Journal of Biomedical Informatics 147 (2023) 104510

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

MAGNETO: Cell type marker panel generator from single-cell transcriptomic data

Andrea Tangherloni [a,b,f,*], Simone G. Riva [c], Brynelle Myers [d], Francesca M. Buffa [a,b,e], Paolo Cazzaniga [f,g,**]

[a] Department of Computing Sciences, Bocconi University, Via Guglielmo Röntgen 1, Milan, 20136, Italy
[b] Bocconi Institute for Data Science and Analytics, Bocconi University, Via Guglielmo Röntgen 1, Milan, 20136, Italy
[c] Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford, Headley Way, Oxford, OX3 9DS, United Kingdom
[d] Wellcome Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford, OX3 7BN, United Kingdom
[e] Department of Oncology, University of Oxford, Old Road Campus Research Building, Oxford, OX3 7DQ, United Kingdom
[f] Department of Human and Social Sciences, University of Bergamo, Piazzale S. Agostino 2, Bergamo, 24129, Italy
[g] Bicocca Bioinformatics, Biostatistics, and Bioimaging Centre - B4, Via Follereau 3, Vedano al Lambro, 20854, Italy

ARTICLE INFO

Dataset link: https://gitlab.com/andrea-tango/magneto

Keywords:
Single-cell RNA-seq
Marker gene selection
Marker panels
Bioinformatics
Multi-objective optimization

ABSTRACT

Single-cell RNA sequencing experiments produce data useful to identify different cell types, including uncharacterized and rare ones. This enables us to study the specific functional roles of these cells in different microenvironments and contexts. After identifying a (novel) cell type of interest, it is essential to build succinct marker panels, composed of a few genes referring to cell surface proteins and clusters of differentiation molecules, able to discriminate the desired cells from the other cell populations. In this work, we propose a fully-automatic framework called MAGNETO, which can help construct optimal marker panels starting from a single-cell gene expression matrix and a cell type identity for each cell. MAGNETO builds effective marker panels solving a tailored bi-objective optimization problem, where the first objective regards the identification of the genes able to isolate a specific cell type, while the second conflicting objective concerns the minimization of the total number of genes included in the panel. Our results on three public datasets show that MAGNETO can identify marker panels that identify the cell populations of interest better than state-of-the-art approaches. Finally, by fine-tuning MAGNETO, our results demonstrate that it is possible to obtain marker panels with different specificity levels.

1. Introduction

During the past decade, we have seen a continuous increase in the production of high-throughput sequencing data in single cells from a variety of biological models, samples, and experiments, including *ex vivo* or *in vivo* [1–8]. This has provided high-resolution information enabling the discovery of novel and rare cell populations, and revolutionizing the biomedical research field as never before. In addition, the combination of such information with tailored and powerful computational approaches has helped obtain a better description of cellular identity and the functional roles of cell populations, based on both the context and environment where they are [9]. For instance, single-cell data have been used to elucidate the molecular processes that drive both cell development and progression in different pathologies [10,11].

After the initial discovery, precise isolation of the cell populations is fundamental in characterizing each population's specific functional role in different contexts and microenvironments. Furthermore, as the therapy of complex diseases (cancer, metabolic, immune, and heart diseases) moves away from monotherapies towards combinatorial approaches targeting different cell populations, it is increasingly necessary to accurately identify and characterize these cells.

Flow cytometry and imaging for either physical isolation or spatial characterization of these cells are examples of techniques where panels of marker genes can be used to isolate the cell type of interest from the other cell populations, aiming at elucidating different questions of biomedical relevance [12–14]. Succinct marker panels, composed of genes referring to cell surface proteins and clusters of differentiation (CD) molecules [15,16], should be built to enable the identification,

* Corresponding author at: Department of Computing Sciences, Bocconi University, Via Guglielmo Röntgen 1, Milan, 20136, Italy.
** Corresponding author at: Department of Human and Social Sciences, University of Bergamo, Piazzale S. Agostino 2, Bergamo, 24129, Italy.
E-mail addresses: andrea.tangherloni@unibocconi.it (A. Tangherloni), paolo.cazzaniga@unibg.it (P. Cazzaniga).

