MitoSort: robust demultiplexing of pooled single-cell genomics data using endogenous germline mitochondrial variants

Zhongjie Tang[1, #], Weixing Zhang[1, #], Peiyu Shi[1, #], Sijun Li[1], Xinhui Li[1], Yicong Xu[1], Yaqing Shu[2], Jin Xu[1, *]

[1]State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-Sen University, Guangzhou, China

[2]Department of Neurology, The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China

[#] These authors contributed equally to the work.

[*]Correspondence should be addressed to:

Jin Xu, PhD. School of Life Sciences, Sun Yat-Sen University, No. 135 Xingang Xi Road, Guangzhou, Guangdong, China. Email: xujin7@mail.sysu.edu.cn.

Abstract

Multiplexing across donors has emerged as a popular strategy to increase throughput, reduce costs, overcome technical batch effect and improve doublet detection in single-cell genomic studies. Using endogenous genetic barcodes eliminates the need for additional experimental processing steps. The unique features of mtDNA make mtDNA variants a more computationally efficient and robust choice for endogenous barcodes compared to genome variants. Here we developed MitoSort, a method that uses mtDNA germline variants to assign cells to their donor of origin and identify cross-genotype doublets robustly and efficiently. Our study validated MitoSort's performance by both in silico data and cell hashing data and demonstrated its use in various single-cell sequencing approaches.

28 **Main**

29 The advent of single-cell sequencing techniques has expanded the repertoire of
30 detectable modalities within a single cell(1, 2). For instance, DOGMA-seq(3) enables
31 the simultaneous measurement of gene expression, chromatin accessibility, and protein
32 levels in individual cells, providing a powerful tool for the comprehensive exploration
33 of biological processes. However, the high costs of multimodal single-cell techniques
34 limit their application in large cohort studies. To meet the growing demand for profiling
35 large-scale samples, sample multiplexing has become a popular strategy that not only
36 reduces costs but also overcomes technical batch effects and improves doublet
37 detection(4). To avoid additional experimental processing steps, several computational
38 tools have been proposed to harness genome variants for demultiplexing cells from
39 different individuals and identifying cross-genotype doublets(5–9). This process
40 typically requires significant computational resources, even with reference genotypes,
41 due to the large size of the human genome. In contrast, mtDNA variants offer a more
42 computationally efficient and robust option, given their hundreds of copies per cell and
43 the small size of the mtDNA genome(10, 11). Moreover, mtDNA variants can be
44 captured with high coverage at single-cell resolution in multimodal single-cell
45 sequencing techniques such as mtscATAC-seq(12), ASAP-seq(3), DOGMA-seq(3), as
46 well as full-length single-cell RNA-seq (e.g., Smart-seq(13)). Despite the benefits of
47 using mtDNA variants for demultiplexing samples and identifying doublets, few
48 software options are currently available. To fill the gap, we developed MitoSort, a
49 software that exploits mtDNA germline variants exclusively to demultiplex samples
50 from different individuals and detect cross-genotype doublets efficiently and accurately
51 (Fig.1a).

52 The MitoSort pipeline is mainly composed of four steps: MT reads realignment, variant
53 calling, doublet identification and demultiplexing (Fig.1b). Briefly, to reduce false-
54 positive variants resulting from STAR(14) alignments artifacts, MT reads are realigned
55 with GATK(15) to facilitate accurate variant calling. VarScan(16) is then used to

identify putative single-nucleotide polymorphisms (SNPs) in mtDNA. An SNP-by-cell matrix was generated after filtering low-frequency variants. By the initial K-means clustering, sample-specific germline variants are identified, allowing for the precise clustering of cells based on unique mitochondrial haplotypes. To detect doublets, we model a cell's allele counts as being drawn from a beta-binomial distribution, with parameters derived from one or two clusters. After removing doublets, cells are assigned to their donor of origin by K-means clustering again. Additionally, our model estimates the most likely number of pooled individuals, which is helpful when clinical information is lost (Methods). It only takes 10 minutes for demultiplexing a mixture of ~10000 cells by MitoSort (the step of MT reads realignment was not involved in the estimated runtime).

We first used *in silico* data with known cell origins to evaluate our method. To create synthetic datasets, we collected eight published mtscATAC-seq datasets from genetically distinct donors (detailed information is shown in Extended Data Table1). The impact of factors affecting the classification performance of MitoSort was assessed. Among these factors, the number of pooled samples, the number of cells for each sample, and the percentage of doublets had little effect on the classification performance of MitoSort (Fig.1c-e, Extended Data Fig.1a-f). We consistently obtained high accuracy of origin assignment and doublet identification for 2- to 8-mixed samples (500 cells per sample, 2000 mtDNA reads per cell, and 8% of cells as doublets) (Fig.1c). Besides, when we synthetically mixed 8 samples (2000 mtDNA reads per cell and 8% of cells as doublets) and varied the number of cells from 100 to 600 per sample, we found the true positive rate (TPR) of both donor assignment and doublet identification was consistently achieved at a rate of 1, even when the number of cells was as low as 100 (Extended Data Fig.1e). Only a small number of singlets were erroneously classified as doublets, constituting less than 0.3% of all cases (Extended Data Fig.1f). In addition, MitoSort achieved perfect performance regardless of the percentage of simulated doublets (Fig.1e Extended Data Fig.1g-h). We found that the classification

84 performance of MitoSort was only sensitive to the sequencing depth of mitochondrial
85 genome (Fig.1f). Specifically, an increase in the number of mapped reads (ranging from
86 250 to 2000) resulted in a higher true positive rate (TPR) and a lower false discovery
87 rate (FDR), leading to improved accuracy (Fig.1f, Extended Data Fig.1g-h). It is worth
88 noting that when the number of reads mapped to the mitochondrial genome reached
89 approximately 1000 (roughly four-fold coverage), MitoSort was able to achieve perfect
90 cell assignment to their respective donors and accurately detect cross-genotype doublets.
91 Collectively, these data show that MitoSort is a highly accurate and scalable method
92 for demultiplexing samples from different individuals and identifying cross-genotype
93 doublets using mtDNA germline variants.
94 We next evaluated our method by two published datasets that used cell-hashing
95 technology to mark sample origins of cells prior to multiplexing. The DOGMA-seq
96 dataset consisted of activated and stimulated T cells from two healthy donors(17), while
97 the ASAP-seq dataset comprised NK cells from four donors(18). MitoSort was able to
98 assign cells to their respective donors accurately in both datasets, demonstrating high
99 consistency with the known cell hashing tags (Fig. 2a-b). MitoSort achieved high
100 accuracy of origin assignment (>= 98%) when the sequencing depth of mitochondrial
101 genome was not less than 4X (Extended Data Fig.2). These results further suggest the
102 robustness of MitoSort in assigning cells from different donors using mtDNA germline
103 variants.
104 Compared to droplet-based single-cell RNA sequencing techniques that typically
105 sequence the 3' ends of RNAs, full-length single-cell RNA-seq protocols offer
106 improved coverage of mitochondrial transcripts and facilitate variant calling for larger
107 parts of the mtDNA genome (Extended Data Fig.3a). Consequently, it is feasible to
108 demultiplex full-length single-cell RNA-seq data using MitoSort. We evaluated the
109 performance of MitoSort using the Smart-seq2 dataset comprising human cells from
110 four donors and the HEK293T cell line(19). We also benchmarked our method against
111 souporcell(8), one of the best computational tools for demultiplexing pooled single-cell

