Published online 31 December 2022

NAR Genomics and Bioinformatics, 2022, Vol. 4, No. 4 **1**
https://doi.org/10.1093/nargab/lqac095

scAVENGERS: a genotype-based deconvolution of individuals in multiplexed single-cell ATAC-seq data without reference genotypes

Seungbeom Han[1], Kyukwang Kim[1], Seongwan Park[1], Andrew J. Lee[1], Hyonho Chun[2] and Inkyung Jung[1,*]

[1]Department of Biological Sciences, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea and [2]Department of Mathematical Sciences, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea

Received May 31, 2022; Revised November 09, 2022; Editorial Decision December 01, 2022; Accepted December 11, 2022

ABSTRACT

Genetic differences inferred from sequencing reads can be used for demultiplexing of pooled single-cell RNA-seq (scRNA-seq) data across multiple donors without WGS-based reference genotypes. However, such methods could not be directly applied to single-cell ATAC-seq (scATAC-seq) data owing to the lower read coverage for each variant compared to scRNA-seq. We propose a new software, scATAC-seq Variant-based EstimatioN for GEnotype ReSolving (scAVENGERS), which resolves this issue by calling more individual-specific germline variants and using an optimized mixture model for the scATAC-seq. The benchmark conducted with three synthetic multiplexed scATAC-seq datasets of peripheral blood mononuclear cells and prefrontal cortex tissues showed outstanding performance compared to existing methods in terms of accuracy, doublet detection, and a portion of donor-assigned cells. Furthermore, analyzing the effect of the improved sections provided insight into handling pooled single-cell data in the future. Our source code of the devised software is available at GitHub: https://github.com/kaistcbfg/scAVENGERS.

INTRODUCTION

The degree of chromatin accessibility is one of the major epigenetic factors used to decipher the functional role of non-coding regions in gene regulation. Assays combined with high-throughput sequencing technologies have been developed to obtain a genome-wide profile of accessible chromatin regions (1). Among these methods, Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) (2) has become very popular because of its simple and time-saving protocol. These advantages enabled ATAC-seq to be easily ported to single-cell level experiments (3,4), which allowed single-cell ATAC-seq to become one of the most successfully established single-cell profiling methods after being combined with droplet-based technologies (5).

Although the development of sequencing technology has resulted in increased efficiency, the cost and batch effect still hinder scaling up single-cell experiments to a large number of samples. To overcome such limitations, an experimental design that multiplexes samples with diverse genetic backgrounds and deconvolves them with detected genetic variants in sequenced reads was proposed in single-cell RNA-seq. A method named Demuxlet (6) which measures the likelihood of observing RNA-seq reads containing single-nucleotide polymorphisms (SNPs) sets using a statistical model, was developed and successfully applied to single-cell multi-omics integrative analyses (7). However, a reference genotype database for individuals must be created. This was addressed by calling genetic variants directly from the scRNA-seq reads and utilizing them for the demultiplexing. Methods such as souporcell (8), Vireo (9) and scSplit (10) were developed based on this concept.

However, there are hurdles to apply this approach directly to the scATAC-seq. The scATAC-seq reads are sampled directly from genomic DNA, which has lower copy numbers compared to RNA molecules. Furthermore, the genomic regions where the reads are sampled are broader compared to those in scRNA-seq since scATAC-seq collects reads from accessible chromatin regions while scRNA-seq collects reads from only exonic regions. As the loci to be covered by the cell barcode per drop increases and the amount of genetic material in each locus decreases, the read coverage per variant in scATAC-seq becomes lower than those of scRNA-seq. These inherent differ-

*To whom correspondence should be addressed. Tel: +82 42 350 7315; Fax: +82 42 350 2610; Email: ijung@kaist.ac.kr

ences might cause the algorithms used in demultiplexing of pooled scRNA-seq to produce inappropriate results, which indicates that several modifications that consider the characteristics of scATAC-seq are needed, rather than directly applying the demultiplexing method for scRNA-seq. In response, we devised a scAVENGERS pipeline, which introduces an appropriate read alignment tool, variant caller, and mixture model to appropriately process the demultiplexing of scATAC-seq data. Benchmarks showed that scAVENGERS successfully demultiplexes samples in human peripheral blood mononuclear cell (PBMC) and human prefrontal cortex scATAC-seq datasets, achieving better performance compared to the conventional method.

MATERIALS AND METHODS

Single-cell ATAC sequencing dataset preparation

Single-nucleus ATAC sequencing (snATAC-seq) datasets from two human PBMCs (*in-house* generated and public) and public human prefrontal cortices were prepared for the benchmark (Supplementary Table 1). For the *in-house* dataset, blood acquisition and isolation of five donor samples were conducted by referring to Lee *et al.*'s protocol (11). Following the isolation, snATAC-seq libraries were generated using the ChromiumTM Single Cell ATAC Library & Gel Bead Kit v1 (10X genomics, Pleasanton, CA) and sequenced by DNBSEQ-G400 (MGI, Shenzhen, China). The public data were acquired by You *et al.*'s ten PBMC data (12) and Morabito *et al.*'s 12 human prefrontal cortex data (13) from Sequence Read Archive (SRA, accession number: PRJNA718009 and PRJNA729525, respectively). Individual samples were processed by the Cell Ranger ATAC pipeline (14) with '-p -t 4 -M -R' options and hg38 reference genome (refdata-cellranger-atac-GRCh38-1.2.0). Generated BAM files of each dataset were merged by the 'samtools merge' command to make synthetic multiplexed snATAC-seq datasets with known donor labels. After merging, read group information was removed by the 'samtools addreplacerg' command. In this process, the snATAC-seq barcode information was also merged, using the overlapping barcodes as the synthetic doublet labels. Subsampling cell barcodes were done by Linux command 'shuf | head -n N', where N is the number of cell barcodes to sample. For making a subset of BAM file by these subsampled cell barcodes, we used subset-bam program.

Variant calling and processing

Strelka2 software (15) was used as a main variant caller for the scAVENGERS. The 'configureStrelkaGermlineWorkflow.py' script in strelka2 package was first executed with the –bam, –referenceFasta, and –runDir options. Then, 'runWorkflow.py' script generated in the path provided as an input to –runDir was executed with '-m local -j 20' parameters. Default values were used for the other input parameters. The output Variant Call Format (VCF) file from strelka2 was further processed by the Linux AWK command to extract SNPs (awk '{if ($0~"#" || length($4)==1 && length($5)==1) {print}'). Further filtration based on the VCF FILTER tag is provided by using the bcftools. 'bcftools view -f PASS -Ob' command was used to extract the variants with a certain quality (PASS or lowGQX) from the raw VCF file.

The scAVENGERS pipeline also supports freebayes software for the variant calling process. The freebayes parameters of '-iXu -C 2 -q 20 -n 3 -E 1 -m 30 –min-coverage 20 –pooled-continuous' were used. After the variant call, the variants with a Phred-scaled quality score (QUAL field) <100 were filtered by using the bcftools.

The prepared variant file was converted to ref.mtx and alt.mtx file using VarTrix software to express reference/alternative allele counts on each locus for each cell barcode.

Benchmark against existing methods

We compared the performance of scAVENGERS against souporcell (8), demuxlet (6) and scSplit (10). For all these programs, we used default or recommended settings in the documentations of each program. For souporcell, we used a default pipeline, which uses freebayes as variant caller and VarTrix as allele count matrix generator. We remapped the reads by minimap2 in default.

Before running demuxlet, we obtained reference-genotypes by calling variants from each donor using Strelka2. Then, we ran demuxlet using these reference-genotypes for individual donors, alignment and cell barcodes generated by Cell Ranger ATAC-seq pipeline. After we acquired results, we interpreted ambiguous cell barcodes marked with prefix 'AMB' as unassigned cell barcodes.

We performed variant calling for scSplit (10) by using freebayes with parameters '-iXu -C 2 -q 1', which is an option recommended in documentation of scSplit. Then, we ran 'scSplit count' in default option to create allele count matrix. The matrices are used for demultiplexing by 'scSplit run' command.

Implementation details

The overall pipeline was implemented by using the Python programming language. For the likelihood computation and parameter update by maximum likelihood, the just-in-time (JIT) compiler for Python (Numba package) was used for faster computation. Since the barcode-variant count matrix is a very sparse matrix, scAVENGERS uses Scipy's sparse matrix structure to enable large data processing with a memory-efficient structure. By adopting the troublet method for the doublet detection part, Rust programming language was also used.