https://doi.org/10.1016/j.jbi.2023.104510
Received 28 February 2023; Received in revised form 12 September 2023; Accepted 29 September 2023
Available online 4 October 2023

isolation, and validation of the new populations in follow-up studies and experiments [15]. For example, Ranzoni et al. tested and manually refined different marker panels to better delineate intermediate steps and reconstruct the whole differentiation process in hematopoiesis [12]. Similarly, other studies focused on the isolation of the cell population of interest by means of flow cytometry assay, which allowed them to isolate the cells of interest physically and, at the same time, quantify the cell populations or the expression of the markers [17]. Thus, flow cytometry, together with other independent methods [15], can help us perform comprehensive functional studies to validate and further characterize the transcriptomic observations in both research and clinical applications [18]. Besides flow cytometry, multiplex immunohistochemistry (IHC) imaging can be fruitfully used to measure protein abundance at a cellular level in tissue cross-sections, allowing for the identification of the cells in a spatial context [18]. In addition, relying on specific and IHC-compatible antibodies, IHC can be performed using small marker panels [19].

Existing technologies allow fast and affordable acquisition of multiple markers, both in the clinic and in laboratory settings. These markers can identify cell populations from biological specimens and isolate them for phenotypical and functional studies; however, significant challenges remain. Firstly, existing technologies are limited to a relatively small number of markers. Secondly, current panels have emerged heuristically, and there are often significant discrepancies between laboratories. For example, even though the identification of marker panels is a crucial step to enable the functional exploration and characterization of novel cell populations, the standard approach to finding a suitable marker panel consists of a manual screen of hundreds or thousands of genes, sometimes even the full transcriptome, carried out by domain experts. On the one hand, curated databases can be exploited to find information regarding marker genes. For instance, Zhang et al. proposed the CellMarker database to provide the community with a comprehensive annotated resource of possible markers for different cell types in human and mouse tissues [20]. Unfortunately, CellMarker is unsuitable for building panels for novel and rare cell types. Moreover, it cannot be used to analyze other well-studied species (e.g., zebrafish, Drosophila, and Caenorhabditis elegans). Finally, it does not allow building panels able to distinguish the same cell population in different cell states.

Single-cell RNA sequencing (scRNA-Seq) acquires the expression of thousands of genes at a single-cell resolution; thus, methods exploiting this complexity have tremendous potential to address these problems. Therefore, annotated scRNA-Seq data could be used to find new marker panels. Currently, the mainstream methods to build a candidate marker panel from scRNA-Seq data rely on a list of ranked genes, obtained from differential analysis of the gene expression in the desired cell type of interest (see, e.g., [1–4,7,8,14]). This can be a good approach in the first instance but the downside of such approaches is that they do not account for gene combinations. It is worth mentioning that the combination of genes refers to the co-expression of the so-called positive and negative genes, where positive (negative) genes are expressed (not expressed) by a specific cell type. In addition, this strategy requires an extensive, time-consuming, and often subjective manual curation of the ranked list of genes to evaluate *(i)* if the top-ranked genes are suitable for the isolation of the cell population of interest, and *(ii)* their ability to pair with each other to build more selective multi-gene marker panels (see [15] and references therein). Finally, all these methods do not directly test the ability of the genes to isolate the desired cell population from the background (i.e., the other cell types). On the contrary, they only assess with standard statistical tests if the expression of each gene significantly differs from the given background.