112 RNA-seq data using genome variants. The classification results obtained by MitoSort
113 were highly consistent with those produced by souporcell (Fig.2c). Cells that were
114 assigned as to the same donors shared the same mitochondrial haplotype (Extended
115 Data Fig.3b). Furthermore, MitoSort was faster than souporcell in processing the same
116 dataset (Fig.2d). These results also demonstrate that MitoSort can accurately
117 demultiplex pooled datasets and applied to variant single-cell genomics libraries,
118 provided that the mitochondrial sequence can be readout.
119 In the end, we conducted a case study in which we multiplexed B cells from four normal
120 individuals and four patients with auto-immune diseases to profile gene expression,
121 chromatin accessibility, and mitochondrial DNA (Figure 2e). The scATAC-seq library
122 contained an average of 25% mtDNA reads per cell, which allowed accurate mtDNA
123 variant calling and demultiplexing (Extended Data Fig.4a). Using MitoSort, we could
124 clearly classify 16472 cells into eight singlet populations, as well as doublet groups
125 based on the allele frequency of unique mtDNA germline variants for each individual
126 (Figure 2f). Doublet is another important issue to resolve in multiplexing datasets,
127 which become increasingly prevalent with the growing number of captured cells.
128 Traditional methods for doublet detection, which rely on the transcriptional profiles(20),
129 may not effectively address this issue when the pooled cells lack cell type diversity. In
130 contrast, mitochondrial variant-based approaches hold promise as an effective strategy
131 for identifying cross-genotype doublets and mitigating the effects of doublets in
132 multiplexed datasets, particularly when dealing with a large number of pooled samples.
133 With MitoSort, we detected 6.4% of cell barcodes as cross-genotype doublets involved
134 more than one MT genotype (Figure 2f, Extended Data Fig.4b-c). Further quantification
135 metrics showed an expected higher number of RNA molecules and chromatin
136 fragments in doublets (Extended Data Fig.4d). Following the removal of low-quality
137 cells and cross-genotype doublets, we obtained 12943 high-quality B cell singlets.
138 High-dimensional clustering of multi-omic profiles identified 5 clusters, which could
139 be annotated as naïve B cells and memory B cells based on known markers (Fig.2g,

140 Extended Data Fig.4e-f). Using mtDNA SNP information, we can match donors with
141 MitoSort assigned labels. Among the four patients with auto-immune diseases, two
142 received treatments with rituximab (RTX). We found that most memory B cells were
143 depleted in these two patients treated with RTX (Figure 2h-i), which was consistent
144 with the function of RTX(21). These data show that MitoSort can be used to accurately
145 demultiplex samples of individuals under different pathologies, which facilitates
146 downstream analyses without technical batch effect.
147 In conclusion, the unique properties of mitochondrial DNA (mtDNA) make it a more
148 efficient and robust choice for endogenous genetic barcodes in sample demultiplexing
149 and the detection of cross-genotype doublets. However, only a few software options are
150 currently available for this purpose. MitoSort fills this gap by providing an accurate and
151 scalable method for demultiplexing pooled single-cell genomics data (mtscATAC-seq,
152 ASAP-seq, DOGMA-seq, Smart-seq) across donors and identifying cross-genotype
153 doublets. Nonetheless, the matrilineal inheritance of mitochondria poses a limitation to
154 this approach, as it is challenging to distinguish individuals with haploidentical mtDNA.
155

156 **Methods**

157 **Realignment for Mitochondrial reads**

158 To correct potential mapping errors around indels, the reads mapped to mitochondrial
159 genome were realigned locally using GATK version 3.5(15).
160

161 **Variant calling and germline mutation selection**

162 VarScan2(16) was employed with the parameters "–min-var-freq 0.01," "min-depth 8,"
163 and "–min-reads2 2" to call variants on the mitochondrial genome using BAM files
164 containing read pairs realigned to the mitochondrial genome. Since mitochondrial reads
165 from the BAM file were contributed by multiple individuals, we retained variants with
166 allele frequencies ranging from 1% to 99% as potential germline variants for each
167 individual (1% < bulk VAF < 99%).To determine the number of reference alleles and
168 alternate alleles in potential germline variants for each cell, we developed an in-house

Python script (available at https://github.com/tangzhj/MitoSort) to generate two matrices, namely, Ref.matrix and Alt.matrix. For each potential germline variant per cell, the allele frequency was calculated as the ratio of alternate counts to all counts and stored in the Frequency.matrix. The Frequency.matrix contained the allele frequency for each potential germline variant, including sample-specific germline variants, common germline variants, and widely spread somatic variants.

To ensure that only germline variants were included in our analysis, a straightforward filtering criterion was employed. Specifically, we selected variants where the proportion of cells having a variant allele frequency (VAF) above 99% among those with a non-zero VAF was higher than 50% . This criterion was chosen to ensure the selection of high-confidence germline variants with a high VAF in most cells, while simultaneously excluding low-coverage germline variants and somatic variants with low VAF.

Identification of specific germline variants for each individual

To identify sample-specific germline variants while excluding common germline variants that could potentially interfere with subsequent doublet detection, we utilized the robust information on germline variants in each sample. First, we performed initial clustering using the k-means algorithm from the sklearn library, with parameters n_clusters = pooled_sample_number, max_iter=1000, tol=1e-5, algorithm='full', and n_init = 10, based on the information obtained from the Frequency.matrix. Within each cell cluster, variants where the ratio between cells with VAF>99% and cells with VAF>0 was higher than 0.5 were assigned as a variant set. Subsequently, common variants that occurred in more than two sets were removed to obtain the final set of specific germline variants for each individual. The approach we described above allowed us to identify the cell clusters of each individual, as well as the sample-specific germline variants for the first time, thereby providing an efficient and robust method for downstream analysis such as doublet detection.

Doublet detection

Definitions:

K: the number of clusters defined by the above steps.

C: the number of cells. Lower-case c is used for indexing and referring to a specific cell barcode.

