Stoeckius et al. Genome Biology (2018) 19:224
https://doi.org/10.1186/s13059-018-1603-1

Genome Biology

METHOD Open Access

Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics

Marlon Stoeckius[1†], Shiwei Zheng[2†], Brian Houck-Loomis[1], Stephanie Hao[1], Bertrand Z. Yeung[3], William M. Mauck III[2], Peter Smibert[1*] and Rahul Satija[2*]

Abstract

Despite rapid developments in single cell sequencing, sample-specific batch effects, detection of cell multiplets, and experimental costs remain outstanding challenges. Here, we introduce Cell Hashing, where oligo-tagged antibodies against ubiquitously expressed surface proteins uniquely label cells from distinct samples, which can be subsequently pooled. By sequencing these tags alongside the cellular transcriptome, we can assign each cell to its original sample, robustly identify cross-sample multiplets, and "super-load" commercial droplet-based systems for significant cost reduction. We validate our approach using a complementary genetic approach and demonstrate how hashing can generalize the benefits of single cell multiplexing to diverse samples and experimental designs.

Introduction

Single cell genomics offers enormous promise to transform our understanding of heterogeneous processes and to reconstruct unsupervised taxonomies of cell types [1, 2]. As studies have progressed to profiling complex human tissues [3, 4] and even entire organisms [5, 6], there is a growing appreciation of the need for massively parallel technologies and datasets to uncover rare and subtle cell states [7–9]. While the per-cell cost of library prep has dropped, routine profiling of tens to hundreds of thousands of cells remains costly both for individual labs and for consortia such as the Human Cell Atlas [10].

Broadly related challenges also remain, including the robust identification of artifactual signals arising from cell multiplets or technology-dependent batch effects [11]. In particular, reliably identifying expression profiles corresponding to more than one cell remains an unsolved challenge in single-cell RNA-seq (scRNA-seq) analysis, and a robust solution would simultaneously improve data quality and enable increased experimental throughput. While multiplets are expected to generate higher complexity libraries compared to singlets, the strength of this signal is not sufficient for unambiguous identification [11]. Similarly, technical and "batch" effects have been demonstrated to mask biological signal in the integrated analysis of scRNA-seq experiments [12], necessitating experimental solutions to mitigate these challenges.

Recent developments have poignantly demonstrated how sample multiplexing can simultaneously overcome multiple challenges [13, 14]. For example, the demuxlet [13] algorithm enables the pooling of samples with distinct genotypes together into a single scRNA-seq experiment. Here, the sample-specific genetic polymorphisms serve as a fingerprint for the sample of origin and therefore can be used to assign each cell to an individual after sequencing. This workflow also enables the detection of multiplets originating from two individuals, reducing non-identifiable multiplets at a rate that is directly proportional to the number of multiplexed samples. While this elegant approach requires pooled samples to originate from previously genotyped individuals, in principle, any approach assigning sample fingerprints that can be measured alongside scRNA-seq would enable a similar strategy. For instance, sample multiplexing is frequently utilized in flow and mass cytometry by labeling distinct

* Correspondence: psmibert@nygenome.org; rsatija@nygenome.org
†Marlon Stoeckius and Shiwei Zheng contributed equally to this work.
[1]Technology Innovation Lab, New York Genome Center, New York, NY, USA
[2]NYU Center for Genomics and Systems Biology, New York Genome Center, New York, NY, USA
Full list of author information is available at the end of the article

samples with antibodies to the same ubiquitously expressed surface protein but conjugated to different fluorophores or isotopes, respectively [15–17].

We recently introduced CITE-seq [18], where oligonucleotide-tagged antibodies are used to convert the detection of cell surface proteins into a sequenceable readout alongside scRNA-seq. We reasoned that a defined set of oligo-tagged antibodies against ubiquitous surface proteins could uniquely label different experimental samples. This enables us to pool these together and use the barcoded antibody signal as a fingerprint for reliable demultiplexing. We refer to this approach as Cell Hashing, based on the concept of hash functions in computer science to index datasets with specific features; our set of oligo-derived hashtags equally define a "lookup table" to assign each multiplexed cell to its original sample. We demonstrate this approach by labeling and pooling eight human PBMC samples and running them simultaneously in a single droplet-based scRNA-seq run. Cell hashtags allow for robust sample multiplexing, confident multiplet identification, and discrimination of low-quality cells from ambient RNA. In addition to enabling "super-loading" of commercial scRNA-seq platforms to substantially reduce costs, this strategy represents a generalizable approach for multiplet identification and multiplexing that can be tailored to any biological sample or experimental design.

Results

Hashtag-enabled demultiplexing based on ubiquitous surface protein expression

We sought to extend antibody-based multiplexing strategies [16, 17] to scRNA-seq using a modification of our CITE-seq method [18]. We initially chose a set of monoclonal antibodies directed against ubiquitously and highly expressed immune surface markers (CD45, CD98, CD44, and CD11a), combined these antibodies into eight identical pools (pool A through H), and subsequently conjugated each pool to a distinct Hashtag oligonucleotide (henceforth referred to as HTO, Fig. 1a; "Methods" section). The HTOs contain a unique 12-bp barcode that can be sequenced alongside the cellular transcriptome, with only minor modifications to standard scRNA-seq protocols. We utilized an improved and simplified conjugation chemistry compared to our previous approach [18], by using iEDDA click chemistry to covalently attach oligonucleotides to antibodies [19] ("Methods" section).

We designed our strategy to enable CITE-seq and Cell Hashing to be performed simultaneously, but to generate separate sequencing libraries. Specifically, the HTOs contain a different amplification handle than our standard CITE-seq antibody-derived tags (ADT) ("Methods" section). This allows HTOs, ADTs, and scRNA-seq libraries to be independently amplified and pooled at desired quantities. Notably, we have previously observed robust recovery of antibody signals from highly expressed epitopes due to their extremely high copy number. This is in contrast to the extensive "dropout" levels observed for scRNA-seq data and suggests that we can faithfully recover HTOs from each single cell, enabling assignment to sample of origin with high fidelity.

To benchmark our strategy and demonstrate its utility, we obtained peripheral blood mononuclear cells (PBMCs) from eight separate human donors (referred to as donors A through H) and independently stained each sample with one of our HTO-conjugated antibody pools, while simultaneously performing a titration experiment with a pool of seven immunophenotypic markers ("Methods" section) for CITE-seq. We subsequently pooled all cells together in equal proportion, alongside an equal number of unstained HEK293T cells (and 3% mouse NIH-3T3 cells) as negative controls, and ran the pool in a single lane on the 10x Genomics Chromium Single Cell 3′ v2 system. Following the approach in Kang et al. [13], we "super-loaded" the 10x Genomics instrument, loading cells at a significantly higher concentration with an expected yield of 20,000 single cells and 5000 multiplets. Based on Poisson statistics, 4365 multiplets should represent cell combinations from distinct samples and can potentially be discarded, leading to an unresolved multiplet rate of 3.1%. Notably, achieving a similar multiplet rate without multiplexing would yield ~ 4000 singlets. As the cost of commercial droplet-based systems is fixed per run for sample preparation, multiplexing therefore allows for the profiling of ~ 400% more cells for the same cost.

