


Comprehensive evaluation and practical guideline of gating methods for high-dimensional cytometry data: manual gating, unsupervised clustering, and auto-gating

Peng Liu^{1,†}, Yuchen Pan^{2,†}, Hung-Ching Chang^{1,†}, Wenjia Wang¹, Yusi Fang¹, Xiangning Xue¹, Jian Zou¹, Jessica M. Toothaker^{3,4}, Oluwabunmi Olaloye⁴, Eduardo Gonzalez Santiago⁴, Black McCourt⁴, Vanessa Mitsialis^{5,6}, Pietro Presicce⁷, Suhas G. Kallapur⁷, Scott B. Snapper^{5,6}, Jia-Jun Liu^{8,9}, George C. Tseng ^{1,10,*}, Liza Konnikova^{4,7,11,12,13,14,15,*}, Silvia Liu ^{8,9,10,16,17,*}

¹Department of Biostatistics, School of Public Health, University of Pittsburgh, 130 De Soto St., Pittsburgh, PA 15261, US

²Department of Bioinformatics and Computational Biology, University of Texas MD Anderson Cancer Center, 1400 Pressler St., Houston, TX 77030, US

³Department of Immunology, University of Pittsburgh, 5051 Centre Avenue, Pittsburgh, PA 15213, US

⁴Department of Pediatrics, Yale University, 15 York Street New Haven, CT 06510, US

⁵Department of Pediatrics, Division of Gastroenterology, Hepatology, and Nutrition, Boston Children's Hospital and Department of Pediatrics, Harvard Medical School, 300 Longwood Ave., Boston, MA 02115, US

⁶Department of Medicine, Division of Gastroenterology, Hepatology, and Endoscopy, Brigham & Women's Hospital and Department of Medicine, Harvard Medical School, 300 Longwood Ave., Boston, MA 02115, US

⁷Division of Neonatology and Developmental Biology, David Geffen School of Medicine at the University of California Los Angeles, 757 Westwood Plaza, Los Angeles, CA 90095, US

⁸Drug Discovery Institute, School of Medicine, University of Pittsburgh, 700 Technology Dr, Pittsburgh, PA 15219, US

⁹Pittsburgh Liver Research Center, School of Medicine, University of Pittsburgh, 200 Lothrop Street, Pittsburgh, PA 15261, US

¹⁰Computational and Systems Biology, School of Medicine, University of Pittsburgh, 3420 Forbes Avenue, Pittsburgh, PA 15213, US

¹¹Department of Obstetrics, Gynecology and Reproductive Sciences, Yale University, 333 Cedar Street, New Haven, CT 06510, US

¹²Department of Immunobiology, Yale University, 300 Cedar Street, New Haven, CT 06520, US

¹³Program in Human and Translational Immunology, Yale University, 300 Cedar Street, New Haven, CT 06520, US

¹⁴Program in Translational Biomedicine, Yale University, 300 Cedar Street, New Haven, CT 06520, US

¹⁵Center for Systems and Engineering Immunology, Yale University, 100 College St., New Haven, CT 06510, US

¹⁶Department of Pharmacology and Chemical Biology, School of Medicine, University of Pittsburgh, 200 Lothrop St., Pittsburgh, PA 15261, US

¹⁷Hillman Cancer Center, University of Pittsburgh, 5150 Centre Ave., Pittsburgh, PA 15232, US

*Corresponding authors. Silvia Liu. E-mail: shl96@pitt.edu; Liza Konnikova. E-mail: liza.konnikova@yale.edu; George C. Tseng. E-mail: ctseng@pitt.edu

†Peng Liu, Yuchen Pan and Hung-Ching Chang to be the co-first authors.

Abstract

Cytometry is an advanced technique for simultaneously identifying and quantifying many cell surface and intracellular proteins at a single-cell resolution. Analyzing high-dimensional cytometry data involves identifying and quantifying cell populations based on their marker expressions. This study provided a quantitative review and comparison of various ways to phenotype cellular populations within the cytometry data, including manual gating, unsupervised clustering, and supervised auto-gating. Six datasets from diverse species and sample types were included in the study, and manual gating with two hierarchical layers was used as the truth for evaluation. For manual gating, results from five researchers were compared to illustrate the gating consistency among different raters. For unsupervised clustering, 23 tools were quantitatively compared in terms of accuracy with the truth and computing cost. While no method outperformed all others, several tools, including PAC-MAN, CCAST, FlowSOM, flowClust, and DEPECHE, generally demonstrated strong performance. For supervised auto-gating methods, four algorithms were evaluated, where DeepCyTOF and CyTOF Linear Classifier performed the best. We further provided practical recommendations on prioritizing gating methods based on different application scenarios. This study offers comprehensive insights for biologists to understand diverse gating methods and choose the best-suited ones for their applications.

Keywords: cytometry; manual gating; unsupervised clustering; auto-gating

Received: August 13, 2024. Revised: November 13, 2024. Accepted: November 25, 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Introduction

Cytometry is a powerful single-cell assay that allows for high-dimensional profiling of diverse cell populations in suspension [1–3]. This technique has been widely applied to clinical diagnosis, immunology and cancer research, and the pharmaceutical industry [3–5]. Flow cytometry (FCM) and mass cytometry (or cytometry by time-of-flight, CyTOF) are two major techniques that employ labeled antibodies to quantify the cell surface and intracellular proteins. FCM labels the markers by fluorescence and measures the fluorescence emitted per cell as they pass individually through a laser beam. The traditional FCM technique can detect ~8–10 markers, while the recent study has developed a 43-color flow cytometry panel [6]. As an advanced technique, CyTOF utilizes antibodies chelated with heavy metal isotopes to identify cell surface and intracellular proteins. The metal isotopes are primarily from the lanthanide series of elements, making them neither biologically derived nor radioactive. Once cell suspensions are stained and introduced into the mass cytometer, they are nebulized into droplets containing individual cells. The droplets are then ionized with argon plasma to release the metal isotopes attached to the antibodies in each droplet. The ions are separated based on mass such that the lower mass biologically derived atoms are removed and those of higher mass enter a time-of-flight chamber to measure their mass-to-charge ratios, allowing for the quantification of relative isotope abundance in each droplet [7–9]. Since mass cytometry has minimal background and high specificity, CyTOF allows for simultaneous combined measurement of up to 40–50 different antibodies, enabling the identification of a large number of cellular populations from an individual sample.

Analysis of cytometry data, although a powerful technique, brings challenges to the community. One of the major obstacles is to properly identify cell populations among thousands to millions of cells based on their high-dimensional markers [10]. In this study, we performed a comprehensive quantitative evaluation of available gating methods for high-dimensional cytometry data. Figure 1 provides a flow chart demonstrating the cytometry data format, composition of marker tables, and the evaluation pipeline of three gating categories in this study: manual gating, unsupervised clustering, and supervised auto-gating. The first manual 2D gating is the most traditional method used by biologists to identify cell populations. As illustrated in Fig. 1, in the “Manual gating” block, pairwise markers are selected based on prior knowledge (known markers of interest) and applied to identify a subset of cells. This subset of cells can be further selected and grouped by other markers to identify cell subpopulations. Eventually, hierarchical layers of cell populations are built according to different markers applied. Manual gating can be performed using FlowJo (BD Life Sciences), Cytobank (Beckman Coulter), or other software with friendly graphical user interfaces. It has advantages in identifying different cell populations of interest straightforwardly and flexibly. However, the gating process is experience-based, time-consuming, and relies on prior knowledge and arbitrary cutoffs to assign cell populations.

In addition to manual gating, computer-aided unbiased algorithms have been developed to identify cell populations in a more automated manner [11], including the second and third categories: unsupervised clustering and supervised auto-gating (which further includes semisupervised or supervised gating). For clustering methods, as shown in the “Unsupervised Clustering” block of Fig. 1, cells are grouped into clusters based on marker intensities without human intervention. Cell type characteristics

of the identified clusters are, however, unknown and rely on researchers to further annotate. The supervised auto-gating methods (illustrated in the “Supervised Auto-gating” block of Fig. 1) not only group cells into clusters based on the marker intensities but also additionally curate the cell populations identified by assigning labels to each cell cluster based on prespecified cell-type marker tables. Compared to manual gating, computer-aided methods are faster, can simultaneously analyze multiple datasets in a highly efficient and reproducible manner, and do not rely on prior knowledge to cluster the cellular populations. However, these methods sacrifice the flexibility afforded by manual gating.

As shown in Fig. 1, three types of gating methods (manual gating, unsupervised clustering, and supervised auto-gating) have been developed with increasing automation and decreasing human intervention. With the advanced popularity and development of cytometry technology, several review and evaluation papers have been published in the last 10 years. Most review literature provides descriptive introductions of clustering and visualization tools, along with conceptual guidelines for users, but offers little to no quantitative comparisons to support the conclusion [11–20]. To our knowledge, only two papers have performed numerical evaluation. Weber et al. [21] compared 18 unsupervised clustering algorithms on four CyTOF and two FCM datasets. Liu et al. [22] evaluated seven unsupervised clustering methods and two semisupervised methods across six datasets to provide guidelines on choosing clustering algorithms for cytometry data. However, a more comprehensive evaluation of tools, especially in manual gating and supervised auto-gating, and extensive panels of evaluation criteria are lacking to conclude a solid guideline for users. In contrast to the limited scope of existing papers, we performed comprehensive investigation and evaluation in all three categories in this paper (see Supplementary Table 1 for a side-by-side comparison of existing literature and the current paper). Firstly, in manual gating, we collected gating results from five raters in three different labs and evaluated gating consistency across raters. Two hierarchical layers of gating results further served as ground truth to evaluate the gating performance of the other computer-aided algorithms. Secondly, in unsupervised clustering, we attempted 32 unsupervised clustering tools previously reviewed by Liu et al. [11] and successfully implemented and compared 23 tools across six datasets. Based on the truth from manual gating, we expanded evaluation criteria (adjusted Rand index [ARI] and F-measure) and computing benchmarks to provide an evaluation panel for prioritizing overall tool performance. We also evaluated the tools’ ability to detect rare populations, which is critical in many biological or clinical applications. Finally, in auto-gating, we successfully implemented 4 out of the 6 auto-gating (supervised or semi-supervised) methods [11] to provide guidelines on the application of automatic cell population identification.

The innovation and merits of this evaluation paper compared to existing review papers [11, 14, 15, 20–24] are highlighted below (Supplementary Table 1).

Gating methods

Manual gating (5 raters), unsupervised clustering (23 tools), and auto-gating (4 tools) methods were systematically reviewed and compared. To the best of our knowledge, such a comprehensive evaluation has not been performed before, particularly as previous assessments largely lacked evaluations of both manual gating and auto-gating.

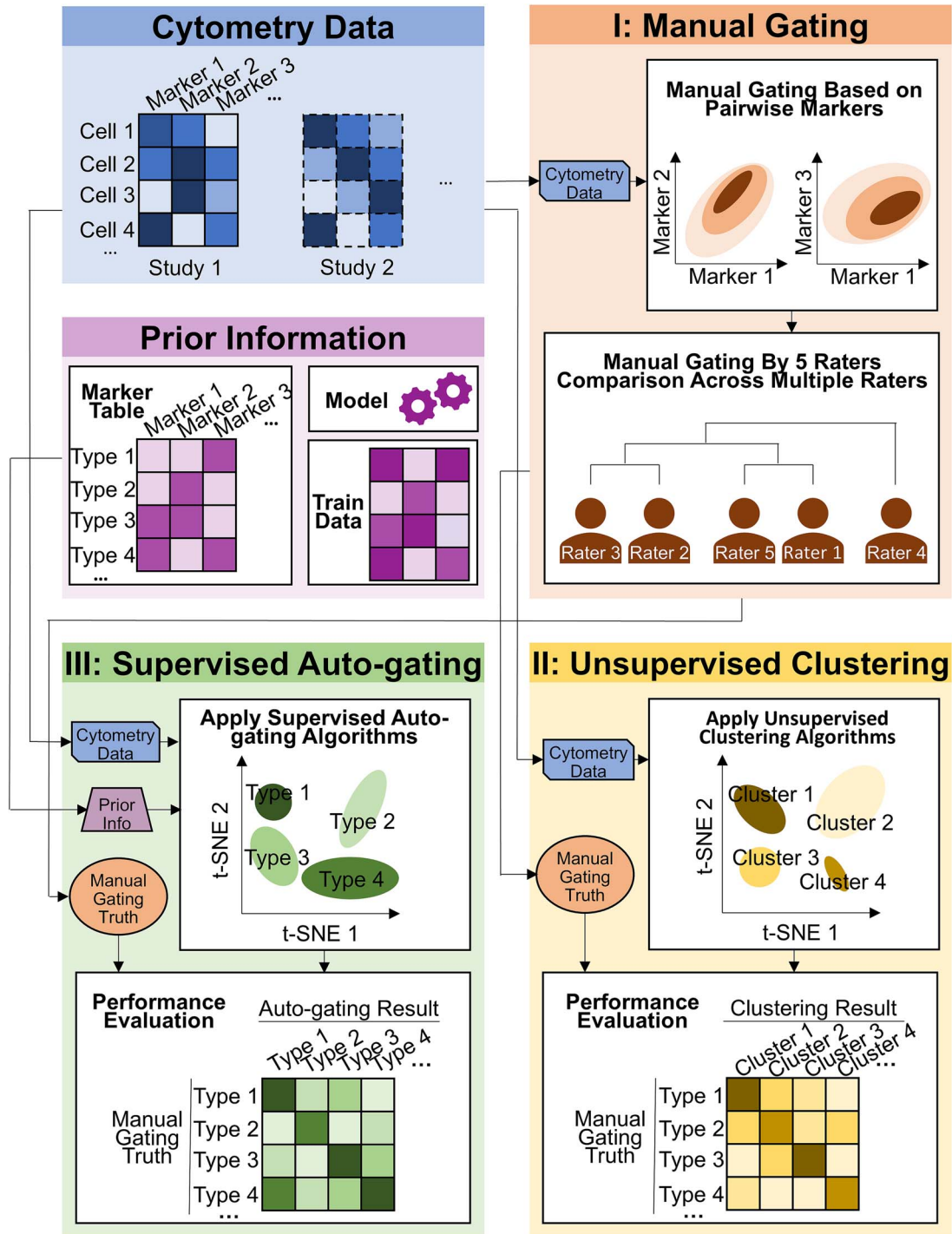


Figure 1. Workflow of the cytometry data analysis pipeline. The cell-by-marker intensity table per study was used as the input. For the (I) manual gating method, scatter plots on pairwise markers were drawn to define the cell populations. In this paper, manual gating was performed by five raters independently, and their gating results were compared. For the (II) unsupervised clustering methods, cytometry data were input into multiple clustering algorithms. Their results were compared to the manual gating and evaluated. For the (III) supervised auto-gating methods, both cytometry data and prior knowledge (such as marker table, model, and training data) were used as input. Multiple auto-gating algorithms were applied and evaluated based on the manual gating truth.