G: the number of sample-specific germline variants sets. Lower-case *g* is used to represent all the specific germline variants in a cluster, equal to G_i (*i* means cluster *i*). *g,c* means these cluster-specific germline variants with observed data in cell *c*.

A: allele counts. *Ag, c* is a tuple of size 2 with the first number representing the number of reference alleles and the second number representing the number of alternate alleles observed in the cluster-specific germline variants in cell *c*.

We already defined the initial clusters of cells and their corresponding cluster-specific germline variants. Our next objective was to identify doublets between each cluster.

For cluster-specific germline variants in each single cell, the observation of alternative allele counts follows a binomial distribution, with *p* being the probability of alternative allele frequency, $A_{g,c,1}$ being the alternative allele counts and *N* being the total counts for cluster-specific germline variants.

Equation (1):

$$P\{X = A_{g,c,1}\} = \binom{N}{A_{g,c,1}} p^{A_{g,c,1}} (1-p)^{A_{g,c,0}}$$

For a given cell, the probability of alternative alleles (*p*) or BAF at sample-specific germline variants can be modeled by a beta prior.

The Beta prior model for *p* can be tuned to reflect the relative prior plausibility of each *p*.

Equation (2): prior model:

$$f(p) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1} (1-p)^{\beta-1}$$

Upon observing data X=x where x∈{0, 1,…, n}, the likelihood function of p, obtained by plugging x into the Binomial pmf equation (1), we used equation (3):

$$L(p|X = A_{g,c,1}) = \binom{N}{A_{g,c,1}} p^{A_{g,c,1}} (1-p)^{A_{g,c,0}}$$

Via Bayes' Rule, the conjugate Beta prior combined with the Binomial data model produce a Beta posterior model for posterior probability of cells belong to each cluster(p). Putting equation (2) and equation (3) together, the posterior pdf is descried by equation (4):

$$f(p|X = A_{g,c,1}) \propto f(p)L(p|X = A_{g,c,1}) \propto p^{(\alpha + A_{g,c,1})-1}(1-p)^{(\beta + A_{g,c,0})-1}$$

For each cluster, α and β can be estimated using the allele counts across all the sample-specific germline variants in one cluster.

α = 1 + mean of alternate counts across variants in G_i of K_i, and β = 1. (for cluster).

For a cell c belongs to first-fitted cluster i (K_i), and g=G_i, the p_i was described by equation (5):

$$f(p_i|X = A_{g,c,1}, A_{g,c,1} + A_{g,c,0}, c \in K_i, g = G_i) \propto B(\alpha + A_{g,c,1}, \beta + A_{g,c,0})$$

When the cell c belongs to first-fitted cluster i (K_i) but g=G_j, p_j was described by equation (6):

$$f(p_j|X = A_{g,c,1}, A_{g,c,1} + A_{g,c,0}, c \in K_i, g = G_j) \propto B(\alpha + A_{g,c,1}, \beta + A_{g,c,0})$$

Then the P_i was p_1 for the first-fitted cluster and the highest P_j was selected as p_2 for the second fitted cluster. We estimate p as the mode of the posterior distribution.

When the cell is a singlet, p_1 will be dramatically higher than any P_j.

To give a reasonable threshold, we treated p_1, p_2 as two-dimensional data. We treated p_1>0.99 and p_2<0.01 as "Singlet" and treated p_1<0.8 or p_2>0.2 as "Doublet" at the beginning, this cutoff can be set by users. To predict the class of unlabeled cells, we

use the 'KNeighborsClassifier' method, which calculates the distance between the labeled cell point and all other unlabeled cell points in the dataset. The k=5 nearest neighbors are identified based on the distance metric, and the majority class of the k=5 nearest neighbors is used to predict the class of the unlabeled cells. It should be noted that if p_1 is greater than 0.99, we recommend not considering p_2, as high values of p_2 may be caused by potential somatic mutations.

Demultiplexing pooled samples

After removing doublets that were identified in the previous step, we further clustered the remaining singlets using the k-means clustering from the sklearn library with parameters n_clusters = pooled_sample_number, max_iter=1000, tol=1e-5, algorithm='full', and n_init = 10. The value of n_clusters parameter was set to correspond to the number of pooled samples. In cases where the exact number of pooled samples was unknown, we also provided the silhouette method to estimate the optimal number of clusters. The silhouette score reaches its global maximum at the optimal number of clusters, allowing for demultiplexing of pooled samples accurately.

Simulation of *in silico* data

In this study, we simulated mixtures of mtscATAC-seq data obtained from the 10x Genomics platform using custom Python scripts. Following the alignment of reads to the human reference genome hg38 using 10x Genomics Cell Ranger ATAC(22), mitochondrial reads were merged into a bam file with cell barcodes prefixed to distinguish individuals. Next, reads were realigned using GATK(15). In total, we had 52283 cells across 8 samples. To assess the performance of MitoSort systematically, we generated synthetic mixtures into BAM files by random sampling using a range of parameter choices, including the number of reads mapped to mitochondrial genome (ranging from 250 to 2000), the number of pooled samples (ranging from 2 to 8), the number of cells for each sample (ranging from 100 to 600), and the percentage of

doublets (1%, 5%, 10%, and 15%). Doublets were simulated by randomly choosing two cell barcodes and merging their read counts as a new cell. For each parameter set, simulations were repeated five times to account for variability. The MitoSort pipeline, including variant calling, doublet identification and demultiplexing, was ran on each simulated BAM file. We evaluated the performance of MitoSort by comparing its classification results with the true labels, and computed accuracy, true positive rate (TPR), and false discovery rate (FDR).

Evaluation of MitoSort using cell hashing data

For multiplexed DOGMA-seq dataset comprising activated and stimulated T cells from two healthy donors(17), we first downloaded its HTO count data (GSM6032896) and ATAC raw data (SRX14779147). HTO count data containing hashtag reads were used to assign each cell a hashtag ID by the function HTODemux() from Seurat(23) package. Raw reads in ATAC data were aligned to the human reference genome (hg38) using Cell Ranger ATAC(22). Based on the generated BAM file, we ran MitoSort pipeline, including MT reads realignment, variant calling, doublet identification and demultiplexing. Cell barcodes classified as singlets by hashtag and MitoSort were compared to evaluate the consistency of assignment results obtained from both approaches.

For multiplexed ASAP-seq dataset comprising NK cells from four donors(18), we first downloaded mitochondrial genotyping data and hashtag-based sample identity information (GSM6413442). We then used MitoSort to identify doublets and demultiplex samples based on the mitochondrial genotyping data. Finally, we compared the donor assignment results obtained from MitoSort with those derived from the hashtag data.