We performed partitioning and reverse transcription according to the standard protocols, utilizing only a slightly modified downstream amplification strategy ("Methods" section) to generate transcriptome, HTO, and ADT libraries. We pooled and sequenced these on an Illumina HiSeq2500 (two rapid run flowcells), aiming for a 90%:5%:5% contribution of the three libraries in the sequencing data. Additionally, we performed genotyping of all eight PBMC samples and HEK293T cells with the Illumina Infinium CoreExome array, allowing us to utilize both HTOs and sample genotypes (assessed by demuxlet [13]) as independent demultiplexing approaches.

When examining pairwise expression of two HTO counts, we observed relationships akin to "species-mixing" plots (Fig. 1b), suggesting mutual exclusivity of HTO signal between singlets. Extending beyond pairwise analysis, we developed a statistical model to classify each barcode as "positive" or "negative" for each HTO ("Methods" section). Briefly, we modeled the "background" signal for each HTO independently as a

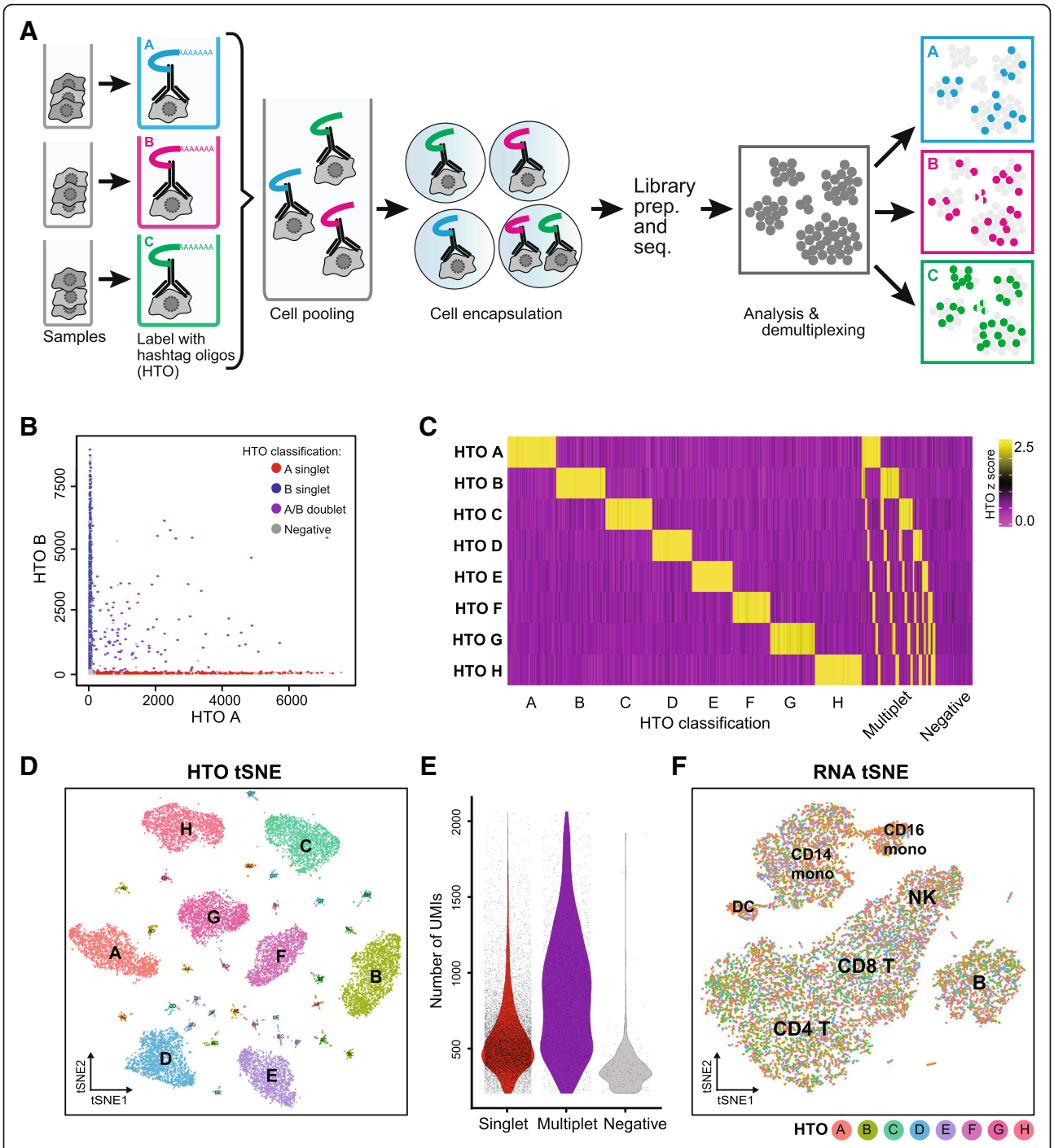

Fig. 1 Sample multiplexing using DNA-barcoded antibodies. **a** Schematic overview of sample multiplexing by Cell Hashing. Cells from different samples are incubated with DNA-barcoded antibodies recognizing ubiquitous cell surface proteins. Distinct barcodes (referred to as hashtag-oligos, HTO) on the antibodies allow pooling of multiple samples into one scRNA-seq experiment. After sequencing, cells can be assigned to their sample of origin based on HTO levels ("Methods" section). **b** Representative scatter plot showing raw counts for HTO A and HTO B across all cell barcodes. Both axes are clipped at 99.9% quantiles to exclude visual outliers. **c** Heatmap of scaled (*z*-scores) normalized HTO values based on our classifications. Multiplets express more than one HTO. Negative populations contain HEK293T and mouse NIH-3T3 cells that were spiked into the experiments as negative controls. **d** tSNE embedding of the HTO dataset. Cells are colored and labeled based on our classifications. Eight singlet clusters and all 28 cross-sample doublet clusters are clearly present. **e** Distribution of RNA UMIs per cell barcode in cells that were characterized as singlets (red), multiplets (violet) or negatives (grey). **f** Transcriptome-based clustering of single-cell expression profiles reveals distinct immune cell populations interspersed across donors. B, B cells; T, T cells; NK, natural killer cells; mono, monocytes; DC, dendritic cells. Cells are colored based on their HTO classification (donor ID), as in **d**

negative binomial distribution, estimating background cells based on the results of an initial *k*-medoids clustering of all HTO reads ("Methods" section). Barcodes with HTO signals above the 99% quantile for this distribution were labeled as "positive," and barcodes that were "positive" for more than one HTO were labeled as multiplets. We classified all barcodes where we detected at least 200 RNA UMI, regardless of HTO signal.