Datasets

Tools were evaluated by both in-house and public datasets, including multiple species (human, mouse, and nonhuman primates) and cell types (peripheral blood mononuclear cells [PBMCs], placental villi, and bone marrow). For in-house data, two hierarchical layers of manual gating were used as ground truth, where the first layer included major populations, and the second

layer contained a more detailed identification of subpopulations. Data for two rare populations were also used to check the tool's ability to detect small clusters of cells.

Evaluation benchmarking

Multiple evaluation measurements were employed, including F-measure, ARI, and Cohen's kappa index. In addition, performance

in different settings for the number of clusters was compared and discussed. Computing time was evaluated on a grid of cell numbers to estimate the scalability of the tools for ultra-large cell number applications in the future.

Tool recommendation

Based on comprehensive comparisons of a wide range of computer-assisted tools, our overall recommendations align and further extend from the previous publications, suggesting tools based on criteria such as accuracy, runtime, cluster-setting capabilities, graphical visualization, and performance in detecting rare populations.

Practical guideline

All programming scripts for tool implementation and comparison were made available on GitHub (https://github.com/hung-ching-chang/GatingMethod_evaluation/), on Zenodo (<https://zenodo.org/records/13851548>), and attached as Supplementary Files. During our research, we were surprised to find that implementing many published tools was not straightforward; even after contacting the original authors and making multiple attempts by several co-authors, many tools proved infeasible to use. This paper thus enables future users in the field to easily apply and compare various tools on their datasets. The GitHub and Zenodo resources also provide an evaluation platform when a new gating method is developed in the future.

Materials and methods

Cytometry by time-of-flight experimental pipeline for rhesus macaque non-human primate (NHP) samples

'Placental samples' were collected from the California Primate Center at the University of California–Davis as described in our published study [25] and from which CyTOF experiment data were extracted. Briefly, pregnant macaques were injected with either lipopolysaccharide (LPS) or saline solution; their offspring were delivered via Cesarean section 16 h following injection, and their placental biopsies were collected following delivery. For cryogenic storage, each placental layer sample was stored and processed according to a previously published protocol [26]. Fresh tissue samples were cut to 1 mm size and stored in 1 ml of freezing media (10% dimethyl sulfoxide (DMSO, Sigma) and 90% fetal bovine serum (FBS, Gibco)) by slow-freezing in a Nalgene Mr. Frosty freezing container (Sigma). For experimental processing, fresh or cryopreserved samples were made into single-cell suspensions by digesting overnight with DNase and collagenase diluted 1:5000 in digestion media (Hank's Balanced Salt Solution (HBSS) w/o Ca++ and Mg++, containing 5 mM ethylenediaminetetraacetic acid (EDTA) and 10 mM 2-[4-(2-hydroxyethyl)piperazin-1-yl]ethanesulfonic acid (HEPES)) on an orbital shaker. Single-cell suspensions underwent staining for CyTOF as described below. Details for this dataset were described in our published manuscript [25].

'PBMCs' were isolated from rhesus macaque blood via a Ficoll gradient. Blood was diluted 1:1 with Phosphate-buffered saline (PBS) in a conical tube, and an equal volume of Ficoll was added below the blood layer with a Pasteur pipette. The tube underwent a 30-min spin with low acceleration and no brake to separate the PBMCs from the rest of the blood contents. The PBMC layer was pipetted out and washed with PBS twice before resuspension in freezing media and slow freezing in a Mr. Frosty for cryopreservation. They were then thawed and underwent staining for CyTOF as

described below. Data for this set of samples are publicly available on Cytobank (Beckman Coulter).

Cytometry by time-of-flight experimental pipeline for human samples

'Placental samples' were collected through the University of Pittsburgh Biospecimen Core as described in Toothaker et al. [27] and from which CyTOF data were extracted. Human placental biopsies were separated by layers for long-term cryogenic storage or immediate experimental processing. They were stored cryogenically and prepared into single-cell suspensions in the same manner as described above for the NHP tissue samples. Single-cell suspensions underwent staining for CyTOF as described below. Details for this dataset were described in our published manuscript [27].

'PBMCs' were isolated from human blood draws via Ficoll gradient and cryopreserved as described above for the NHP blood samples. Human patients were recruited from Boston Children's Hospital (BCH) under the BCH Institutional Review Board protocol number P00000529. They were then thawed and underwent staining for CyTOF as described below. CyTOF data for this dataset are publicly available on Cytobank (Beckman Coulter).

'CyTOF' staining was performed according to the previously published protocol in Stras et al. [28]. Briefly, single-cell suspensions were washed in cell-staining buffer (CSB) composed of PBS with 0.5% bovine serum albumin (Sigma) and 0.02% sodium azide. Viability was assessed with Rh103 (Fluidigm) DNA intercalator. After an additional wash, cells were stained with their respective surface-staining antibody cocktails. For intracellular staining, cells were washed and fixed utilizing FoxP3 fixation and permeabilization kit (Invitrogen). After fixation, cells were washed with the FoxP3 wash buffer and then incubated in their respective intracellular antibody cocktails. Cells were then washed with CSB again and fixed with 1.6% paraformaldehyde (Sigma). After storage in CSB overnight, cells were incubated with 191Ir/193Ir DNA Intercalator (Fluidigm) in Maxpar Permeabilization Buffer (Standard Biotech) for cellular identification. On the day of analysis, cells were washed with MilliQ water and resuspended in normalization beads at a 1:10 dilution (Fluidigm). Data collection for the samples was done on a Fluidigm mass cytometer, and data were exported as Flow Cytometry Standard (FCS) files.

Data description

To evaluate the gating methods, we collected datasets generated by cytometry technology from both our in-house samples and publicly available sources. As shown in Table 1, six datasets were used in this study: human PBMCs, rhesus macaque PBMCs, human placental villi [27], rhesus macaque placental villi [25], human bone marrow [29], and mouse bone marrow [30]. Among these, the first four datasets were generated from our in-house libraries as described above. The human bone marrow [29] and mouse bone marrow [30] datasets were collected from the public databases from previous studies. For all six datasets, manual gating and cell population annotation were available (from the original papers for public data or by our manual gating for in-house data) that can serve as the ground truth for performance evaluation of computer-aided gating methods. Detailed descriptions of each dataset, cell population, and marker table are summarized in Table 1, Supplementary Tables 2 and 3. Down-sampled datasets to 20 000 cells were generated for tools that could not finish running the full datasets within 3 h.

For the human PBMC study, libraries for four treatments were included: fresh unstimulated (–) T cell, fresh stimulated (+) T cell, frozen T cell –, and frozen T cell +. As shown in Fig. 2A

Table 1. Data summary.

Study (citation)	Sample ID	Manual gating results	Manual gating from five experts	Clustering evaluation	Rare population identification	Auto-gating evaluation	Time benchmark evaluation	# Clusters (Layer 1, 2, rare)	# Markers (Layer 1, 2, rare)	# cells	# study/sample	Species	Tissue
Human PBMC (CytoBank)	Fresh_Tcell-	Y	Y	Y	Y	Y	Y	7, 14, 3	6, 9, 12	190 931	1	Human	PBMC
	Fresh_Tcell+	Y	N	Y	Y	Y	N	7, 14, 3	6, 9, 12	171 297	1	Human	PBMC
	Frozen_Tcell-	Y	N	Y	Y	Y	N	7, 14, 3	6, 9, 12	90 382	1	Human	PBMC
	Frozen_Tcell+	Y	N	Y	Y	Y	N	7, 14, 3	6, 9, 12	87 459	1	Human	PBMC
	Pooled 6-sample	Y	N	Y	N	Y	N	6, 13	6, 9	1252	6	Rhesus	PBMC
Rhesus PBMC (CytoBank)													
Human Villi [27]	ID1042	Y (pooled gating)	N	Y	N	Y	N	7, 14	6, 9	8334	1	Human	Villi
	ID1130	Y (pooled gating)	N	Y	N	Y	N	7, 14	6, 9	6485	1	Human	Villi
	Pooled 12-sample	Y	N	Y	N	Y	N	7, 14	6, 9	42 414	12	Human	Villi
	ID430	Y (pooled gating)	N	Y	N	Y	N	7, 14	6, 9	70 170	1	Rhesus	Villi
	ID437	Y (pooled gating)	N	Y	N	Y	N	7, 14	6, 9	66 726	1	Rhesus	Villi
Human bone marrow [29]	Pooled 14-sample	Y	N	Y	N	Y	N	7, 14	6, 9	345 886	14	Rhesus	Villi
	Set1	Y	N	Y	N	N	N	8, 20	13	167 039	1	Human	bone marrow
	Set2.H1	Y	N	Y	N	N	N	7, 10	32	72 017	1	Human	bone marrow
	Set2.H2	Y	N	Y	N	N	N	7, 10	32	31 324	1	Human	bone marrow
	S01	Y	N	Y	N	N	N	7, 24	39	53 173	1	Mouse	bone marrow
Mouse bone marrow [30]	Pooled 10-sample	Y	N	Y	N	N	N	7, 24	39	514 386	10	Mouse	bone marrow

and [Supplementary Tables 2 and 3](#), the original datasets were based on staining for 52 markers. To serve as ground truth, manual gating was performed at two hierarchical layers ([Fig. 2B](#)). The first layer employed six markers (CD3, CD19, CD4, CD8a, CD38, and CD14) to identify seven major cell populations (CD14–innate, CD38– B cells, monocytes, Natural killer T (NKT) cells, CD4 T cells, CD8 T cells, and CD38+ B cells). The second layer further divided the major populations and eventually identified 14 cell types (CD14– HLADR– innate, CD38– B cells, CD4 central memory T cells, CD4 effector memory T cells, CD4 effector T cells, CD4 naïve T cells, CD8 central memory T cells, CD8 effector memory T cells, CD8 effector T cells, CD8 naïve T cells, Dendritic cells (DCs), monocytes, NKT, and CD38+ B cells) using nine markers (CD3, CD19, CD4, CD8a, CD38, CD14, CCR7, CD45RA, and HLA-DR) [31]. To evaluate the tool's ability to identify rare populations, innate lymphoid cells (ILCs) and regulatory T cells (Tregs) (each with <3% of the total cell numbers) were selected by the manual gating based on 12 markers (CD3, CD19, CD4, CD8a, CD38, and CD14, CCR7, CD45RA, HLA-DR, CD25, FoxP3, and CD127).

In addition to the human PBMCs, the other three in-house datasets were gated into two hierarchical layers. For the NHP PBMC dataset, six samples were used and pooled together for the evaluation. Similar to the human PBMC dataset, the first layer included six major populations (CD14–innate, monocytes, NKT cells, CD8 T cells, CD4 T cells, and CD38+ B cells) using six markers (CD19, CD8a, CD14, CD38, CD4, and CD3). The second layer contained 13 clusters (DCs, monocytes, NKT, CD14– HLADR–innate, CD8 central memory T cells, CD8 effector memory T cells, CD8 effector T cells, CD8 naïve T cells, CD4 central memory T cells, CD4 effector memory T cells, CD4 effector cells, CD4 naïve T cells, and CD38+ B cells) gated by nine markers (CD19, HLA-DR, CD8a, CD14, CD45RA, CD38, CCR7_CD197, CD4, and CD3). For the human placental villi study, a total of 12 samples were analyzed. Two samples (ID1042 and ID1130) with the highest number of cells were selected as representative data. A pooled library merging all 12 samples was also used for evaluation. Similarly, for the NHP villi datasets, a total of 14 samples were analyzed. The top two samples (ID430 and ID437) and the pooled library were compared in this paper. For both human and NHP villi studies, the first layer includes 7 major populations (CD14–innate, monocytes, NKT cells, CD38– B cells, CD8 T cells, CD4 T cells, and CD38+ B cells) gated by six markers (CD18, CD8a, CD38, CD14, CD3, and CD4). The second layer contained 14 cell populations (CD4 central memory T cells, CD4 effector T cells, CD4 effector memory T cells, CD4 naïve T cells, CD8 central memory T cells, CD8 effector T cells, CD8 effector memory T cells, CD8 naïve T cells, DCs, monocytes, NKT cells, CD38– B cells, CD14– HLADR–innate, and CD38+ B cells) identified by nine markers (CD19, HLA-DR, CD8a, CD38, CD14, CD45RA, CD3, CCR7, and CD4). [Supplementary Tables 2 and 3](#) describe the cell population and marker table in more detail.