Analysis of multiplexed Smart-seq2 data

301 We obtained raw Smart-seq2 data for the Human Cell Atlas (HCA) benchmarking
302 sample(19) from SRR9621775 and processed it using zUMIs
303 (https://github.com/sdparekh/zUMIs), which involved reads filtering, mapping to the
304 human reference genome (hg38), and quantification of gene expression. Based on
305 species annotation from the original paper, human cells including PBMCs from four
306 donors and the HEK293T cell line were retained for downstream analysis. Next, we
307 used MitoSort and soupercell, respectively for demultiplexing human cells. The donor
308 assignment results and run time of MitoSort and soupercell were compared to evaluate
309 their performance.
310

311 **Isolation of PBMCs and B cells**

312 The study protocol was approved by the local ethics committee of Sun Yat-sen
313 University (#2015/521), and written informed consent was obtained from all donors.
314 All experiments and protocols for human studies involving human samples were
315 performed in accordance with the relevant guidelines and regulations. To obtain
316 peripheral blood mononuclear cells (PBMCs), the peripheral blood was diluted 1:1.5 in
317 0.9% physiological saline and cells were isolated with a Lymphoprep™ density
318 gradient (density 1.077 g/mL) according to the manufacturer's instructions
319 (STEMCELL). B cells were isolated using Miltenyi CD19 MicroBeads (#130-050-301)
320 as described by the manufacturer. Cells were counted using Trypan Blue exclusion
321 under a hemocytometer.
322

323 **Single cell multi-omics library construction and sequencing**

324 Single-cell multi-omics libraries were generated using the 10X Chromium Controller
325 and the Chromium Next GEM Single Cell Multiome ATAC + Gene Expression kit
326 according to the manufacturer’s instructions. DNA LoBind tubes (1.5 ml; Eppendorf)
327 were used to wash cells in PBS and in downstream processing steps. To retain mtDNA,
328 permeabilization was performed using LLL conditions, as described in mtscATAC-

329 seq(3, 12) with 10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% NP40 and
330 1% BSA. After washing, cells were fixed with 1% formaldehyde (ThermoFisher no.
331 28906) in PBS for 10 minutes at room temperature, followed by quenching with glycine
332 solution to a final concentration of 0.125 M. After washing twice in PBS via
333 centrifugation at 400*g*, 5 min, 4 °C, cells were treated with lysis buffer (10 mM Tris-
334 HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% NP40, 1% BSA) for 3 minutes, followed
335 by adding 1 ml of chilled wash buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM
336 $MgCl_2$, 1% BSA) and inversion before centrifugation at 400*g*, 5 min, 4 °C. The
337 supernatant was discarded, and cells were diluted in 1× Diluted Nuclei buffer (10X
338 Genomics) before counting using Trypan Blue and a Countess II FL Automated Cell
339 Counter. Briefly, after tagmentation, the cells were loaded on a Chromium Controller
340 Single-Cell instrument to generate single-cell Gel Bead-In-Emulsions (GEMs)
341 followed by PCR as described in the 10X protocol. The final libraries were quantified
342 using a Qubit dsDNA HS Assay kit (Invitrogen) and a High Sensitivity DNA chip run
343 on a Bioanalyzer 2100 system (Agilent). Libraries were sequenced on illumina
344 NovaSeq 6000 at Berry Genomics.
345
346 **Analysis of multi-omics data of multiplexed B cells**
347 Raw data of multiome were preprocessed and aligned to the human reference genome
348 (hg38) using 10x Genomics Cell Ranger arc. The resulting BAM file from ATAC data
349 was subjected to the MitoSort pipeline for doublet identification and sample
350 demultiplexing. The cell-by-gene and cell-by-peak count matrices were further
351 processed using the R package Seurat(23) and Signac(24). High-quality cells with less
352 than 5000 RNA molecules, mitochondrial RNA content <10%, 200-3500 expressed
353 genes, 2000-50000 fragments, TSS enrichment score >= 5, nucleosome signal <1, and
354 the percentage of reads in peaks >= 50 were retained for downstream analysis.
355 Subsequently, weighted nearest neighbor (WNN) analysis was performed on both RNA
356 and ATAC-seq modalities, resulting in 10 clusters. Cells were then annotated into four

cell types using the following marker genes: B cells (BANK1, MS4A1); Plasma (MZB1); T cells (CD3D, CD3E); monocytes (CD14, S100A9). After removing cross-genotype doublets assigned by MitoSort, we obtained a total of 12943 B cells. Next, WNN analysis was applied to the subclustering analysis of multiome data of B cells, which yielded 5 clusters. B cells were further divided into two subtypes using the following marker genes: naive B cells (TCL1A, IL4R); memory B cells (CD27). Based on the donor origins of cells obtained from MitoSort, the percentage of each cluster in each sample was calculated.

Code availability

MitoSort is freely available at https://github.com/tangzhj/MitoSort.

Contributions

J. X. designed and conceived the study. Z.T., P.S. collected and analyzed the data. X.L., S.L. and Y.X. supported the data analysis. Y.S. and W.Z. collected samples. W.Z. generated the single cell multi-omic library. J. X., Z.T. and P.S. wrote the manuscript with inputs from all authors. All authors read and approved the final manuscript.

Funding

This work was supported by National Key R&D Program of China (2021YFA1102100), National Natural Science Foundation of China (32070644), Guangdong Basic and Applied Basic Research Foundation (2019A1515110387, 2019B1515130004) to J.X.

Competing interests

The authors declare no conflict of interest

Fig.1 Workflow and evaluation of MitoSort.

a, Schematic showing the purpose of MitoSort.

385 **b**, Overview of MitoSort pipeline. Firstly, mitochondrial reads are realigned and
386 putative single-nucleotide polymorphisms (SNPs) are called. Subsequently, count
387 matrices of reference and alternate alleles are generated for each cell, where each row
388 represents a cell and each column represents a variant. Following that, sample-specific
389 germline variants are identified for each individual. Cross-genotype doublets are then
390 detected by a beta-binomial model. Lastly, singlets are demultiplexed for downstream
391 analyses.
392 **c-f**, Systematic assessment of MitoSort performance on simulated data using a range of
393 parameter choices, five replicate runs are used to ensure the consistency of the results.
394 **c**, Classification accuracy of MitoSort on simulated datasets comprising 2-8 mixed
395 samples, each with 500 cells, an 8% doublet rate, and 2000 MT reads per cell. **d**,
396 Classification accuracy of MitoSort on simulated datasets containing eight pooled
397 samples, each with varying numbers of cells, an 8% doublet rate and 2000 MT reads
398 per cell. **e**, Classification accuracy of MitoSort on simulated datasets containing eight
399 pooled samples, each with 500 cells, 2000 MT reads per cell and varying doublet rates.
400 **f**, Classification accuracy of MitoSort on simulated datasets containing eight pooled
401 samples, each with 500 cells, an 8% doublet rate, and varying numbers of mitochondrial
402 (MT) reads per cell.
403
404 **Fig.2 Application of MitoSort to real multiplexed samples**
405 **a,** Percentage of barcodes shared between hashtag-based (rows) and MitoSort-based
406 (columns) assignments in multiplexed DOGMA-seq data consisting of two donors.
407 **b**, Percentage of barcodes shared between hashtag-based (rows) and MitoSort-based
408 (columns) assignments in multiplexed ASAP-seq data consisting of four donors.
409 **c**, Percentage of barcodes shared between soupercell-based (rows) and MitoSort-based
410 (columns) assignments in multiplexed Smart-seq2 data consisting of human cells from
411 four donors and the HEK293T cell line.