Our classifications (visualized as a heatmap in Fig. 1c) suggested a clear identification of 8 singlet populations, as well as multiplet groups. We also identified barcodes with negligible background signal for any of the HTOs (labeled as "negatives"), consisting primarily (86.5%) of HEK293T and mouse cells. We removed all HEK293T and mouse cells from downstream analyses ("Methods" section), with the remaining barcodes representing 14,002 singlets and 2974 identifiable multiplets, in line with expectations. Our classifications were also fully concordant with a tSNE embedding, calculated using only the 8 HTO signals, which enabled the clear visualization not only of the 8 groups of singlets (donors A through H) but also the 28 small groups representing all possible doublet combinations (Fig. 1d). Moreover, we observed a clear positive shift in the distribution of RNA UMI per barcode for multiplets, as expected (Fig. 1e), while the remaining negative barcodes expressed fewer UMIs and may represent failed reactions or "empty" droplets containing only ambient RNA. These results strongly suggest that HTOs successfully assigned each barcode into its original sample and enabled robust detection of cross-sample multiplets. The large dynamic range of RNA UMI per cell barcode in multiplets (Fig. 1e) illustrates the difficulty of unambiguous multiplet assignment based on higher UMI counts. , and we observe the same challenges with total HTO signal (Additional file 1: Figure S1A). Performing transcriptomic clustering of the classified singlets enabled clear detection of seven hematopoietic subpopulations, which were interspersed across all 8 donors (Fig. 1f).

Genotype-based demultiplexing validates Cell Hashing

We next compared our HTO-based classifications to those obtained by demuxlet [13]. Overall, we observed a strong concordance between the techniques, even when considering the precise sample mixture in called doublets (Fig. 2a). Exploring the areas of disagreement, we identified 871 barcodes that were classified based on HTO levels as singlets but were identified as "ambiguous" by demuxlet. Notably, the strength of HTO classification for these discordant barcodes (represented by the number of reads assigned to the most highly expressed HTO) was identical to the barcodes that were classified as singlets by both approaches (Fig. 2b). However, discordant barcodes did have reduced RNA UMI counts (Fig. 2c). We conclude that these barcodes likely could not be genetically classified at our relatively shallow sequencing depth (~ 24,115 reads per cell), which is below the recommended depth for using demuxlet, but likely represent true single cells based on our HTO classifications.

In addition, we also observed 2528 barcodes that received discordant singlet/doublet classifications between the two techniques (Fig. 2d). We note that this does reflect a minority of barcodes (compared to 13,421 concordant classifications) and that in these discordant cases, it is difficult to be certain which of these methods is correct. However, when we examined the UMI distributions of each classification group, we observed that only barcodes classified as doublets by both techniques exhibited a positive shift in transcriptomic complexity (Fig. 2d). This suggests that these discordant calls are largely made up of true singlets and represent conservative false positives from both methods, perhaps due to ambient RNA or HTO signal. Consistent with this interpretation, when we restricted our analysis to cases where demuxlet called barcodes as doublets with > 95% probability, we observed a 75% drop in the number of discordant calls (Fig. 2e). Demuxlet requires sufficient numbers of reads and SNPs to unequivocally classify a cell to a donor, and as expected, discordantly classified cells had lower numbers of sequencing reads and SNPs (Additional file 1: Figure S2A-D).

Finally, we also observed a rare number of cases where both Cell Hashing and demuxlet classified cells as singlets but with discordant (216/11,464; 1.9%) donor classifications. To investigate further, we took advantage of the fact that all donors (A–G) except one (H) were also stained with CITE-seq antibodies, and therefore, donor H cells should not contain ADT reads. However, in 40 instances where demuxlet, but not Cell Hashing, classified cells as donor H, we observed robust (> 1000) ADT counts in 37 cases, suggesting that these discordant calls are misclassification errors from demuxlet (Additional file 1: Figure S2E), in line with demuxlet's estimated error rate of 1–2% [13].

To further ensure that background binding levels did not lead to incorrectly demultiplexed samples, we performed a separate experiment where we mixed four cell lines (HEK293T, THP1, K562, and KG1) together, each independently labeled with three distinct Cell Hashing oligos. After demultiplexing, to assign each barcode to a cell line of origin, we clustered cells on the basis of their RNA expression levels, obtaining four transcriptomic clusters (as expected). Comparing our transcriptomic clusters with the demultiplexing results, we observed nearly perfect

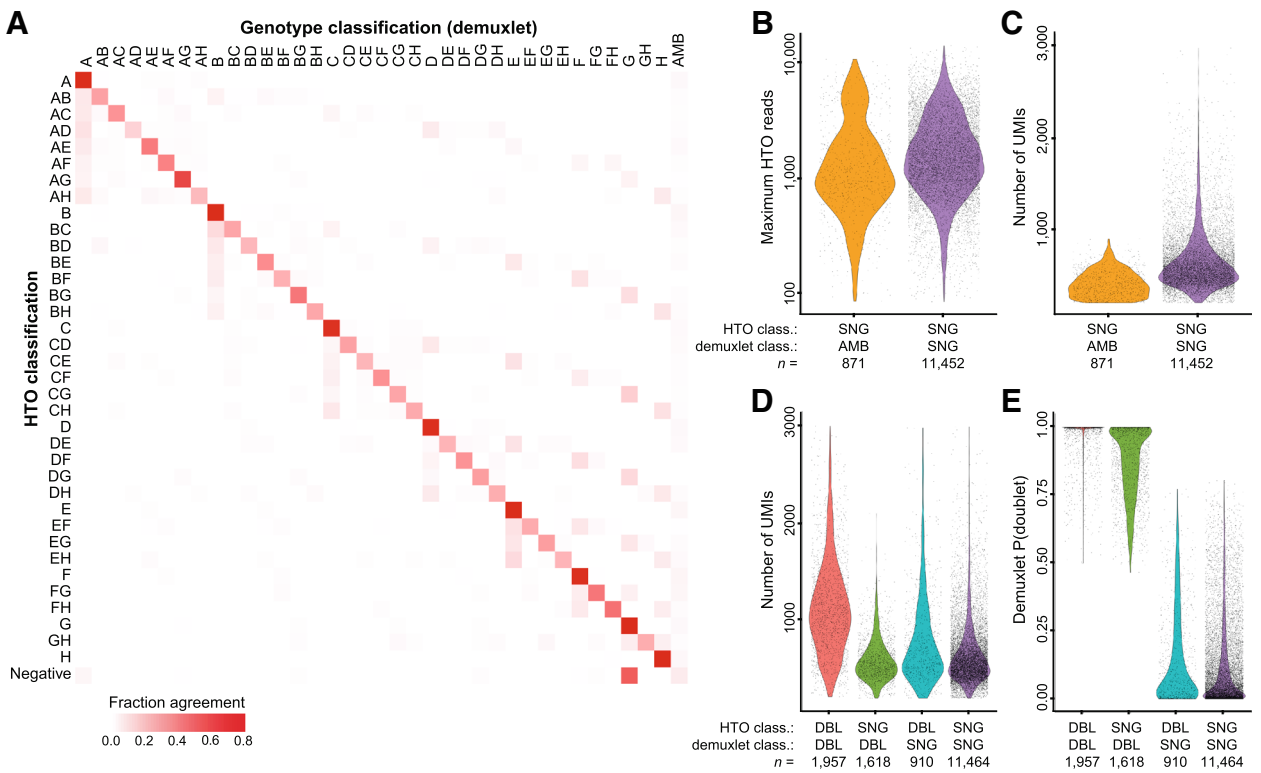

Fig. 2 Validation of Cell Hashing using demuxlet. **a** Row-normalized "confusion matrix" comparing demuxlet and HTO classifications. Each value on the diagonal represents the fraction of barcodes for a given HTO classification that received an identical classification from demuxlet. **b** Count distribution of the most highly expressed HTO for groups of concordant and discordant singlets. Both groups have identical classification strength based on Cell Hashing. **c** Discordant singlets have lower UMI counts, suggesting that a lack of sequencing depth contributed to "ambiguous" calls from demuxlet. **d** RNA UMI distributions for discordant and concordant multiplets. Only concordant multiplets exhibit increased molecular complexity, suggesting that both methods are conservatively overcalling multiplets in discordant cases. **e** In support of this, demuxlet assigns lower multiplets posterior probabilities to discordant calls