Human and mouse bone marrow datasets were generated from previous studies and downloaded from a public database [29, 30]. For the human bone marrow study [29], one healthy donor was measured in the first dataset with 13 markers ([Supplementary Table 3](#)). Manual gating was available with 25 cell populations. To avoid rare populations, several subpopulations with fewer cells were merged or removed for our analysis. Eventually, 8 and 20 populations derived from manual gating and merging were used as the first and second layers of ground truth, respectively ([Supplementary Table 2](#)). In addition, two healthy donor samples were available for the second dataset where 32

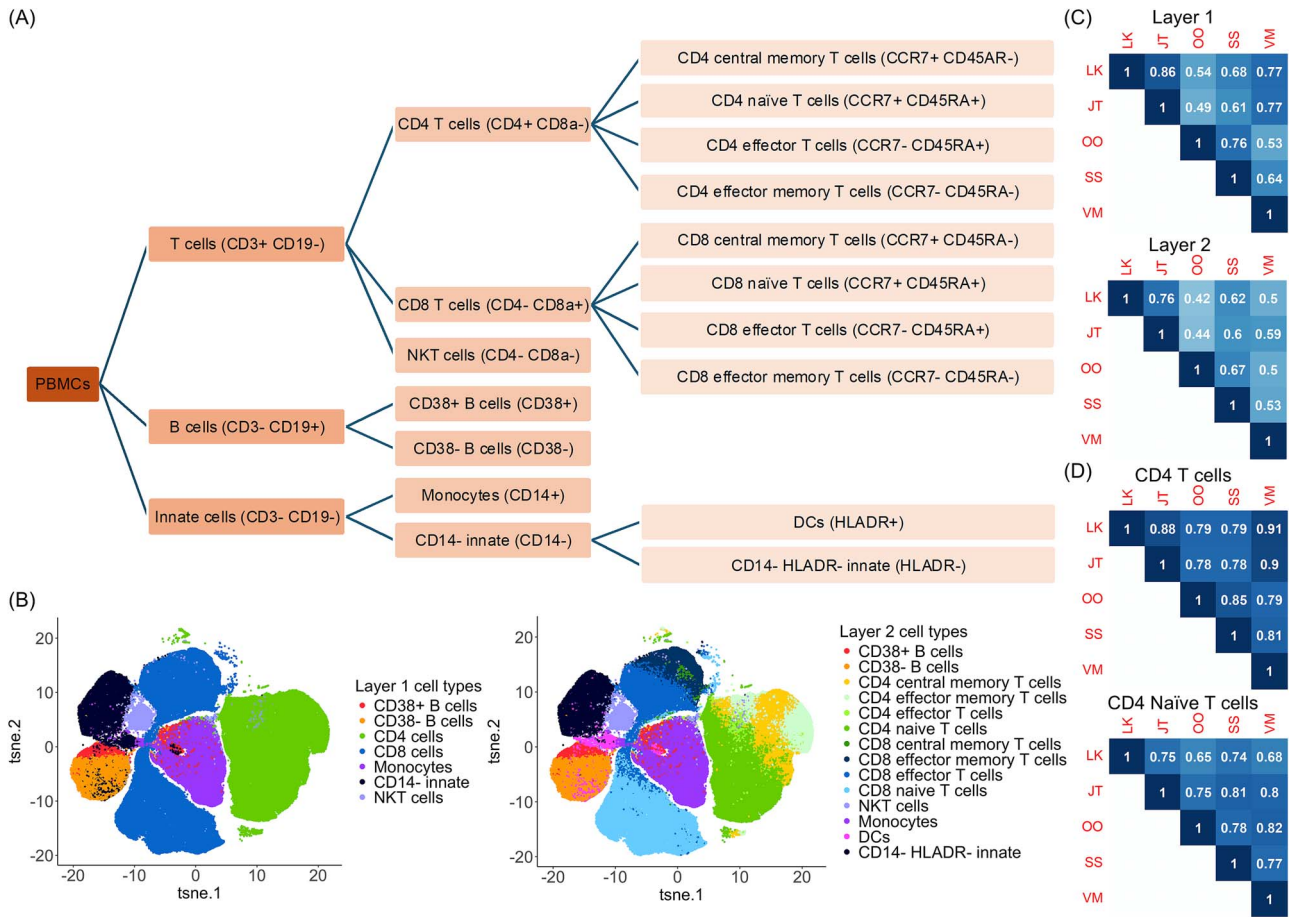


Figure 2. Coherence of manual gating across five raters. (A) A hierarchical layer of the human PBMC populations gated by the selected markers. (B) t-SNE figures indicating the manual gating cell population in Layer 1 and Layer 2 by rater L.K. (C) Average pairwise kappa index among five raters with hierarchical clustering to illustrate the similarity of the raters based on their gating assignments. (D) Pairwise kappa index among five raters on CD4 T cells and CD4 naive T cells.

markers were measured. On top of the original manual gating, we further removed or merged several smaller populations. Finally, we applied 7 and 10 populations as the first and second layers of underlying truth. For the mouse bone marrow datasets [30], 10 samples were available in total. We selected two mice (S01 and S02) and a pooled library of 10 samples as representative for the following study. This dataset measured 39 markers and manually gated cells into 7 and 24 populations for the first and second layers. The details of the cellular populations and markers used for gating and identification are described in Table 1 and Supplementary Tables 2 and 3.

Data format, preprocessing, and parameter setting

Cytometry data are commonly stored in FCS format, containing information on both metadata and marker expression. The associated metadata table generally describes experimental and channel information, such as marker name, marker description, and range information. The expression file is in an array or matrix format where each row represents an individual cell, and each column stands for a marker/channel. These channels correspond to fluorescent markers or heavy metals in flow or mass cytometry data, which have been described in the metadata table [13, 32, 33].

As the data preprocessing steps, all the negative intensities were trimmed at zero or very small random numbers close to zero (if the algorithms report an error when using multiple zeros as

input). Cytometry data were further scaled and inverse hyperbolic sine-transformed ($X_{new} = \text{asinh}(a + b * X_{old}) + c$), with $a = 0$, $b = 0.2$, and $c = 0$) [34]. Tools were first applied to the full data with all cells. If the tool could not complete the run within 3 h, down-sampled data with 20 000 cells were used as an alternative. All the tools were run by default parameter settings except for the number of clusters. If the tool allowed for the specification of the number of clusters to be generated, the true number of clusters was used as input. For detailed scripts for preprocessing and running the tools, please refer to the script files deposited to GitHub (https://github.com/hung-ching-chang/GatingMethod_evaluation/) and Zenodo (<https://zenodo.org/records/13851548>).

Manual gating collected from five raters

The “fresh T-cells -” library from the human PBMC study was manually gated on the Cytobank [35] platform by five raters (L.K., J.T., O.O., S.S., and V.M.) who were asked to manually gate the immune cell populations based on their experience independently. No computational algorithm was allowed. Eventually, commonly identified cell populations by all the raters were selected for further evaluation. Based on the hierarchical gating structure in Fig. 2A, these cell populations were categorized into two layers, containing 7 and 14 populations for the first and second layers, respectively. The cell populations gated by LK in both layers were visualized by t-distributed stochastic neighbor embedding (t-SNE) [36] plots in Fig. 2B, and the gating results by the other four raters are shown in Supplementary Fig. 1. To quantitatively

Table 2. Rank-sum for unsupervised clustering methods.

Data	Major layer 1	Major layer 2	Major layer 1	Major layer 2	Rare	Overall ranking	Overall ranking
Measurement	ARI	ARI	F-measure	F-measure	F-measure	Mean rank	Overall rank
ACCENSE	22	22	22	15	13	18.8	21
CCAST	4	1	8	4	1	3.6	2
ClusterX	19	20	20	21	2	16.4	19
CosTaL	14	6	9	10	20	11.8	12
Cytometree	10	16	6	5	19	11.2	9
densityCUT	9	15	19	17	17	15.4	17
DensVM	13	7	10	10	11	10.2	7
DEPECHE	1	1	5	13	3	4.6	4
FLOCK	5	5	3	7	7	5.4	6
flowClust	2	9	1	6	5	4.6	4
FlowGrid	7	12	14	20	5	11.6	11
flowMeans	15	19	18	16	10	15.6	18
flowPeaks	11	17	17	19	8	14.4	15
FlowSOM	6	3	2	2	9	4.4	3
immunoClust	20	21	15	18	21	19	22
k-means	8	10	11	9	18	11.2	9
PAC-MAN	3	4	3	1	4	3	1
PhenoGraph	17	13	7	3	12	10.4	8
Rclusterpp	15	8	13	8	16	12	13
SamSPECTRAL	18	18	21	22	14	18.6	20
SPADE	21	10	15	14	15	15	16
X-shift	11	14	12	12	22	14.2	14

Note: Tool SWIFT can only be applied to human PBMC data, so it's not included in the overall evaluation.

evaluate the magnitude of agreement between the raters, both kappa index and ARI measurements were employed, as described in detail in the [Methods](#) section. A follow-up hierarchical clustering analysis was performed to group the raters with similar gating results based on cellular annotation in the two layers, respectively.

List of unsupervised clustering algorithms evaluated

Unsupervised clustering algorithms group cells into clusters based solely on their marker intensities, lacking the ability to assign the resulting clusters to known cell populations ([Fig. 1](#)). In a previous publication [11], we summarized 32 tools for unsupervised clustering, which were either specifically designed for cytometry data or for general clustering. Based on the number of input libraries, clustering methods are categorized by the tools that can work on individual samples and tools that need multiple libraries as input. This comparison study only focused on the former ones. As such, we quantitatively compared a total of 23 clustering methods ([Table 2](#)): ACCENSE [37], CCAST [38], ClusterX [32], CosTaL [39], Cytometree [40], densityCUT [41], DensVM [42], DEPECHE [43], FLOCK [44], flowClust [45], FlowGrid [46], flowMeans [47], flowPeaks [48], FlowSOM [49, 50], immunoClust [51], k-means [52], PAC-MAN [53], PhenoGraph [29], Rclusterpp [54], SamSPECTRAL [55], SPADE [56], SWIFT [57], and X-shift [30]. These tools have free publicly available software and are compatible with our parameter settings (see [Data Format, Preprocessing, and Parameter Setting](#) section). Manual gating was employed as the underlying truth to evaluate the performance of these unsupervised clustering tools.

List of auto-gating algorithms evaluated

Unsupervised clustering methods can identify cell clusters with distinct marker characteristics. However, without further annotation of cell populations, these clusters provide no biological

insight. Manual annotation of cell clusters is time-consuming and biased. To overcome this problem, several automated annotation and cell-type identification algorithms have been developed. These auto-gating algorithms are designed to identify the resulting clusters from clustering algorithms based on either the prior knowledge of the relationship between lineage markers and the identity of cellular populations or learned from training datasets ([Fig. 1](#)). Our previous publication [11] summarized six auto-gating algorithms: DeepCyTOF [58], CyTOF linear classifier [59], ACDC [60], MP [61], OpenCyto [62], and flowLearn [63]. DeepCyTOF [58] utilizes deep learning techniques and training data to assign cells to known cell types. Likewise, the CyTOF liner classifier [59] applies training data to predict cell types based on linear discriminant analysis. In addition, ACDC [60] uses a prespecified marker matrix to guide the grouping of cells based on a semisupervised learning approach, and MP [61] employs a marker matrix to predict cell types through a Bayesian model. In this paper, we compared the performance of the above four methods, while OpenCyto and flowLearn were not evaluated since both of them are not fully automated and require user supervision. As the cell types could be further divided, the underlying truth for cell type identification is controversial and needs to be adjusted depending on the problem at hand. Therefore, to perform a comprehensive evaluation, two hierarchical layers of manually gated references were used in the manuscript.

Methods for performance evaluation

Definition

Cell population identification by manual gating is used as truth to evaluate the performance of unsupervised and supervised algorithms. For each individual or pooled library, a set of n cells $S = \{o_1, o_2, \dots, o_n\}$ can be grouped into two partitions with r populations and c populations, defined as $X = \{X_1, X_2, \dots, X_r\}$ and $Y = \{Y_1, Y_2, \dots, Y_c\}$. For the pairwise comparison between X and Y , the contingency table can be defined as

	Y_1	Y_2	\dots	Y_c	Sum
X_1	n_{11}	n_{12}	\dots	n_{1c}	a_1
X_2	n_{21}	n_{22}	\dots	n_{2c}	a_2
\dots	\dots	\dots	\dots	\dots	\dots
X_r	n_{r1}	n_{r2}	\dots	n_{rc}	a_r
Sum	b_1	b_2	\dots	b_c	

For $i \in \{1, 2, \dots, r\}$ and $j \in \{1, 2, \dots, c\}$, $n_{ij} = |X_i \cap Y_j|$ represents the number of overlapping cells between partition X_i and Y_j , and a_i and b_j indicate $|X_i|$ and $|Y_j|$, respectively.

When focusing on a certain population i from partition X and population j from partition Y , the 2-by-2 confusion matrix can be written as

		Prediction	
		Y_j	$\neg Y_j$
Truth	X_i	TP (true positive)	FN (false negative)
	$\neg X_i$	FP (false positive)	TN (true negative)

In this paper, we employed the following measurements to indicate the agreement of the two partitions.

The 'ARI' aims to measure the agreement of two clustering partitions without cell population identification [64]. A value close to 1 indicates high consistency, while a value close to zero or even negative means decreased similarity. Assuming X represents the true populations identified by manual gating and Y the predicted groupings generated by clustering or auto-gating methods, the ARI is defined as

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}}{\frac{1}{2} \left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2} \right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}}$$

The 'Kappa index', also known as Cohen's kappa coefficient, is an indicator that measures the inter-rater reliability [65]. For a binary scenario, the kappa index can be defined as

$$\kappa = \frac{2 \times (TP \times TN - FN \times FP)}{(TP + FP) \times (FP + TN) + (TP + FN) \times (FN + TN)}$$

'F-measure' (or F1 score) is an accuracy evaluation measurement to balance the precision and recall in binary conditions [66]. Assuming X represents the true population and Y the predicted groupings, the precision, recall, and F-measure are defined as follows:

$$\begin{aligned} \text{precision} &= \frac{TP}{TP + FP} \\ \text{recall} &= \frac{TP}{TP + FN} \\ F - \text{measure} &= 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \end{aligned}$$

Computing time evaluation

To benchmark the computing time, the "fresh T cells -" library from the human PBMC dataset was randomly down-sampled to 1000, 2000, 4000, 8000, 16 000, 32 000, 64 000, and 128 000 cells. Twenty-three unsupervised clustering and four supervised

auto-gating methods were applied to these gradient cell numbers to evaluate both the performance and computing time. For tools where the number of clusters can be set as input, the true number of clusters was set. Otherwise, all the tools were run based on their default parameter settings (supplementary script files submitted to GitHub and Zenodo). Both clustering and auto-gating tools were run on the same Windows computer (Intel Core i7-8700 CPU @ 3.20GHz 3.19GHz, 16GB RAM, 64-bit operating system, x64-based processor) to reduce machine variability.

Results

Manual gating: consistency across multiple raters

Manual gating has been widely applied to cytometry data to group and annotate individual cells into populations of interest, where researchers have the flexibility to choose the markers and set up the cutoff to define cell populations. However, when performing manual gating, researchers often classify cells into populations based on their personal experience and preferred markers. Additionally, it's rather arbitrary when one draws a line to split the populations. To test the consistency of manual gating across different raters, we invited five researchers from three different labs to perform manual gating independently and compare their performances on the "Fresh T cells - library" from the human PBMC dataset. When excluding the "other cells," cells from the human PBMC study were grouped into 7 major clusters in Layer 1 and then further grouped into 14 clusters in Layer 2 (Fig. 2A). Figure 2B and Supplementary Fig. 1 illustrated the t-SNE plots of the cell populations gated by rater L.K. and all the other four researchers (J.T., O.O., S.S., and V.M.). To assess the gating similarity among the five raters, the kappa index was used as the evaluation method, where $\kappa = 1$ indicates complete agreement and $\kappa = 0$ means no agreement. As shown in Fig. 2C, the pairwise kappa index ranged from 0.44 to 0.86, showing a significant level of variation in manual gating by different experts. The highest kappa index was between LK and JT, 0.86 in Layer 1 and 0.76 in Layer 2. The lowest kappa index was the one between JT and OO in Layer 1 (0.49) and the one between LK and OO in Layer 2 (0.42). The kappa indexes in Layer 1 were higher than that in Layer 2 for the same pair, likely because the extra steps needed for the manual gating for Layer 2 introduced the variation. Furthermore, CD4 T cells and CD4 naïve T cells were selected to illustrate the agreement for individual cell populations between the raters in Fig. 2D. The same patterns were observed, though the actual numbers varied.