412 **d**, Bar plot showing running time of MitoSort and soupercell when processing Smart-
413 seq2 data.
414 **e**, Schematic of experimental design. Peripheral blood mononuclear cells (PBMCs)
415 from four normal individuals and four patients with antoimmune diseases were
416 collected and pooled. Next, B cells were enriched to profile gene expression, chromatin
417 accessibility, and mitochondrial DNA.
418 **f**, Heatmap showing the allele frequency of aggregated sample-specific mitochondrial
419 germline mutations (rows) across cells (columns) sorted by MitoSort assignment. The
420 number in the parentheses of each row name represents the number of sample-specific
421 mitochondrial germline mutations for this sample.
422 **g**, UMAP of weighted-nearest neighbor (WNN) graph for RNA and ATAC-seq
423 modalities showing cell type annotation for each cell.
424 **h**, UMAP of weighted-nearest neighbor (WNN) graph for RNA and ATAC-seq
425 modalities showing cells per individual colored by cell number densities. Disease status
426 and clinical information for each individual are shown in subtitles.
427 **i**, Bar plot showing the percentage of each cell cluster in each individual.
428
429 **Extended Data Fig.1 Evaluation of MitoSort using in silico data.**
430 **a-b,** The true positive rate (TPR) and false discovery rate (FDR) of donor assignment
431 and doublet identification by MitoSort on simulated datasets containing eight pooled
432 samples, each with 500 cells, an 8% doublet rate, and varying numbers of mitochondrial
433 (MT) reads per cell.
434 **c-d**, The true positive rate (TPR) and false discovery rate (FDR) of donor assignment
435 and doublet identification by MitoSort on simulated datasets comprising 2-8 mixed
436 samples, each with 500 cells, an 8% doublet rate, and 2000 MT reads per cell.
437 **e-f**, The true positive rate (TPR) and false discovery rate (FDR) of donor assignment
438 and doublet identification by MitoSort on simulated datasets containing eight pooled

samples, each with varying numbers of cells, an 8% doublet rate and 2000 MT reads per cell.

g-h, The true positive rate (TPR) and false discovery rate (FDR) of donor assignment and doublet identification by MitoSort on simulated datasets containing eight pooled samples, each with 500 cells, 2000 MT reads per cell and varying doublet rates.

Extended Data Fig.2 Evaluation of MitoSort using cell hashing data.

a, Heatmap showing the allele frequency of sample-specific mitochondrial germline mutations (columns) across cells (rows) with different sequencing depth of mitochondrial genome (2-4X, 4-8X and >8X) in multiplexed DOGMA-seq data.

b, Percentage of barcodes with different sequencing depth of mitochondrial genome (2-4X, 4-8X and >8X) shared between hashtag-based (rows) and MitoSort-based (columns) assignments in multiplexed DOGMA-seq data.

c, Heatmap showing the allele frequency of sample-specific mitochondrial germline mutations (columns) across cells (rows) with different sequencing depth of mitochondrial genome (5-10X, 10-20X and >20X) in multiplexed ASAP-seq data.

d, Percentage of barcodes with different sequencing depth of mitochondrial genome (5-10X, 10-20X and >20X) shared between hashtag-based (rows) and MitoSort-based (columns) assignments in multiplexed ASAP-seq data.

Extended Data Fig.3 Application of MitoSort to multiplexed Smart-seq2/3 data

a, Mean sequencing coverage along the mitochondrial genome per cell for the Human Cell Atlas (HCA) benchmarking data profiled by 10x 3' scRNA-seq and Smart-seq2.

b, Heatmap showing the allele frequency of sample-specific mitochondrial germline mutations (columns) across cells (rows) in multiplexed Smart-seq2 data comprising human cells from four donors and the HEK293T cell line.

Extended Data Fig.4 Analysis of multi-omics data of multiplexed B cells

467 **a,** Violin plot showing distribution of percentage of mtDNA reads per cell in the
468 scATAC-seq library.
469 **b**, tSNE plot of cell barcodes based on mitochondrial germline mutation profiles. Each
470 dot represents a cell barcode and is colored based on the singlet/doublet classification
471 by MitoSort.
472 **c**, Bar plot showing proportion of cell barcodes that are classified as singlet and doublet
473 by MitoSort.
474 **d**, Violin plots showing the distribution of the number of RNA molecules (top) and
475 ATAC fragments (bottom) per cell in cell barcodes classified as singlet and doublet.
476 **e**, UMAP of weighted-nearest neighbor (WNN) graph for RNA and ATAC-seq
477 modalities. Each dot represents a cell colored by its cluster.
478 **f**, Dot plot showing expression of marker genes in each cluster. Dot size indicates the
479 proportion of cells expressing the gene and dot color represents the scaled average
480 expression level.
481
482 **Reference**

