incorrect results in datasets with high doublet rate, revealing a critical caveat of this method. We applied SNP-profiling to dataset 6 that is comprised of B cells and T cells from nine human donors and used cell hashing as a quality control for demultiplexing. Shockingly, we observed that each genotype cluster identified by SNP profiling expressed all HTOs evenly instead of exclusively expressing one HTO, indicating that SNP profiling failed to demultiplex dataset 6 (Figure 1-G, Table S1). Further investigation on single cell gene expression revealed that a significant number of cells expressed both B cell (CD19, MS4A1) and T cell (CD3E, IL7R) markers and therefore were most likely doublets with B and T cells in the same droplet (n = 2419, Figure 1-H). Notably, cell hashing also identified a higher-than-normal doublet rate in dataset 6 (3,910 doublets in 14,366 cells), supporting our observation that high doublet rate interferes with the formation of genotype clusters in SNP profiling(Figure 1-I).

**HTOreader, an improved demultiplexing pipeline for cell hashing that increases calling accuracy and cell recovery**

The key step of cell hashing-based demultiplexing approach is to determine the cutoffs that distinguish true positive from background for each individual hashtag. The existing demultiplexing algorithm in Seurat package often generates inaccurate cutoffs for datasets with highly imbalanced sample sizes, resulting in increased mislabeling and decreased recovery rates. To address this problem, we developed a finite-mixture-modeling-based method that increases cutoff calling accuracy for all types of dataset. The pipeline is comprised of four steps: 1) normalize raw counts for each individual hashtag; 2) perform a mixture model to fit normalized data into two Gaussian distributions, representing background and true positive groups; 3) determine the cutoff based on means and standard divisions of the two groups; 4) determine the sample identities of each cell (singlet, doublet, or negative) according to their binding status for each individual hashtag (Figure 2-A). We have implemented this pipeline in R, calling it HTOreader, and have made it compatible with Seurat data so that all Seurat users can easily integrate it into their own pipelines (https://github.com/WilsonImmunologyLab/HTOreader).

We set out to test HTOreader with the perfectly balanced benchmark dataset 1 and compare its performance to the demultiplexing function in Seurat(Stoeckius et al., 2018). Results demonstrated that the performance of HTOreader was comparable to Seurat (Figure S1). Next, we compared their performances on two unpublished datasets (dataset 2 and 3) containing B and T cell data with imbalanced sample sizes (Figure 2-B,E). For dataset 2, Seurat alone failed to determine a proper cutoff for hashtag1 and incorrectly assigned a sizeable amount of hashtag1-labeled cells (n = 1322) into the negative group, resulting in a 13.22% cell loss (Figure 2-B). In contrast, by generating accurate cutoffs with HTOreader, the cell recovery rate (singlet rate) increased from 79.57% to 88.71% (Figure 2-C). For dataset 3, as we expected, Seurat incorrectly assigned a fair amount of hashtag3-labeled cells (n = 435) into negative group, resulting in a 5.98% cell loss (Figure 2-D,E). Again, HTOreader was able to increase the cell recovery rate from 80.31% to 84.74% (Figure 2-F,G). To further demonstrate that this issue commonly exists, we performed the comparison on a published dataset (dataset 4) from our lab (Figure 2-H,I)(Dugan et al., 2021). Unsurprisingly, Seurat mislabeled a significant amount of hashtag2_R2-labeled cells (n = 210, 21.23% of total cell) and HTOreader increased the cell recovery rate from 62.29% to 82.10% (Figure 2-H). To verify the accuracy of HTOreader, we performed SNP profiling to dataset 4 via Souporcell, a highly accurate, genotype-based demultiplexing pipeline (Heaton et al., 2020). Results showed that almost all hashtag2_R2-labeled cells retrieved from negative group by HTOreader were kept in "singlet0" in Souporcell, demonstrating the superior accuracy of HTOreader (Figure 2-I). Therefore, compared to Seurat, HTOreader has similar performance for well-balanced datasets and much higher accuracy and cell recovery rate when it comes to imbalanced, real-world datasets.

### A hybrid demultiplexing strategy that combines cell hashing with SNP profiling increases cell recovery rate and calling accuracy at lower cost

Although HTOreader has greatly improved the calling accuracy and cell recovery rate of cell hashing, the cell loss caused by poor staining quality and high false-doublet rate remain unsolved. To solve these problems, we exploit the recent advances in computational demultiplexing algorithms, such as Souporcell, which clusters single cells based on genetic variations obtained from their transcriptomic data. Here, we propose a hybrid demultiplexing strategy that integrates results from HTOreader (cell hashing) and Souporcell (SNP profiling). Using this hybrid strategy, sample identities of poorly stained cells (negative) and single cells stained by multiple hashtags (false doublet) can be determined by clustering with singly-hashtagged cells (singlet) according to their SNP profiles. Since sample identities have been linked to hashtags, there is no need to generate SNP references separately. Therefore, by performing SNP clustering via Souporcell and integrating the results with cell hashing, both cell recovery and calling accuracy can be increased with no additional cost.

We applied this hybrid strategy to real-world datasets we recently generated to examine its performance. In this dataset, a small aliquots of PBMCs from each of eight donors were stained with hashtag individually and pooled together to sort carrier cells that contained CD19+ B cells and CD4+ T cells (dataset 9), whereas the rest of PBMCs from the eight donors were pooled together to sort antigen-specific B cells (dataset 8) (Figure S2-A,B). We applied our hybrid strategy to demultiplex this integrated dataset by running all cells in dataset 8 and 9 through Souporcell and cells in dataset 9 through HTOreader. Results showed that HTOreader grouped all cells in dataset 9 into ten clusters, including 10,000 cells in eight singlet clusters (77.6%), 2,795 cells in doublet cluster (21.69%) and 127 cells in negative cluster (0.71%) (Figure 3-A, Figure S2-C,D, Table 1). On the other hand, souporcell also groups all cells in dataset 8 and 9 into ten genotype clusters, specifically, 13,837 cells in eight singlet clusters (2,285 in dataset 8 and 11,552 in dataset 9), 1,742 cells in doublet cluster (416 in dataset 8 and 1,326 in dataset 9) and 71 cells in unassigned cluster (63 in dataset 8 and 8 in dataset 9). As expected, the singlet rate of dataset 9 by souporcell reached 89.65%, which is 12.05% higher than that of HTOreader (Figure 3-B, Table 1). To take a closer look, 99.39% of singlets identified by HTOreader were also registered as singlet in souporcell, validating the accuracy of both methods (Figure 3-B). Furthermore, the extra 12.05% singlets recovered in souporcell were almost exclusively from the doublet group in HTOreader (1,541 out of 1664), consistent with observation of high false doublet rate in cell hashing (Figure 3-B, Table 1). Finally, since datasets 8 and 9 shared the same group of donors with the same SNP profiles, cells in dataset 8 that were not labeled with hashtag can be demultiplexed in the same way as dataset 9 and our hybrid strategy achieved an overall cell recovery rate of 88.42% with all cells in the 2 datasets combined (Figure 3-C). Therefore, only a small portion of cells from each donor is required to be properly stained with hashtag when using our strategy, which increases experimental flexibility and decreases reagent cost without sacrificing performance.

We have demonstrated the hybrid demultiplexing strategy increases cell recovery rate by moving false doublet calls into the singlet group that share the same SNP profile. Another benefit of this strategy is to detect errors in souporcell caused by a high true doublet rate by validating SNP profiling results with cell hashing (Figure 3-D). As previously mentioned, we applied the hybrid strategy to dataset 7 and observed extremely poor correlation on singlet calls between HTOreader and souporcell (Figure 1-D, Table S1). We suspected there were unexpectedly high amounts of true doublets that interfere with the unsupervised clustering process of souporcell and make it unable to correctly distinguish genotype clusters. To test our hypothesis, we removed cell clusters consisting of true doublets that express marker genes of both B and T cells and found correlation greatly improved to 54.52% (Figure 3-E, Table S2). We then removed all doublets assigned by cell hashing (Figure 1-H) and obtained 96.06% of correlation (Figure 3-F, Table S3). Our results suggested that souporcell (and most likely other methods that use unsupervised clustering algorithms) is not accurate when applied to cell populations with high percentages of true doublets, while the hybrid strategy can identify and remove doublets to completely recover accuracy.