Detailed information of used external software is listed below.

Cell Ranger ATAC-seq: Release 1.2.0. https://github.com/10XGenomics/cellranger-atac

samtools: Release 1.14. https://github.com/samtools/samtools/releases/tag/1.14

subset-bam: Release 1.1.0. https://github.com/10XGenomics/subset-bam

BWA: Release 0.7.17. https://github.com/lh3/bwa/releases/tag/v0.7.17
Strelka2: Release 2.9.2. https://github.com/Illumina/strelka/releases/tag/v2.9.2
freebayes: Release 1.3.5. https://github.com/freebayes/freebayes/releases/tag/v1.3.5
VarTrix: Release 1.1.22. https://github.com/10XGenomics/vartrix/releases/tag/v1.1.22
troublet (souporcell): Release 2.0. https://github.com/wheaton5/souporcell/releases/tag/2.0
Rust (souporcell): Release 1.55.0. https://www.rust-lang.org/

Design of scATAC-seq specific mixture mode

The generative model in scAVENGERS aims to estimate the probability of each cell being annotated to a specific donor. Unlike scRNA-seq, the read coverage of each variant in scATAC-seq has a value of 0, 1 or 2 in the diploid genome. The probability of scATAC-seq read being captured is determined by the sampling of accessible chromatin regions with attachment of cell barcodes and the sampling of PCR amplified DNA sequences, which can be modeled by hypergeometric distribution and binomial distribution, respectively. Descriptions for each parameter are summarized in Table 1.

Complete log-likelihood. The objective is to maximize the expected value of the complete log likelihood that is given below:

$$\log \mathcal{L} = \log\left(\prod_{c\in\{1,\ldots C\}} \mathcal{L}_{c,k_c}\right) = \sum_{c\in\{1,\ldots,C\}} \log \mathcal{L}_{c,k_c}$$

Because k_c is unknown, we will use the expected complete log-likelihood as follows:

$$E_{k_c|\mathbf{a}_c,\mathbf{r}_c,\boldsymbol{\alpha}}[\log\mathcal{L}] = \sum_{c\in\{1,\ldots,C\}}\sum_{k\in\{1,\ldots,K\}} P(k_c = k|\mathbf{a}_c,\ \mathbf{r}_c, \boldsymbol{\alpha}) \log(\mathcal{L}_{c,k})$$

where,

$$\mathcal{L}_{c,k} = P(\mathbf{a}_c, \mathbf{r}_c, k_c = k|\boldsymbol{\alpha}) = P(k_c = k|\boldsymbol{\alpha})P(\mathbf{a}_c, \mathbf{r}_c\ |k_c = k, \boldsymbol{\alpha})$$
$$= P(k_c = k|\boldsymbol{\alpha}) \prod_{l\in\{1,\ldots,L\}} P(a_{c,l}, r_{c,l}|k_c = k, \boldsymbol{\alpha})$$

which is the likelihood of each cell and donor being annotated.

$P(k_c = k|\mathbf{a}_c,\ \mathbf{r}_c, \boldsymbol{\alpha})$ is the probability of a specific cell barcode being assigned to certain donor. In detail, it is a probability such that k_c is assigned to donor k, given observed reference and alternative allele counts $\mathbf{a}_c$, $\mathbf{r}_c$ and the actual alternative allele counts for every donor and locus $\boldsymbol{\alpha}$. If the genotype is known, $P(k_c = k|\mathbf{a}_c,\ \mathbf{r}_c, \boldsymbol{\alpha})$ has a value of 0 or 1, which labels the genotype for each cell. We did not model $P(a_{c,l}, r_{c,l}|k_c = k, \boldsymbol{\alpha})$ in case when the reference and alternative read counts are both zeros. Instead, we coerced $P(a_{c,l} = 0, r_{c,l} = 0|k_c = k, \boldsymbol{\alpha})$ into 1 since most of variant loci from scATAC-seq data are covered by zero read count, which reduces the computational cost.

Likelihood for each cell, loci and cluster. The mixture model describes observed reference and alternative read counts from the actual alternative read counts. The model is divided into two parts as presented below, and each part is intended to model the PCR amplification and the attachment of reads to the barcode sequences, respectively.

From the DNA sequences of accessible chromatin regions, whose actual alternative read counts are denoted as $\alpha_{k,l}$, a fraction of reads is selected by the attachment of cell barcodes. Because the counts of these selected alleles are not observable due to the following sampling process after PCR amplification, we denoted them 'latent allele count.' The latent count of reference and alternative alleles for these reads are denoted as $r'_{c,l}$ and $a'_{c,l}$ each. The sampling of reads after PCR amplification stage then leads to observable reference and alternative allele read counts, denoted as $r_{c,l}$ and $a_{c,l}$ respectively.

$$P(a_{c,l}, r_{c,l}|k_c = k, \boldsymbol{\alpha})$$
$$= \sum_{a'_{c,l}, r'_{c,l},\ \text{where } a'_{c,l}+r'_{c,l}\le n} P(a_{c,l}, r_{c,l}|a'_{c,l}, r'_{c,l}, k_c = k, \alpha_{k,l})$$
$$P(a'_{c,l}, r'_{c,l}|k_c = k, \alpha_{k,l})$$

Attachment of reads to cell barcode sequences. The model below describes the latent reference and alternative read counts, given the genotype. The latent reads are originated from the template DNA sequences in each cell, not from the duplicate DNA sequences from PCR amplification. To note, the notation of random variables is different from how they are usually denoted. The detailed explanation is in Supplementary Information.

$$P(a'_{c,l}, r'_{c,l}|k_c = k, \alpha_{k,l}) = P(a'_{c,l}, a'_{c,l} + r'_{c,l}|k_c = k, \alpha_{k,l})$$
$$= P(a'_{c,l}|a'_{c,l} + r'_{c,l}, k_c = k, \alpha_{k,l})P(a'_{c,l} + r'_{c,l}|k_c = k, \alpha_{k,l})$$

$P(a'_{c,l}|a'_{c,l} + r'_{c,l}, k_c = k, \alpha_{k,l})$ follows a hypergeometric distribution since the sampling of template DNA sequences by cell barcode attachment is done without replacement.

$$P(a'_{c,l}|a'_{c,l} + r'_{c,l}, k_c = k, \alpha_{k,l}) = \frac{\binom{\alpha_{k,l}}{a'_{c,l}}\binom{n-\alpha_{k,l}}{r'_{c,l}}}{\binom{n}{a'_{c,l}+r'_{c,l}}}$$

$P(a'_{c,l} + r'_{c,l}|k_c = k, \alpha_{k,l})$ is expressed as a binomial distribution, which models the process of taking reads from a cell.

$$P(a'_{c,l} + r'_{c,l}|k_c = k, \alpha_{k,l}) = \binom{n}{a'_{c,l}+r'_{c,l}} p^{a'_{c,l}+r'_{c,l}}(1-p)^{n-a'_{c,l}-r'_{c,l}}$$

Since most variants are covered by zero read count, $P(a'_{c,l} + r'_{c,l} = 0|k_c = k, \alpha_{k,l})$ approximates to $P(a_{c,l} + r_{c,l} = 0|k_c = k, \alpha_{k,l})$ well, which is utilized to obtain p.

$$p = 1 - P(a'_{c,l} + r'_{c,l} = 0|k_c = k, \alpha_{k,l})^{\frac{1}{n}}$$
$$\approx 1 - P(a_{c,l} + r_{c,l} = 0|k_c = k, \alpha_{k,l})^{\frac{1}{n}}$$

PCR amplification. From latent reference and alternative read counts, the probability of occurrence of observed ref-