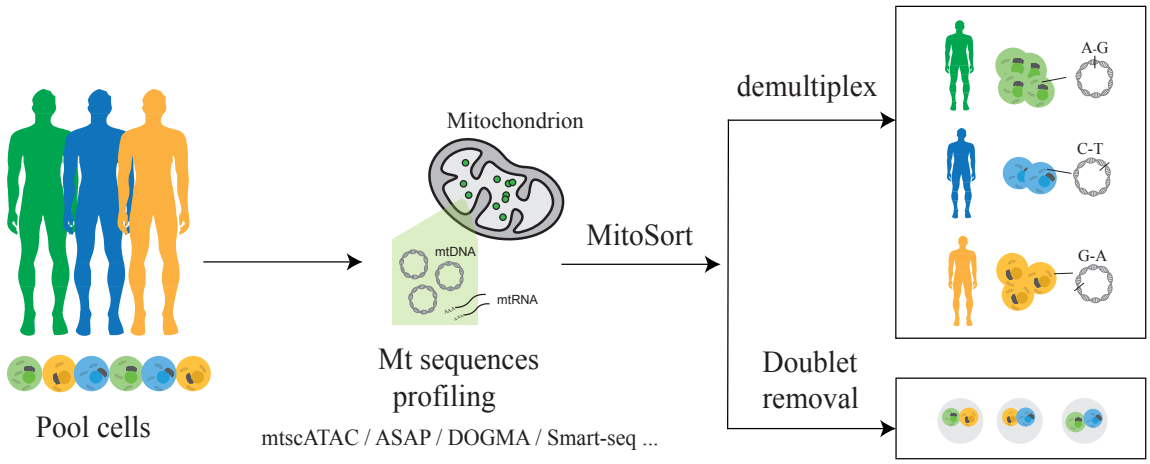
483 1. Vandereyken,K., Sifrim,A., Thienpont,B. and Voet,T. (2023) Methods and
484 applications for single-cell and spatial multi-omics. *Nat Rev Genet*,
485 10.1038/s41576-023-00580-2.
486 2. Dimitriu,M.A., Lazar-Contes,I., Roszkowski,M. and Mansuy,I.M. (2022) Single-
487 Cell Multiomics Techniques: From Conception to Applications. *Front Cell Dev*
488 *Biol*, **10**.
489 3. Mimitou,E.P., Lareau,C.A., Chen,K.Y., Zorzetto-Fernandes,A.L., Hao,Y.,
490 Takeshima,Y., Luo,W., Huang,T.S., Yeung,B.Z., Papalexi,E., *et al.* (2021)
491 Scalable, multimodal profiling of chromatin accessibility, gene expression and
492 protein levels in single cells. *Nat Biotechnol*, **39**.
493 4. Zhang,Y., Xu,S., Wen,Z., Gao,J., Li,S., Weissman,S.M. and Pan,X. (2022) Sample-
494 multiplexing approaches for single-cell sequencing. *Cellular and Molecular Life*
495 *Sciences*, **79**.
496 5. Kang,H.M., Subramaniam,M., Targ,S., Nguyen,M., Maliskova,L., McCarthy,E.,
497 Wan,E., Wong,S., Byrnes,L., Lanata,C.M., *et al.* (2018) Multiplexed droplet
498 single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol*, **36**.
499 6. Xu,J., Falconer,C., Nguyen,Q., Crawford,J., McKinnon,B.D., Mortlock,S.,
500 Senabouth,A., Andersen,S., Chiu,H.S., Jiang,L., *et al.* (2019) Genotype-free
501 demultiplexing of pooled single-cell RNA-seq. *Genome Biol*, **20**.

502 7. Huang,Y., McCarthy,D.J. and Stegle,O. (2019) Vireo: Bayesian demultiplexing of
503 pooled single-cell RNA-seq data without genotype reference. *Genome Biol*, **20**.
504 8. Heaton,H., Talman,A.M., Knights,A., Imaz,M., Gaffney,D.J., Durbin,R.,
505 Hemberg,M. and Lawniczak,M.K.N. (2020) Souporcell: robust clustering of
506 single-cell RNA-seq data by genotype without reference genotypes. *Nat*
507 *Methods*, **17**.
508 9. Han,S., Kim,K., Park,S., Lee,A.J., Chun,H. and Jung,I. (2022) scAVENGERS: a
509 genotype-based deconvolution of individuals in multiplexed single-cell ATAC-
510 seq data without reference genotypes. *NAR Genom Bioinform*, **4**.
511 10. Penter,L., Gohil,S.H. and Wu,C.J. (2022) Natural Barcodes for Longitudinal
512 Single Cell Tracking of Leukemic and Immune Cell Dynamics. *Front Immunol*,
513 **12**.
514 11. Kulhankova,L., Montiel González,D., Bindels,E., Kling,D., Kayser,M. and
515 Mulugeta,E. (2023) Single-cell transcriptome sequencing allows genetic
516 separation, characterization and identification of individuals in multi-person
517 biological mixtures. *Commun Biol*, **6**.
518 12. Lareau,C.A., Ludwig,L.S., Muus,C., Gohil,S.H., Zhao,T., Chiang,Z., Pelka,K.,
519 Verboon,J.M., Luo,W., Christian,E., *et al.* (2021) Massively parallel single-cell
520 mitochondrial DNA genotyping and chromatin profiling. *Nat Biotechnol*, **39**.
521 13. Ramsköld,D., Luo,S., Wang,Y.C., Li,R., Deng,Q., Faridani,O.R., Daniels,G.A.,
522 Khrebtukova,I., Loring,J.F., Laurent,L.C., *et al.* (2012) Full-length mRNA-Seq
523 from single-cell levels of RNA and individual circulating tumor cells. *Nat*
524 *Biotechnol*, **30**.
525 14. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P.,
526 Chaisson,M. and Gingeras,T.R. (2013) STAR: Ultrafast universal RNA-seq
527 aligner. *Bioinformatics*, **29**.
528 15. McKenna,A., Hanna,M., Banks,E., Sivachenko,A., Cibulskis,K., Kernytsky,A.,
529 Garimella,K., Altshuler,D., Gabriel,S., Daly,M., *et al.* (2010) The genome
530 analysis toolkit: A MapReduce framework for analyzing next-generation DNA
531 sequencing data. *Genome Res*, **20**.
532 16. Koboldt,D.C., Zhang,Q., Larson,D.E., Shen,D., McLellan,M.D., Lin,L.,
533 Miller,C.A., Mardis,E.R., Ding,L. and Wilson,R.K. (2012) VarScan 2: Somatic
534 mutation and copy number alteration discovery in cancer by exome sequencing.
535 *Genome Res*, **22**.
536 17. Xu,Z., Heidrich-O'Hare,E., Chen,W. and Duerr,R.H. (2022) Comprehensive
537 benchmarking of CITE-seq versus DOGMA-seq single cell multimodal omics.
538 *Genome Biol*, **23**.
539 18. Rückert,T., Lareau,C.A., Mashreghi,M.F., Ludwig,L.S. and Romagnani,C. (2022)
540 Clonal expansion and epigenetic inheritance of long-lasting NK cell memory.
541 *Nat Immunol*, **23**.