Given the raters were only instructed to gate the cell population by their own experience, the raters not only gated common populations differently but also focused on different sets of subpopulations. For example, LK gated 39 subpopulations in total, while OO gated 24 subpopulations, where FoxP3⁺ CD4 T cells and GATA-3⁺ CD4 T cells were gated only by LK, while Th1 CD4 T cells and Th17 CD4 T cells were gated only by OO (Supplementary Table 4). We observed that the lines drawn to split the populations were subjective, and the raters had their own preferences in splitting the populations. These results gave rise to the need for more reproducible and less labor-intensive gating strategies, which motivated us to review the clustering and auto-gating tools in the following sections.

Clustering: detection of major cell populations

Unsupervised clustering refers to the computational method for cell grouping without population annotation. Thirty-two unsupervised clustering tools were reviewed and discussed in

our previous study [11]. When we attempted to evaluate them comprehensively, 23 tools could be applied to individual datasets and installed successfully on our end (Table 2). To evaluate the performance of unsupervised clustering algorithms, we applied them to six datasets as described in Table 1. For each study, the major populations were identified hierarchically in two layers: Layer 1 represented top-level major populations, and Layer 2 included lower-level and detailed populations. Both the ARI and F-measure (described in the Methods section) were applied to evaluate the agreement between the clustering results for each tool and the manual gating truth. The ARI values for each dataset and method were illustrated in heatmaps with rows representing the tools and columns indicating datasets (Fig. 3A and B), where ARI = 1 indicated an exact match between the two groups, while an ARI value close to zero meant a lack of consistency. As illustrated by the heatmap color for ARI, results for both Layers 1 and 2 indicated overall high performance for the human PBMC and human bone marrow and mouse bone marrow studies, while comparatively lower performance for the rhesus PBMCs, human placental villi, and rhesus placental villi were observed. This was potentially due to human antibodies' suboptimal staining of rhesus samples or the need for altered phenotyping in rhesus samples. Additionally, staining within the tissue (placental villi) as compared to blood resulted in less distinct marker differences than that obtained from PBMCs.

Many clustering algorithms include random initial steps to learn the grouping and subsequently cluster cells, which will result in distinct clustering outcomes for different runs. To check the robustness/consistency across multiple runs, all the tools were applied to the human PBMC dataset for 10 independent runs. As shown in Fig. 3C, the bar plot indicated the average and SD of the 10 repeated runs. Among these clustering algorithms, tools such as ACCENSE, CosTaL, FlowSOM, immunoClust, k-means, and SPADE presented larger variations among multiple runs, while the remaining tools generated comparatively consistent or even identical results.

When comparing the tool's performance in detecting top-layer major populations and lower-layer detailed populations, we evaluated their performance against two hierarchical truths (described in the Methods section). For each truth layer, the rank-sum of the tool over multiple studies was calculated and averaged (Supplementary Table 5). Eventually, the average rank-sum per layer was compared (Fig. 3D), and the overall ranking was summarized in Table 2, where lower rank indicated higher consistency with the truth. When comparing tool performance between Layer 1 and Layer 2, Fig. 3D showed high agreement for most tools (close to the diagonal line), except for CCAST, CosTAL, DensVM, Rclusterpp, and SPADE. These five tools presented comparatively lower performance in Layer 1 but higher performance in Layer 2 (higher rank in Layer 1 than Layer 2). These tools tended to provide a larger number of clusters in their default parameter settings, which resulted in better detection of more detailed subcell populations than the major ones.

In addition to the ARI measure, the F-measure was also applied in our study as an alternative evaluation benchmark (see Methods section). Per tool-predicted cluster, the highest F-measure was recorded as the best match between the predicted cluster and the true cluster. Eventually, averaged F-measures were calculated to indicate the overall performance of the clustering algorithms. Supplementary Fig. 2 illustrates tool performance quantified by F-measure, and Table 2 summarizes their overall performance ranking. When comparing the performance consistency between multiple measurements, Fig. 3E showed the average rank-sum of

the ARI and F-measure. The majority of the tools showed high agreement between the two measurements, while some tools showed comparatively higher performance for one measurement than the other. For example, CosTAL, PhenoGraph, Cytometree, and SPADE resulted in better performance by the F-measure but lower performance by the ARI measure. This may be due to the mapping between tool-predicted clusters with the ground truth. Generally, a tool yielding a larger number of clusters than the truth will tend to perform better by F-measure, which will choose the best cluster to match the truth. However, this tool will receive lower performance by ARI for punishing a large number of clusters. In conclusion, when considering both layer and measurement effects, PAC-MAN, FlowSOM, CCAST, flowClust, FLOCK, and DEPECHE performed overall the best for the selected datasets (Table 2 and Supplementary Table 5).

Clustering: setting the number of clusters

Setting an appropriate number of clusters is crucial for clustering cytometry data. Among the 23 tested algorithms, only eight tools allow users to directly set the number of clusters, namely, CCAST, DEPECHE, flowClust, flowMeans, FlowSOM, k-means, PAC-MAN, and SPADE. Other tools adjust clustering indirectly via parameters like resolution, settings for nearest neighbors, or other cell distance measures. To ensure a fair comparison, we only evaluated the performance of these eight tools using the "human PBMC fresh T cell -" library, with 7 and 14 manually gated clusters as the ground truth for the two hierarchical layers. For each algorithm, we set the gradient numbers of clusters close to the ground truth to evaluate their performances. As shown in Fig. 3F and G, most tools performed optimally at the true cluster number and had comparable performance under other settings. Generally, tools k-means, flowClust, and CCAST present larger variations of the performance across different numbers of clusters. Tool flowMeans presents a smaller variation in Layer 1 but a larger variation in Layer 2. As shown in Supplementary Fig. 2F and G, the F-measure was less sensitive to the cluster number than the ARI measure because it will select the optimal cluster to match the truth. In practice, without knowing the true number of cell populations, we suggest users set a cluster number close to the anticipated populations they want to detect. Alternatively, tools like INFLECT [67] have been developed specifically for FlowSOM to optimize the number of clusters. Borrowing techniques from tools designed to estimate the number of cell types in single-cell RNA-seq data [68] may also be beneficial.

Clustering: computing time benchmarking

Besides clustering accuracy, computational cost is another factor in evaluating these tools. In this study, we selected the "human PBMC fresh T cells - library" as an example for benchmarking the computing time. Specifically, we subsampled cells from this dataset by log2 gradient numbers: 1k, 2k, 4k, 8k, 16k, 32k, 64k, and 128k. All the clustering algorithms were applied to these eight subsets based on default parameter settings to record their computing times. Figure 4A demonstrates the computing time (average of 10 runs) per tool per data subset. Among these tools, k-means, FlowGrid, and densityCUT overall consumed a shorter time for these datasets. To predict the running time for a larger number of cells, we fitted the computing time to logistic regression in terms of the number of cells (Fig. 4B and Supplementary Fig. 3). Except for flowPeaks (Fig. 4B), which consumed constant computing time across different datasets, all the other tools (for example, DensVM in Fig. 4B) resulted in longer computing time with an increasing number of cells. However,

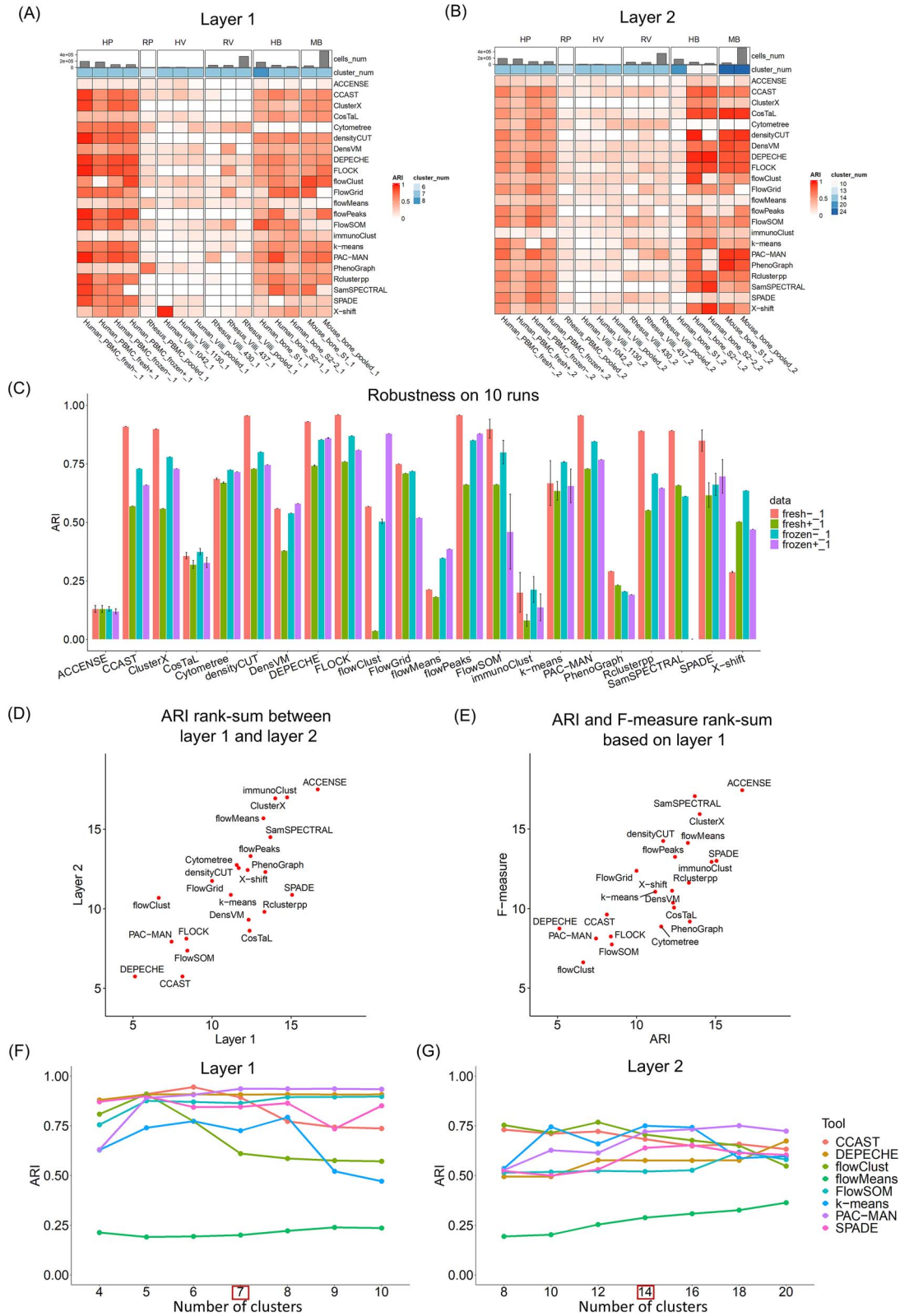


Figure 3. Major population detection by unsupervised clustering algorithms. (A) Performance of the clustering algorithms based on Layer 1 cell populations using the ARI measure. (B) Performance of the clustering algorithms based on Layer 2 cell populations using the ARI measure. (C) Robustness of the clustering algorithms on 10 runs using the ARI measure. (D) Rank-sum between Layer 1 and Layer 2 using the ARI measure. (E) Rank-sum between the ARI and F-measure based on Layer 1 data. (F) Performance of the clustering algorithms across different settings for the number of clusters on Layer 1 cell populations using the ARI measure. (G) Performance of the clustering algorithms across different settings for the number of clusters on Layer 2 cell populations using the ARI measure.

these tools yielded different increases in computing time for additional cells, which were reflected by the slope of the regression model. Tools such as FLOCK, FlowSOM, and FlowGrid achieved comparatively shorter computing costs for additional cells (with slope to be 0.12, 0.25, and 0.43 in [Supplementary Fig. 3](#)), indicating better compatibility to analyze larger datasets. In contrast, tools such as ClusterX and CCAST increased the amount of computing time fast (with slope to be 1.89 and 1.52 in [Supplementary Fig. 3](#)), which may result in much longer computing time for larger libraries.

For an overall evaluation of the tools, [Fig. 4C and D](#) visualized the tool performance and the computing time simultaneously. In [Fig. 4C](#), tools PAC-MAN, k-means, FlowSOM, densityCUT, FLOCK, and FlowGrid have higher ARI with comparatively shorter computing time, while in [Fig. 4D](#), tools k-means, FlowSOM, PAC-MAN, PhenoGraph, densityCUT, X-shift, and FLOCK yielded the best performance quantified by higher F-measure and lower computing cost.

Clustering: detection of rare populations

We referred to a “rare population” as a cell type with fewer cells compared with the major populations. Since each rare population only represents <3% of the whole library, it is easily missed when all the cells are clustered. To evaluate the capability of the clustering algorithms to identify rare populations, we selected two rare populations from the human PBMC dataset to illustrate ([Fig. 5A](#)): innate lymphoid cells (ILCs) and regulatory T cells (Tregs). The number of cells within each population and the markers applied are listed in [Supplementary Tables 2 and 3](#). Their distributions to the whole libraries are shown in the t-SNE plot in [Fig. 5A](#). When aiming for rare population detection, parameters of the clustering tools were adjusted to detect a larger number of small clusters, rather than big clusters for the major population. The detailed settings are shown in the supplementary script files. F-measure was applied to evaluate the consistency between the clustering algorithms and the manual gating truth. As shown in [Fig. 5B](#), the heatmap visualized the performance of 23 clustering algorithms in detecting these two rare populations across the four human PBMC libraries. In general, many tools resulted in overall high and robust performance across multiple libraries and for both of the rare populations, including ACCENSE, CCAST, ClusterX, DensVM, DEPECHE, FLOCK, flowClust, FlowGrid, flowMeans, flowPeaks, FlowSOM, PAC-MAN, and PhenoGraph.