Table 1. Lists of parameters used in scAVENGERS probability model

symbol	Aggregated symbol	Description	Additional description
n	-	ploidy	$n = 2$ for typical human cells. This is defined by users.
C	-	number of the cells	
c	-	label for an arbitrary cell	A single cell is represented as $c \in [1, C]$.
K	-	number of the donors	This is defined by users.
k, κ		label for an arbitrary donor	A single donor is represented as $k \in [1, K]$ or $\kappa \in [1, K]$.
k_c	-	donor label corresponding to a certain cell c	scAVENGERS estimates this label.
L	-	number of loci	
l	-	label for an arbitrary locus	A single locus is represented as $l \in [1, L]$
$a_{c,l}$, $r_{c,l}$	$\mathbf{a}_c = [a_{c,l}]_{l\in[1,L]}$ $\mathbf{r}_c = [r_{c,l}]_{l\in[1,L]}$	number of observed reference and alternative alleles for a specific cell c and locus l	Number of reference and alternative alleles for reads selected after PCR amplification. A vector of every observed reference and alternative allele counts for a cell are represented as $\mathbf{a}_c$ and $\mathbf{r}_c$. This is given in variant-barcode matrix as an input.
$a'_{c,l}$, $r'_{c,l}$	$\mathbf{a}'_c = [a'_{c,l}]_{l\in[1,L]}$ $\mathbf{r}'_c = [r'_{c,l}]_{l\in[1,L]}$	number of latent reference and alternative alleles for a specific cell c and locus l	Number of alternative alleles for reads after selection by barcode attachment.
$\alpha_{k,l}$	$\boldsymbol{\alpha} = [\alpha_{k,l}]_{k\in[1,K],l\in[1,L]}$	number of actual alternative alleles for a specific cell c and locus l	Number of alternative alleles for reads in a cell. A matrix of every actual alternative alleles are denoted $\boldsymbol{\alpha}$. $\boldsymbol{\alpha}$ in the t'th iteration in the EM algorithm is denoted $\boldsymbol{\alpha}^{(t)}$. The generative model aims to find maximum likelihood estimate of this value.
ϵ	-	correction factor	Correction factor to prevent taking logs of zeros. This is defined by users.

erence and alternative read counts is calculated as given below. A detailed explanation of how random variables are denoted is in Supplementary Information.

$$\begin{aligned}
&P(a_{c,l}, r_{c,l}|a'_{c,l}, r'_{c,l}, k_c = k, \alpha_{k,l}) \\
&= P(a_{c,l}, a_{c,l} + r_{c,l}|a'_{c,l}, r'_{c,l}, k_c = k, \alpha_{k,l}) \\
&= P(a_{c,l}|a_{c,l} + r_{c,l}, a'_{c,l}, r'_{c,l}, k_c = k, \alpha_{k,l}) \\
&= P(a_{c,l} + r_{c,l}|a'_{c,l}, r'_{c,l}, k_c = k, \alpha_{k,l})
\end{aligned}$$

$P(a_{c,l}|a_{c,l} + r_{c,l}, a'_{c,l}, r'_{c,l}, k_c = k, \alpha_{k,l})$ follows a binomial distribution, under the assumption that the PCR amplification is sufficient to regard the process of taking reference and alternative reads as sampling with replacement.

$$\begin{aligned}
&P(a_{c,l}|a_{c,l} + r_{c,l}, a'_{c,l}, r'_{c,l}, k_c = k, \alpha_{k,l}) \\
&= \binom{a_{c,l} + r_{c,l}}{a_{c,l}} \left(\frac{a'_{c,l}}{a'_{c,l} + r'_{c,l}}\right)^{a_{c,l}} \left(\frac{r'_{c,l}}{a'_{c,l} + r'_{c,l}}\right)^{r_{c,l}}
\end{aligned}$$

$P(a_{c,l} + r_{c,l}|a'_{c,l}, r'_{c,l}, k_c = k, \alpha_{k,l})$ is calculated from the observed aligned read counts given locus l in cell c under the assumption that it does not depend on the cell, locus, and genotypes.

Correction factor. Because the Expectation-Maximization (EM) algorithm used in scAVENGERS conveys the process of selecting actual alternative allele counts α to maximize the expected value of total log-likelihood, a probability value of zero inevitably appears during the calculation of likelihood for suboptimal α. The correction factor ϵ defined by users prevents taking logs of zeros.

$$(1 - \epsilon)\, P(a_{c,l}, r_{c,l}|k_c = k, \boldsymbol{\alpha}) + \epsilon$$

Deterministic anneal expectation-maximization algorithm for parameter optimization

Expectation step. The probability of each cell being assigned to a specific donor is defined as follows.

$$\begin{aligned}
P\left(k_c = k|\mathbf{a}_c, \mathbf{r}_c, \boldsymbol{\alpha}^{(t)}\right) &= \frac{P(\mathbf{a}_c, \mathbf{r}_c|k_c = k, \boldsymbol{\alpha}^{(t)})P(k_c = k|\boldsymbol{\alpha}^{(t)})}{P(\mathbf{a}_c, \mathbf{r}_c|\boldsymbol{\alpha}^{(t)})} \\
&= \frac{P(\mathbf{a}_c, \mathbf{r}_c, k_c = k|\boldsymbol{\alpha}^{(t)})}{\sum_{\kappa\in\{1,\ \ldots,K\}} P(\mathbf{a}_c, \mathbf{r}_c, k_c = \kappa|\boldsymbol{\alpha}^{(t)})} \\
&\equiv \frac{e^{\frac{\log \mathcal{L}^{(t)}_{c,k}}{T}}}{\sum_{\kappa\in\{1,\ldots,K\}} e^{\frac{\log \mathcal{L}^{(t)}_{c,\kappa}}{T}}}
\end{aligned}$$

The deterministic annealing variant of the Expectation Maximization (EM) algorithm (16) uses the temperature parameter to adjust the probability of donor assignment over multiple iterations (step = t). The temperature parameter T is initialized to the average value of the sum of total read counts in a cell.

Given the above probability, the expected log-likelihood is calculated as given below:

$$
\begin{aligned}
&Q(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(t)}) \equiv E_{\mathbf{k}|\mathbf{A},\mathbf{R},\boldsymbol{\alpha}^{(t)}}\left[\log(\mathcal{L})\right] \\
&= \sum_{c\in\{1,\ldots,C\}} \sum_{k\in\{1,\ldots,K\}} P\left(k_c = k|\mathbf{a}_c,\ \mathbf{r}_c, \boldsymbol{\alpha}^{(t)}\right) \log\left(\mathcal{L}_{c,k}\right) \\
&= \sum_{c\in\{1,\ldots,C\}} \sum_{k\in\{1,\ldots,K\}} P\left(k_c = k|\mathbf{a}_c,\ \mathbf{r}_c, \boldsymbol{\alpha}^{(t)}\right) \log\left(P(k|\boldsymbol{\alpha}) \prod_{l\in\{1,\ldots,L\}} P(a_{c,l}, r_{c,l}|k, \boldsymbol{\alpha})\right) \\
&= \sum_{c\in\{1,\ldots,C\}} \sum_{k\in\{1,\ldots,K\}} P\left(k_c = k|\mathbf{a}_c,\ \mathbf{r}_c, \boldsymbol{\alpha}^{(t)}\right) P(k|\boldsymbol{\alpha}) \\
&\quad + \sum_{c\in\{1,\ldots,C\}} \sum_{k\in\{1,\ldots,K\}} \sum_{l\in\{1,\ldots,L\}} P\left(k_c = k|\mathbf{a}_c,\ \mathbf{r}_c, \boldsymbol{\alpha}^{(t)}\right) \log P(a_{c,l}, r_{c,l}|k, \boldsymbol{\alpha})
\end{aligned}
$$

Maximization step. For the maximization of the expected log-likelihood, the term where $\alpha_{k,l}$ is involved is maximized.

$$
\begin{aligned}
&\boldsymbol{\alpha}^{(t+1)} = \mathrm{argmax}_{\alpha} Q(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(t)}) \\
&= \mathrm{argmax}_{\boldsymbol{\alpha}} \sum_{c\in\{1,\ldots,C\}} \sum_{k\in\{1,\ldots,K\}} \sum_{l\in\{1,\ldots,L\}} P\left(k_c = k|\mathbf{a}_c,\ \mathbf{r}_c, \boldsymbol{\alpha}^{(t)}\right) \log P(a_{c,l}, r_{c,l}|k, \boldsymbol{\alpha}) \\
&= \mathrm{argmax}_{\boldsymbol{\alpha}} \sum_{k\in\{1,\ldots,K\}} \sum_{l\in\{1,\ldots,L\}} \sum_{c\in\{1,\ldots,C\}} P\left(k_c = k|\mathbf{a}_c,\ \mathbf{r}_c, \boldsymbol{\alpha}^{(t)}\right) \log P(a_{c,l}, r_{c,l}|k, \boldsymbol{\alpha}) \\
&\alpha_{k,l}^{(t+1)} = \mathrm{argmax}_{\alpha_{k,l}} \left(\sum_{c\in\{1,\ldots,C\}} P\left(k_c = k|\mathbf{a}_c,\ \mathbf{r}_c, \boldsymbol{\alpha}^{(t)}\right) \log P(a_{c,l}, r_{c,l}|k, \alpha_{k,l}) \right)
\end{aligned}
$$

In the maximization step, an appropriate $\alpha_{k,l}^{(t+1)}$ is chosen among integers from 0 to n, and then the process is repeated from the E step until convergence. If the difference between $Q(\boldsymbol{\alpha}^{(t+1)}|\boldsymbol{\alpha}^{(t+1)})$ and $Q(\boldsymbol{\alpha}^{(t)}|\boldsymbol{\alpha}^{(t)})$ is not bigger than the stop criterion, which is 0.1 by default, the algorithm declares convergence. After convergence, the iteration goes on with the halved temperature until 1. After convergence at the final temperature 1, the entire iteration stops. If the iteration in the EM algorithm does not converge, it is forced to stop when the number of iterations exceeds a certain stop criterion for each temperature step.