Together, by integrating results from cell hashing and SNP profiling, the hybrid demultiplexing strategy corrects errors, decreases reagent cost, plus increases accuracy and recovery rate without extra experimental efforts.

### Comparative analysis of the multiple demultiplexing strategy for different types of single-cell datasets

Like detailed before, each demultiplexing strategy has its own application range, strengths and limitations. Here, we compare all available strategies and provide suggestions for demultiplexing different types of single-cell datasets (Figure 4-A).

To demultiplex mice samples or samples from a single donor (e.g. different time points) that share the same genetic background, cell hashing is the only strategy available. For such datasets lacking genetic variants, using HTOreader, the improved demultiplexing algorithm for cell hashing, would increase both cell recovery and accuracy, especially under the condition of bad staining quality. To demultiplex samples from different human donors, cell hashing is optional since both cell hashing and SNP profiling are available. For such datasets, demultiplexing with cell hashing alone leads to reduced cell recovery rate, and demultiplexing with SNP profiling alone requires generation of genetic variant references which leads to extra reagent cost and experimental efforts. Therefore, the best way to demultiplex these datasets is the hybrid strategy that combines results from cell hashing and SNP profiling, ensuring high cell recovery rate and accuracy at low cost and labor. In addition, the hybrid strategy is especially useful for demultiplexing large-scale and complicated datasets. For an integrated analysis of multiple datasets from the same group of donors (e.g. multiple cell populations), the hybrid strategy only requires users to label a small fraction of cells or cells in one dataset with hashtags to link individual hashtag with donors' identities. Then cells from all datasets that do or do not have hashtag profile are clustered together based on their SNP profile where donor identities of each cluster can be determined by the small fraction of hashtagged cells within them. We have made a decision tree of demultiplexing method selection that users can follow and determine the demultiplexing method that works best for their experimental setting and budget (Fig 4-B).

## Discussion

Despite being more and more accessible, single-cell sequencing remains relatively unpopular largely because of the high per-cell cost that is not affordable for many research labs. Thus, methods that allow users to pool multiple samples into one lane and demultiplex in downstream analysis have been aggressively studied. To date, there are 2 major demultiplexing approaches, cell hashing that labels each individual cell from one sample with a unique barcode and SNP profiling that group cells from each subject based on their SNPs. However, each demultiplexing approach has its own limitations and have a lot of room for improvement. Here, we propose a hybrid demultiplexing strategy for single-cell sample pooling and super loading. By integrating results of both cell hashing and SNP profiling, we successfully complement the two approaches with each other and hugely improve their weaknesses. We used this hybrid strategy to demultiplex several real-word datasets and found it increased cell recovery rate and accuracy at lower cost and with less bench work, compared to demultiplexing with either approach alone. Therefore, we developed a hybrid demultiplexing strategy that has better performance and cost-effectiveness than the existing demultiplexing approaches.

R and Python are two major coding languages in single-cell analysis. In the R research community, Seurat package is the most widely used pipeline for downstream analysis because of its versatility, rich documentations, usability, and compatibility with multi-omics. As we demonstrated above, the default demultiplexing algorithm for cell hashing in Seurat performs poorly in terms of cell recovery for datasets with imbalanced sample sizes (Fig 1B-D). We found that Seurat often generates inaccurate cutoffs for each hashtag when staining quality or number of cells varies across samples, causing a significant number of cells with low hashtag signal to be classified as negative. To improve this, we developed a toolkit, HTOreader, to conveniently demultiplex cell hashing data for R users. Compared to Seurat, we demonstrated that HTOreader performs equally well on benchmark dataset and has much higher accuracy and cell recovery rate on real-world datasets with imbalanced sample sizes. The HTOreader is also compatible with the Seurat data format and can be seamlessly integrated into any single-cell analysis workflow in R.

One major caveat of cell hashing is that the cell recovery rate decreases as the number of hashtags being used increases due to the accumulation of false doublets. Based on our experience, it is completely normal to lose 25-40% of cells when using more than 8 hashtags. On the other hand, recent advances in bioinformatics allow users to cluster cells from each human donor according to their SNP profiles obtained from scRNA-seq data. This SNP-based demultiplexing approach is highly accurate and can recovery up to 90% of cells. However, this approach is often considered complicated and expensive because it requires users to generate SNP references for each donor via bulk DNA- or RNA-seq to link donors' identities to cell genotype clusters. Therefore, we devised a hybrid demultiplexing strategy that integrates results of cell hashing and SNP-based clustering so that the two approaches can validate and complement with each other to increase both accuracy and cell recovery. Our results showed that this hybrid strategy increases cell recovery to 90% regardless of staining quality of hashtag. Since only a small fraction of singlet determined by cell hashing is sufficient to link donor identities with each SNP cluster, the generation of SNP references is no longer needed. Additionally, this approach is extremely useful to demultiplex multiple pooled samples from the same group of donors since only cells from one sample need to be hashtagged and all cells can be demultiplexed by SNP-clustering (e.g., dataset 6 + dataset 7, dataset 8 + dataset 9). Lastly, this strategy allows users to use N hashtags to label N+1 donors since cells from unlabeled donor will form a SNP cluster without any hashtag signal.

Sample super-loading and demultiplexing will continuously be important topics for single-cell sequencing. Recently, the 10x Genomics 3' CellPlex, a lipid-based cell hashing technology, provides a promising alternative for super-loading and demultiplexing that might solve some of the issues of antibody-based cell hashing approach. So far, CellPlex is only available for 3' sequencing, which is not compatible with single cell immune profiling and some other applications. However, as single cell sequencing becomes more available, more demultiplexing techniques, both computational and experimental, are going to be available in the near future to greatly benefit biological and medical research.

## Material and Methods

### Datasets

**Dataset 1** A single-cell dataset from human peripheral blood mononuclear cells (PBMCs). This dataset is comprised of 8 individual donors that are uniquely labeled by 8 cell hashtags. This dataset has been published with cell hashing original paper(Stoeckius et al., 2018). Dataset is available from https://www.dropbox.com/sh/ntc33ium7cg1za1/AAD_8XIDmu4F7lJ-5sp-rGFYa?dl=0. More details of this dataset can also be found from Seurat website: https://satijalab.org/seurat/articles/hashing_vignette.html .

**Dataset 2** A novel single-cell dataset generated for this paper, labeled by subject ID, 3V007. In this dataset, mRNA, B Cell Receptor (BCR) repertoire, surface protein expression (CD27 and CD79b), and binding affinity of 14 antigen-probes, including HA proteins of several endemic influenza strains, were measured. Two groups of cells (from same donor), antigen-specific B cell (influenza HA specific) and carrier cells (T cells and B cells) were uniquely labeled by 2 cell hashtags, hashtag1 and hashtag2, respectively. These three hashtags were also sequenced with those antigen-probes. This dataset will be uploaded to public repository upon publication.

**Dataset 3** A novel single-cell dataset generated for this paper, labeled by subject ID, S414. In this dataset, mRNA, B Cell Receptor (BCR) repertoire, surface protein expression (CD27 and CD79b), and binding affinity of 8 antigen-probes, including HA proteins of several endemic influenza strains, were measured. Four groups of cells (from same donor) were uniquely labeled by 4 cell hashtags, hashtag3, hashtag4, hashtag5, and hashtag6. We splitted carrier cells (T cells with a few B cells) into two groups, labeled them with hashtag3 and hashtag4; and splitted antigen-specific B cells into two groups, labeled them with hashtag5 and hashtag6. These four hashtags were also sequenced with those antigen-probes. This dataset will be uploaded to public repository upon publication.

**Dataset 4** This dataset (dataset ID is R125) is from a published single-cell dataset from a previous publication(Dugan et al., 2021). In this dataset, mRNA, B Cell Receptor (BCR) repertoire, and binding affinity of 17 antigen-probes, including Spike, NP, ORF8 and RBD protein of endemic and pandemic COVID strain, HA protein of influenza virus, and interferon alpha and omega, were measured. Cells from 3 individual human donors were uniquely labeled by 3 cell hashtags, hashtag1-R1, hashtag2-R2, and hashtag3-R5. These three hashtags were also sequenced with these antigen-probes. This dataset (R125) is available from Mendeley Data: https://doi.org/10.17632/3jdywv5jrv.3 .