So far, some computational studies have been proposed to identify the most promising marker panels efficiently [15–17,21–25]. These tools can be used to explore the high gene coverage provided by scRNA-Seq to both refine the panels of well-known cell types and delineate the cell markers of novel cell types, detected and annotated via cluster analysis of the scRNA-Seq data, to isolate the cell populations better using subsequent antibody-based flow cytometry or IHC imaging. However, in such a context, providing an effective and efficient computational approach that can help researchers identify the most promising marker panels, from a list of ranked marker panels, remains a difficult task for different reasons. This is inherently combinatorial, which makes it an NP-hard problem [15], and the possible combinations of genes that should be tested become extraordinarily high even considering only the CD molecules (e.g., $\sim 10^{10}$ possible combinations considering panels composed of only 4 genes) [15]. In addition, the availability of experimental reagents (e.g., antibodies for flow cytometry or *in situ* staining) is still limited, forcing the computational approaches to propose a (generally long) list of possibly ranked candidate marker panels, which must be then assessed and validated for both their accuracy and the availability of the reagents by expert biologists.

Problem	Building optimal marker panels starting from a single-cell gene expression matrix and a cell type identity for each cell
What is Already Known	Computational tools based on statistical strategies and optimization algorithms can be used to identify the most promising marker panels
What this Paper Adds	A fully-automatic framework that builds effective marker panels solving a tailored bi-objective optimization problem. + MAGNETO implements different multi-objective optimization algorithms. + It exploits a novel weighted fitness function based on a parameter α, able to balance the contributions of the two components of the fitness function (objective 1), to set the marker panel's specificity. + It implements three binarization strategies to deal with different (input) gene expression matrices. + It can read and handle different input files. + It reports the calculated marker panels ranked by their capability to isolate the desired cell population. + It can analyze multiple clusters in parallel, allowing for building the marker panels for all the clusters of small scRNA-Seq datasets in parallel. + It is modular, meaning that it can be easily extended to accommodate new objectives.

In this study, we propose a novel fully-automatic framework called marker panel generator with multi-Objective optimization (MAGNETO), which is based on the computational strategy presented in [16]. MAGNETO can help the researchers to identify the most promising marker panels, starting from a gene expression matrix with single-cell resolution data, and a cell type identity for each cell. MAGNETO encodes the marker panel construction problem as a bi-objective optimization problem, where the first objective regards the identification of the genes able to isolate a specific cell type, while the second objective concerns the minimization of the total number of genes included in the panel. Such a bi-objective problem is then solved with a Multi-objective Evolutionary Algorithm (MOEA) called Adaptive Geometry Estimation based on MOEA (AGEMOEA) [26]. At the end of the optimization, MAGNETO provides the user with a ranked list of the identified marker panels. We compare the performance of our framework against the

state-of-the-art approaches, including RANKCORR [23], SMaSH [27], Hypergate [22], sc2marker [17], and COMET [15] on three public datasets, namely: Peripheral Blood Mononuclear Cells (PBMC), Pancreatic Islet Cells (PIC), and Human Fetal Cells (HFC). Our results highlight that MAGNETO can identify marker panels that better characterize the cell populations of interest, compared to the other approaches, in all tested datasets. Moreover, different settings of MAGNETO result in the definition of marker panels with different specificity levels, allowing the user to fine-tune the tool according to their need. The ability of MAGNETO to build effective marker panels is fundamental as these panels can drastically increase the capability of isolating "pure" cell populations to obtain a thorough analysis and characterization of them. Considering that it can be used by novice and expert researchers, it can be easily applied in clinical contexts where rare and transient cell populations need to be isolated. MAGNETO can thus be a starting point for future investigations, and it could be included in studies regarding the characterization of tumor microenvironments, therapeutic resistance, and regenerative medicine. Thanks to its unique features, MAGNETO outperforms the existing state-of-the-art approaches.

The paper is organized as follows, Section 2 briefly describes the computational works related to MAGNETO. Section 3 first presents the datasets employed in this work and introduces the computational problem related to the identification of marker panels, then it describes MAGNETO in detail. Section 5 reports the outcomes obtained with MAGNETO and compared to the state-of-the-art approaches; finally, Section 6 reports some conclusive remarks and directions for future extensions of this work.