Donor-cluster matching for the evaluation procedure

The clustering by mixture model assigns the given cell barcodes into *n* clusters according to the similarity of the set of their variants. However, the relationship between donor and cluster cannot be specified as no reference genotypes are given. Therefore, the accuracy was measured after rematching the donors and clusters. In the case of $n = 5$, A-E donors and doublet were sequentially assigned for six clusters (5+ doublet), and then a confusion matrix was generated. By using the Python numpy package's argmax function, the cluster that matches the most for each donor is assigned. When the argmax values overlap, the combination of donor and cluster that maximizes the macro-average F1-score was assigned. As we generated a synthetic pooled scATAC-seq dataset by merging multiple independent scATAC-seq results, each cell barcode was assigned to the corresponding donor. If the cell barcode was assigned to multiple donors, we defined them as a doublet. For all the classified cell barcodes, we evaluated the performance based on whether the classified barcodes match the actual donor (precision) and whether the barcodes from a certain donor are correctly classified (recall). Python scikit-learn package was used for the evaluation.

RESULTS

Workflow of scAVENGERS

The core design principle of scAVENGERS is mainly composed of four steps: read alignment, variant calling, genotype clustering, and doublet assignment for final demultiplexing (Figure 1A). The overall architecture is similar to the read variant-based demultiplexing methods for scRNA-seq, such as souporcell. However, different software and mixture model were used according to the characteristics of scATAC-seq reads in alignment and clustering procedures.

In the case of alignment, Burrows-Wheeler Aligner (BWA) (17) with MEM option was used for efficient mapping of ATAC-seq reads to achieve proper variant calling and ultimately more accurate clustering. Followed by the changes in the read mapping method, variant caller software was upgraded to extract proper individual-specific variants to distinguish donors in the pooled sample. In the existing method (8), freebayes was used while scAVENGERS incorporated strelka2 (15), which showed superior speed and accuracy in the benchmark (18). The option for post-filtering variants based on the quality score provided by strelka2 was also implemented, which was not offered in freebayes-based pipelines.

The mixture model was also modified for the scATAC-seq reads (Figure 1B). Unlike scRNA-seq reads, which have multiple copies, scATAC-seq reads have limited copies as they are sampled from the diploid genome. The barcoding and PCR amplification processes under the limited condition of initial genetic materials were modeled by combining hypergeometric and binomial distribution. The EM algorithm was used to estimate the likelihood of the cell barcodes for being assigned to each donor-specific cluster. 'Troublet' of the souporcell pipeline was applied to the final process of the clustering, which yields the final donor assignment (including the unassigned cells and doublets) of each barcode as a final output.

Benchmark on human PBMC scATAC-seq datasets

For the performance evaluation, we used synthetic multiplexed human PBMC scATAC-seq datasets for the benchmark. A synthetic set with source-labeled barcodes was generated by merging individually sequenced scATAC-seq BAM files from multiple donors. The barcodes overlapping between donors were used as synthetic doublets (Figure 1C). Two datasets were created by applying the synthetic procedure to the *in-house* generated dataset consisting of five donors and the published dataset consisting of ten donors (12).

For the quantitative comparison, donor-wise precision/recall, doublet detection accuracy, and a fraction of donor assigned barcodes were measured. In the *in-house* PBMC datasets, scAVENGERS achieved an average of 0.997 precision and 0.986 recall (macro-average), while souporcell achieved an average precision of 0.702 and recall of 0.739 (Figure 2A left). In doublet detection, precision

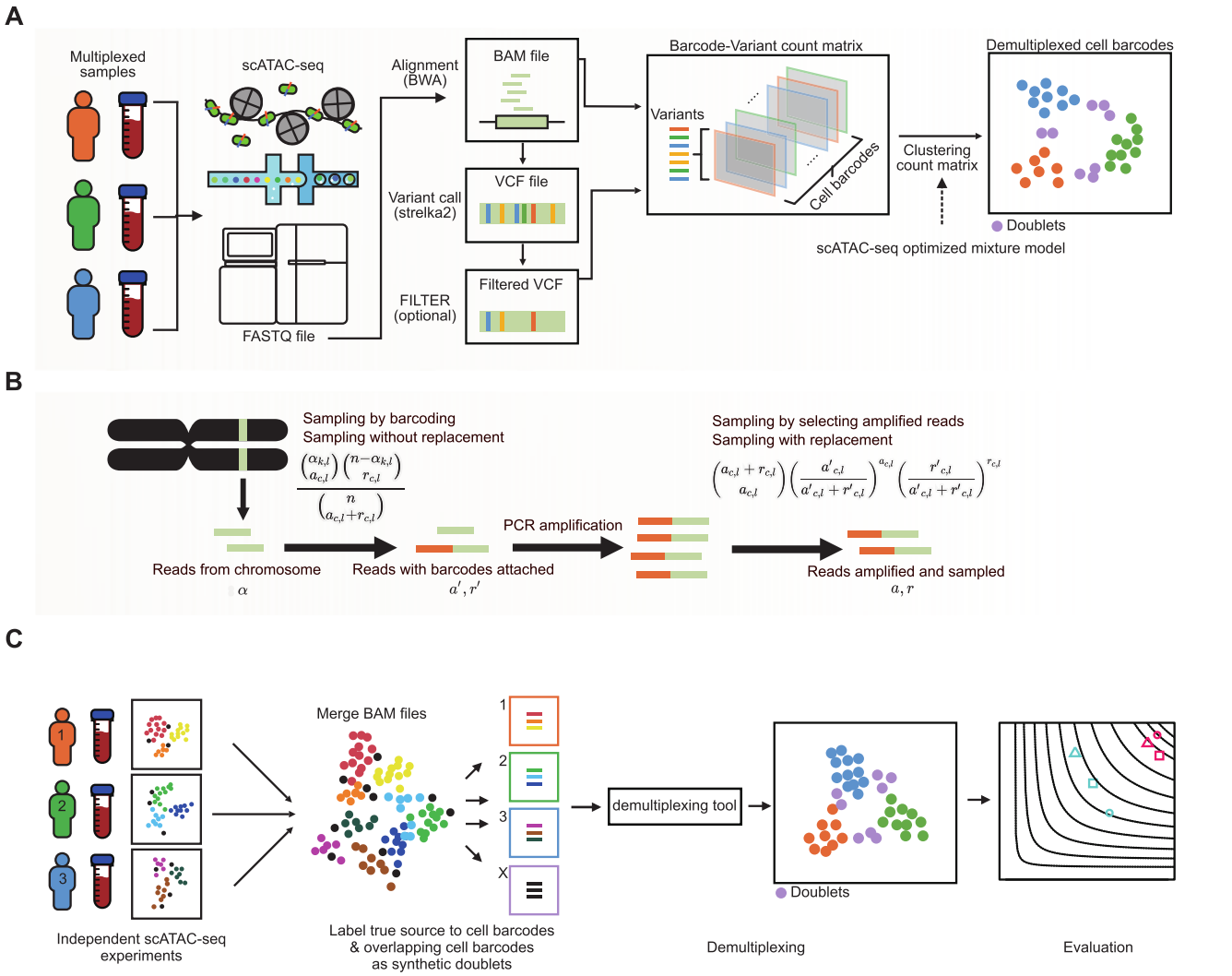

Figure 1. (**A**) Overall workflow of scAVENGERS pipeline. (**B**) A schematic of probability model for scATAC-seq read generation. (**C**) Schematic showing benchmark process using synthetic pooled datasets.

of 0.105 and 0.482 and recall of 0.451 and 0.825 were observed, respectively, by souporcell and scAVENGERS, which shows more than two-fold performance improvement in scAVENGERS (Figure 2A middle). The ratio of unassigned barcodes to a specific donor owing to the ambiguous likelihood values was also significantly reduced from 7% to 1% (Figure 2A right).