**Dataset 5** This dataset (dataset ID is R6) is from a published single-cell dataset from a previous publication(Dugan et al., 2021). In this dataset, mRNA, B Cell Receptor (BCR) repertoire, and binding affinity of 17 antigen-probes, including Spike, NP, ORF8 and RBD protein of endemic and pandemic COVID strain, HA protein of influenza virus, and interferon alpha and omega, were measured. Cells from 2 time points of an individual human donor were uniquely labeled by 2 cell hashtags, hashtag3-early and hashtag4-late. These three hashtags were also sequenced with these antigen-probes. This dataset (R6) is available from Mendeley Data: https://doi.org/10.17632/3jdywv5jrv.3 .

**Dataset 6** A novel single-cell dataset generated for this paper, labeled as 9pool-CA (carrier). In this dataset, we sorted B cells and T cells from nine subjects and pooled them together. mRNA, and surface protein expression panel were measured. Cells from nine individual human donors were uniquely labeled by nine cell hashtags, hashtag1 to hashtag9. These eight hashtags were also sequenced with those surface proteins. This dataset will be uploaded to public repository upon publication.

**Dataset 7** A novel single-cell dataset generated for this paper, labeled as 9pool-AS (antigen-specific). In this dataset, we sorted antigen-specific B cells from nine human donors (as same as dataset 6) and pooled them together. mRNA, B Cell Receptor (BCR) repertoire, surface protein expression (CD27 and CD79b), and binding affinity of 18 antigen-probes, including HA proteins of several endemic influenza strains, were measured. This dataset will be uploaded to public repository upon publication.

**Dataset 8** A novel single-cell dataset generated for this paper, labeled as 8pool-AS (antigen-specific). In this dataset, we sorted antigen-specific B cells from eight subjects and pooled them together. mRNA, B Cell Receptor (BCR) repertoire, surface protein expression (CD27 and CD79b), and binding affinity of several antigen-probes, including HA proteins of several endemic influenza strains, were measured. This dataset will be uploaded to public repository upon publication.

**Dataset 9** A novel single-cell dataset generated for this paper, labeled as 8pool-CA (carrier). In this dataset, we sorted B cells and T cells from eight subjects (as same as dataset 8) and pooled them together. mRNA, and surface protein expression panel were measured. Cells from eight individual human donors were uniquely labeled by eight cell hashtags, hashtag1 to hashtag8. These eight hashtags were also sequenced with those surface proteins. This dataset will be uploaded to public repository upon publication.

## Cell hashing demultiplexing methods

We developed and introduced an improved demultiplexing approach for single-cell cell hashing, called HTOreader. To accurately determine the hashtag identity for each individual cell, we developed a cutoff calling method that precisely distinguishes true positive from background. Specifically, the distributions of normalized counts for each hashtag were first fitted into two Gaussian distributions, representing background and true positive groups. Then a cutoff value that distinguishes the two groups was calculated based on means and standard divisions of these two Gaussian distributions. Finally, the identity of each individual cell was determined according to the hashtags they were positive of.

**Data normalization** Two normalization methods: Centered Log-Ratio (CLR) and Log ($log1p$) normalization are available. CLR method is more common in normalization of CITE-seq protein expression and hashtags(Aitchison, 1982, Hao et al., 2021). For a given raw counts vector $W$ of a hashtag, the CLR normalization will be:

$$CLR(W) = [log\frac{w_1}{g(w)}, log\frac{w_2}{g(w)}, \dots, log\frac{w_n}{g(w)}]$$

Where $n$ is the length of vector $W$, and $g(w) = (\prod_{i=1}^{n} w_i)^{1/n}$ denotes the geometric mean of $W$. The conventional Log normalization works well in some datasets. We use $log1p$ to avoid the undefined log[0]. For a given raw counts vector $W$ of a hashtag, the Log normalization will be:

$$Log(W) = [\log(w_1 + 1), \log(w_2 + 1), \dots, \log(w_n + 1)]$$

**Mixture modeling** Mixture modeling has been extensively used in single-cell data pre-processing, such as estimation of the drop-out rate, determination of effective sequencing depth and amplification noise(Fan et al., 2016, Kharchenko et al., 2014). We adopted an mixture modeling approach implemented in the Flexmix package to fit two Gaussian distributions from a vector of normalized hashtag counts(Leisch, 2004). In this step, we fit normalized data of each hashtag into two Gaussian distributions indicating one positive group, representing background and true positive groups, and calculate the means and standard deviations of these two groups respectively.

**Cutoff determination** For two Gaussian distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, $\mu_1 < \mu_2$, we determine the cutoff to distinguish true positive and background using the following equation:

$$Cutoff = \mu_1 + \frac{\sqrt[n]{\sigma_1}}{\sqrt[n]{\sigma_1} + \sqrt[n]{\sigma_2}}(\mu_2 - \mu_1)$$

Where $n$ is the rank of the model, the recommended rank is 2 in most cases. Please see the supplementary material for details.

**Sample identity assignment** For each cell, we assign their sample identities based on their binding status of each hashtag(Hao et al., 2021). If a cell is deemed positive for only one hashtag, it will be labeled as a singlet for that corresponding hashtag; if it's deemed positive to multiple hashtags, it will be labeled as doublet; if it's deemed background for all hashtags, it will be labeled as negative. Sample identities of every cell labeled as singlet will be assigned according to their hashtag identities.

**Genomic signature demultiplexing**

Genomic signature demultiplexing method is an essential part of this hybrid demultiplexing strategy, and to date many computational methods are available, for example demuxlet and souporcell. In this paper, we applied souporcell, which has been widely used in the community, onto our strategy to demonstrate the effectiveness of this workflow. As most SNP-based demultiplexing methods do, souporcell aligns all short reads against a reference genome to get the SNPs for each cell, then groups cells into multiple clusters according to their genotypes (SNP signatures) using an unsupervised learning algorithm. The number of genotype clusters is pre-defined by users according to the number of subjects in the pooled sample. For a sample pooled from N subjects (individual human donors), there will be N+2 distinct genotype clusters identified, in which one "doublet" cluster containing cells fit in more than one genotypes, one "negative" cluster containing cells whose SNP signatures are not sufficient therefore cannot be fit into any genotype, and N singlet clusters indicating cells from N individual donors.

### Flow cytometry staining and cell sorting

Flow staining and cell sorting were performed as previously described (Dugan et al., 2021). Briefly, human PBMCs were thawed in 10% FBS RPMI1640 medium and enriched by negative selection using a pan-B cell isolation kit according the manufacturer's instruction (StemCell, Cat#. 19554) prior to staining with the following antibodies and flurorescently oligonucleotide-labeled streptavidin-antigen tetramers (Biolegend) : anti-huCD19-PE-Cy7, anti-huCD3-BB515, anti-huCD4-BB515, anti-huIgD-BB515, TotalSeq-C anti-human hashtag antibodies, antigen-PE or-APC, and at 4 degree for 30 mins. Cells were subsequently washed three times with 2% FBS PBS buffer supplemented with 2mM D-biotin. Finally, cells were adjusted at a maximum of 2 million cells per ml in washing buffer, stained with DAPI and subjected to sorting by either MACSQuantTyto (Miltenyi) or BD Melody (BD). Cells that were viable/CD19$^+$/antigen-PE$^+$ and antigen-APC$^+$ or viable/CD4$^+$ were sorted for downstream 10X Genomics processing.

### 10X Genomics libraries construction and Next Generation Sequencing

5' gene expression, VDJ and surface protein feature libraries were prepared using the 10X genomics platform as per the manufacturer's instructions (Chromium Next GEM Single Cell 5' (HT) Reagent Kits v2 (Dual Index)). Three libraries were quantified by real-time quantitative PCR using KAPA Library Quanitification Kits (Roche) and pooled at recommended ratio and sequenced using NextSeq1000 (Illumina) with 26 cycles for read 1, 10 cycles for i7/i5 index, 150 cycles for read 2.