2. Related works

The first computational strategies used to identify possible marker genes are the statistical methods included in the Seurat and Scanpy toolkits [28–31]. Such strategies however merely compare the expression distribution of every gene between two different groups of cells (e.g., the cell type of interest and the other cell types). In such a case, the calculated p-values can be combined with the expression fold-change to evaluate the possible marker genes. MAST [32] is a statistical framework that explores a hurdle model specifically adapted for scRNA-Seq data. It is a two-part generalized linear model designed to simultaneously model the rate of the expression over the background of various transcripts and the positive expression mean. MAST assumes that the logarithmic expression follows a normal distribution and, under this assumption, it identifies the genes that are differentially expressed between two groups of cells. As for the other statistical tests, the obtained p-value can be used to rank the candidate genes. CombiROC [21,33] identifies a list of possible markers and computes the sensitivity and specificity values for each possible marker combination. Then, the Receiver Operating Characteristic curves are calculated for each marker and for the selected combinations of markers, previously filtered by using sensitivity and specificity thresholds, to evaluate and rank the calculated panels. Delaney et al. proposed COMET (COmbinatorial Marker dEtection from single-cell Transcriptomics) [15], which tries to build candidate marker panels that allow for distinguishing the desired cell population from the background by using statistical tests proposed for other contexts and purposes. In particular, COMET directly works on a gene-by-cell expression matrix (raw counts or normalized values) and requires a cell type identity for each cell, as well as a list of genes over which to conduct the search to identify the best candidate marker panels for the desired cell type. Hypergate [22] exploits a non-parametric score based on true positives, false positives, false negatives, and true negatives values, specifically combined to take into account both purity (sensitivity) and yield (specificity), to pinpoint the most promising marker panels. RANKCORR [23] is a method that first performs multi-class marker selection by ranking the mRNA counts and considering sparse binomial regression models. It is worth reminding that ranking the genes is the most straightforward non-parametric approach to analyzing and comparing count data. Then, it tries to linearly separate and recognize a small number of marker genes from the ranked values. RANKCORR can be used to build marker panels for the desired cell type by setting the multi-class marker selection as a two-class classification problem. SMaSH [27], instead, employs Machine Learning (ML) approaches to discriminate and rank, in a fast and efficient way, a list of marker genes starting from a cell-by-gene expression matrix and the given cell identity (e.g., cell type or tissue-specific). Based on the selected ML approach, the extracted markers are then ranked according to the gene importance score assigned by the Gini criterion [34] or the Shapley value [35]. Finally, Li et al. proposed sc2marker [17], a feature selection approach that combines the maximum margin index and a database of proteins with antibodies to identify the most promising marker genes for flow cytometry or IHC imaging.

Considering that the identification of candidate marker panels is a combinatorial NP-hard problem, it can also be tackled by exploiting linear programming approaches or global optimization techniques as long as it is reformulated as an optimization problem. On the one hand, Dumitrascu et al. first redefined the identification of candidate marker panels as a liner program problem [24], then they proposed an approach called scGeneFit to solve it. scGeneFit is an efficient method for the marker selection of scRNA-Seq data having a hierarchical partition of the cell types. Specifically, scGeneFit selects the marker genes that jointly optimize the cell type recovery using cell type-aware compressive classification methods. These methods allow for finding a projection of the marker to a low-dimensional subspace where cells having the same annotation are closer to each other than the cells differently annotated. To increase the selection of suitable marker genes, scGeneFit constrains the projection so that each subspace dimension is aligned to a coordinate axis in the original space. Thanks to this specific constraint, each dimension in the low-dimensional space directly captures a single marker, which corresponds to a single gene, and not a weighted linear combination of different genes. On the other hand, PanGA is the first framework that has been designed to solve this problem, reformulated as a single-objective optimization problem, by relying on the optimization capabilities of Genetic Algorithms [25], which were previously applied to tackle different combinatorial problems in biomedical applications (see, e.g., [36–38]). Like COMET, PanGA requires a cell-by-gene expression matrix (raw counts or normalized values) and a cell type assignment for each cell, along with the cell type of interest. Then, it evolves a population of candidate marker panels evaluated according to the designed fitness function. Finally, PanGA provides the user with a list of ranked marker panels, according to their fitness value (the higher the better). In [16], the optimization problem proposed in PanGA has been re-redefined as a bi-objective problem with a modified fitness function to minimize the total number of genes included in the panel. Both these works focused on the comparison of different representations for the candidate solutions, evaluating their impact on the optimization capabilities of the tested objective optimization algorithms.