Auto-gating: prior knowledge preparation and performance evaluation

Supervised or semisupervised auto-gating methods refer to the algorithms that take in prior knowledge or training sets to train the model and then perform cell population identification based on these parameters. In this paper, we quantitatively reviewed four auto-gating methods: DeepCyTOF [58], CyTOF linear classifier [59], ACDC [60], and MP [61]. To check the agreement between these auto-gating algorithms and the manual gating truth, both ARI and F-measure were applied to evaluate the performance of these tools on the PBMC and the placental villi data. As shown in [Fig. 6A](#), the heatmap indicated the ARI value for each tool (row) across each dataset (column) when using the Layer 1 manual gating as truth. Across all the datasets, DeepCyTOF and CyTOF Linear Classifier had an overall better performance, while ACDC and MP only performed well with the human PBMC data. The potential reason might be that the human PBMC dataset had a higher number of cells to cover a more robust set of different immune cell populations when compared with the other datasets. In addition

to Layer 1, Layer 2 truth was applied and demonstrated similar patterns ([Fig. 6B](#)). The rank-sum comparison between Layers 1 and 2 is shown in [Fig. 6C](#), where these four tools showed a high agreement of performance when applied to the two hierarchical truths, and DeepCyTOF and CyTOF Linear Classifier achieved high ARI consistently. To evaluate the robustness of these tools, we ran all the algorithms 10 times on the human PBMC data. As shown in [Fig. 6D](#), all four auto-gating tools presented low variations across multiple runs. Besides the ARI measurement, a similar performance evaluation was measured by the F score, and the same conclusion can be drawn ([Supplementary Fig. 4](#)). [Table 3](#) summarizes the rank-sum per measurement and their overall ranking, where DeepCyTOF and CyTOF Linear Classifier were the top two algorithms recommended based on our evaluation.

In the next step, we aimed to check the computing cost of these auto-gating tools. [Figure 6E](#) indicates the computing time across a gradient number of cells for each algorithm (similar pipeline as was done for the clustering methods). DeepCyTOF achieved a low computing time and the lowest increasing slope (0.46) for the regression model among the four tools. CyTOF Linear Classifier resulted in the overall shortest computing time while maintaining a linear slope (0.99) for increasing time. When considering the tool's computing time and performance at the same time, [Fig. 6F and G](#) illustrates that both DeepCyTOF and CyTOF Linear Classifier achieved the highest performance and shortest computing cost.

Discussion

This paper comprehensively investigated and compared three main categories of methods for analyzing cytometry data: manual gating, unsupervised clustering, and supervised auto-gating ([Fig. 1](#)). Among them, manual gating involves visually inspecting multidimensional plots of the data and drawing boundaries (gates) around populations of interest. This gating method is widely applied by expert researchers and can be operated with flexibility and transparency. However, manual gating has limitations when dealing with high-dimensional and large-scale datasets, and the results are subject to the researcher's experience. In contrast to manual gating, unsupervised clustering and auto-gating methods are computer-assisted algorithms, which have the advantages of automation and reproducibility for large datasets, but face the limitations in transparency and parameter sensitivity. When distinguishing these two computational gating methods, clustering algorithms aim to group similar cells into clusters without predefined populations, while auto-gating tools are designed to mimic the manual gating process to identify and gate cell populations. Although many automated gating algorithms have been developed, manual gating and unsupervised clustering followed by manual annotation are the most widely used pipelines for cell population identification.

Overall recommendations

In this manuscript, we systematically evaluated 23 unsupervised clustering algorithms and 4 auto-gating tools (supervised or semisupervised). [Tables 2 and 3](#) summarize the rank of these tools when applied to six cytometry datasets by both ARI and F-measure benchmarks and based on both Layers 1 and 2 truth. [Figure 7](#) presents a workflow and comparison for an overall recommendation of the algorithms. Among all the computer-assisted tools, if no prior knowledge nor training data were available, unsupervised clustering tools were suggested. Among them, tools with higher performance and shorter computing time

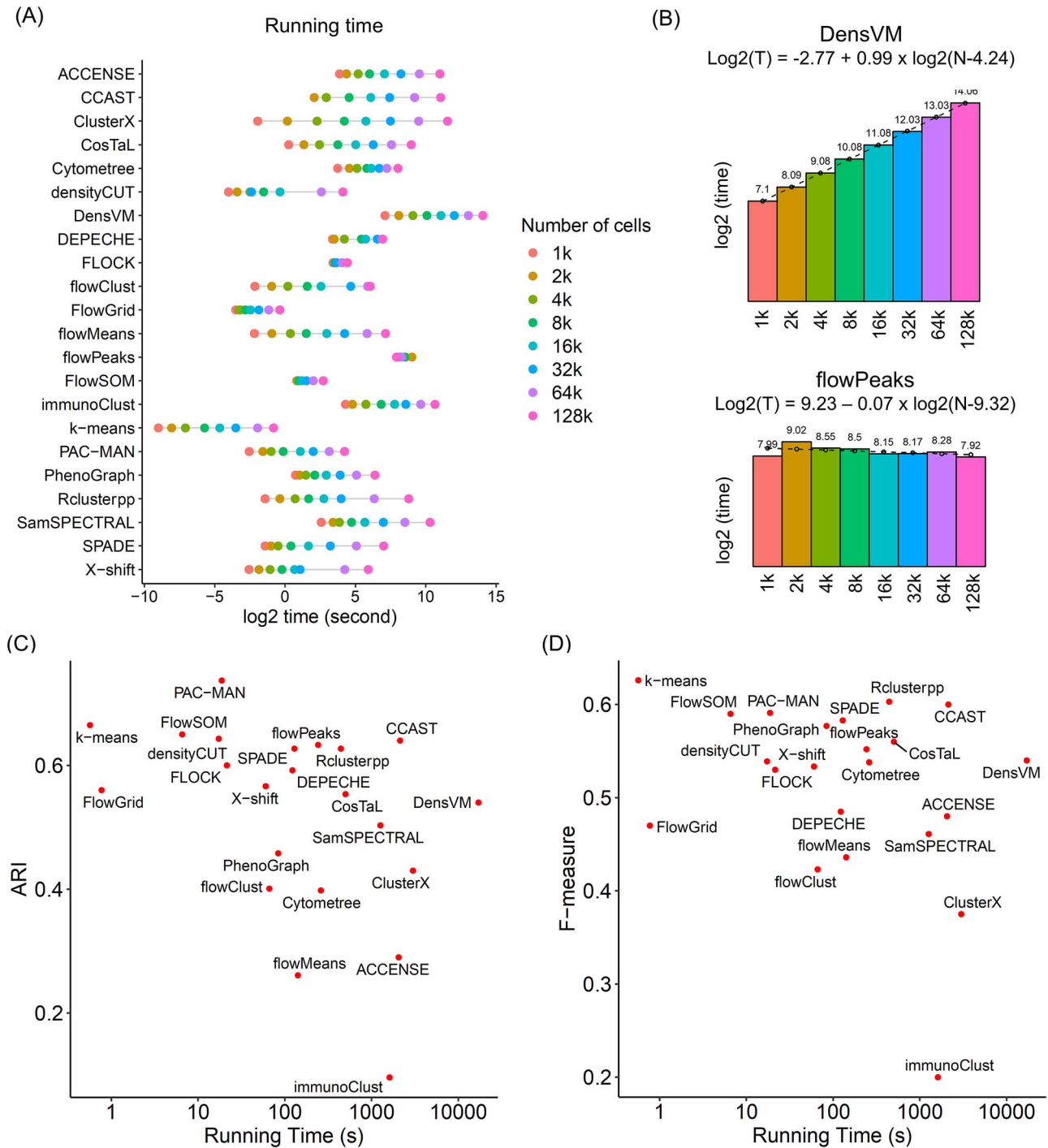


Figure 4. Time benchmark for unsupervised clustering algorithms. (A) Time benchmark for clustering algorithms on gradient number of cells. (B) Computing time for two representative algorithms with linear (DensVM) and flat (flowPeaks) increasing speed. Bar graph represents the real running time, and the dashed line is for the predicted running time when fitting into the regression model. (C) The comparison between computing time and ARI performance. (D) The comparison between computing time and F-measure performance.

were recommended. Other specific recommendations were made as well. For example, if the users wanted to specify the number of clusters, tools such as DEPECHE, flowClust, FlowSOM, k-means, and PAC-MAN would be good options. If the researchers were interested in graphical visualization of the clustering results, tools such as Cytometree, FlowSOM, PhenoGraph, and X-shift could yield figures for hierarchical trees or gating networks. If the rare populations were the major focus of the study, we would recommend DEPECHE, flowClust, FlowGrid, and PAC-MAN.

When considering all these factors and balancing accuracy and computing time, PAC-MAN, CCAST, and FlowSOM were the top three recommended tools. In addition to clustering algorithms, auto-gating methods were suggested for studies with prior knowledge of cell populations. In general, our evaluation study suggested DeepCyTOF and CyTOF Linear Classifier as they had the highest accuracy and shortest computing time.

Depending on different research interests, some studies only focus on major cell types, while other studies aim to explore

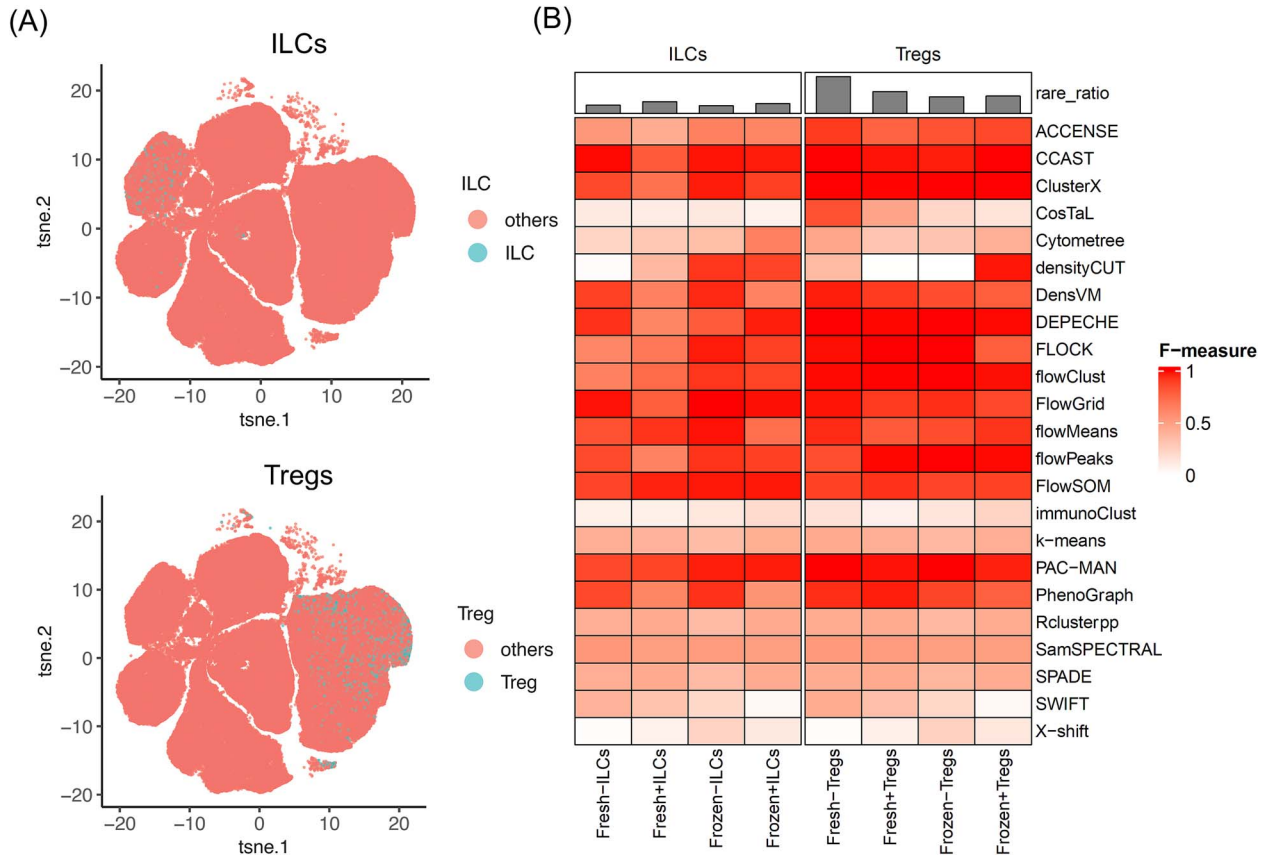


Figure 5. Rare population detection by clustering algorithms. (A) Manual gating for the selection of the two rare populations: ILCs and Tregs. (B) Heatmap for the performance of the clustering algorithms on the rare population based on the F-measure.

Table 3. Rank-sum for supervised clustering methods.

Data	Major layer 1	Major layer 2	Major layer 1	Major layer 2	Overall ranking	Overall ranking
Measurement	ARI	ARI	F-measure	F-measure	Mean rank	Overall rank
ACDC	3	3	3	3	3	3
CytoF Linear Classifier	2	2	1	1	1.5	1
DeepCyTOF	1	1	2	2	1.5	1
MP (Mondrian)	4	4	4	4	4	4

detailed subpopulations or even rare populations. In this study, we provided two hierarchical layers of truth: Layer 1 for major populations and Layer 2 for detailed types. When comparing the consistency of the gating results between the two layers, most of the tools had high agreement (Figs 3D and 6C), while tools that generally yield a higher number of clusters tended to favor Layer 2 to detect more detailed populations. As an overall suggestion, if the researchers are interested in subpopulations or even rare populations, tools that can specify the number of clusters or perform well for rare populations are recommended (Figs 3 and 7). Meanwhile, selecting appropriate biomarkers is crucial for the detection of sub- and rare populations (Supplementary Table 3).

Among the six datasets, diverse samples were employed for tool comparison, including multiple tissue types (PBMCs, bone marrows, and placental villi), fresh and frozen samples, and samples from different species (human, nonhuman primate, and mouse). In general, PBMC and bone marrow samples outperformed the placental villi. This is likely due to marker expression on cells within the tissue being less distinct than in the

bone marrow and the peripheral blood. As such, auto-gating for tissue samples is not recommended. In contrast, the difference between fresh and frozen PBMC samples was trivial, which indicated that this treatment wouldn't influence the antibody capture in the cytometry experiment [26]. In addition, samples from three species were included in this study: human, rhesus, and mouse. The human samples performed the best, likely secondary to more optimized antibodies and better phenotyping.