concordance (99.7%), demonstrating a low rate of misassignment for this experiment (Additional file 1: Figure S3A, B).

Finally, we attempted to estimate the false-negative rates for Cell Hashing, representing true single cells that do not receive sufficient Cell Hashing signal to be classified as singlets. To do this, we examined all HTO-classified "singlet" and "negative" barcodes from the PBMC experiment and performed clustering based on transcriptome data. As expected, we found that "negative" cells predominantly formed a distinct cluster from singlets. However, we did observe 117 barcodes originally classified as negatives, but whose transcriptomic profiles clustered across PBMC singlet subtypes. These barcodes likely represent single cells that were incorrectly classified from Cell Hashing, representing a false-negative rate of 0.9% (Additional file 1: Figure S4), but have negligible effects on cell type proportion estimates. Taken together, our results validate that Cell Hashing enables robust and accurate sample classification across diverse systems.

Cell Hashing enables the efficient optimization of CITE-seq antibody panels

Our multiplexing strategy not only enables pooling across donors but also the simultaneous profiling of multiple experimental conditions. This is widely applicable to the simultaneous profiling of diverse environmental and genetic perturbations, but we reasoned that we could also efficiently optimize experimental workflows, such as the titration of antibody concentrations for CITE-seq experiments. In flow cytometry, antibodies are typically run individually over a large dilution series to assess signal-to-noise ratios and identify optimal concentrations [20]. While such experiments would be extremely cost prohibitive if run as individual 10x Genomics lanes, we reasoned that we could multiplex these experiments together using Cell Hashing.

We therefore incubated the PBMCs from different donors with a dilution series of antibody concentrations ranging over three orders of magnitude ("Methods" section). Concentrations of CITE-seq antibodies were staggered between the different samples to keep the total amount of antibody and oligo consistent in each sample.

After sample demultiplexing, we examined ADT distributions across all concentrations for each antibody (examples in Fig. 3a–c) and assessed signal-to-noise ratio by calculating a staining index similar to commonly used metrics for flow cytometry optimization (Fig. 3d) ("Methods" section).

All antibodies exhibited only background signal in the negative control conditions and very weak signal-to-noise at 0.06 μg/test. We observed that the signal-to-noise ratio for most antibodies began to saturate within the concentration range of 0.5 to 1 μg/test, comparable to the recommended concentrations for flow cytometry (Fig. 3d). This experiment was meant as a proof of concept; an ideal titration experiment would use cells from the same donor for all conditions and a larger range of concentrations but clearly demonstrates how Cell Hashing can be used to rapidly and efficiently optimize experimental workflows.

Cell Hashtags enable the discrimination of low-quality cells from ambient RNA

Our cell hashtags can discriminate single cells from doublets based on the clear expression of a single HTO, and we next asked whether this feature could also distinguish low-quality cells from ambient RNA. If so, this would enable us to reduce our UMI "cutoff" (previously set at 200) and would allow for the possibility that certain barcodes representing ambient RNA may express more UMI than some true single cells. Most workflows set stringent UMI cutoffs to exclude all ambient RNA, biasing scRNA-seq results against cells with low RNA content and likely skewing proportional estimates of cell type.