This trend was also reproduced in the benchmark conducted with the public dataset. Overall accuracy (precision: 0.732–0.985 and recall: 0.57–0.989, macro-average), doublet detection accuracy (precision: 0.095–0.753 and recall: 0.607–0.731), and genotype assigned barcode ratio (45–99%) all increased in the scAVENGERS processed result (Figure 2B). Principal component analysis (PCA) of the log-likelihood profile between barcodes and donors was also conducted to intuitively demonstrate the demultiplexing performance. On the PCA plot of the *in-house* dataset, the pooled samples showed clear, distinguished clusters where the doublets are located between the clusters (Figure 2C). The PCA plot of the public dataset was not well differentiated as the PCA of the *in-house* dataset, but the distinction between samples could be observed (Figure 2D).

We further confirmed the performance of scAVENGERS using *in-house* PBMC datasets in comparison to demuxlet (6), a method with reference-genotypes (Supplementary Figure 1A). We found that scAVENGERS showed a comparable or slightly better performance compared to demuxlet in terms of donor-wise precision/recall, doublet detection accuracy, and a fraction of donor assigned barcodes. We also conducted benchmark by taking 500 cell barcodes from each donor of *in-house* PBMC datasets to compare another scRNA-seq demultiplexing method without reference genotypes, scSplit (10), due to its memory-inefficient storage of allele counts. Again, scAVENGERS outperformed scSplit in terms of precision, recall, and doublet detection rate (Supplementary Figure 1B). Our results also indicated that the low number of cell barcodes itself did not degrade the performance of scAVENGERS much by taking only 500 cell barcodes from each donor. The benchmark results with different methods and conditions strongly

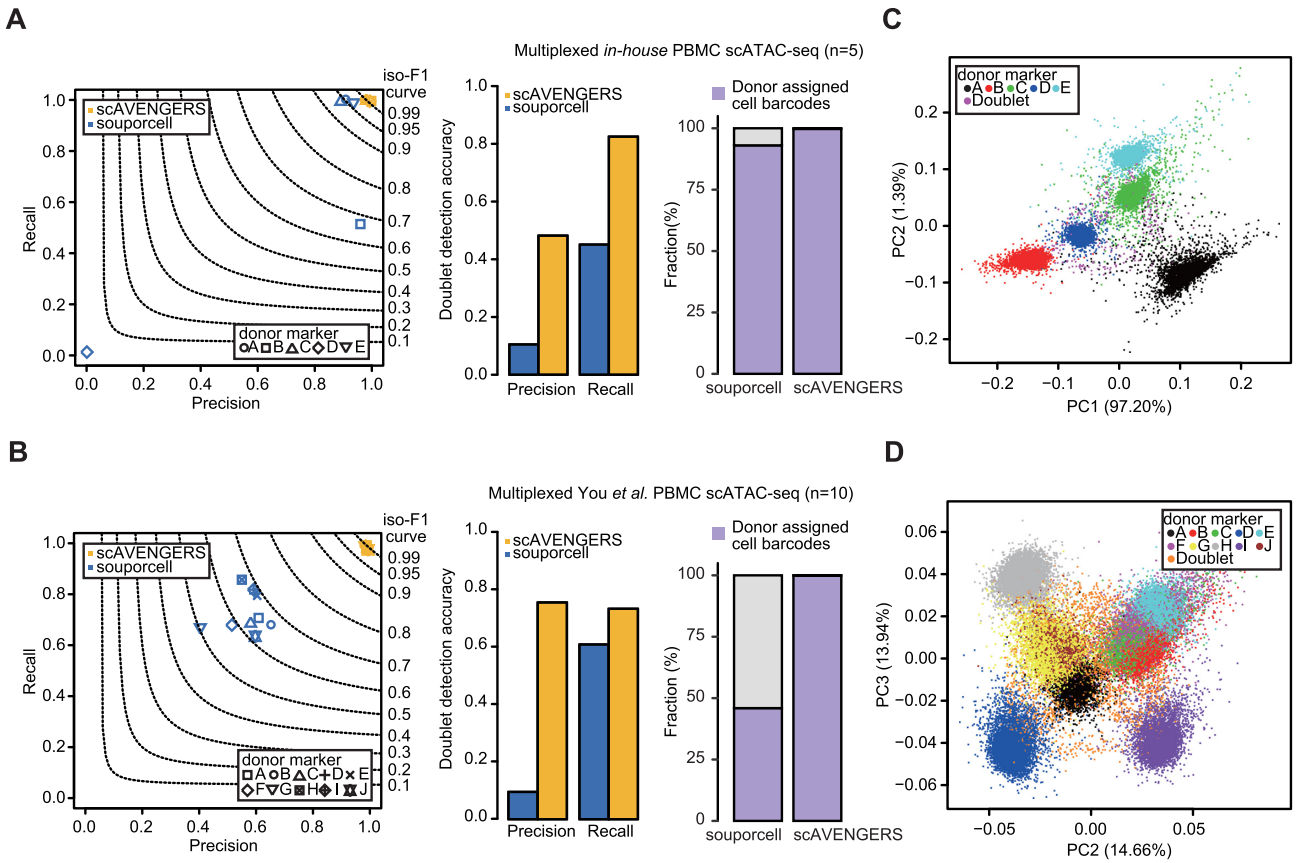

Figure 2. (A and B) Benchmark results of the *in-house* dataset (**A**) and You *et al.*'s dataset (**B**) showing sample demultiplexing accuracy (left), doublet detection accuracy (middle), and the fraction of clusters assigned barcodes (right). The color indicates the results from the scAVENGERS (yellow) and souporcell (blue). The purple portion of the stacked bar chart indicates cells assigned to a certain donor. (**C** and **D**) Plots showing PCA applied to the barcodes' log-likelihood profile for each donor after processing benchmark datasets (C: *in-house* and D: You *et al.*) by scAVENGERS. Color markers indicate the true labels (source donor and synthetic doublet) of the dataset's barcodes.

supported the performance improvement and robustness of scAVENGERS.

Effect of variant discovery and quality on performance

Given the outstanding performance of scAVENGERS, we investigated how the improved factors affected the performance. We first examined the effect of the read mapping and variant calling steps. In *in-house* as well as public datasets, more variants were called when the alignment software was switched from minimap2 to BWA if the variant caller was fixed to freebayes (Figures 3A and B for left). Change of variant caller to strelka2 yielded more than 1.5-fold of variants compared to freebayes in both datasets. As the donor assignment for each cell is estimated with the variant combinations obtained from individual cells, the sufficient discovery of the variants through the improvement of alignment and variant caller may have brought an increase in the performance of scAVENGERS.

Strelka2 provides the variant quality information by 'PASS' for high-quality variants and 'lowGQX' for low-quality variants in the FILTER fields. A total of 49% and 47% of the variants were classified as PASS in *in-house* and public datasets, respectively (Figures 3A and B for right). To check the relationship between performance and variant quality, performance evaluation was conducted by dividing the variants into two sets based on quality. Interestingly, the variant set only consisting of the lowGQX showed comparable performance to the entire dataset while the PASS-only set showed performance degradation (Figures 4A–C). This result is presumed to occur as the germline variant caller was applied to the pooled sample. In general, germline callers assume variants on the diploid genome and make calls from the supporting reads. However, the pooled data violates this assumption. A unique variant that can distinguish a specific individual from the others is rather likely to be considered as low quality as the allele frequency is diluted by the reads from the other donors.