3. Material and methods

3.1. Datasets

3.1.1. Peripheral Blood Mononuclear Cells

This specific PBMC dataset consists of three thousand cells collected from a Healthy Donor and is freely available from 10× Genomics (https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc3k). The cell type annotation can be performed following Scanpy's tutorial *"Preprocessing and clustering 3k PBMCs"* (https://scanpy-tutorials.readthedocs.io/en/latest/pbmc3k.html).

3.1.2. Pancreatic Islet Cells

The PIC dataset is composed of four different experiments sequenced independently using four different platforms: CEL-Seq [39] (1004 cells), CEL-Seq2 [40] (2285 cells), Fluidigm C1 [41] (638 cells), and Smart-Seq2 [42] (2394 cells). The count matrices can be downloaded from Seurat's tutorial *"Integration and Label Transfer"* (https://satijalab.org/seurat/v3.0/integration.html). We analyzed all the 13 major cell types identified in [29].

3.1.3. Human Fetal Cells

The HFC dataset contains 15 independent scRNA-Seq experiments of human fetal liver and bone marrow [12] samples, sequenced using the SmartSeq2 protocol [43]. The filtered count gene expression matrix, composed of 4504 cells and 29680 genes, was obtained using the downstream analysis proposed in [12], which exploits the Scanpy and scAEspy toolkits [31,44,45]. We grouped some cell subtypes to analyze the major cell types, namely: B cells, Endothelial cells, Erythrocytes, Granulocytes, Hematopoietic Stem Cells/Multipotent Progenitors (HSC/MPPs), Lympho–Myeloid Progenitors (LMPs), Megakaryocyte-Erythroid-Mast Progenitors (MEMPs), Mast cells, Megakaryocytes, Monocytes, Natural Killer (NK) cells, Unspecified, and Plasmacytoid Dendritic Cells (pDCs).

3.2. Multi-objective optimization

Multi-objective Evolutionary Algorithms (MOEAs) are used to tackle multi-objective problems, which are characterized by two or more objective functions that must be simultaneously optimized [46]. In particular, MOEAs aim to determine a set of non-dominated candidate solutions, that is, solutions that cannot improve a single objective without affecting the others. MOEAs make use of a population of candidate solutions evolved to approximate the so-called Pareto optimal set; indeed, there is not a single solution able to simultaneously optimize all the objectives. On the contrary, there exist a set of trade-off solutions, which are known as Pareto optimal solutions.

There are several MOEAs, like the Non-dominated Sorting Genetic Algorithm II (NSGA-II) [47], which follows the general scheme of Genetic Algorithms and uses a crowding distance to explicitly preserve the diversity in the population, or the S metric selection evolutionary multiobjective optimization algorithms (SMSEMOA) [48] that exploits a hypervolume measure to assess the quality of the candidate solutions. MAGNETO employs the AGEMOEA, which can be seen as a modified version of NSGA-II exploiting a survival score that combines both the diversity and proximity of the non-dominated fronts instead of the crowding distance [26]. The first front computed by AGEMOEA is exploited to normalize the objective space, as well as to estimate the Pareto front geometry, by using a fast procedure with reduced computational complexity compared to the majority of the MOEA approaches. In addition, AGEMOEA adapts diversity and proximity metrics according to the estimated front geometry. In particular, the closest solution to the middle of the first front is used to estimate the parameters of the Minkowski p-norm, which is then used to compute the survival score, a measure that combines the distance from the neighbors and proximity to the ideal point (i.e., the origin of the axes). An extension of AGEMOEA was presented in [49]. This method, called AGEMOEA-II, introduces a novel method to model the Pareto front's geometry and calculate the diversity of the candidate solutions.