## Acknowledgements

This project was funded in part by the National Institute of Allergy and Infectious Diseases (NIAID; National Institutes of Health grant numbers U19AI082724 (PCW), U19AI109946 (PCW), U19AI057266 (PCW), the NIAID Centers of Excellence for Influenza Research and Surveillance (CEIRS) grant number HHSN272201400005C (PCW), and the NIAD Centers of Excellence for Influenza Research and Response (CEIRR) grant number 75N93019R00028 (PCW). This work was also partially supported by the NIAID Collaborative Influenza Vaccine Innovation Centers (CIVIC; 75N93019C00051). In addition, this work was also partially funded by the NIAID Centers of Excellence for Influenza Research and Surveillance (CEIRS, contract # HHSN272201400008C), the NIAID Centers of Excellence for Influenza Research and Response (CEIRR, contract# 75N93021C00014) and by anonymous donors. JJCM was supported by a CIHR Banting Postdoctoral Fellowship (BPF-186528).

## Author Contribution

LL designed the model and workflow, implemented the package, performed computational analyses of single-cell data, and wrote the manuscript. JS and YF collected samples, designed the workflow, performed experiments, and wrote the manuscript.SC and JJCM collected samples, helped in workflow design, and revised the manuscript. PCW supervised the work, designed the workflow, and wrote the manuscript.

## Declaration of interests

The authors declare no competing interests.

## References

AITCHISON, J. 1982. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological),* 44**,** 139-160.

CAO, J., PACKER, J. S., RAMANI, V., CUSANOVICH, D. A., HUYNH, C., DAZA, R., QIU, X., LEE, C., FURLAN, S. N., STEEMERS, F. J., ADEY, A., WATERSTON, R. H., TRAPNELL, C. & SHENDURE, J. 2017. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science,* 357**,** 661-667.

CHANGROB, S., FU, Y., GUTHMILLER, J. J., HALFMANN, P. J., LI, L., STAMPER, C. T., DUGAN, H. L., ACCOLA, M., REHRAUER, W. & ZHENG, N.-Y. 2021. Cross-Neutralization of Emerging SARS-CoV-2 Variants of Concern by Antibodies Targeting Distinct Epitopes on Spike. *Mbio,* 12**,** e02975-21.

DUGAN, H. L., STAMPER, C. T., LI, L., CHANGROB, S., ASBY, N. W., HALFMANN, P. J., ZHENG, N.-Y., HUANG, M., SHAW, D. G. & COBB, M. S. 2021. Profiling B cell immunodominance after SARS-CoV-2 infection reveals antibody evolution to non-neutralizing viral targets. *Immunity,* 54**,** 1290-1303. e7.

FAN, J., SALATHIA, N., LIU, R., KAESER, G. E., YUNG, Y. C., HERMAN, J. L., KAPER, F., FAN, J.-B., ZHANG, K. & CHUN, J. 2016. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nature methods,* 13**,** 241-244.

HAO, Y., HAO, S., ANDERSEN-NISSEN, E., MAUCK III, W. M., ZHENG, S., BUTLER, A., LEE, M. J., WILK, A. J., DARBY, C. & ZAGER, M. 2021. Integrated analysis of multimodal single-cell data. *Cell,* 184**,** 3573-3587. e29.

HEATON, H., TALMAN, A. M., KNIGHTS, A., IMAZ, M., GAFFNEY, D. J., DURBIN, R., HEMBERG, M. & LAWNICZAK, M. K. 2020. Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nature methods,* 17**,** 615-620.

KANG, H. M., SUBRAMANIAM, M., TARG, S., NGUYEN, M., MALISKOVA, L., MCCARTHY, E., WAN, E., WONG, S., BYRNES, L. & LANATA, C. M. 2018. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature biotechnology,* 36**,** 89-94.

KHARCHENKO, P. V., SILBERSTEIN, L. & SCADDEN, D. T. 2014. Bayesian approach to single-cell differential expression analysis. *Nature methods,* 11**,** 740-742.

KLEIN, ALLON M., MAZUTIS, L., AKARTUNA, I., TALLAPRAGADA, N., VERES, A., LI, V., PESHKIN, L., WEITZ, DAVID A. & KIRSCHNER, MARC W. 2015. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell,* 161**,** 1187-1201.

LEISCH, F. 2004. Flexmix: A general framework for finite mixture models and latent glass regression in R.

LI, L., DUGAN, H. L., STAMPER, C. T., LAN, L. Y.-L., ASBY, N. W., KNIGHT, M., STOVICEK, O., ZHENG, N.-Y., MADARIAGA, M. L. & SHANMUGARAJAH, K. 2021. Improved integration of single-cell transcriptome and surface protein expression by LinQ-View. *Cell Reports Methods,* 1**,** 100056.

MACOSKO, EVAN Z., BASU, A., SATIJA, R., NEMESH, J., SHEKHAR, K., GOLDMAN, M., TIROSH, I., BIALAS, ALLISON R., KAMITAKI, N., MARTERSTECK, EMILY M., TROMBETTA, JOHN J., WEITZ, DAVID A., SANES, JOSHUA R., SHALEK, ALEX K., REGEV, A. & MCCARROLL, STEVEN A. 2015. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell,* 161**,** 1202-1214.

PICELLI, S., BJORKLUND, A. K., FARIDANI, O. R., SAGASSER, S., WINBERG, G. & SANDBERG, R. 2013. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods,* 10**,** 1096-8.

STOECKIUS, M., ZHENG, S., HOUCK-LOOMIS, B., HAO, S., YEUNG, B. Z., MAUCK, W. M., SMIBERT, P. & SATIJA, R. 2018. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biology,* 19**,** 224.

STUART, T. & SATIJA, R. 2019. Integrative single-cell analysis. *Nature Reviews Genetics,* 20**,** 257-272.

SVENSSON, V., VENTO-TORMO, R. & TEICHMANN, S. A. 2018. Exponential scaling of single-cell RNA-seq in the past decade. *Nature protocols,* 13**,** 599.

TANG, F., BARBACIORU, C., WANG, Y., NORDMAN, E., LEE, C., XU, N., WANG, X., BODEAU, J., TUCH, B. B., SIDDIQUI, A., LAO, K. & SURANI, M. A. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods,* 6**,** 377-382.