To test this hypothesis, the pooled sample and each donor's variant call results were compared. As a variant becomes more common among the donors (denote as common in *n*-individuals), it was more likely to be considered as PASS in the pooled sample (Figure 4D). In contrast, the majority of individual-specific variants were assigned as lowGQX. For this reason, it is estimated that the lowGQX-only set has better discrimination ability than the PASS-only set. We further examined whether donor-specific variants can be distinguished from true low-quality variants based on quality scores and read depth. We found that the distributions of genotype quality score (GQ score) and

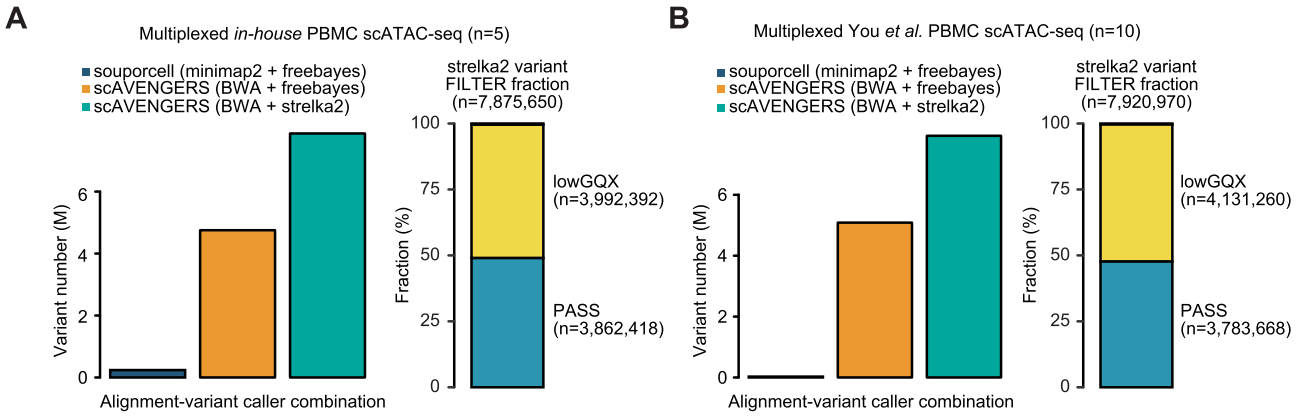

Figure 3. (**A** and **B**) Bar plot and stacked bar plot pairs showing the number of called variants obtained from different combinations of alignment and variant caller (color markers) and the fraction of variant quality in strelka2-called result in two benchmark datasets (A: *in-house* and B: You *et al.*).

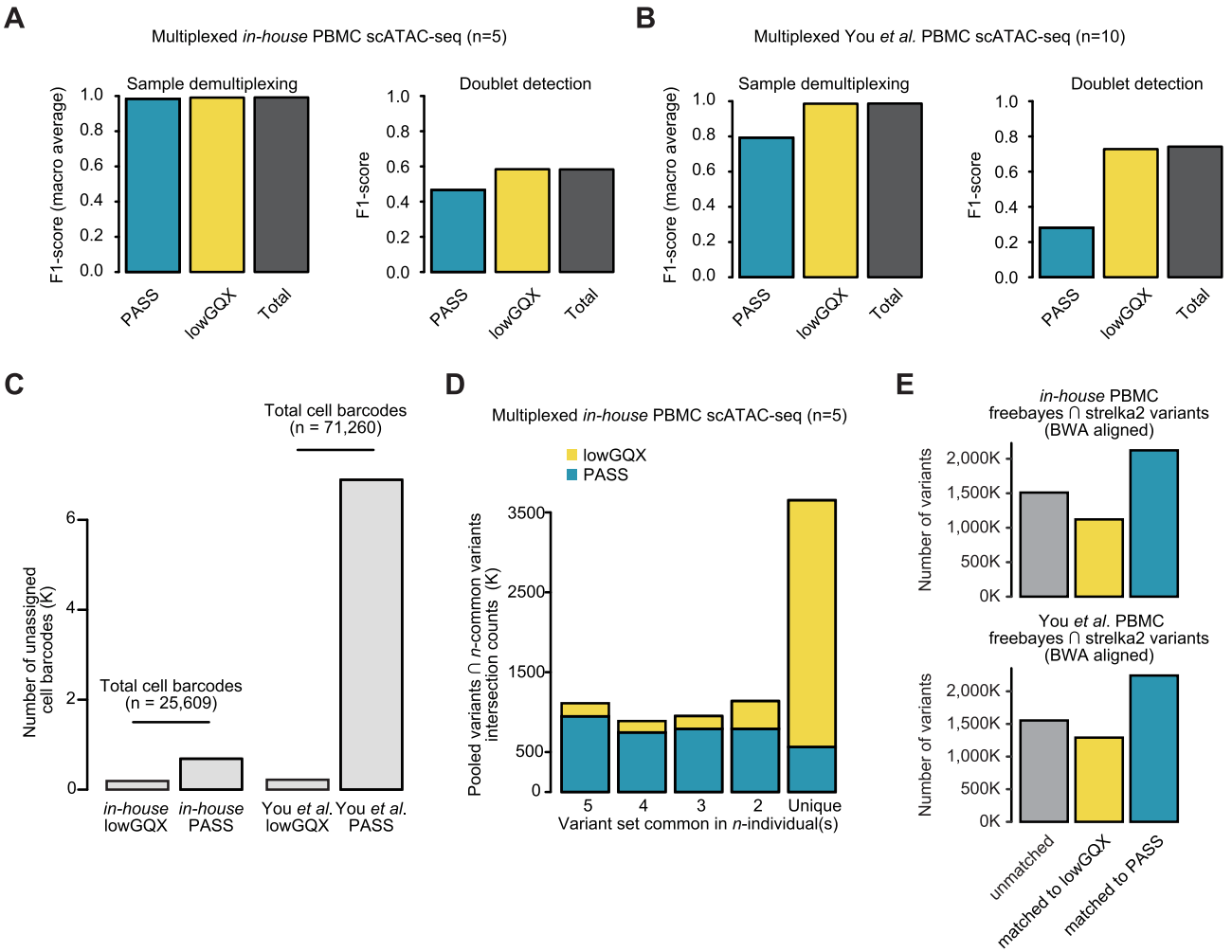

Figure 4. (**A** and **B**) Performance evaluation of PASS only variants (jade green) set and lowGQX-only variants set (yellow) compared to the total dataset (orange). Macro-average was used to summarize multiple donors' results (A: *in-house* and B: You *et al.*). (**C**) Barplot showing the number of unassigned cell barcodes for diverse variant input sets. (**D**) Barplot showing the number of PASS (jade green) and lowGQX (yellow) variants in the intersection set between the pooled and *n*-common variant sets. (**E**) Bar plots showing quality (Strelka2's) distribution of freebayes-called variants reproduced in the strelka2 results (top: *in-house* and bottom: You *et al.*). Unmatch: the number of variants uniquely identified by freebays, matched to lowGQX: the number of variants identified by freebays and annotated as lowGQX by Strelka2, and matched to PASS: the number of variants identified by freebays and annotated as PASS by Strelka2.