Compared with existing review papers [11, 14, 15, 20–24] with a limited number of clustering tools and benchmarks, we provided a more systematic review and overall recommendations. For clustering, PAC-MAN ranks high in our study, but this tool was rarely discussed by the current review papers. Moreover, we evaluated the tools from the views of data types, multiple hierarchical layers, rare populations, settings for the number of clusters, and graphical visualizations. In addition, auto-gating highlights the promising future of cytometry data analysis, where our manuscript is the first attempt to comprehensively benchmark four cutting-edge tools in this field.

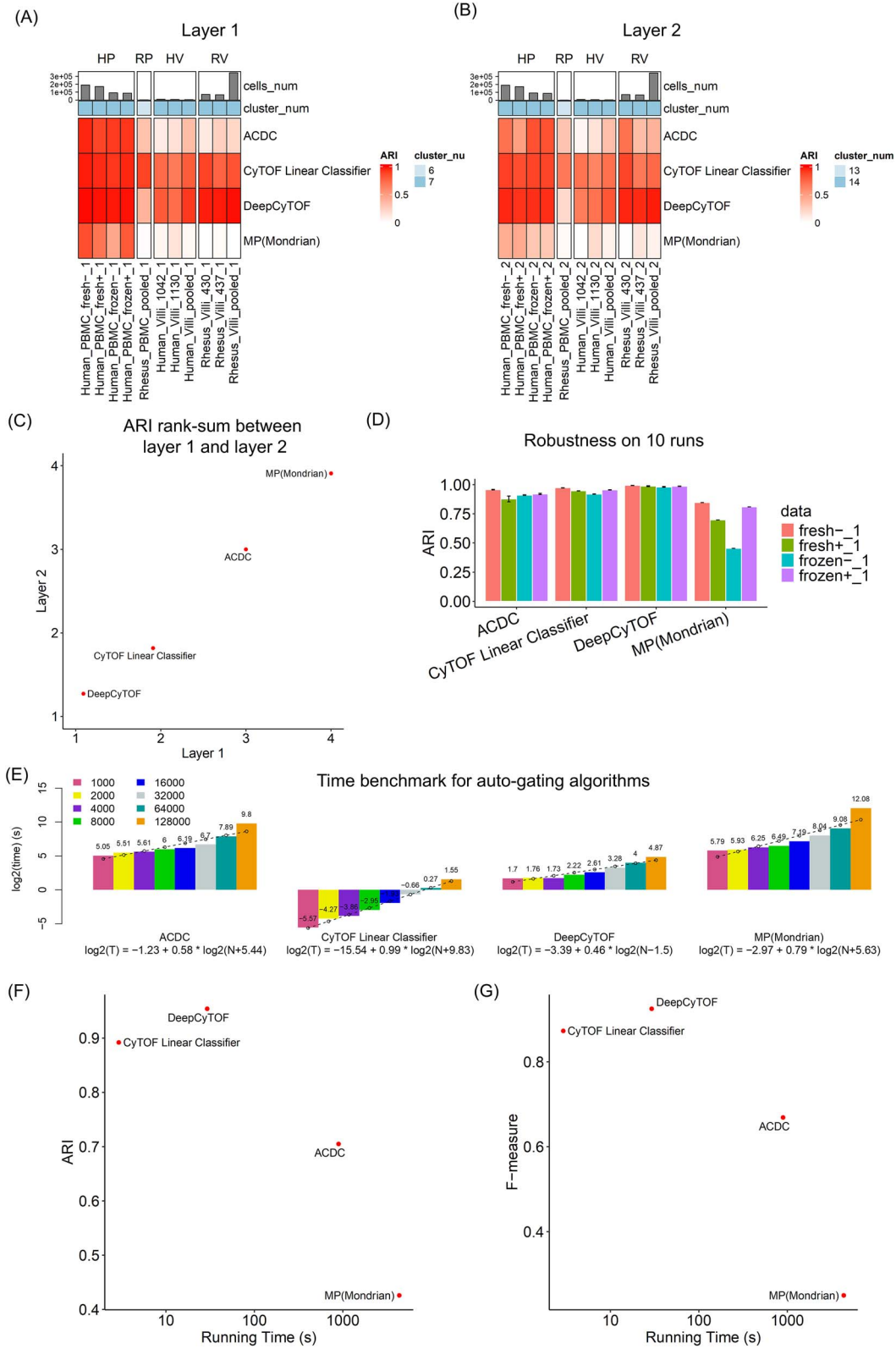


Figure 6. Population detection by supervised auto-gating algorithms. (A) Performance of the auto-gating algorithms based on Layer 1 cell populations using the ARI measure. (B) Performance of the auto-gating algorithms based on Layer 2 cell populations using the ARI measure. (C) Rank-sum between Layer 1 and Layer 2 using the ARI measure. (D) Robustness of the auto-gating algorithms on 10 runs using the ARI measure. (E) Time benchmark for auto-gating algorithms on gradient number of cells. Bar graph represents the real computing time, and the dashed line is for the predicted computing time when fitting into the regression model. (F) the comparison between computing time and ARI performance. (G) The comparison between computing time and F-measure performance.

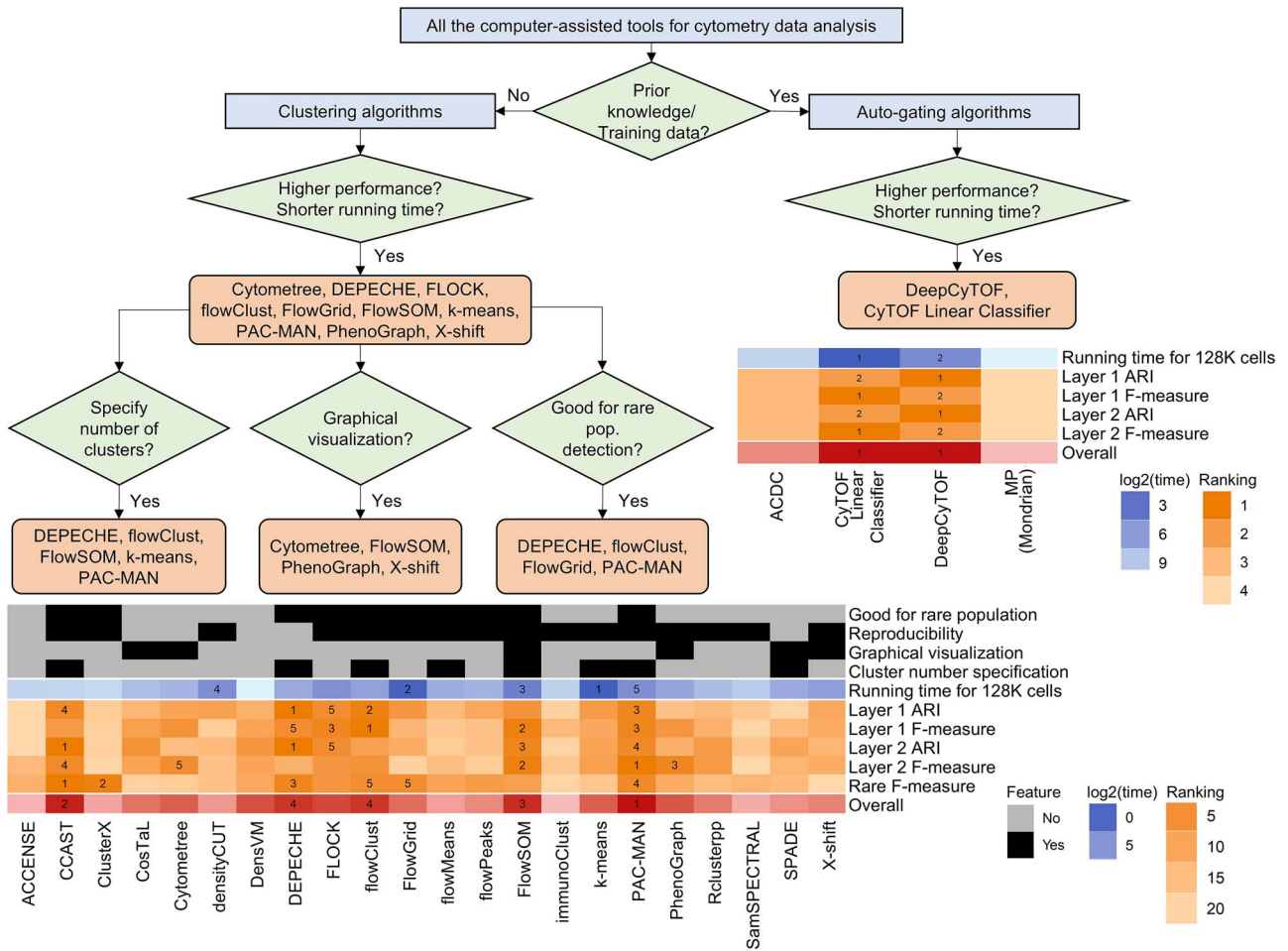


Figure 7. Overall recommendation for the selection of computer-assisted algorithms. Workflow: recommendation of the cytometry tools for different applications. Heatmap for clustering tools: including tool features, rankings for computing time, accuracy evaluation, and overall accuracy ranking. Heatmap for auto-gating tools: rankings for computing time, accuracy evaluation, and overall accuracy ranking.

While this study aimed to provide a comprehensive evaluation of cytometry gating methods, several limitations remain to be addressed. First, given that clustering and auto-gating tools update quickly, we've tried our best to apply the newest version of the software. Tools that were no longer maintained, failed to install on our computer, or could not complete running on some datasets were not included in the final results. Second, although we attempted to include diverse datasets, the comparison of the results may differ by the specific datasets applied. In this study, both our in-house data and public datasets were employed to provide an unbiased evaluation. Third, the truth was generated by the manual gating of the most experienced research in our evaluation. As such, we only focused on the well-defined immune cell populations for the purpose of this study. Lastly, the comparison of auto-gating methods on rare populations was not performed, given the limitation of prior knowledge and training data.

Practical guidelines for computational cytometry data analysis

The experimental design for research and clinical applications of cytometry data was well summarized by previous research [4, 69, 70]. In general, the computational analysis of cytometry data involves three main steps: data preprocessing and normalization, gating and visualization, and downstream analysis. In preprocessing, tools like the R package flowClust [45] are used to import

raw FCS files, followed by quality control (QC) to filter low-quality data, remove dead cells and outliers, and impute missing values. To scale the cytometry data, transformation, normalization, and batch effect correction will then be applied. All these preprocessing steps can be customized by many well-established QC and integration tools, including flowClust [45], PeacoQC [71], ANPELA [72], CATALYST [73], swiftReg [74], CytoNorm [75], Cytotfn [76], and flowAI [77].

After preprocessing, gating will be performed to annotate the cell types of the single-cell cytometry data. This project has comprehensively reviewed manual-gating, 23 clustering, and 4 auto-gating approaches. In addition to the tools evaluated in this study, clustering tools such as Citrus [78] and CellCnn [79] are only compatible with multiple datasets. To optimize the clustering results, tools such as Phenograph and X-shift have the embedded algorithm to determine the optimal number of clusters. Meanwhile, tools such as INFLECT [67] are developed to identify the best number of clusters for FlowSOM. Based on QC and gating, multiple algorithms can be employed to visualize the high-dimensional cytometry data. For example, principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP) are widely applied for dimension reduction. Tools such as CosTaL, Cytometree, FlowSOM, PhenoGraph, SPADE, and X-shift embed graphical learning and visualization algorithms.

Following this, multiple downstream analyses can be applied based on the research purpose and experimental design. For instance, differential abundance and state analysis aim to discover differentially expressed markers across different platforms, experimental conditions, or sample types within cell populations [80]. These analyses can be performed using tools Cydar [81], diffcyt [82], CytoGLMM [83], and CyEMD [80] following the single-cell clustering. Moreover, phenotypic comparisons of cell populations can be achieved by CytoCompare [84]. In addition to the tools that are focused on a specific step of cytometry data analysis, several pipelines have been well established for comprehensive analyses, including platforms such as CytoBank and FlowJo, and computational software like cytofkit [32], PICAFlow [85], CRUSTY [86], ImmunoCluster [87], Cyclone [88], CytoPipeline [89], CYANUS [80], and Cytofast [90].

Future opportunities for cytometry data analysis

Determining the number of clusters is a key challenge in unsupervised learning. In cytometry data analysis for single-cell clustering annotation, certain tools incorporate algorithms to optimize cluster numbers. For example, SPADE allows users to set a desired number of cluster k , while the tool selects the optimal number within the range of $k/2$ and $3k/2$ [56]. Tool INFLECT [67] is specifically designed to optimize the number of clusters for FlowSOM. However, few tools are tailored to general clustering algorithms for cytometry data. By leveraging the clustering algorithms developed for estimating the number of cell types in single-cell RNA-seq data [68], future methods specifically designed for cytometry data clustering will be developed to optimize the number of cell populations.

Graph learning has already played a crucial role in cytometry data analysis. Among the 23 clustering tools, CosTaL, Cytometree, FlowSOM, PhenoGraph, SPADE, and X-shift provide graphical visualizations of the clustering results (Fig. 7). For instance, FlowSOM clusters cytometry data using self-organizing maps (SOMs) [49]. The algorithm first creates an SOM to assign cells to the closest nodes, followed by constructing a minimal spanning tree to connect the nodes into a graph. By integrating advanced graph neural network (GNN) algorithms—successfully applied in single-cell RNA sequencing [91–93] and spatial transcriptomics data [94]—more GNN-based clustering and auto-gating methods are anticipated for cytometry data analysis.

Large language models (LLMs) are artificial intelligence (AI) models designed to process and generate human language. Using LLMs for single-cell analysis is a growing and promising frontier in bioinformatics. For example, several machine learning models based on transformer architectures (like bidirectional encoder representations from transformers (BERT) [95] and generative pre-trained transformer (GPT) [96]) are already being explored for cell-type annotation in single-cell RNA-seq analysis. Moreover, while LLMs specifically for single-cell analysis are still evolving, they show great application potential for cellular interaction and trajectory inference analysis [97, 98]. When evaluating auto-gating methods in this project, state-of-the-art deep learning algorithms have demonstrated their effectiveness in gating cytometry data and automating cell-type annotation. LLMs and other AI models [99] are expected to play a significant role in future systematic cytometry data analysis, potentially integrating with traditional clustering methods and auto-gating algorithms to enhance cell-type annotation.