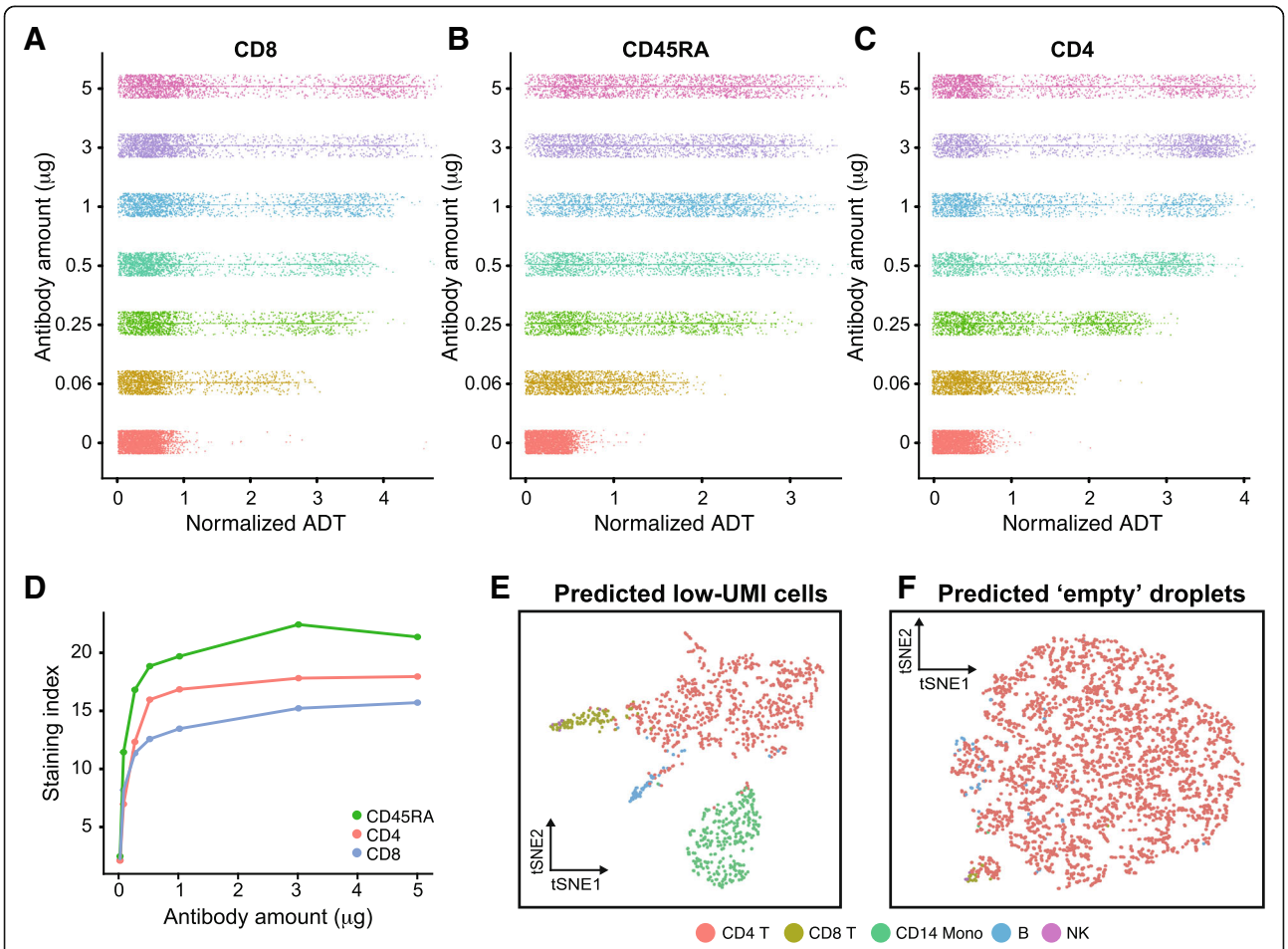

Fig. 3 Cell Hashing enables efficient experimental optimization and identification of low-quality cells. **a**–**c** We performed a titration series to assess optimal staining concentrations for a panel of CITE-seq immunophenotyping antibodies. Normalized ADT counts for CD8 (**a**), CD45RA (**b**), and CD4 (**c**) are depicted for the different concentrations used per test. **d** Titration curve depicting the staining index (SI; "Methods" section) for these three antibodies across the titration series. The signal/noise ratio for these antibodies begins to saturate at levels similar to manufacturer recommended staining concentrations typical for flow cytometry antibodies. **e** Cells with low UMI counts can be distinguished from ambient RNA using HTO classifications. Classified singlets group into canonical hematopoietic populations. **f** Barcodes classified as "negative" do not group into clusters and likely represent "empty" droplets containing only ambient RNA

Indeed, when considering 4344 barcodes containing 50–200 UMI, we recovered 1110 additional singlets based on HTO classifications, with 3108 barcodes characterized as negatives. We classified each barcode as one of our previously determined 7 hematopoietic populations ("Methods" section; Fig. 1F) and visualized the results on a transcriptomic tSNE embedding, calculated independently for both "singlet" and "negative" groups. For predicted singlets, barcodes projected to B, NK, T, and myeloid populations which were consistently separated on tSNE, suggesting that these barcodes represent true single cells (Fig. 3e). In contrast, "negative" barcodes did not separate based on their forced classification, consistent with these barcodes reflecting ambient RNA mixtures that may blend multiple subpopulations. We therefore conclude that by providing a readout of sample identity that is independent of the transcriptome, Cell Hashing can help recover low-quality cells and/or cells with very low RNA content that can otherwise be difficult to distinguish from ambient RNA (Fig. 3f).

Towards a universal Cell Hashing antibody reagent

For our proof of principle experiments, we used a pool of antibodies directed against highly expressed immune surface markers (CD45, CD98, CD44, and CD11a). To enable multiplexing of any cell type and sample, we decided to redesign our panel to target more ubiquitously expressed surface markers. MHC class I complex (beta-2-microglobulin) and the sodium-potassium ATPase-subunit (CD298) are among the most broadly expressed surface proteins in human tissues [21]. Using a pool of antibodies directed against both proteins would allow us to multiplex virtually any cell type in one experiment. While this manuscript was under revision, the same antibody combination was demonstrated by Hartmann and colleagues to be a universal multiplexing reagent for CyTOF [22]. The extremely high expression levels of both markers should enable robust HTO demultiplexing, but in principle could label cells with an overwhelming number of single-stranded polyA oligos that might compete with polyadenylated cellular mRNAs, resulting in lower gene and/or UMI counts per cell. To investigate this potential competition, we stained Jurkat cells with a dilution series of Cell Hashing antibodies, ran a lane of 10x Chromium single cell 3′ v2 alongside a lane with non-hashed cells, and sequenced the resulting transcriptome libraries. Transcriptomic complexity levels, as indicated by the relationship between sequencing reads and UMI counts per cell, were indistinguishable from non-hashed cells over all tested concentrations of Cell Hashing antibodies, illustrating no disadvantages when multiplexing samples (Additional file 1: Figure S5). Taken together, these results demonstrate how Cell Hashing can be easily applied to virtually any human sample with readily available commercial reagents and without a loss of transcriptomic complexity.

Discussion

Here, we introduce a new method for scRNA-seq multiplexing, where cells are labeled with sample-specific "hashtags" for downstream demultiplexing and multiplet detection. Our approach is complementary to pioneering genetic multiplexing strategies, with each having unique advantages. Genetic multiplexing does not utilize exogenous barcodes and therefore does not require alterations to existing workflows prior to or after sample pooling. In contrast, Cell Hashing requires incubation with antibodies against ubiquitously expressed surface proteins but can multiplex samples with the same genotype. Both methods do slightly increase downstream sequencing costs, due to the increased depth or read length needed to identify SNPs (genetic approaches) or sequencing of HTO libraries (Cell Hashing; approximately 2–5% of transcriptome sequencing costs). We believe that researchers will benefit from both approaches, enabling multiplexing for a broad range of experimental designs. In particular, we envision that our method will be most useful when processing genetically identical samples subjected to diverse perturbations (or experimental conditions/optimizations, as in our titration experiment) or to reduce the doublet rate when running cells from a single sample.

By enabling the robust identification of cell multiplets, both Cell Hashing and genetic multiplexing allow the "super-loading" of scRNA-seq platforms. We demonstrate this in the context of the 10x Genomics Chromium system, but this benefit applies to any single cell technology that relies on Poisson loading for cell isolation. The per-cell cost savings for library preparation can therefore be significant, approaching an order of magnitude as the number of multiplexed samples increases. Notably, Cell Hashing enables even a single sample to be highly multiplexed, as cells can be split into an arbitrary number of pools. As clearly discussed in Kang et al. [13], savings in library prep are partially offset by reads originating from multiplets, which must be sequenced and discarded. Still, as sequencing costs continue to drop, and experimental designs seek to minimize technology-driven batch effects, multiplexing should facilitate the generation of large scRNA-seq and CITE-seq datasets. Informatic detection of multiplets based on transcriptomic data also remains an important challenge for the field, for example, to identify doublets originating from two cells within the same sample.

While we used a pool of antibodies directed against highly and ubiquitously expressed lymphocyte surface proteins as the vehicle for our HTOs in our

proof-of-principle experiments, we also introduced a more universal pool of antibodies directed against two ubiquitously expressed markers (B2M and CD298) to be used as a Cell Hashing reagent for studies beyond the hematopoietic system. Using a pool of antibodies mitigates the possibility that stochastic or cell type variation in the expression of any one marker would introduce bias in HTO recovery. We however caution that there can be instances when a cell type of interest does not express these virtually ubiquitous surface proteins, which would result in failure to successfully label and demultiplex these cells. With the increasing interest in single nucleus sequencing [23, 24], an additional set of hashing reagents directed against nuclear proteins would further generalize this approach. Beyond antibody/epitope interactions, cell or nucleus hashing could also be performed using alternative means of attaching an oligo to a cell or nucleus, including aptamers [25] or direct chemical conjugation of oligos to cells or nuclei. Indeed, recently described approaches accomplish similar goals through transient transfection of oligos [26], direct oligo to cell conjugation based on NHS chemistry [27], lipid membrane intercalating oligos [28], and viral integration-based genomic barcoding [30]. These improvements will further enable multiplexing strategies to generalize to diverse experiments regardless of species, tissue, or technology.