4. MAGNETO: The proposed pipeline

Fig. 1 shows the workflow of MAGNETO, which requires a gene expression matrix **E** and a cell type assignment (i.e., label) for each cell present in **E**. Differently from PanGA and the work proposed in [16], MAGNETO can read and manage `h5ad` files, which are modified `hdf5` files specifying how AnnData/Scanpy objects have to be stored [31]; it can also directly handle AnnData/Scanpy objects. In these specific cases, the cell type assignment can be a column of the provided AnnData/Scanpy object. Finally, the gene expression and the cell type assignment can be provided as two distinct Pandas dataframes [50,51], TSV files, or CSV files.

We assume here that **E** has C cells and G genes, whose elements $E_{c,g} \in \mathbb{N}$, with $g = 1, \ldots, G$ and $c = 1, \ldots, C$, are greater than zero when the gene g is expressed in cell c; otherwise, $E_{c,g} = 0$. We also assume that the cells are partitioned in L distinct clusters (i.e., cell types) and labeled using L different labels (i.e., $l_1, \ldots, l_L$). A cell type of interest $C \in \{l_1, \ldots, l_L\}$ (i.e., C must be one of the L distinct cell types) can be optionally given as input to MAGNETO. In the case that C is not provided, MAGNETO automatically identifies the optimal marker panels for all the L distinct cell types. Finally, a CSV file containing the identified marker panels, ranked according to their fitness value, is saved.

4.0.1. Individuals encoding and marker panel construction

MAGNETO is an improved version of the evolutionary approach proposed in [16], which exploits AGEMOEA to identify the most promising marker panels. In particular, MAGNETO solves a modified version of the multi-objective optimization problem proposed in [16]. In fact, we introduce here a user-defined parameter α to balance the contributions of the two components of the fitness function defined in Section 4.0.2. Moreover, MAGNETO can use three different binarization strategies to binarize the gene expression matrix and then solve the multi-objective optimization problem with the modified and improved fitness function with respect to that used in [16]. Finally, MAGNETO exploits the Python multiprocessing package to analyze all the clusters of small scRNA-Seq datasets in parallel (this feature is currently available only on Linux-based systems). In the multi-objective formulation of the problem, each individual I of the MOEA represents a candidate solution (i.e., a marker panel) and is encoded as a vector $\mathbf{x} = (x_1, \ldots, x_\rho)$, where ρ is a user-defined parameter used to limit the number of genes composing the solution. The parameter ρ is introduced as most of the existing marker panels include only a few genes [12,52] (e.g., up to 10 genes). This constraint is imposed by the current flow cytometry equipment that does not allow for the validation and use of large panels [52]. However, MAGNETO is able to build large panels, which could potentially provide high-resolution analyses of the desired cell type [52].

A candidate solution $\mathbf{x}$ can be decoded into a set $\mathcal{P}$ of positive genes and a set $\mathcal{N}$ of negative genes, which are then used to evaluate the designed fitness functions. In this context, the terminology positive and negative genes is used to indicate whether a certain gene is expressed by a certain cell or not. To be more precise, positive genes are expressed by most of the cells of type C, while the cells of the other populations might only lowly express this gene. A negative gene instead is expressed by only a few cells of type C but the cells of other cell types highly express this gene. To decode $\mathbf{x}$ into the sets $\mathcal{P}$ and $\mathcal{N}$, each component x_j of $\mathbf{x}$, with $j = 1, \ldots, \rho$, is evaluated. A component x_j can assume values in $\{-G, \ldots, -1, 0, 1, \ldots, G\}$ and is assigned to $\mathcal{P}$ ($\mathcal{N}$) if $x_j > 0$ ($x_j < 0$); the elements of $\mathbf{x}$ whose value is 0 are not taken into account (i.e., they are included in neither $\mathcal{P}$ nor $\mathcal{N}$). Considering that $|\mathbf{x}| = \rho$, it results that $|\mathcal{P}| + |\mathcal{N}| \leq \rho$. Consequently, an intrinsic upper bound to the number of selected genes is imposed on all the individuals, and the resulting marker panels can be composed of no more than ρ genes.

4.0.2. Fitness function definition and evaluation

The first step towards the fitness calculation is binarizing **E** to obtain a binary matrix **B**. The binarization step can be performed using different strategies, and MAGNETO comprises the following three.