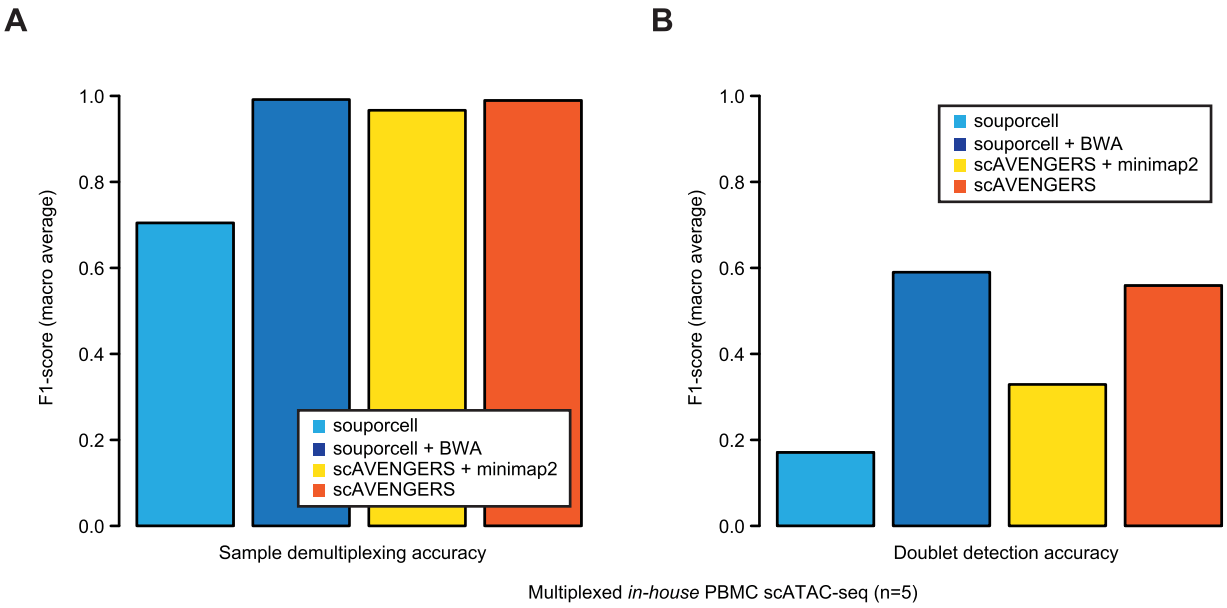

Figure 5. Mixture model performance evaluation result using the *in-house* dataset. (**A**) Sample demultiplexing accuracy. (**B**) Doublet detection accuracy. Macro-average of *F*1-score was used. The color marker indicates the original (light blue and orange) and variant set exchanged clustering result (dark blue and red.

read depth for donor-specific variants were not sufficient to clearly distinguish donor-specific variants from low-quality variants (Supplementary Figure 2).

Analysis of the variant quality distribution also provides an explanation for the performance difference between freebayes and strelka2. Among reproduced variants between two callers, we checked the strelka2-assigned quality (Figure 4E). Approximately 68% of freebayes variants were reproduced in strelka2 variants, and ∼64% of them were assigned to the PASS quality. Enrichment of freebayes results in the PASS set shows that freebayes-called variants are biased towards common variants with less distinguishing ability, which explains the lower performance when the freebayes is used.

Discovering unique variants is thought to be less problematic in the demultiplexing of scRNA-seq as RNA molecules exist in multiple copies. By contrast, this issue is likely to be intensified in scATAC-seq, where the amount of genetic material is limited, and eventually decreases the coverage for a specific variant. These results support the requirements for scATAC-seq specific pipeline development, again.

The robust performance of scAVENGERS with a limited number of variants

Next, we also evaluated the mixture model of scAVENGERS in terms of performance robustness. Owing to the alignment and variant calling software difference between souporcell and scAVENGERS, we measured the performance variability according to the variant call results by exchanging the set of freebayes-called variants of souporcell and scAVENGERS in the *in-house* dataset. When a sufficient number of variants obtained from the BWA-using pipeline was provided, souporcell's original mixture model for scRNA-seq also showed enhanced performance similar to scAVENGERS in terms of the donor assignment accuracy and doublet detection accuracy (darkblue and red bars in Figures 5A and B). In contrast, scAVENGERS 's scATAC-seq optimized model showed robustness as the performance was not significantly degraded even when very few variants from the minimap2-freebayes were given (yellow and red bars in Figures 5A and B). Thus, it is thought that both models produce similar results in an optimal situation where there are sufficient variants, but the scAVENGERS's optimized model performs better when the number of variants is insufficient.

Other factors affecting the performance of scAVENGERS

Although scAVENGERS outperformed scRNA-seq demultiplexing tools for every benchmarking result and showed robust performance with a limited number of variants, performance variations between public and *in-house* PBMC datasets remain to be explained. To identify factors that drive lower accuracy in public PBMC datasets, we compared multiple properties including the number of multiplexed samples, the number of cells per donor, and the number of identified variants per donor of both public and *in-house* PBMC scATAC-seq results.

We first excluded the factor of a number of variants since scAVENGERS is highly robust against the number of identified variants (Figure 5) and we identified a similar number of variants between public and *in-house* PBMC scATAC-seq datasets. Next, to test the effect of the number of multiplexed donors, we measured the demultiplexing performance by increasing the number of multiplexed samples from 3 to 12 donors with the human prefrontal cortex tissue scATAC-seq dataset. We excluded the factor of a number of multiplexed donors since there is no

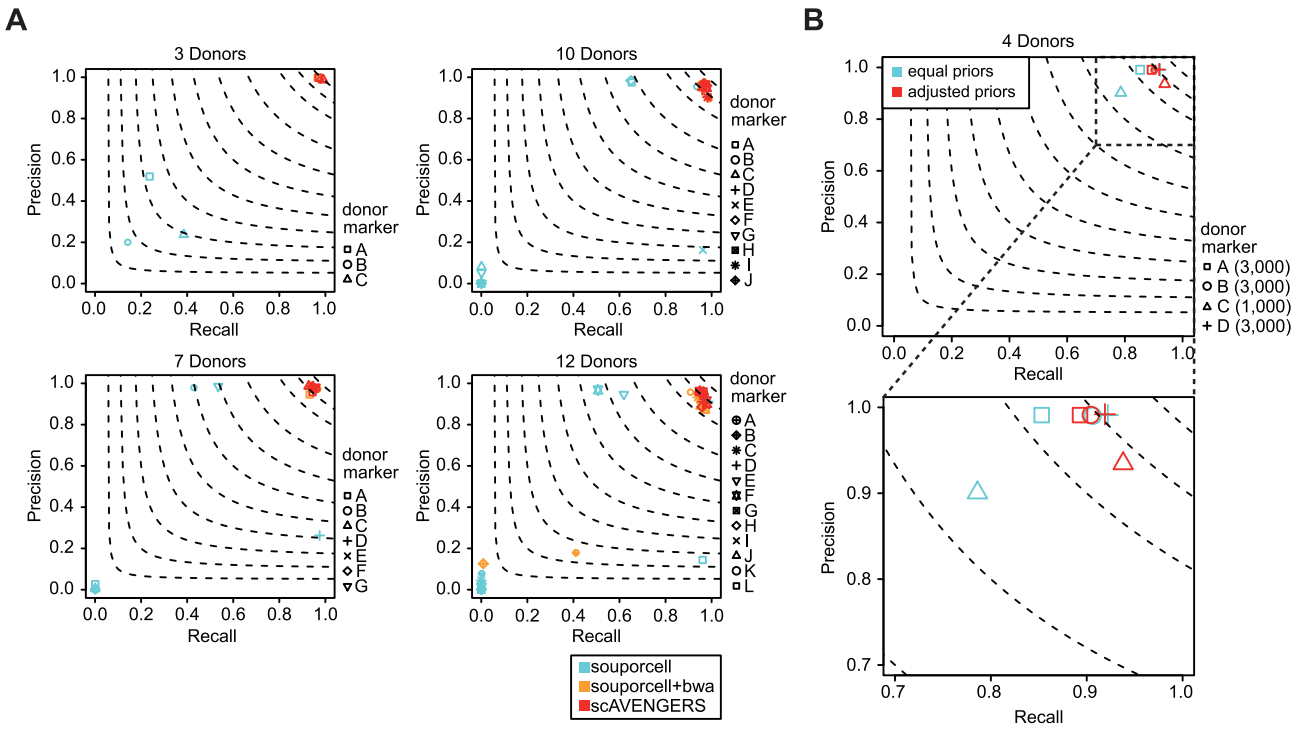

Figure 6. (**A**) Benchmark results with different numbers of multiplexed donors from human prefrontal cortex scATAC-seq. (**B**) Performance evaluation with an unbalanced pooled sample with equal priors (blue) and with adjusted priors (red). The number of cell barcodes for each donor is shown together in each donor marker.

critical performance degradation in scAVENGERS unlike souporcell (Figure 6A). Lastly, we tested the effect of the number of cells per donor since the public PBMC dataset showed more deviation in the number of cells per donor compared to those in the *in-house* dataset (Supplementary Table 2). To further examine the association between the imbalanced number of cells per donor and the lowered demultiplexing performance, we made a synthetic mixture by sampling a certain number of cell barcodes per donor with human prefrontal cortex tissue scATAC-seq. We took 1000 cell barcodes from one donor, and 3000 cell barcodes from three other donors. When we applied default parameters, we observed slightly lower macro-average precision and recall with 0.969 and 0.867, respectively (Figure 6B). Such performance degradation was mainly caused by the donor with 1000 cell barcodes, showing the lowest accuracy with 0.902 and 0.786 precision and recall, respectively.