Spatial tissue cytometry is an advanced imaging technology that analyzes the spatial organization of cells and their molecular properties within tissue samples [100–103]. As a cutting-edge

method for spatial proteomics [104], it merges traditional flow and mass cytometry techniques with high-resolution imaging to preserve the spatial context of cells within tissues. However, current algorithms for spatial tissue cytometry analysis are limited. Models designed for spatial transcriptomics cannot fully capture the unique characteristics of single-cell cytometry data, and the clustering and auto-gating methods evaluated in this project are not equipped to account for spatially resolved cell interactions. Therefore, new computational approaches are urgently needed to address cellular heterogeneity, cell–cell interactions, and tissue microenvironments. In addition, leveraging the technology revolution for single-cell and spatial multi-omics [105, 106], computational integration of multimodalities across diverse platforms holds great promise in unraveling molecular mechanisms and therapeutic targets for precision medicine [107, 108].

Key Points

- This study provided a quantitative review and comparison of various ways to phenotype cellular populations within the cytometry data. Manual gating (5 raters), unsupervised clustering (23 tools), and auto-gating (4 tools) methods were systematically reviewed and compared. To the best of our knowledge, such a comprehensive evaluation has not been performed before.
- Tools were evaluated by both in-house and public datasets, including multiple species (human, mouse, and nonhuman primates) and cell types (peripheral blood mononuclear cells, placental villi, and bone marrow).
- Multiple evaluation measurements (F-measure, adjusted Rand index, Cohen's kappa index) were employed. Computing time was evaluated on a grid of cell numbers to estimate the scalability of the tools for ultra-large cell number applications in the future.
- All programming scripts for tool implementation and comparison were made available on GitHub and Zenodo.
- We further provided practical recommendations on prioritizing gating methods based on different application scenarios. This study offers comprehensive insights for biologists to understand diverse gating methods and choose the best-suited ones for their applications.

Acknowledgements

We thank Stephanie Stras for helping with manual gating.

Author contributions

Conceptualization: P.L., G.C.T., L.K. and S.L.; Data curation: J.T., L.K., E.G.S., B.M., V.M., O.O., Stephanie S, P.P., S.B.S. and S.K.; Formal analysis: P.L., Y.P., H.C., W.W., Y.F., X.X., J.Z., J.L., and S.L. Funding acquisition: S.L.; Investigation: All; Methodology: All; Project administration: S.B.S, P.P., S.K., G.C.T., L.K. and S.L.; Resources: J.T., L.K., E.G.S., B.M., V.M., O.O., Stephanie S, P.P. and S.B.S.; Software: P.L., Y.P., H.C., W.W., Y.F., X.X., J.Z., J.L., and S.L.; Supervision: G.C.T., L.K. and S.L.; Validation: Y.P., H.C. and S.L.; Visualization: P.L., Y.P., H.C., W.W. and S.L.; Writing: All.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Funding

This research was supported in part by the Competitive Medical Research Fund (CMRF) of the UPMC Health System (to S.L.) and the University of Pittsburgh Center for Research Computing with HTC cluster (NIH award number S10OD028483).

Data availability

All animals analyzed in this study are stored in accordance with Institutional Animal Care and Use Committee (IACUC) guidelines at the University of California Davis and endorsed by the University of California, Los Angeles (protocol #20330 and #22121). Requests can be directed to Liza Konnikova (Liza.konnikova@chp.edu). Human blood samples were collected at Boston Children's Hospital under Institutional Review Board (IRB) protocol IRB-P00000529, and placental villi samples were collected under IRB approval at the University of Pittsburgh. Cytometry data for our in-house datasets are publicly available on Cytobank (Beckman Coulter). All the script files are available at GitHub (https://github.com/hung-ching-chang/GatingMethod_evaluation), Zenodo (<https://zenodo.org/records/13851548>), and supplementary files.

References

- Bandura DR, Baranov VI, Ornatsky OI. et al. Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal Chem* 2009;**81**:6813–22. <https://doi.org/10.1021/ac901049w>.
- McKinnon KM. Flow cytometry: an overview. *Curr Protoc Immunol* 2018;**120**:5.1.1–11. <https://doi.org/10.1002/cpim.40>.
- Adan A, Alizada G, Kiraz Y. et al. Flow cytometry: basic principles and applications. *Crit Rev Biotechnol* 2017;**37**:163–76. <https://doi.org/10.3109/07388551.2015.1128876>.
- Manohar SM, Shah P, Nair A. Flow cytometry: principles, applications and recent advances. *Bioanalysis* 2021;**13**:181–98. <https://doi.org/10.4155/bio-2020-0267>.
- Hartmann FJ, Bendall SC. Immune monitoring using mass cytometry and related high-dimensional imaging approaches. *Nat Rev Rheumatol* 2020;**16**:87–99. <https://doi.org/10.1038/s41584-019-0338-z>.
- Sahir F, Mateo JM, Steinhoff M. et al. Development of a 43 color panel for the characterization of conventional and unconventional T-cell subsets, B cells, NK cells, monocytes, dendritic cells, and innate lymphoid cells using spectral flow cytometry. *Cytometry A* 2020;**105**:404–10. <https://doi.org/10.1002/cyto.a.24288>.
- Bendall SC, Nolan GP, Roederer M. et al. A deep profiler's guide to cytometry. *Trends Immunol* 2012;**33**:323–32. <https://doi.org/10.1016/j.it.2012.02.010>.
- Spitzer MH, Nolan GP. Mass cytometry: single cells, many features. *Cell* 2016;**165**:780–91. <https://doi.org/10.1016/j.cell.2016.04.019>.
- Simoni Y, Chng MHY, Li S. et al. Mass cytometry: a powerful tool for dissecting the immune landscape. *Curr Opin Immunol* 2018;**51**:187–96. <https://doi.org/10.1016/j.coi.2018.03.023>.
- Pedersen CB, Olsen LR. Analysis of Mass Cytometry Data. *Mass Cytometry: Methods and Protocols*, 2019. 267–79. <https://link.springer.com/book/10.1007/978-1-4939-9454-0>.
- Liu P, Liu S, Fang Y. et al. Recent advances in computer-assisted algorithms for cell subtype identification of cytometry data. *Front Cell Dev Biol* 2020;**8**:234. <https://doi.org/10.3389/fcell.2020.00234>.
- Mair F, Hartmann FJ, Mrdjen D. et al. The end of gating? An introduction to automated analysis of high dimensional cytometry data. *Eur J Immunol* 2016;**46**:34–43. <https://doi.org/10.1002/eji.201545774>.
- Kimball AK, Oko LM, Bullock BL. et al. A Beginner's guide to Analyzing and visualizing mass cytometry data. *J Immunol* 2018;**200**:3–22. <https://doi.org/10.4049/jimmunol.1701494>.
- Saeyns Y, Van Gassen S, Lambrecht BN. Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nat Rev Immunol* 2016;**16**:449–62. <https://doi.org/10.1038/nri.2016.56>.
- Todorov H, Saeyns Y. Computational approaches for high-throughput single-cell data analysis. *FEBS J* 2019;**286**:1451–67. <https://doi.org/10.1111/febs.14613>.
- Palit S, Heuser C, de Almeida GP. et al. Meeting the challenges of high-dimensional single-cell data analysis in immunology. *Front Immunol* 2019;**10**:1515. <https://doi.org/10.3389/fimmu.2019.01515>.
- Nowicka M, Krieg C, Weber LM. et al. CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets. *F1000Res* 2017;**6**:748. <https://doi.org/10.12688/f1000research.11622.1>.
- Chester C, Maecker HT. Algorithmic tools for mining high-dimensional cytometry data. *J Immunol* 2015;**195**:773–9. <https://doi.org/10.4049/jimmunol.1500633>.
- Montante S, Brinkman RR. Flow cytometry data analysis: recent tools and algorithms. *Int J Lab Hematol* 2019;**41** Suppl 1:56–62. <https://doi.org/10.1111/ijlh.13016>.
- Mair F. Gate to the future: computational analysis of Immunophenotyping data. *Cytometry A* 2019;**95**:147–9. <https://doi.org/10.1002/cyto.a.23700>.
- Weber LM, Robinson MD. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry A* 2016;**89**:1084–96. <https://doi.org/10.1002/cyto.a.23030>.
- Liu X, Song W, Wong BY. et al. A comparison framework and guideline of clustering methods for mass cytometry data. *Genome Biol* 2019;**20**:297. <https://doi.org/10.1186/s13059-019-1917-7>.
- Cheung M, Campbell JJ, Whitby L. et al. Current trends in flow cytometry automated data analysis software. *Cytometry A* 2021;**99**:1007–21. <https://doi.org/10.1002/cyto.a.24320>.
- Cheung M, Campbell JJ, Thomas RJ. et al. Assessment of automated flow cytometry data analysis tools within cell and gene therapy manufacturing. *Int J Mol Sci* 2022;**23**:3224. <https://doi.org/10.3390/ijms23063224>.
- Toothaker JM, Presicce P, Cappelletti M. et al. Immune cells in the placental villi contribute to intra-amniotic inflammation. *Front Immunol* 2020;**11**:866. <https://doi.org/10.3389/fimmu.2020.00866>.
- Konnikova L, Boschetti G, Rahman A. et al. High-dimensional immune phenotyping and transcriptional analyses reveal robust recovery of viable human immune and epithelial cells from frozen gastrointestinal tissue. *Mucosal Immunol* 2018;**11**:1684–93. <https://doi.org/10.1038/s41385-018-0047-y>.
- Toothaker JM, Olaloye O, McCourt BT. et al. Immune landscape of human placental villi using single-cell analysis. *Development* 2022;**149**:dev200013. <https://doi.org/10.1242/dev.200013>.

28. Stras SF, Werner L, Tothaker JM. et al. Maturation of the human intestinal immune system occurs early in Fetal development. *Dev Cell* 2019;**51**:357–373.e5. <https://doi.org/10.1016/j.devcel.2019.09.008>.
29. Levine JH, Simonds EF, Bendall SC. et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* 2015;**162**:184–97. <https://doi.org/10.1016/j.cell.2015.05.047>.
30. Samusik N, Good Z, Spitzer MH. et al. Automated mapping of phenotype space with single-cell data. *Nat Methods* 2016;**13**:493–6. <https://doi.org/10.1038/nmeth.3863>.
31. Maecker HT, McCoy JP, Nussenblatt R. Standardizing immunophenotyping for the human immunology project. *Nat Rev Immunol* 2012;**12**:191–200. <https://doi.org/10.1038/nri3158>.
32. Chen H, Lau MC, Wong MT. et al. Cytofkit: a Bioconductor package for an integrated mass cytometry data analysis pipeline. *PLoS Comput Biol* 2016;**12**:e1005112. <https://doi.org/10.1371/journal.pcbi.1005112>.
33. Rybakowska P, van Gassen S, Quintelier K. et al. Data processing workflow for large-scale immune monitoring studies by mass cytometry. *Comput Struct Biotechnol J* 2021;**19**:3160–75. <https://doi.org/10.1016/j.csbj.2021.05.032>.
34. Hahne F, LeMeur N, Brinkman RR. et al. flowCore: a Bioconductor package for high throughput flow cytometry. *BMC Bioinformatics* 2009;**10**:1–8. <https://doi.org/10.1186/1471-2105-10-106>.
35. Kotecha N, Krutzik PO, Irish JM. Web-based analysis and publication of flow cytometry experiments. *Curr Protoc Cytom* 2010;**53**:Unit10 17. <https://doi.org/10.1002/0471142956.cy1017.s53>.
36. van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;**9**:2579–605.
37. Shekhar K, Brodin P, Davis MM. et al. Automatic classification of cellular expression by nonlinear stochastic embedding (ACCENSE). *Proc Natl Acad Sci USA* 2014;**111**:202–7. <https://doi.org/10.1073/pnas.1321405111>.
38. Anchang B, do MT, Zhao X. et al. CCAST: a model-based gating strategy to isolate homogeneous subpopulations in a heterogeneous population of single cells. *PLoS Comput Biol* 2014;**10**:e1003664. <https://doi.org/10.1371/journal.pcbi.1003664>.
39. Li Y, Nguyen J, Anastasiu DC. et al. CosTaL: an accurate and scalable graph-based clustering algorithm for high-dimensional single-cell data analysis. *Brief Bioinform* 2023;**24**:bbad157. <https://doi.org/10.1093/bib/bbad157>.
40. Commenges D, Alkhassim C, Gottardo R. et al. Cytometree: a binary tree algorithm for automatic gating in cytometry analysis. *Cytometry A* 2018;**93**:1132–40. <https://doi.org/10.1002/cyto.a.23601>.
41. Ding J, Shah S, Condon A. densityCut: an efficient and versatile topological approach for automatic clustering of biological data. *Bioinformatics* 2016;**32**:2567–76. <https://doi.org/10.1093/bioinformatics/btw227>.
42. Becher B, Schlitzer A, Chen J. et al. High-dimensional analysis of the murine myeloid cell system. *Nat Immunol* 2014;**15**:1181–9. <https://doi.org/10.1038/ni.3006>.
43. Theorell A, Bryceson YT, Theorell J. Determination of essential phenotypic elements of clusters in high-dimensional entities-DEPECHE. *PLoS One* 2019;**14**:e0203247. <https://doi.org/10.1371/journal.pone.0203247>.
44. Qian Y, Wei C, Eun-Hyung Lee F. et al. Elucidation of seventeen human peripheral blood B-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data. *Cytometry B Clin Cytom* 2010;**78B**:S69–82. <https://doi.org/10.1002/cyto.b.20554>.
45. Lo K, Hahne F, Brinkman RR. et al. flowClust: a Bioconductor package for automated gating of flow cytometry data. *BMC Bioinformatics* 2009;**10**:145. <https://doi.org/10.1186/1471-2105-10-145>.
46. Ye X, Ho JWK. Ultrafast clustering of single-cell flow cytometry data using FlowGrid. *BMC Syst Biol* 2019;**13**:35. <https://doi.org/10.1186/s12918-019-0690-2>.
47. Aghaeepour N, Nikolic R, Hoos HH. et al. Rapid cell population identification in flow cytometry data. *Cytometry A* 2011;**79A**:6–13. <https://doi.org/10.1002/cyto.a.21007>.
48. Ge Y, Sealfon SC. flowPeaks: a fast unsupervised clustering for flow cytometry data via K-means and density peak finding. *Bioinformatics* 2012;**28**:2052–8. <https://doi.org/10.1093/bioinformatics/bts300>.
49. Van Gassen S, Callebaut B, Van Helden MJ. et al. FlowSOM: using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry A* 2015;**87**:636–45. <https://doi.org/10.1002/cyto.a.22625>.
50. Quintelier K, Couckuyt A, Emmaneel A. et al. Analyzing high-dimensional cytometry data using FlowSOM. *Nat Protoc* 2021;**16**:3775–801. <https://doi.org/10.1038/s41596-021-00550-0>.
51. Sorensen T, Baumgart S, Durek P. et al. immunoClust—an automated analysis pipeline for the identification of immunophenotypic signatures in high-dimensional cytometric datasets. *Cytometry A* 2015;**87**:603–15. <https://doi.org/10.1002/cyto.a.22626>.
52. Murphy RF. Automated identification of subpopulations in flow cytometric list mode data using cluster analysis. *Cytometry* 1985;**6**:302–9. <https://doi.org/10.1002/cyto.990060405>.
53. Li YH, Li D, Samusik N. et al. Scalable multi-sample single-cell data analysis by partition-assisted clustering and multiple alignments of networks. *PLoS Comput Biol* 2017;**13**:e1005875. <https://doi.org/10.1371/journal.pcbi.1005875>.
54. Linderman M. Rclusterpp: Linkable C++ ClusteringR package version 0.22013. 3. <https://rdrr.io/cran/Rclusterpp/man/Rclusterpppackage.html>.
55. Zare H, Shooshtari P, Gupta A. et al. Data reduction for spectral clustering to analyze high throughput flow cytometry data. *BMC Bioinformatics* 2010;**11**:403. <https://doi.org/10.1186/1471-2105-11-403>.
56. Qiu P, Simonds EF, Bendall SC. et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat Biotechnol* 2011;**29**:886–91. <https://doi.org/10.1038/nbt.1991>.
57. Mosmann TR, Naim I, Rebhahn J. et al. SWIFT-scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, part 2: biological evaluation. *Cytometry A* 2014;**85**:422–33. <https://doi.org/10.1002/cyto.a.22445>.
58. Li H, Shaham U, Stanton KP. et al. Gating mass cytometry data by deep learning. *Bioinformatics* 2017;**33**:3423–30. <https://doi.org/10.1093/bioinformatics/btx448>.
59. Abdelaal T, van Unen V, Höllt T. et al. Predicting cell populations in single cell mass cytometry data. *Cytometry A* 2019;**95**:769–81. <https://doi.org/10.1002/cyto.a.23738>.
60. Lee HC, Kosoy R, Becker CE. et al. Automated cell type discovery and classification through knowledge transfer. *Bioinformatics* 2017;**33**:1689–95. <https://doi.org/10.1093/bioinformatics/btx054>.
61. Ji D, Nalisnick E, Qian Y, Scheuermann RH, Smyth P. Bayesian trees for automated cytometry data analysis. *Proceedings of*