- A global threshold value θ equal for all the genes:

$$B_{c,g} = \begin{cases} 1 \text{ if } E_{c,g} > \theta, \\ 0 \text{ otherwise.} \end{cases} \quad (1)$$

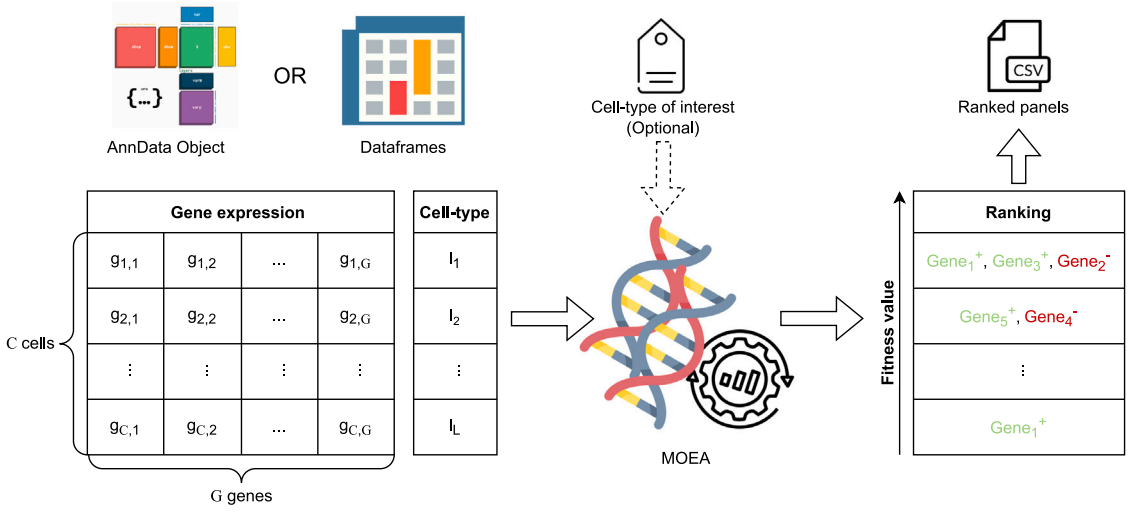


Fig. 1. Workflow of MAGNETO. Given a gene expression matrix (raw counts or normalized values) and a cell type assignment for each cell, MAGNETO identifies the most promising marker panels for either the cell type of interest (if provided) or for all the cell types present in the dataset. The marker panels are ranked according to their fitness value (the higher the better) and a CSV file, containing the ranked marker panels, is saved.

- A specific threshold for each gene. In this case, the first quartile $Q1_g$ is calculated for each gene g and then used to binarize the expression values of the gene g, as follows:

$$B_{c,g} = \begin{cases} 1 \text{ if } E_{c,g} > Q1_g, \\ 0 \text{ otherwise.} \end{cases} \tag{2}$$

- An interactive k-means procedure for each gene. In particular, the k-means algorithm is applied T times to identify 2^t classes, with $t = 1, \dots, T$. At the end of the procedure, the two identified classes are used to binarize the values. In particular, the elements belonging to the class associated with the greater centroid are set to 1, while the other elements are set to 0.

We compared the effect of the binarization strategies on the percentage of cells expressing the best marker panels identified by MAGNETO for all the cell populations of the three datasets. Our results showed that the first quartile $Q1$ strategy is the most robust, allowing for obtaining better results among all the tested cell populations (see Supplementary Figs. 1, 2, and 3); as a consequence, we opted to use this strategy in all the tests shown hereafter. For each cell type of interest $\mathcal{C}$, the corresponding optimal marker panels can be built by identifying the best sets $\mathcal{P}$ and $\mathcal{N}$ that allow for selecting as many cells in $\mathbf{C}$ as possible, while reducing the number of cells in $\mathbf{S}$, where $\mathbf{C} \subset \mathbf{B}$ is the submatrix of $\mathbf{B}$ comprising the cells of type $\mathcal{C}$, while the submatrix $\mathbf{S} = \mathbf{B} \backslash \mathbf{C}$ includes all the other cells.