Methods

PBMC genotyping

Peripheral blood mononuclear cells were obtained from AllCells (USA). Genomic DNA was purified using the AllPrep kit (Qiagen, USA) and genotyped using the Infinium CoreExome 24 array (Illumina, USA) according to the manufacturer's instructions.

Cell culture

HEK293T (human) and NIH-3T3 (mouse) cells were maintained according to the standard procedures in Dulbecco's modified Eagle's medium (Thermo Fisher, USA) supplemented with 10% fetal bovine serum (Thermo Fisher, USA) at 37 °C with 5% CO_2.

Antibody-oligo conjugates

Antibody-oligo conjugates directed against CD8 [clone: RPA-T8], CD45RA [clone: HI100], CD4 [clone: RPA-T4], HLA-DR [clone: L243], CD3 [clone: UCHT1], CCR7 [clone: G043H7], and PD-1 [clone: EH12.2H7] were provided by BioLegend (USA) containing 1–2 conjugated oligos per antibody on average.

First generation Cell Hashing antibodies were conjugated in-house. Antibodies were obtained as purified, unconjugated reagents from BioLegend (CD45 [clone: HI30], CD98 [clone: MEM-108], CD44 [clone: BJ18], and CD11a [clone: HI111]) and were covalently and irreversibly conjugated to HTOs by iEDDA click chemistry as previously described [19]. In short, antibodies were washed into 1X borate buffered saline (50 mM borate, 150 mM NaCl, pH 8.5) and concentrated to 1 mg/ml using an Amicon Ultra 0.5 ml 30 kDa MWCO centrifugal filter (Millipore). Methyltetrazine-PEG4-NHS ester (Click Chemistry Tools, USA) was dissolved in dry DMSO and added at a 30-fold excess to the antibody and allowed to react for 30 min at room temperature. Residual NHS groups were quenched by the addition of glycine and the unreacted label was removed via centrifugal filtration. 5′-Amine HTOs were ordered from Integrated DNA Technologies (USA) and reacted with a 20-fold excess of *trans*-cyclooctene-PEG4-NHS (Click Chemistry Tools, USA) in 1X borate buffered saline supplemented with 20% DMSO for 30 min. Residual NHS groups were quenched by the addition of glycine, and residual label was removed by desalting (Bio-Rad Micro Bio-Spin P6). Antibody-oligo conjugates were formed by mixing the appropriate labeled antibody and HTO and incubating at room temperature for at least 1 h. Residual methyltetrazine groups on the antibody were quenched by the addition of *trans*-cyclooctene-PEG4-acid, and unreacted oligo was removed by centrifugal filtration using an Amicon Ultra 0.5 ml 50 kDa MWCO filter (Millipore, USA). A detailed and regularly updated point-by-point protocol for antibody-oligo conjugation can be found at www.cite-seq.com

Second generation Cell Hashing antibodies consisting of a pool of antibodies directed against B2M [clone: 2 M2] and CD298 [clone: LNH-94] were purchased from BioLegend (USA).

Antibody titration series

To test optimal concentration of antibody-oligo conjugates provided by BioLegend (USA) per CITE-seq experiment, we tested 5 μg, 3 μg, 1 μg, 0.5 μg, 0.25 μg, 0.06 μg, and 0 μg for each conjugate. Titrations were staggered over the different batches to keep the total concentration of antibodies and oligos consistent between conditions (Additional file 2: Table S1).

Sample pooling

PBMCs from different donors were independently stained with one of our HTO-conjugated antibody pools and a pool of seven immunophenotypic markers for CITE-seq at different amounts (see above). All eight PBMC samples were pooled at equal concentration, alongside unlabeled HEK293T and mouse NIH-3T3 as negative controls (see table below) and loaded into the 10x Chromium instrument (Additional file 3: Table S2).

CITE-seq on 10x Genomics instrument

Cells were "stained" with Cell Hashing antibodies and CITE-seq antibodies as described for CITE-seq [18]. "Stained" and washed cells were loaded into 10x Genomics Single Cell 3′ v2 workflow and processed according to the manufacturer's instructions up until the cDNA amplification step (10x Genomics, USA). Two picomoles of HTO and ADT additive oligonucleotides were spiked into the cDNA amplification PCR, and cDNA was amplified according to the 10x Single Cell 3′ v2 protocol (10x Genomics, USA). Following PCR, 0.6X SPRI was used to separate the large cDNA fraction derived from cellular mRNAs (retained on beads) from the ADT- and Cell Hashtag (HTO)-containing fraction (in supernatant). The cDNA fraction was processed according to the 10x Genomics Single Cell 3′ v2 protocol to generate the transcriptome library. An additional 1.4X reaction volume of SPRI beads was added to the ADT/HTO fraction to bring the ratio up to 2.0X. The beads were washed with 80% ethanol, eluted in water, and an additional round of 2.0X SPRI performed to remove excess single-stranded oligonucleotides from cDNA amplification. After final elution, separate PCRs were set up to generate the CITE-seq ADT library (SI-PCR and RPI-x primers) and the HTO library (SI-PCR and D7xx_s). A detailed and regularly updated point-by-point protocol for CITE-seq, Cell Hashing, and future updates can be found at www.cite-seq.com

Oligonucleotide sequences

The following are the oligonucleotide sequences:

Hashtag oligo: GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTxxxxxxxxxxxxBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA*A*A

HTO additive: GTGACTGGAGTTCAGACGTGTGCTC

ADT additive: CCTTGGCACCCGAGAATTCC

SI-PCR: AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGC*T*C

RPI-x: CAAGCAGAAGACGGCATACGAGATxxxxxxxxGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA

D7xx_s: CAAGCAGAAGACGGCATACGAGATxxxxxxxxGTGACTGGAGTTCAGACGTGTGC

x: Barcode or index sequence

B: T,G,C, not A

*: Phosphorothioate bond

Cell Hashing dilution and competition experiment

Jurkat cells were "stained" with decreasing concentrations (1:100, 1:500, 1:1000) of Cell Hashing antibodies (BioLegend, USA; B2M, CD298 pool) as described above and passed through a 10x Genomics Single Cell workflow to yield ~ 2000 cells. As a control, non-hashed cells were passed through a separate 10x Genomics Single Cell lane. Cells from both experiments were subsampled to the same numbers of cells and reads per cell to compare UMI and gene counts.