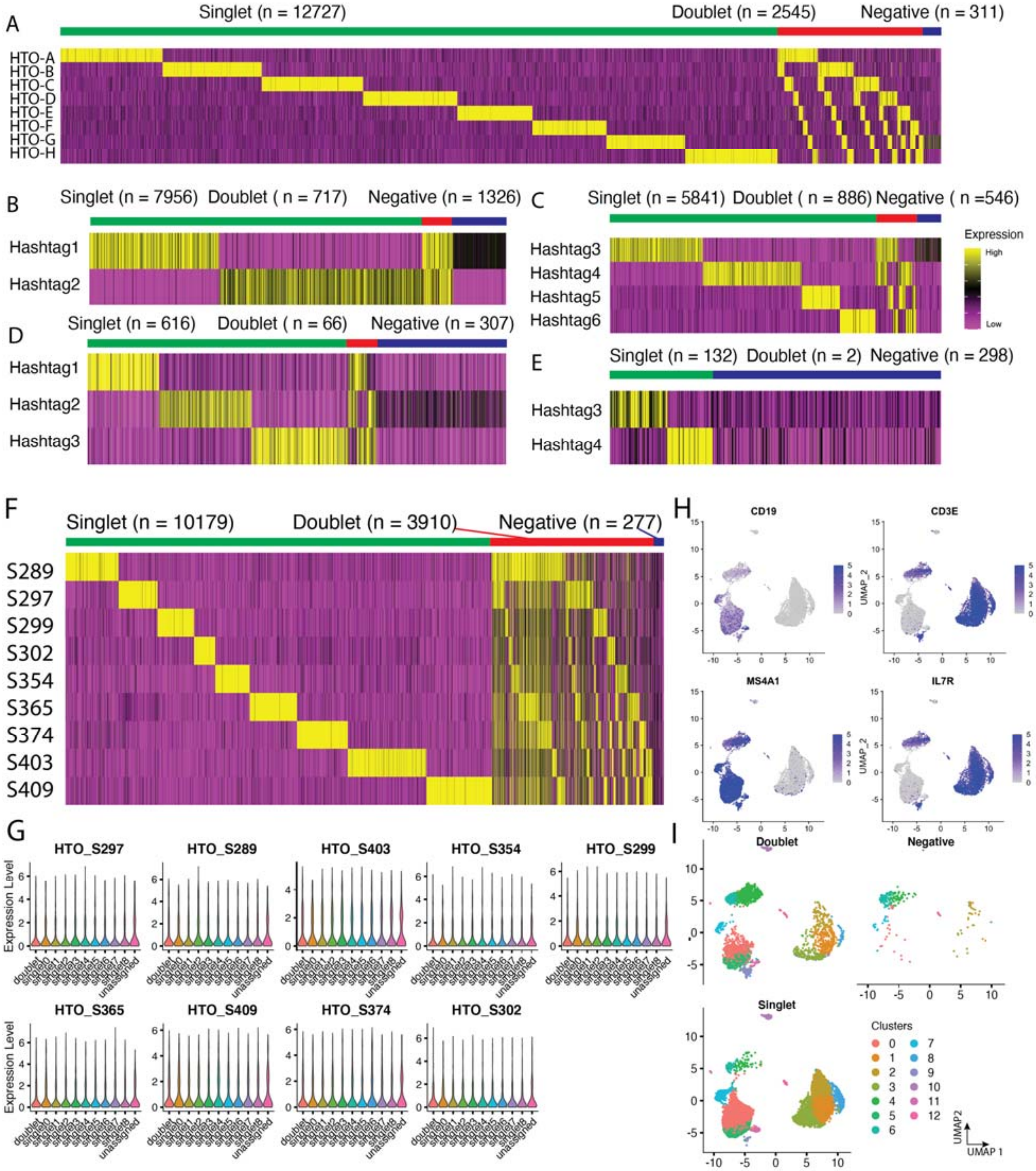

**Figure 1**. Limitations of cell hashing method and SNP calling method revealed by real-world datasets. (**A - E**) Heatmap of expression of hashtags in dataset 1, 2, 3, 4, and 5, respectively. Singlet, doublet, and negative groups were indicated by a color bar with number of cells in each group. (**F**) Expression of nine cell hashtags on dataset 6. Singlet, doublet, and negative groups are indicated on the top of the heatmap. Panels A – F share the same scale bar of heatmap to the right. (**G**) Expression of nine hashtags on genotype clusters generated by SNP calling method. (**H**) Expression of two B cell gene markers (CD19 and MS4A1) and two T cell gene markers (CD3E and IL7R) visualized on a UMAP embedding of carrier cells of dataset 6. (**I**) Doublets,

negative cells, and singlets identified by cell hashing method are visualized on a UMAP individually. Cells are colored by transcriptome clusters.

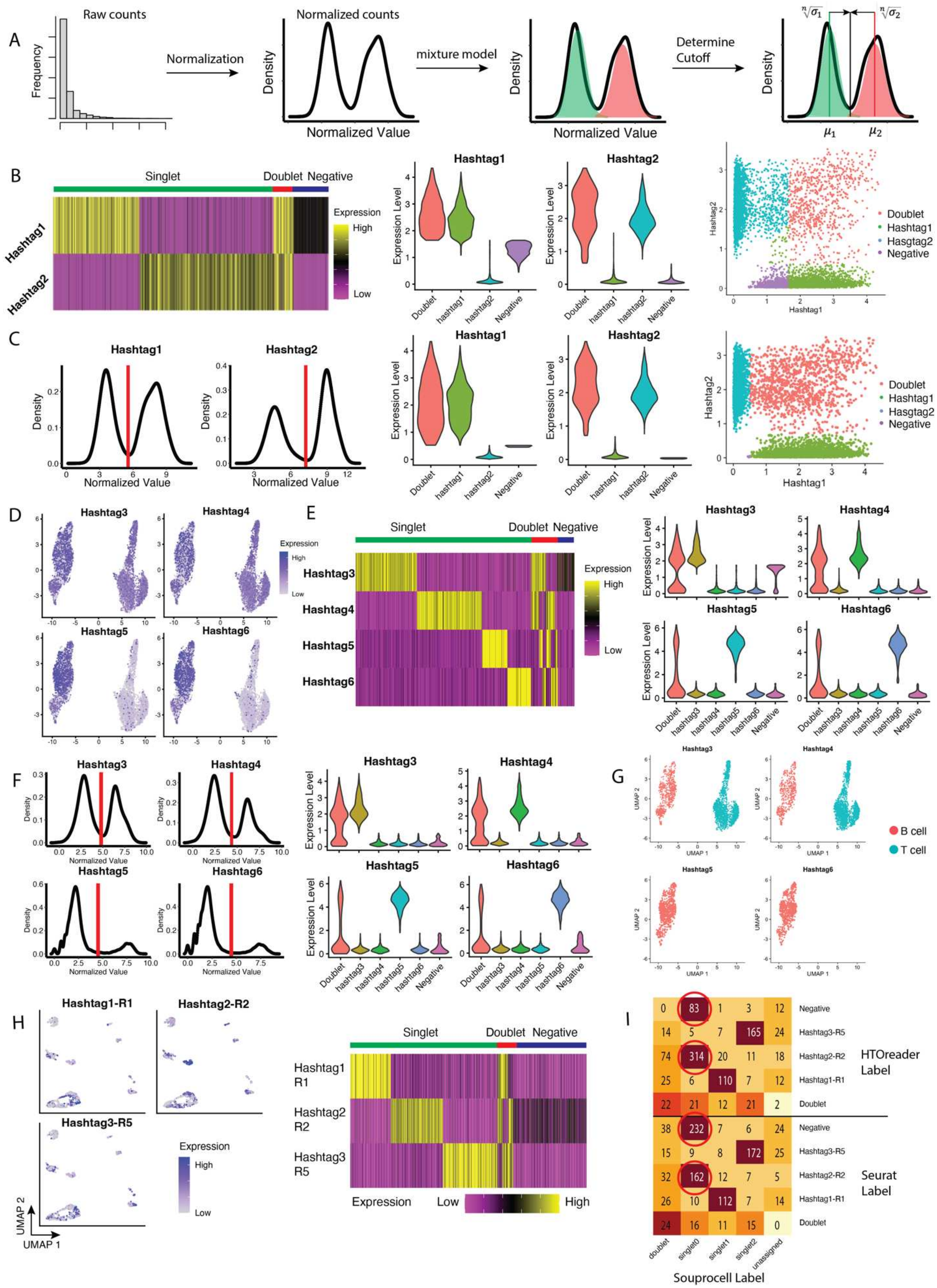
A
Raw counts
Frequency
Normalization
Normalized counts
Density
Normalized Value
mixture model
Density
Normalized Value
Determine Cutoff
$\sqrt[n]{\sigma_1}$
$\sqrt[n]{\sigma_2}$
Density
$\mu_1$
$\mu_2$
B
Singlet
Doublet
Negative
Hashtag1
Hashtag2
Expression
High
Low
Hashtag1
Hashtag2
Expression Level
Doublet
hashtag1
hashtag2
Negative
Hashtag2
Hashtag1
Doublet
Hashtag1
Hasgtag2
Negative
C
Hashtag1
Hashtag2
Density
Normalized Value
Hashtag1
Hashtag2
Expression Level
Doublet
hashtag1
hashtag2
Negative
Hashtag2
Hashtag1
Doublet
Hashtag1
Hasgtag2
Negative
D
Hashtag3
Hashtag4
Hashtag5
Hashtag6
Expression
High
Low
E
Singlet
Doublet
Negative
Hashtag3
Hashtag4
Hashtag5
Hashtag6
Expression
High
Low
Hashtag3
Hashtag4
Hashtag5
Hashtag6
Expression Level
Doublet
hashtag3
hashtag4
hashtag5
hashtag6
Negative
F
Hashtag3
Hashtag4
Hashtag5
Hashtag6
Density
Normalized Value
Hashtag3
Hashtag4
Hashtag5
Hashtag6
Expression Level
Doublet
hashtag3
hashtag4
hashtag5
hashtag6
Negative
G
Hashtag3
Hashtag4
Hashtag5
Hashtag6
UMAP 1
UMAP 2
B cell
T cell
H
Hashtag1-R1
Hashtag2-R2
Hashtag3-R5
Expression
High
Low
UMAP 2
UMAP 1
Singlet
Doublet
Negative
Hashtag1 R1
Hashtag2 R2
Hashtag3 R5
Expression
Low
High
I
0 83 1 3 12 Negative
14 5 7 165 24 Hashtag3-R5
74 314 20 11 18 Hashtag2-R2
25 6 110 7 12 Hashtag1-R1
22 21 12 21 2 Doublet
HTOreader Label
38 232 7 6 24 Negative
15 9 8 172 25 Hashtag3-R5
32 162 12 7 5 Hashtag2-R2
26 10 112 7 14 Hashtag1-R1
24 16 11 15 0 Doublet
Seurat Label
doublet
singlet0
singlet1
singlet2
unassigned
Souprocell Label