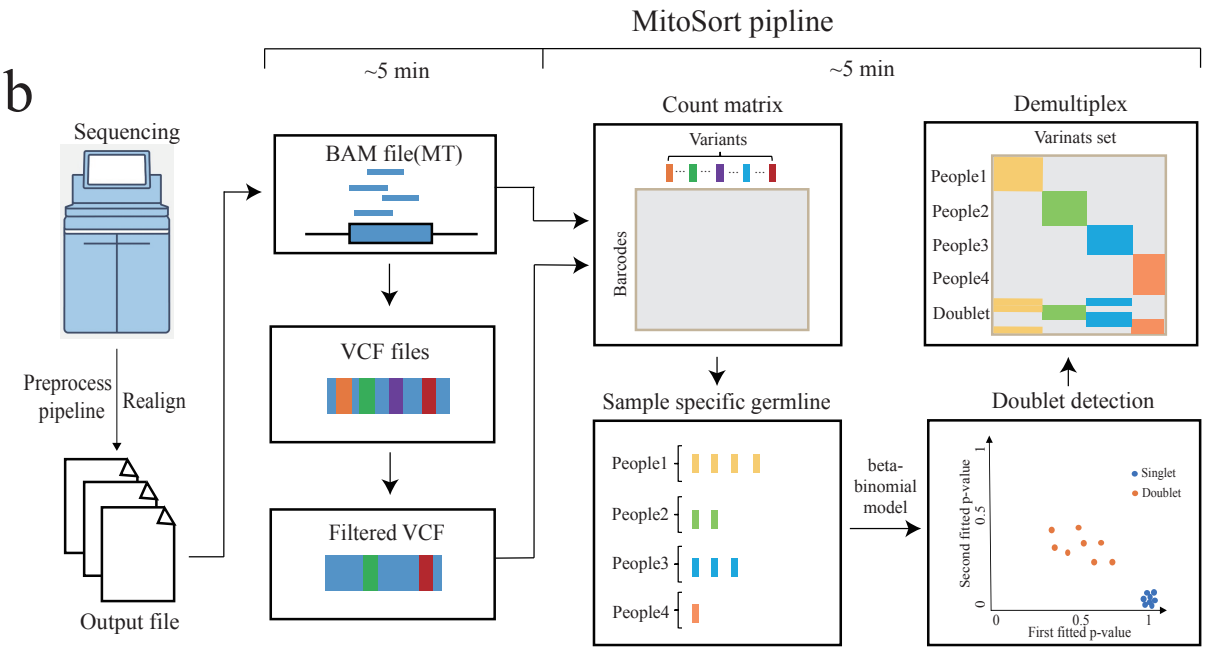
19. Mereu,E., Lafzi,A., Moutinho,C., Ziegenhain,C., McCarthy,D.J., Álvarez-Varela,A., Batlle,E., Sagar, Grün,D., Lau,J.K., *et al.* (2020) Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat Biotechnol*, **38**.
20. Xi,N.M. and Li,J.J. (2021) Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data. *Cell Syst*, **12**.
21. Kneitz,C., Wilhelm,M. and Tony,H.P. (2002) Effective B cell depletion with rituximab in the treatment of autoimmune diseases. *Immunobiology*, **206**.
22. Satpathy,A.T., Granja,J.M., Yost,K.E., Qi,Y., Meschi,F., McDermott,G.P., Olsen,B.N., Mumbach,M.R., Pierce,S.E., Corces,M.R., *et al.* (2019) Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat Biotechnol*, **37**.
23. Hao,Y., Hao,S., Andersen-Nissen,E., Mauck,W.M., Zheng,S., Butler,A., Lee,M.J., Wilk,A.J., Darby,C., Zager,M., *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell*, **184**.
24. Stuart,T., Srivastava,A., Madad,S., Lareau,C.A. and Satija,R. (2021) Single-cell chromatin state analysis with Signac. *Nat Methods*, **18**.

Fig.1

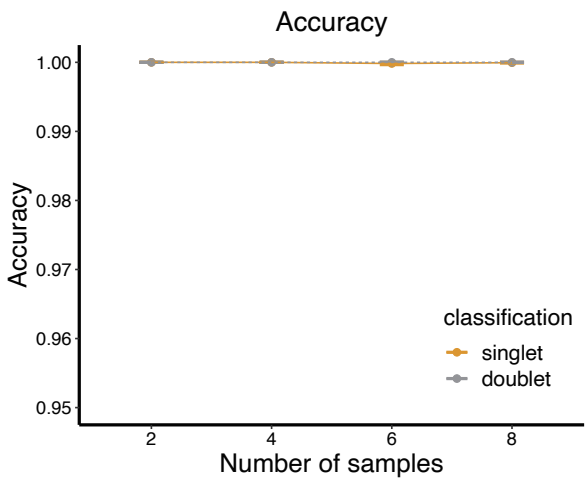
a


Pool cells
Mitochondrion
mtDNA
mtRNA
Mt sequences profiling
mtscATAC / ASAP / DOGMA / Smart-seq ...
MitoSort
demultiplex
A-G
C-T
G-A
Doublet removal

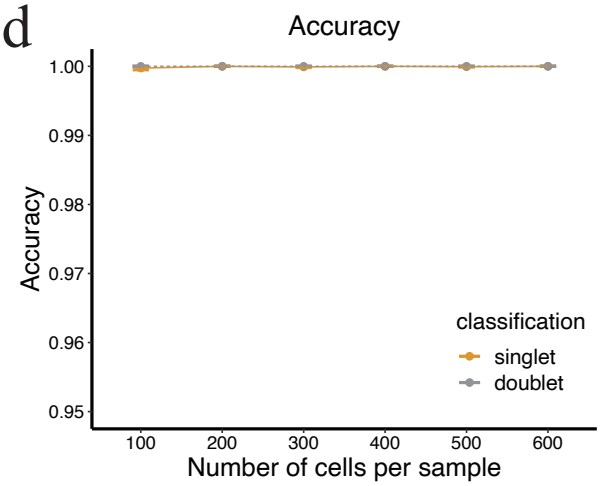
b


MitoSort pipline
~5 min
~5 min
Sequencing
Preprocess pipeline
Realign
Output file
BAM file(MT)
VCF files
Filtered VCF
Count matrix
Variants
Barcodes
Demultiplex
Varinats set
People1
People2
People3
People4
Doublet
Sample specific germline
beta-binomial model
Doublet detection
Singlet
Doublet
Second fitted p-value
First fitted p-value