Computational methods

Single-cell data processing

Fastq files from the 10x Genomics libraries with four distinct barcodes were pooled together and processed using the standard Drop-seq pipeline (Drop-seq tools v1.0, McCarroll Lab). Reads were aligned to the hg19-mm10 concatenated reference, and we included the top 50,000 cell barcodes in the raw digital expression matrix as output from Drop-seq tools. For ADT and HTO quantification, we implemented our previously developed tag quantification pipeline [18] as a python script, available at https://github.com/Hoohm/CITE-seq-Count, and run with default parameters (maximum hamming distance of 1).

Demultiplexing with genotyping data using demuxlet

We first generated a VCF file that contained the individual genotype (GT) from the Infinium CoreExome 24 array output, using the PLINK command line tools (version 1.07). This VCF file (which contained genotype information for the 8 PBMC donors as well as HEK293T cells) and the tagged bam file from Drop-seq pipeline were used as inputs for demuxlet [13], with default parameters.

Single-cell RNA data processing

Normalization and downstream analysis of RNA data were performed using the Seurat R package (version 3.0, Satija Lab [29]) which enables the integrated processing of multi-modal (RNA, ADT, HTO) single cell datasets [31, 32]. We collapsed the joint-species RNA expression matrix to only include the top 100 most highly expressed mouse genes (along with all human genes) using the CollapseSpeciesExpressionMatrix function.

We first considered a set of 20,854 barcodes where we detected at least 200 UMI in the transcriptome data. Since the HEK293T and NIH-3T3 cells were not labeled with HTOs, we identified these cells based on their transcriptomes. We performed a low-resolution pre-clustering by performing PCA on the 500 most highly expressed genes, followed by *k*-medoid clustering on a distance matrix based on the first 2 principal components [33–35]. Based on this clustering, we identified 160 NIH-3T3 cells and 2233 HEK293T cells, with the remainder representing PBMCs.

As a separate test of HEK293T identity, we examined the demuxlet genotype for possible HEK293T cells. We observed 225 barcodes classified as HEK by the demuxlet algorithm but whose transcriptomes clustered with PBMCs. These cells expressed tenfold fewer UMI compared to transcriptomically classified HEK293T cells and did not express HEK293T-specific transcripts (i.e.,

NGFRAP1), both consistent with a PBMC identity. We therefore excluded these barcodes from all further analysis.

Classification of barcodes based on HTO levels

HTO raw counts were normalized using centered log ratio (CLR) transformation, where counts were divided by the geometric mean of an HTO across cells and log-transformed:

$$x_i' = \log \frac{x_i}{\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}}$$

Here, x_i denotes the count for a specified HTO in cell i, n is the total cell number, and *log* denotes the natural log. Pairwise analysis of either normalized or raw HTO counts (Fig. 1B) revealed mutually exclusive relationships, though determining the exact cutoffs for positive and negative signals required further analysis. We reasoned that if we could determine a background distribution for each HTO based on "negative" cells, outliers from this distribution would represent positive signals.

To assist in the unsupervised identification of "negative" cells, we performed an initial *k*-medoids clustering for all cells based on the normalized HTO data. We set $k = 9$ and observed (as expected) that eight of the clusters were highly enriched for expression of a particular HTO, while the ninth cluster was highly enriched for cells with low expression of all HTO. This represents an initial solution to the demultiplexing problem that suggests likely populations of "positive" and "negative" cells for statistical analysis.

Following clustering, we performed the following procedure independently for each of the eight HTOs. We identified the *k*-medoids cluster with the highest average HTO expression and excluded these cells. We next fit a negative binomial distribution to the remaining HTO values, after further excluding the highest 0.5% values as potential outliers. We calculated the $q = 0.99$ quantile of the fitted distribution and thresholded each cell in the dataset based on this HTO-specific value.

We used this procedure to determine an "HTO classification" for each barcode. Barcodes that were positive for only one HTO were classified as singlets. Barcodes that were positive for two or more HTOs were classified as multiplets and assigned sample IDs based on their two most highly expressed HTO. Barcodes that were negative for all eight HTOs were classified as "negative."

We expect that barcodes classified as "singlets" represent single cells, as we detect only a single HTO. However, they could also represent doublets of a PBMC with a HEK293T or NIH-3T3 cell, as the latter two populations were unlabeled and represent negative controls. Indeed, when we analyzed the "HTO classification" of cells that were transcriptomically annotated as HEK293T or NIH-3T3 cells, we found that 60.1% were annotated as "negative," while 32.1% were annotated as singlets, in agreement with our expected ratios in our "superloaded" 10x Genomics experiment. These cells appear in the heatmap in Fig. 1C, but all HEK293T and NIH-3T3 cells were excluded from further analysis.

For 2D visualization of HTO levels (Fig. 1D), we used Euclidean distances calculated from the normalized HTO data as inputs for tSNE. Cells are colored based on their HTO classification as previously described. For visualization and clustering based on transcriptomic data (Fig. 1F), we first performed PCA on the 1000 most highly variable genes (as determined by variance/mean ratio) and used the distance matrix defined by the first 10 principal components as input to tSNE and graph-based clustering in Seurat (Fig. 1E). We annotated the seven clusters based on canonical markers for known hematopoietic populations.

Comparison with demuxlet

Demuxlet classifications were labeled as singlets (SNG), doublets (DBL), or ambiguous (AMB) according to the BEST column in the *.best output file. In Fig. 2e, we plot the posterior probability of a doublet assignment from the PRB.DBL column in the same file.

Calculation of staining index for antibody titrations

To assess the optimal staining efficiency for CITE-seq experiments, we considered ADT levels for cells across a range of antibody concentrations as multiplexed in a titration series. ADT levels were normalized using a CLR transformation of raw counts using an identical approach to the normalization of HTO levels as previously described.

After normalization, we computed a staining index based on standard approaches in flow cytometry, which examine the difference between positive and negative peak medians, divided by the spread (i.e., twice the mean absolute deviation) of the negative peak.

$$\mathrm{SI} = \frac{\mathrm{Pos}_{0.5} - \mathrm{Neg}_{0.5}}{2 \times \mathrm{mad}(\mathrm{Neg})}$$

In order to avoid manual classification of positive and negative peaks, we implemented an automated procedure that can scale to multiple antibodies and concentrations. To approximate the negative peak, we leveraged unstained control cells (donor H). To approximate the positive peak, we clustered the ADT data in each titration experiment (donor A through donor G). To perform clustering, we computed a Euclidean distance

matrix across cells based on normalized ADT levels and used this as input to the FindClusters function in Seurat with default parameters. We examined the results to identify the cluster with the maximally enriched ADT signal and referred to the distribution of ADT levels within this cluster as the positive peak.

Discriminating low-quality cells from ambient RNA

We performed HTO classification of low-quality barcodes (expressing between 50 and 200 UMI), using the previously determined HTO thresholds. For each barcode, we classified its expression as 1 of our previously determined 7 hematopoietic populations using random forests, as implemented in the ranger package in R [36]. We first trained a classifier on the 13,954 PBMCs, using the 1000 most variable genes as input and their clustering identities as training labels. We then applied this classifier to each of the low-quality barcodes. We note that this classifier is guaranteed to return a result for each barcode.