**Figure 2**. HTOreader achieves high accuracy and increases cell recovery for cell hashing datasets. (**A**) The workflow of HTOreader to determine proper cutoff for each cell hashtag. (**B**) Cell demultiplexing by Seurat on dataset2. Left: expression heatmap of both hashtags with Seurat labels annotation. Center: violin plot of expression of both hashtags on Seurat groups. Right: Scatter plot of hashtag expression with Seurat label color coding. (**C**) Cell demultiplexing by HTOreader on dataset2. Left: cutoffs determined by HTOreader based on density of normalized values for both hashtags; Center: violin plot of expression of both hashtags on HTOreader groups; Right: Scatter plot of hashtag expression with HTOreader label color coding. (**D**) Expression of four hashtags on the UMAP embedding of dataset3. (**E**) Cell demultiplexing by Seurat on dataset3. Left: expression heatmap of all hashtags with Seurat labels annotation. Right: violin plot of expression of all hashtags on Seurat groups. (**F**) Cell demultiplexing by HTOreader on dataset3. Left: cutoffs determined by HTOreader based on density of normalized values for all hashtags; Right: violin plot of expression of all hashtags on HTOreader groups. (**G**) Cells labeled by different hashtags demultiplexed by HTOreader. (**H**) Cell demultiplexing by Seurat on dataset4. Left:  Expression of three hashtags on the UMAP embedding of dataset4. Right: expression heatmap of all hashtags with Seurat labels annotation. (**I**) Comparing HTOreader and Seurat using Souprocell as benchmark on dataset4. Numbers of cells are indicated in the heatmap.

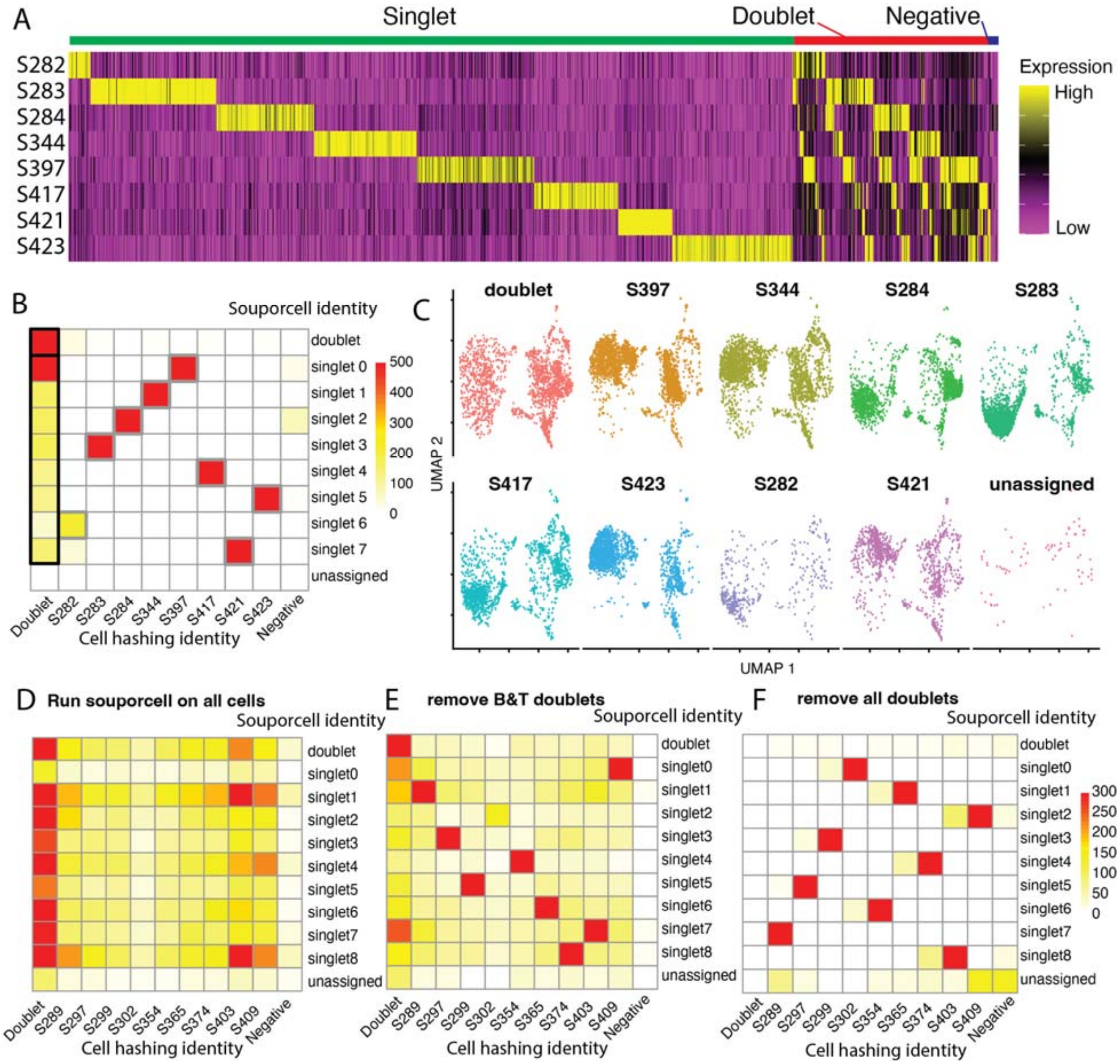

**Figure 3**. The hybrid strategy combines cell hashing technology and genomic signature method and achieves higher recovery rate and calling accuracy on a real-world immune cell dataset. (**A**) Expression of eight cell hashtags on dataset 6. Singlet, doublet, and negative groups are indicated on the top of the heatmap. (**B**) A heatmap of correlation between cell hashing demultiplexing and SNP-based demultiplexing on all cells of dataset 9. True doublets and false doublets were highlighted in thick black border, and genotype cluster-cell hashing pairs were highlighted in thick gray border. (**C**) Cells demultiplexed by this hybrid strategy of dataset 5 and dataset 6 are visualized on an integrated UMAP individually. (**D**) A heatmap of correlation between cell hashing demultiplexing and SNP-based demultiplexing on all cells of dataset 6. (**E**) A heatmap of correlation between cell hashing demultiplexing and SNP-based demultiplexing on all cells of dataset 7 after remove B&T doublets (cells that express both B and T cell markers). (**F**) A heatmap of correlation between cell hashing demultiplexing and SNP-based demultiplexing on all cells of dataset 7 after all potential doublets (cells that express more than one hashtag).