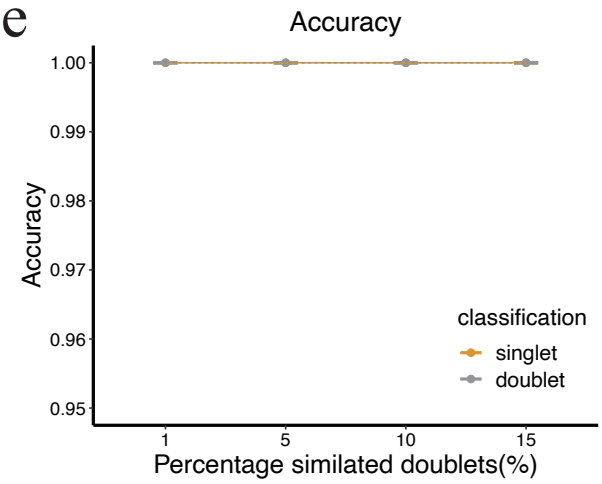
c


Accuracy
Accuracy
classification
singlet
doublet
Number of samples

d


Accuracy
Accuracy
classification
singlet
doublet
Number of cells per sample

e


Accuracy
Accuracy
classification
singlet
doublet
Percentage similated doublets(%)

f

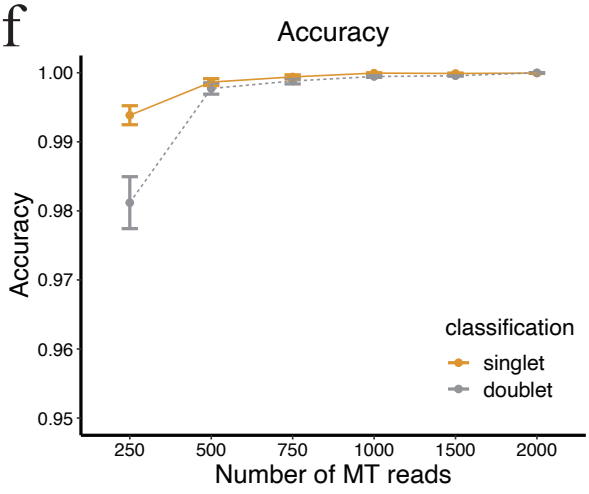
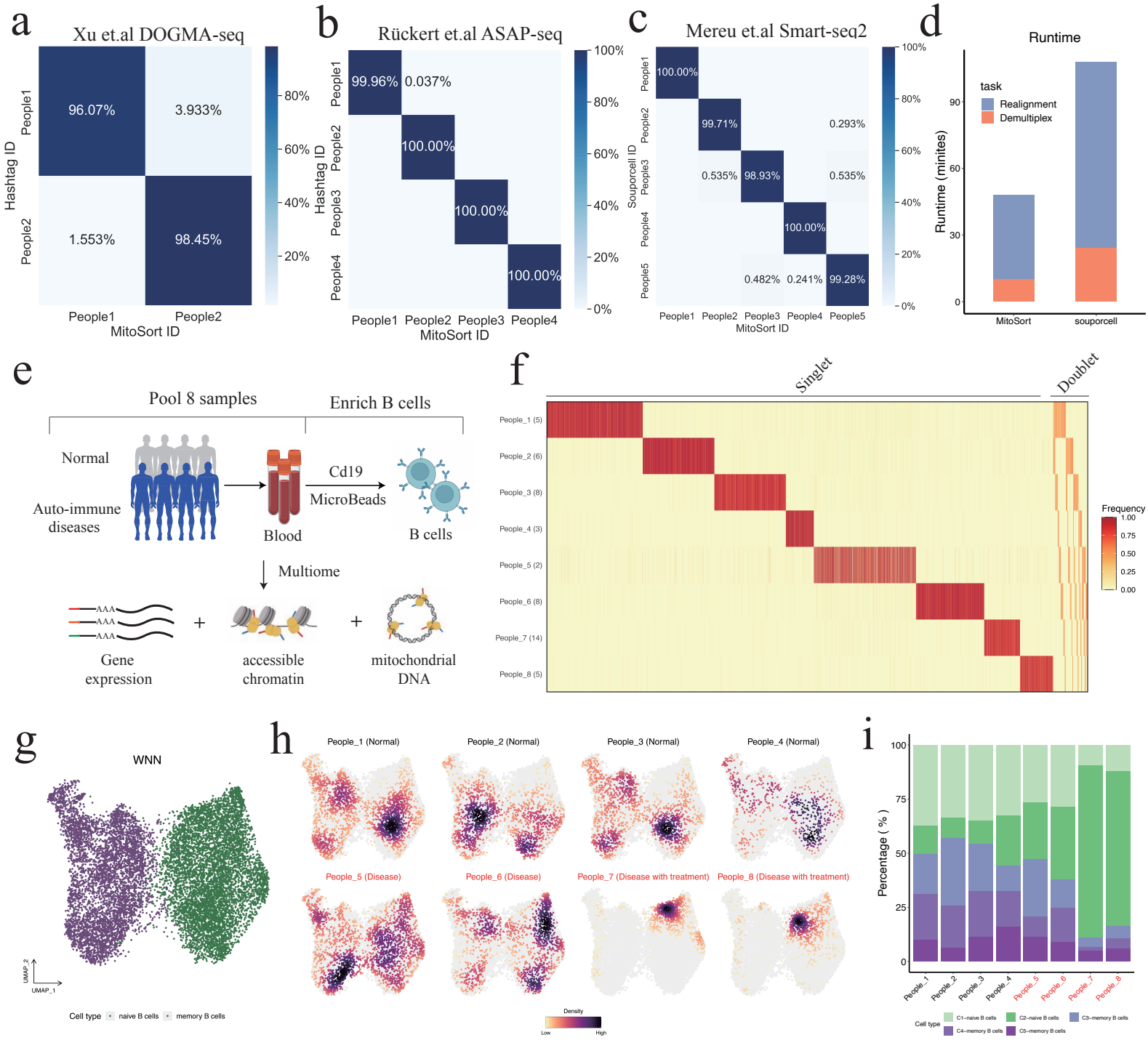

Accuracy
Accuracy
classification
singlet
doublet
Number of MT reads

Fig.2


a
Xu et.al DOGMA-seq
Hashtag ID
People1
People2
96.07%
3.933%
1.553%
98.45%
MitoSort ID
b
Rückert et.al ASAP-seq
99.96%
0.037%
100.00%
100.00%
100.00%
People1 People2 People3 People4
c
Mereu et.al Smart-seq2
Souporcell ID
People1 People2 People3 People4 People5
100.00%
99.71%
0.293%
0.535%
98.93%
0.535%
100.00%
0.482%
0.241%
99.28%
d
Runtime
task
Realignment
Demultiplex
Runtime (minites)
MitoSort
souporcell
e
Pool 8 samples
Enrich B cells
Normal
Auto-immune diseases
Blood
Cd19 MicroBeads
B cells
Multiome
Gene expression
accessible chromatin
mitochondrial DNA
f
Singlet
Doublet
People_1 (5)
People_2 (6)
People_3 (8)
People_4 (3)
People_5 (2)
People_6 (8)
People_7 (14)
People_8 (5)
Frequency
g
WNN
UMAP_1
UMAP_2
Cell type
naive B cells
memory B cells
h
People_1 (Normal)
People_2 (Normal)
People_3 (Normal)
People_4 (Normal)
People_5 (Disease)
People_6 (Disease)
People_7 (Disease with treatment)
People_8 (Disease with treatment)
Density
Low
High
i
Percentage (%)
Cell type
C1–naive B cells
C2–naive B cells
C3–memory B cells
C4–memory B cells
C5–memory B cells