Additional files

Additional file 1: Distribution of HTO UMIs per cell barcode. (PDF 6203 kb)

Additional file 2: Micrograms of antibody used per condition. (XLSX 9 kb)

Additional file 3: Sample composition of experiment referred to by Figures 1-3. (XLSX 9 kb)

Acknowledgements

We acknowledge the members of the NYGC Technology Innovation and Satija Labs for critical discussions and support. We thank M. Coppo, S. Fennessey, B. Baysa, and S. Pescatore at NYGC for the sequencing support.

Funding

This work was supported by the Chan Zuckerberg Initiative (HCA-A-1704-01895, to RS and PS; HCA2-A-1708-02755 to RS), NIHR21-HG-009748 (to PS), NIHR01MH071679-14, OT2OD026673-01, and R35NS097404-02 (RS), and an NIH New Innovator Award (DP2-HG-009623 to RS).

Availability of data and materials

The sequencing data is available in the Gene Expression Omnibus repository, under accession GSE108313 [37]. Computational tools for classifying cells based on HTO levels are freely available as part of the open-source R package Seurat, available at https://github.com/satijalab/seurat [38]. Point-by-point CITE-seq, Cell Hashing, and Antibody-Oligo conjugation protocols are available at www.cite-seq.com

Authors' contributions

MS, RS, and PS conceived and designed the study. MS, BHL, SH, and WMM performed the experiments. SZ and RS designed and contributed the computational analyses with input from SH, MS, and PS. BZY provided input on the reagent composition and interpretation. MS, SZ, PS, and RS interpreted the data. MS, PS, and RS wrote the manuscript with input from all authors. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable for this study.

Competing interests

MS, PS and BHL have filed a patent application based on this work (US provisional patent application 62/515-180). BZY is an employee at BioLegend Inc., which is the exclusive licensee of the New York Genome Center patent application related to this work. All other authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Technology Innovation Lab, New York Genome Center, New York, NY, USA. [2]NYU Center for Genomics and Systems Biology, New York Genome Center, New York, NY, USA. [3]BioLegend Inc., San Diego, CA, USA.

Received: 5 June 2018 Accepted: 4 December 2018
Published online: 19 December 2018

References

1. Stubbington MJT, Rozenblatt-Rosen O, Regev A, Teichmann SA. Single-cell transcriptomics to explore the immune system in health and disease. Science. 2017;358:58–63.
2. Tanay A, Regev A. Scaling single-cell genomics from phenomenology to mechanism. Nature. 2017;541:331–8.
3. Villani A-C, et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science. 2017;356:eaah4573.
4. Velten L, et al. Human haematopoietic stem cell lineage commitment is a continuous process. Nat Cell Biol. 2017;19:271–81.
5. Cao J, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. Science. 2017;357:661–7.
6. Karaiskos N, et al. The Drosophila embryo at single-cell transcriptome resolution. Science. 2017;8:eaan3235–14.
7. Macosko EZ, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161:1202–14.
8. Klein AM, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161:1187–201.
9. Zheng GXY, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:1–12.
10. Regev A, et al. Science forum: the human cell atlas. eLife. 2017;6:e27041.
11. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Publ Group. 2015;16:133–45.
12. Hicks SC, Townes FW, Teng M, Irizarry RA. Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics. 2017. https://doi.org/10.1093/biostatistics/kxx053.
13. Kang HM, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat Publ Group. 2017. https://doi.org/10.1038/nbt.4042.
14. Tung P-Y, et al. Batch effects and the effective design of single-cell gene expression studies. Sci Rep. 2017;7:39921.
15. Mei HE, Leipold MD, Schulz AR, Chester C, Maecker HT. Barcoding of live human peripheral blood mononuclear cells for multiplexed mass cytometry. J Immunol. 2015;194:2022–31.
16. Lai L, Ong R, Li J, Albani S. A CD45-based barcoding approach to multiplex mass-cytometry (CyTOF). Cytometry. 2015;87:369–74.
17. Krutzik PO, Nolan GP. Fluorescent cell barcoding in flow cytometry allows high-throughput drug screening and signaling profiling. Nat Meth. 2006;3:361–8.
18. Stoeckius M, et al. Simultaneous epitope and transcriptome measurement in single cells. Nat Meth. 2017;9:2579–10.
19. van Buggenum JAGL, et al. A covalent and cleavable antibody- DNA conjugation strategy for sensitive protein detection via immuno-PCR. Nat Publ Group. 2016:1–12. https://doi.org/10.1038/srep22675.
20. Hulspas R. Titration of fluorochrome-conjugated antibodies for labeling cell surface markers on live cells. Curr Protoc Cytom. 2010;Chapter 6(Unit 6):29.
21. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C. Tissue-based map of the human proteome. Science. 2015. https://doi.org/10.1126/science.1260419.
22. Hartmann FJ, Simonds EF, Bendall SC. A universal live cell barcoding-platform for multiplexed human single cell analysis. Sci Rep. 2018;8(10770).
23. Lake BB, et al. A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA. Sci Rep. 2017:1–8. https://doi.org/10.1038/s41598-017-04426-w.
24. Habib et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat Methods. 2017;14(10):955–8. PMID: 28846088.

25. Delley CL, liu L, Sarhan MF, Abate AR. Combined aptamer and transcriptome sequencing of single cells. bioRxiv. 2017:1–10. https://doi.org/10.1101/228338.
26. Shin, D., Lee, W., Lee, J. H., bioRxiv, D. B.2018. Multiplexed single-cell RNA-seq via transient barcoding for drug screening. biorxiv.org doi: https://doi.org/10.1101/359851.
27. Gehring J, Park JH, Chen S, Thomson M, Pachter L. Highly multiplexed single-cell RNA-seq for defining cell population and transcriptional spaces. bioRxiv. 2018:1–19. https://doi.org/10.1101/315333.
28. McGinnis, C. S. et al. MULTI-seq: scalable sample multiplexing for single-cell RNA sequencing using lipid-tagged indices 1–34 (2018). doi:https://doi.org/10.1101/387241.
29. Stuart et al. Comprehensive integration of single cell data. biorxiv. 2018. https://doi.org/10.1101/460147.
30. Guo C, Biddy BA, Kamimoto K, Kong W, Morris SA. CellTag indexing: a genetic barcode-based multiplexing tool for single-cell technologies. 2018: 1–20. https://doi.org/10.1101/335547.
31. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015;33:495–502.
32. Butler A, Satija R. Integrated analysis of single cell transcriptomic data across conditions, technologies, and species. bioRxiv. 2017. https://doi.org/10.1101/164889.
33. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech. 2008;2008:P10008.
34. Levine JH, et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell. 2015;162:184–97.
35. Shekhar K, et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. Cell. 2016;166:1308–23.e30.
36. Wright MN, Ziegler A. Ranger: a fast implementation of random forests for high dimensional data in C and R. J Stat Softw. 2017;77:1–17.
37. Stoeckius M, Zheng S, Houck-Loomis B, Hao S, Yeung B, Smibert P, and Satija R. Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE108313 (2018).
38. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Github Repository. https://github.com/satijalab/seurat (2018).