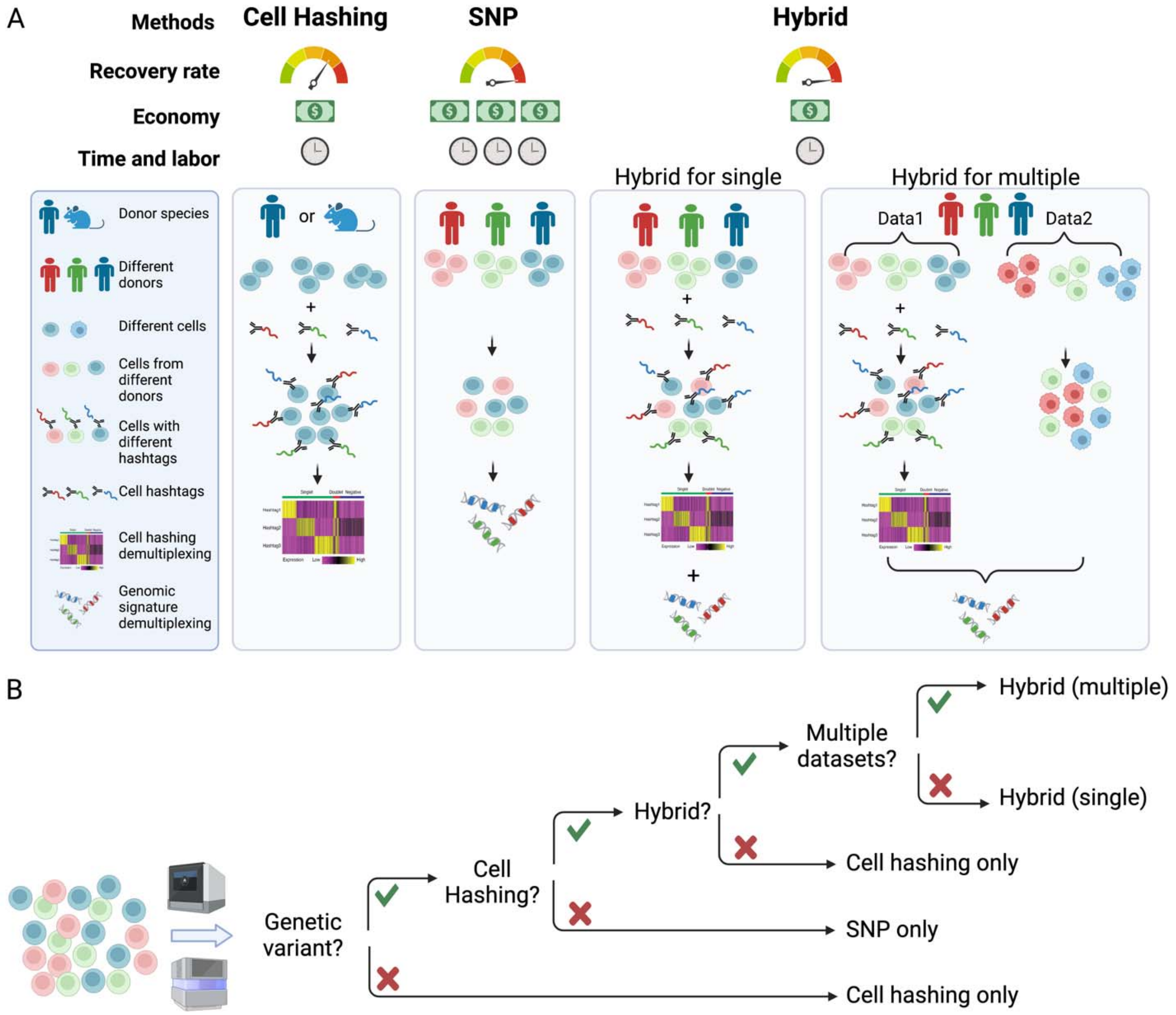

**Figure 4.** Optimized demultiplexing workflows for single-cell datasets with different experimental technics, sample numbers and donor species. (**A**) overall comparison among two existing demultiplexing methods and the hybrid method. The hybrid method performs best in recovery rate, economy, and labor saving among three demultiplexing methods. (**B**) A decision tree for users to select demultiplexing method according to their experiment design.

**Table 1**. Hybrid demultiplexing on dataset 8 and dataset 9.

| | | **SNP demultiplexing** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **doublet** | **singlet 0** | **singlet 1** | **singlet 2** | **singlet 3** | **singlet 4** | **singlet 5** | **singlet 6** | **singlet 7** | **unassigned** |
| **Hashtag demultiplexing on Dataset 9** | **Doublet** | 1213 | 588 | 163 | 167 | 176 | 123 | 125 | 58 | 141 | 5 |
| | **Negative** | 4 | 27 | 1 | 76 | 1 | 5 | 12 | 1 | 0 | 0 |
| | **S282** | 37 | 0 | 0 | 1 | 0 | 0 | 0 | 220 | 43 | 1 |
| | **S283** | 14 | 0 | 0 | 1 | 1682 | 0 | 0 | 1 | 0 | 0 |
| | **S284** | 4 | 0 | 0 | 1385 | 0 | 0 | 0 | 0 | 0 | 0 |
| | **S344** | 18 | 0 | 1450 | 0 | 1 | 0 | 0 | 7 | 0 | 0 |
| | **S397** | 7 | 1627 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| | **S417** | 13 | 1 | 0 | 0 | 1 | 1128 | 1 | 2 | 0 | 2 |
| | **S421** | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 724 | 0 |
| | **S423** | 13 | 0 | 0 | 0 | 0 | 0 | 1612 | 0 | 0 | 0 |
| **Dataset 8** | | 416 | 391 | 868 | 71 | 130 | 447 | 28 | 118 | 232 | 63 |

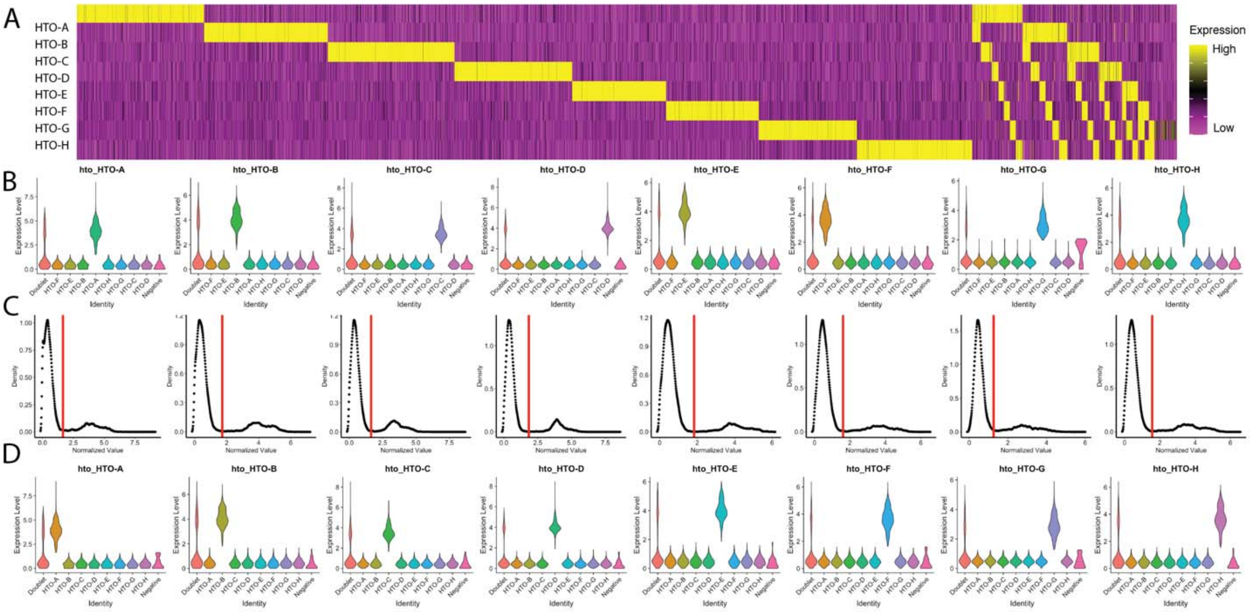

**Figure S1**. Comparing cell demultiplexing on a public benchmark dataset using both Seurat and HTOreader. (**A**) expression heatmap of all eight hashtags ordered by Seurat labels. (**B**) violin plot of expression of all eight hashtags on Seurat groups. (**C**) cutoffs determined by HTOreader based on density of normalized values for all eight hashtags. (**D**) violin plot of expression of all eight hashtags on HTOreader groups.