- the 3rd Machine Learning for Healthcare Conference 2018;**85**: 465–83.
62. Finak G, Frelinger J, Jiang W. et al. OpenCyto: an open source infrastructure for scalable, robust, reproducible, and automated, end-to-end flow cytometry data analysis. *PLoS Comput Biol* 2014;**10**:e1003806. <https://doi.org/10.1371/journal.pcbi.1003806>.
 63. Lux M, Brinkman RR, Chauve C. et al. flowLearn: fast and precise identification and quality checking of cell populations in flow cytometry. *Bioinformatics* 2018;**34**:2245–53. <https://doi.org/10.1093/bioinformatics/bty082>.
 64. Vinh NX, Epps J, Bailey J. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In: *Proceedings of the 26th annual international conference on machine learning*, 2009 **2009** (pp. 1073–1080). Association for Computing Machinery (ACM).
 65. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)* 2012;**22**:276–82. <https://doi.org/10.11613/BM.2012.031>.
 66. Sasaki Y. The truth of the F-measure. *Teach tutor mater*, 1 (5), 1–5. 2007, APPENDICES.
 67. Verhoeff J, Abeln S, Garcia-Vallejo JJ. INFLECT: an R-package for cytometry cluster evaluation using marker modality. *BMC Bioinformatics* 2022;**23**:487. <https://doi.org/10.1186/s12859-022-05018-w>.
 68. Yu L, Cao Y, Yang JYH. et al. Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data. *Genome Biol* 2022;**23**:49. <https://doi.org/10.1186/s13059-022-02622-0>.
 69. Rybakowska P, Alarcon-Riquelme ME, Maranon C. Approaching mass cytometry translational studies by experimental and data curation settings. *Methods Mol Biol* 2024;**2779**:369–94. https://doi.org/10.1007/978-1-0716-3738-8_17.
 70. Flores-Gonzalez J, Cancino-Diaz JC, Chavez-Galan L. Flow cytometry: from experimental design to its application in the diagnosis and monitoring of respiratory diseases. *Int J Mol Sci* 2020;**21**:8830. <https://doi.org/10.3390/ijms21228830>.
 71. Emmaneel A, Quintelier K, Sichien D. et al. PeacoQC: peak-based selection of high quality cytometry data. *Cytometry A* 2022;**101**:325–38. <https://doi.org/10.1002/cyto.a.24501>.
 72. Zhang Y, Sun H, Lian X. et al. ANPELA: significantly enhanced quantification tool for cytometry-based single-cell proteomics. *Adv Sci (Weinh)* 2023;**10**:e2207061. <https://doi.org/10.1002/advs.202207061>.
 73. Crowell HL, Chevrier S, Jacobs A. et al. An R-based reproducible and user-friendly preprocessing pipeline for CyTOF data. *F1000Res* 2020;**9**:1263. <https://doi.org/10.12688/f1000research.26073.1>.
 74. Rebhahn JA, Quataert SA, Sharma G. et al. SwiftReg cluster registration automatically reduces flow cytometry data variability including batch effects. *Commun Biol* 2020;**3**:218. <https://doi.org/10.1038/s42003-020-0938-9>.
 75. Van Gassen S, Gaudilliere B, Angst MS. et al. CytoNorm: a normalization algorithm for cytometry data. *Cytometry A* 2020;**97**: 268–78. <https://doi.org/10.1002/cyto.a.23904>.
 76. Lo YC, Keyes TJ, Jager A. et al. Cytofln enables integrated analysis of public mass cytometry datasets using generalized anchors. *Nat Commun* 2022;**13**:934. <https://doi.org/10.1038/s41467-022-28484-5>.
 77. Monaco G, Chen H, Poidinger M. et al. flowAI: automatic and interactive anomaly discerning tools for flow cytometry data. *Bioinformatics* 2016;**32**:2473–80. <https://doi.org/10.1093/bioinformatics/btw191>.
 78. Bruggner RV, Bodenmiller B, Dill DL. et al. Automated identification of stratifying signatures in cellular subpopulations. *Proc Natl Acad Sci USA* 2014;**111**:E2770–7. <https://doi.org/10.1073/pnas.1408792111>.
 79. Arvaniti E, Claassen M. Sensitive detection of rare disease-associated cell subsets via representation learning. *Nat Commun* 2017;**8**:14825. <https://doi.org/10.1038/ncomms14825>.
 80. Arend L, Bernett J, Manz Q. et al. A systematic comparison of novel and existing differential analysis methods for CyTOF data. *Brief Bioinform* 2022;**23**:bbab471. <https://doi.org/10.1093/bib/bbab471>.
 81. Lun ATL, Richard AC, Marioni JC. Testing for differential abundance in mass cytometry data. *Nat Methods* 2017;**14**:707–9. <https://doi.org/10.1038/nmeth.4295>.
 82. Weber LM, Nowicka M, Soneson C. et al. Diffcyt: differential discovery in high-dimensional cytometry via high-resolution clustering. *Commun Biol* 2019;**2**:183. <https://doi.org/10.1038/s42003-019-0415-5>.
 83. Seiler C, Ferreira AM, Kronstad LM. et al. CytoGLMM: conditional differential analysis for flow and mass cytometry experiments. *BMC Bioinformatics* 2021;**22**:137. <https://doi.org/10.1186/s12859-021-04067-x>.
 84. Platon L, Pejowski D, Gautreau G. et al. A computational approach for phenotypic comparisons of cell populations in high-dimensional cytometry data. *Methods* 2018;**132**:66–75. <https://doi.org/10.1016/j.ymeth.2017.09.005>.
 85. Regnier P, Marques C, Saadoun D. PICAFlow: a complete R workflow dedicated to flow/mass cytometry data, from pre-processing to deep and comprehensive analysis. *Bioinform Adv* 2023;**3**:vbad177. <https://doi.org/10.1093/bioadv/vbad177>.
 86. Puccio S, Grillo G, Alvisi G. et al. CRUSTY: a versatile web platform for the rapid analysis and visualization of high-dimensional flow cytometry data. *Nat Commun* 2023;**14**:5102. <https://doi.org/10.1038/s41467-023-40790-0>.
 87. Opzomer JW, Timms JA, Blighe K. et al. ImmunoCluster provides a computational framework for the nonspecialist to profile high-dimensional cytometry data. *elife* 2021;**10**:e62915. <https://doi.org/10.7554/eLife.62915>.
 88. Patel RK, Jaszcak RG, Im K. et al. Cyclone: an accessible pipeline to analyze, evaluate, and optimize multiparametric cytometry data. *Front Immunol* 2023;**14**:1167241. <https://doi.org/10.3389/fimmu.2023.1167241>.
 89. Hauchamps P, Bayat B, Delandre S. et al. CytoPipeline and CytoPipelineGUI: a Bioconductor R package suite for building and visualizing automated pre-processing pipelines for flow cytometry data. *BMC Bioinformatics* 2024;**25**:80. <https://doi.org/10.1186/s12859-024-05691-z>.
 90. Beyrend G, Stam K, Höllt T. et al. Cytofast: a workflow for visual and quantitative analysis of flow and mass cytometry data to discover immune signatures and correlations. *Comput Struct Biotechnol J* 2018;**16**:435–42. <https://doi.org/10.1016/j.csbj.2018.10.004>.
 91. Gu H, Cheng H, Ma A. et al. scGNN 2.0: a graph neural network tool for imputation and clustering of single-cell RNA-Seq data. *Bioinformatics* 2022;**38**:5322–5. <https://doi.org/10.1093/bioinformatics/btac684>.
 92. Wang J, Ma A, Chang Y. et al. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. *Nat Commun* 2021;**12**:1882. <https://doi.org/10.1038/s41467-021-22197-x>.
 93. Ma A, Wang X, Li J. et al. Single-cell biological network inference using a heterogeneous graph transformer. *Nat Commun* 2023;**14**:964. <https://doi.org/10.1038/s41467-023-36559-0>.

94. Liu T, Fang ZY, Zhang Z. et al. A comprehensive overview of graph neural network-based approaches to clustering for spatial transcriptomics. *Comput Struct Biotechnol J* 2024;**23**: 106–28. <https://doi.org/10.1016/j.csbj.2023.11.055>.
95. Yang F, Wang W, Wang F. et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat Mach Intell* 2022;**4**:852. <https://doi.org/10.1038/s42256-022-00534-z>.
96. Hou W, Ji Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nat Methods* 2024;**21**:1462–5. <https://doi.org/10.1038/s41592-024-02235-4>.
97. Liu J, Yang M, Yu Y. et al. Large language models in bioinformatics: applications and perspectives. *ArXiv* 2024;**2401**:04155v1.
98. Wang Y, Chen X, Zheng Z. et al. scGREAT: transformer-based deep-language model for gene regulatory network inference from single-cell transcriptomics. *iScience* 2024;**27**:109352. <https://doi.org/10.1016/j.isci.2024.109352>.
99. Ng DP, Simonson PD, Tarnok A. et al. Recommendations for using artificial intelligence in clinical flow cytometry. *Cytometry B Clin Cytom* 2024;**106**:228–38. <https://doi.org/10.1002/cyto.b.22166>.
100. Schulz D, Zanotelli VRT, Fischer JR. et al. Simultaneous multiplexed imaging of mRNA and proteins with subcellular resolution in breast cancer tissue samples by mass cytometry. *Cell Syst* 2018;**6**:25–36.e5. <https://doi.org/10.1016/j.cels.2017.12.001>.
101. Patel J, Deng J, Kambala A. et al. Spatial mass cytometry-based single-cell imaging reveals a disrupted epithelial-immune Axis in Prurigo Nodularis. *J Invest Dermatol* 2024;**144**:2501–2512.e4. <https://doi.org/10.1016/j.jid.2024.01.036>.
102. Kuett L, Catena R, Özcan A. et al. Three-dimensional imaging mass cytometry for highly multiplexed molecular and cellular mapping of tissues and the tumor microenvironment. *Nat Can* 2022;**3**:122–33. <https://doi.org/10.1038/s43018-021-00301-w>.
103. Ali HR, Jackson HW, Zanotelli VRT. et al. Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. *Nat Can* 2020;**1**:163–75. <https://doi.org/10.1038/s43018-020-0026-6>.
104. Bressan D, Battistoni G, Hannon GJ. The dawn of spatial omics. *Science* 2023;**381**:eabq4964. <https://doi.org/10.1126/science.abq4964>.
105. Vandereyken K, Sifrim A, Thienpont B. et al. Methods and applications for single-cell and spatial multi-omics. *Nat Rev Genet* 2023;**24**:494–515. <https://doi.org/10.1038/s41576-023-00580-2>.
106. Baysoy A, Bai Z, Satija R. et al. The technological landscape and applications of single-cell multi-omics. *Nat Rev Mol Cell Biol* 2023;**24**:695–713. <https://doi.org/10.1038/s41580-023-00615-w>.
107. Stanojevic S, Li Y, Ristivojevic A. et al. Computational methods for single-cell multi-omics integration and alignment. *Genomics Proteomics Bioinformatics* 2022;**20**:836–49. <https://doi.org/10.1016/j.gpb.2022.11.013>.
108. Adossa N, Khan S, Rytönen KT. et al. Computational strategies for single-cell multi-omics integration. *Comput Struct Biotechnol J* 2021;**19**:2588–96. <https://doi.org/10.1016/j.csbj.2021.04.060>.