The performance degradation of the imbalanced multiplexed dataset is supposed to be associated with the parameter of priors. In the default option, scAVENGERS applies equal priors assuming that cell numbers are similar across donors. However, when we adjusted priors based on the proportion of the number of cells for each donor, scAVENGERS achieved the precision of 0.936 and the recall of 0.938 for the donor with 1000 cell barcodes (Figure 6B). Thus, we can conclude that the higher deviation of the number of cells for each donor will degrade the demultiplexing performance, but scAVENGERS can rescue the performance degradation through the estimation of the proper priors.

DISCUSSION

In this paper, we describe scAVENGERS, a new computational pipeline that can perform genotype-based deconvolution of multiplexed scATAC-seq results. Pooled sequencing and demultiplexing of multiple samples reduce the total cost and batch effect, which gradually resolves issues related to the current single-cell sequencing technology. This advantage is particularly evident in research designs that require multiple cohorts, where it is difficult to sequence all samples. The results of this study provided that demultiplexing can be used appropriately in scATAC-seq by identifying and solving issues in the existing pipelines. The improvements achieved in this study can also be fed back to the existing scRNA-seq demultiplexing techniques without special modification of conventional pipelines.

Careful consideration of the variant calling process showed considerable impact on the demultiplexing performance. We hypothesized that applying germline variant callers in the pooled sample induces variant quality-related bias, which was verified by examining the quality of the variants in demultiplexing performance. It was confirmed that variants with differentiation ability tend to be judged as low quality, and some individual specific variants with low coverage may be lost in this process. By applying a proper variant caller and quality filter, procedures to acquire a proper variant set were established.

The use of an appropriate statistical model was another axis of performance improvement in this study. Unlike RNA-seq, the read generation situation of ATAC-seq, which inevitably has limited coverage per each variant, was

well-simulated. Although scRNA-seq and scATAC-seq are most frequently used, single-cell methods for exploring various genomic features are being developed, and it is challenging to provide proper modeling specific to each genomic feature. Thus, it is necessary to create a model that allows more flexible assumptions about the prior distribution and a clustering algorithm should be built based on it.

For future development, detection of ambient DNA from multiplexed scATAC-seq, demultiplexing of heterogeneous ploidy genome such as cancer genome, and accurate estimation of priors are required. For instance, in the case of souporcell, besides demultiplexing, it also provides a function for detection of ambient RNA released during cell lysis, which is a major confounder of scRNA-seq analysis (8). Therefore, developing a new function to estimate the contamination of ambient DNA in scATAC-seq is necessary. In addition, the incorporation of heterogeneous ploidy in the current scAVENGERS model for the donor estimation would be important for the broad application of our new method in various cellular contexts. Finally, the imbalance of the number of cells among donors is critical in accurate demultiplexing of pooled scATAC-seq results, albeit such an issue can be rescued by providing proper priors (Figure 6B). Thus, development of a new function for an accurate estimation of the number of cells for each donor will be required.

In summary, scAVENGERS achieved the application of demultiplexing from scRNA-seq to scATAC-seq. By using human scATAC-seq data, we confirmed the superior performance of scAVENGERS compared to the pipeline optimized for scRNA-seq, and identified factors affecting the performance for future improvement.

DATA AVAILABILITY

Our source code of the devised software is available at GitHub (https://github.com/kaistcbfg/scAVENGERS) and Zenodo (https://doi.org/10.5281/zenodo.7408717). Published data used for the benchmark is available at the Sequence Read Archive (SRA) under the accession number PRJNA718009. PBMC datasets were deposited in the Gene Expression Omnibus (GEO) under the accession number GSE218918.

ETHICS APPROVALS AND CONSENT TO PARTICIPATE

For the use of blood samples, an expedited review was done by Institutional Review Board (IRB) committee in KAIST and IRB review exemption was obtained (IRB-21-269 [Korea Advanced Institute of Science and Technology]).

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

The authors thank the members of the Jung laboratory for their support and critical suggestions throughout the course of this work.

FUNDING

Ministry of Science and ICT through the National Research Foundation in the Republic of Korea [NRF-2020R1A2C4001464, NRF-2021M3H9A2096767, NRF-2022R1A5A1026413]; Suh Kyungbae Foundation.

Conflict of interest statement. The developed software is registered at the Korea Copyright Commission through KAIST Intellectual Property and Technology Transfer Center.

REFERENCES

1. Boyle,A.P., Davis,S., Shulha,H.P., Meltzer,P., Margulies,E.H., Weng,Z., Furey,T.S. and Crawford,G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.
2. Buenrostro,J.D., Giresi,P.G., Zaba,L.C., Chang,H.Y. and Greenleaf,W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, **10**, 1213–1218.
3. Cusanovich,D.A., Daza,R., Adey,A., Pliner,H.A., Christiansen,L., Gunderson,K.L., Steemers,F.J., Trapnell,C. and Shendure,J. (2015) Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, **348**, 910–914.
4. Buenrostro,J.D., Wu,B., Litzenburger,U.M., Ruff,D., Gonzales,M.L., Snyder,M.P., Chang,H.Y. and Greenleaf,W.J. (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, **523**, 486–490.
5. Lareau,C.A., Duarte,F.M., Chew,J.G., Kartha,V.K., Burkett,Z.D., Kohlway,A.S., Pokholok,D., Aryee,M.J., Steemers,F.J., Lebofsky,R. *et al.* (2019) Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.*, **37**, 916–924.
6. Kang,H.M., Subramaniam,M., Targ,S., Nguyen,M., Maliskova,L., McCarthy,E., Wan,E., Wong,S., Byrnes,L., Lanata,C.M. *et al.* (2018) Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.*, **36**, 89–94.
7. Orchard,P., Manickam,N., Ventresca,C., Vadlamudi,S., Varshney,A., Rai,V., Kaplan,J., Lalancette,C., Mohlke,K.L., Gallagher,K. *et al.* (2021) Human and rat skeletal muscle single-nuclei multi-omic integrative analyses nominate causal cell types, regulatory elements, and snps for complex traits. *Genome Res.*, **31**, 2258–2275.
8. Heaton,H., Talman,A.M., Knights,A., Imaz,M., Gaffney,D.J., Durbin,R., Hemberg,M. and Lawniczak,M.K.N. (2020) Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat. Methods*, **17**, 615–620.
9. Huang,Y., McCarthy,D.J. and Stegle,O. (2019) Vireo: bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome Biol.*, **20**, 273.
10. Xu,J., Falconer,C., Nguyen,Q., Crawford,J., McKinnon,B.D., Mortlock,S., Senabouth,A., Andersen,S., Chiu,H.S., Jiang,L. *et al.* (2019) Genotype-free demultiplexing of pooled single-cell RNA-seq. *Genome Biol.*, **20**, 290.
11. Lee,J.S., Park,S., Jeong,H.W., Ahn,J.Y., Choi,S.J., Lee,H., Choi,B., Nam,S.K., Sa,M., Kwon,J.S. *et al.* (2020) Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19. *Sci. Immunol.*, **5**, eabd1554.
12. You,M., Chen,L., Zhang,D., Zhao,P., Chen,Z., Qin,E.Q., Gao,Y., Davis,M.M. and Yang,P. (2021) Single-cell epigenomic landscape of peripheral immune cells reveals establishment of trained immunity in individuals convalescing from COVID-19. *Nat. Cell Biol.*, **23**, 620–630.
13. Morabito,S., Miyoshi,E., Michael,N., Shahin,S., Martini,A.C., Head,E., Silva,J., Leavy,K., Perez-Rosendahl,M. and Swarup,V. (2021) Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. *Nat. Genet.*, **53**, 1143–1155.
14. 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**, 925–936.

15. Kim,S., Scheffler,K., Halpern,A.L., Bekritsky,M.A., Noh,E., Kallberg,M., Chen,X., Kim,Y., Beyter,D., Krusche,P. *et al.* (2018) Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods*, **15**, 591–594.
16. Ueda,N. and Nakano,R. (1998) Deterministic annealing EM algorithm. *Neural Netw.*, **11**, 271–282.
17. Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.
18. Cooke,D.P., Wedge,D.C. and Lunter,G. (2021) A unified haplotype-based method for accurate and comprehensive variant calling. *Nat. Biotechnol.*, **39**, 885–892.