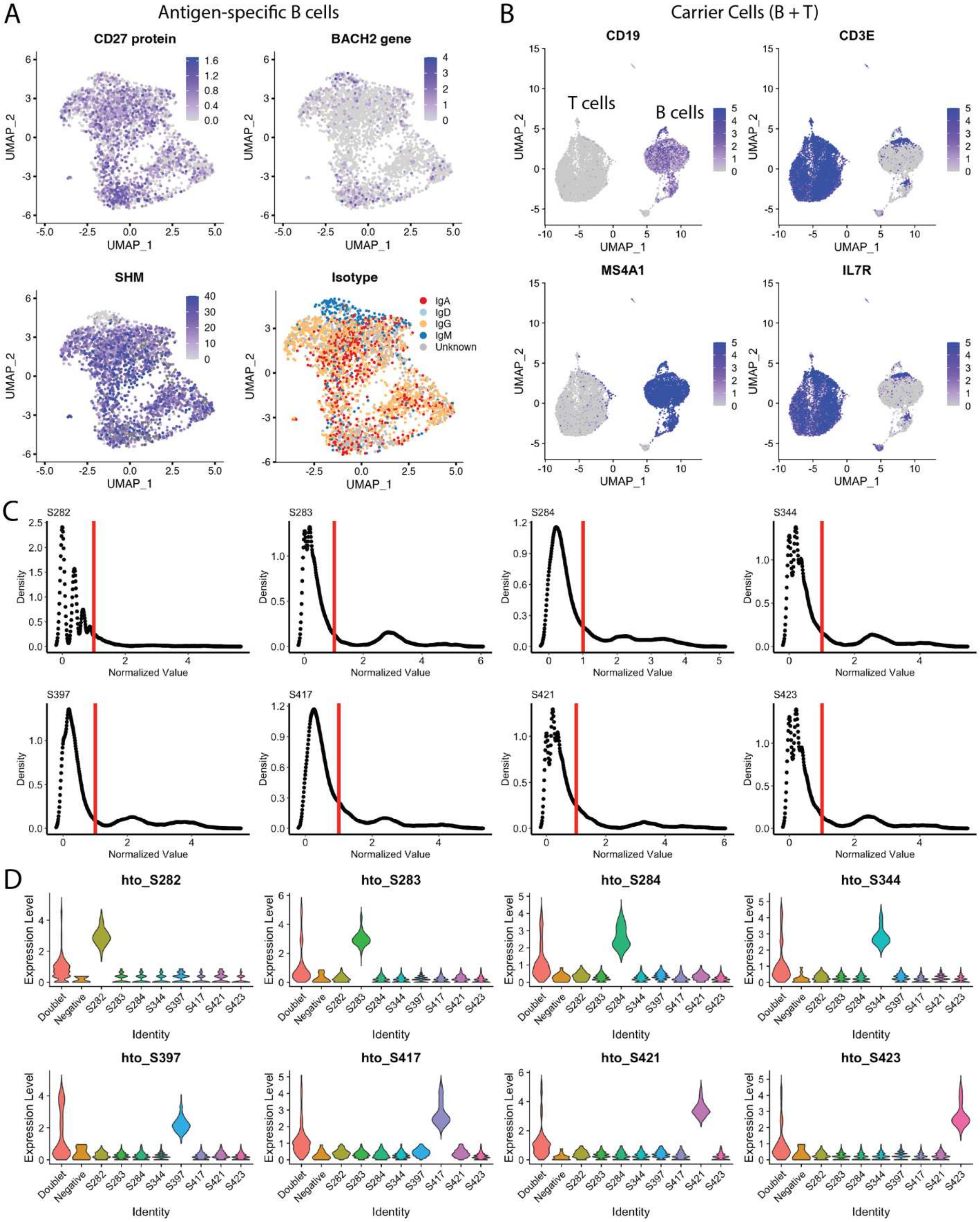

**Figure S2**. Cell demultiplexing on a B cells and T cells pooled dataset (dataset 8 and dataset 9) using hybrid method. (**A**) expression of CD27 protein, expression of BACH2 gene, Somatic hypermutation (SHM) of heavy chain, and isotype of BCR repertoire visualized on a UMAP embedding of antigen-specific B cells of dataset 8. (**B**) Expression of two B cell gene markers

(CD19 and MS4A1) and two T cell gene markers (CD3E and IL7R) visualized on a UMAP embedding of carrier cells of dataset 9. T cell and B cell clusters are roughly indicated by labels (**C**) Cutoffs determined by HTOreader based on density of normalized values for all eight hashtags on dataset 9. (**D**) Violin plot of expression of all eight hashtags on HTOreader demultiplexing groups on dataset 9.

**Table S1**. Hybrid demultiplexing on all cells of dataset 6.

| | | Hashtag demultiplexing | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Doublet | S289 | S297 | S299 | S302 | S354 | S365 | S374 | S403 | S409 | Negative |
| SNP demultiplexing | doublet | 488 | 161 | 102 | 96 | 58 | 104 | 147 | 142 | 222 | 169 | 38 |
| | singlet0 | 134 | 39 | 27 | 34 | 26 | 27 | 42 | 35 | 63 | 58 | 7 |
| | singlet1 | 624 | 195 | 139 | 137 | 78 | 148 | 171 | 200 | 321 | 237 | 60 |
| | singlet2 | 371 | 178 | 62 | 83 | 44 | 60 | 106 | 123 | 164 | 142 | 19 |
| | singlet3 | 261 | 81 | 86 | 72 | 28 | 64 | 75 | 80 | 146 | 93 | 12 |
| | singlet4 | 462 | 129 | 108 | 103 | 60 | 87 | 120 | 146 | 195 | 225 | 34 |
| | singlet5 | 237 | 83 | 55 | 68 | 29 | 61 | 76 | 63 | 114 | 102 | 13 |
| | singlet6 | 372 | 106 | 91 | 82 | 41 | 82 | 90 | 152 | 178 | 126 | 20 |
| | singlet7 | 293 | 98 | 73 | 68 | 34 | 63 | 115 | 94 | 141 | 128 | 22 |
| | singlet8 | 570 | 216 | 141 | 105 | 53 | 105 | 138 | 158 | 355 | 224 | 41 |
| | unassigned | 98 | 25 | 29 | 16 | 15 | 17 | 23 | 27 | 47 | 34 | 11 |

**Table S2**. Hybrid demultiplexing on dataset 6 after remove all B&T doublets.

| | | Hashtag demultiplexing | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Doublet | S289 | S297 | S299 | S302 | S354 | S365 | S374 | S403 | S409 | Negative |
| SNP demultiplexing | doublet | 344 | 50 | 48 | 39 | 18 | 56 | 50 | 47 | 70 | 49 | 6 |
| | singlet0 | 214 | 116 | 55 | 54 | 50 | 41 | 60 | 72 | 96 | 930 | 10 |
| | singlet1 | 185 | 521 | 99 | 74 | 45 | 48 | 80 | 105 | 131 | 88 | 14 |
| | singlet2 | 94 | 74 | 25 | 34 | 145 | 22 | 35 | 42 | 43 | 50 | 4 |
| | singlet3 | 137 | 83 | 388 | 45 | 32 | 30 | 67 | 81 | 60 | 57 | 4 |
| | singlet4 | 80 | 44 | 23 | 31 | 16 | 381 | 38 | 32 | 37 | 20 | 6 |
| | singlet5 | 123 | 51 | 39 | 406 | 32 | 36 | 69 | 57 | 47 | 49 | 1 |
| | singlet6 | 143 | 67 | 63 | 49 | 35 | 52 | 510 | 85 | 66 | 71 | 9 |
| | singlet7 | 258 | 127 | 69 | 57 | 50 | 57 | 62 | 85 | 1262 | 94 | 15 |
| | singlet8 | 168 | 94 | 61 | 63 | 31 | 57 | 90 | 573 | 89 | 77 | 11 |
| | unassigned | 92 | 24 | 25 | 3 | 8 | 31 | 23 | 32 | 30 | 39 | 5 |

**Table S3**. Hybrid demultiplexing on dataset 6 after remove all doublets identified by cell hashing method.

| | | Hashtag demultiplexing | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Doublet | S289 | S297 | S299 | S302 | S354 | S365 | S374 | S403 | S409 | Negative |
| SNP demultiplexing | doublet | 0 | 15 | 12 | 17 | 8 | 13 | 17 | 13 | 28 | 14 | 26 |
| | singlet0 | 0 | 0 | 0 | 32 | 414 | 0 | 0 | 0 | 0 | 0 | 4 |
| | singlet1 | 0 | 0 | 0 | 0 | 0 | 47 | 999 | 0 | 0 | 1 | 6 |
| | singlet2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 91 | 1382 | 26 |
| | singlet3 | 0 | 0 | 23 | 809 | 0 | 0 | 0 | 2 | 0 | 0 | 7 |
| | singlet4 | 0 | 0 | 0 | 0 | 0 | 1 | 59 | 110 | 0 | 0 | 6 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | 0 | | | |
| | singlet5 | 0 | 14 | 852 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| | singlet6 | 0 | 0 | 0 | 0 | 37 | 730 | 0 | 0 | 1 | 0 | 10 |
| | singlet7 | 0 | 1205 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 |
| | singlet8 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 71 | 1798 | 0 | 30 |
| | unassigned | 0 | 77 | 26 | 6 | 7 | 27 | 27 | 33 | 28 | 141 | 148 |