To evaluate the quality of an individual $\mathbf{x}$, which encodes the sets $\mathcal{P}$ and $\mathcal{N}$, and thus the ability of the corresponding marker panels to isolate as many cells of cell type $\mathcal{C}$ as possible, we apply the following fitness function:

$$f(\mathcal{P}, \mathcal{N}, \mathbf{C}, \mathbf{S}, \alpha) = (1 - \alpha) g(\mathcal{P}, \mathcal{N}, \mathbf{C}) - \alpha g(\mathcal{P}, \mathcal{N}, \mathbf{S}),$$

$$g(\mathcal{P}, \mathcal{N}, \mathbf{A}) = \frac{1}{|\mathbf{A}|} \sum_{i=1}^{|\mathbf{A}|} (A_{i,p_1} \wedge \dots \wedge A_{i,p_{|\mathcal{P}|}}) \wedge \neg (A_{i,n_1} \vee \dots \vee A_{i,n_{|\mathcal{N}|}}).$$

$A_{i,p_1}, \dots, A_{i,p_{|\mathcal{P}|}}$ denote the expression or no expression of the positive genes in $\mathcal{P}$ in the cell A_i of the submatrix $\mathbf{A}$, while $A_{i,n_1}, \dots, A_{i,n_{|\mathcal{N}|}}$ indicate the expression or no expression of the negative genes in $\mathcal{N}$ in cell A_i of the submatrix $\mathbf{A}$. Finally, the user-defined parameter $\alpha \in (0, 1)$ is used to balance the two contributions of the two components of the fitness function, which must be simultaneously optimized. Low values of α allow for retrieving more cells of the cell type $\mathcal{C}$ at the cost of including more cells of the other cell populations. On the contrary, high values of α help limit the number of cells of the other cell populations and reduce the number of cells of the cell type $\mathcal{C}$.

It is worth noting that $f(\cdot)$ can only assume values in $[-1, 1]$, where -1 is reached when only the cells in $\mathbf{S}$ are collected, while the value 1 is obtained when $\mathcal{P}$ and $\mathcal{N}$ are able to exclusively isolate the cells in $\mathbf{C}$, avoiding the cells in $\mathbf{S}$.

Considering that marker panels composed of a few genes are more suitable for the current flow cytometry equipment, the identification of the marker panels can be formulated as a constrained bi-objective problem. The first objective is the maximization of the function f while the second objective is the minimization of the number of genes composing the marker panels. In addition, two constraints are introduced to ensure (*i*) that at least one gene is considered within each individual, and (*ii*) that each gene is either used as a positive or negative gene (more formally, if $g \in \mathcal{P}$ then $g \notin \mathcal{N}$, while if $g \in \mathcal{N}$ then $g \notin \mathcal{P}$). Thus, this constrained bi-objective optimization problem is defined as:

$$\begin{aligned} \max \quad & f_1(\mathcal{P}, \mathcal{N}, \mathbf{C}, \mathbf{S}) := f(\mathcal{P}, \mathcal{N}, \mathbf{C}, \mathbf{S}) \\ \min \quad & f_2(\mathcal{P}, \mathcal{N}) := |\mathcal{P}| + |\mathcal{N}| \\ \text{subject to:} \quad & g_1(\mathcal{P}, \mathcal{N}) := |\mathcal{P}| + |\mathcal{N}| \geq 1 \\ & g_2(\mathcal{P}, \mathcal{N}) := \mathcal{P} \cap \mathcal{N} = 0, \end{aligned}$$

4.0.3. Genetic operators

Similar to [16,25], a binary tournament selection is used to select the parent individuals to generate the offspring population. To recombine the information of the pairs of individuals, the *exponential crossover*, which mainly acts as a single-point crossover but occasionally it becomes a two-point crossover, is applied with a crossover rate p_c. An *ad-hoc* mutation strategy is used to increase the variability of the offspring population by altering one or more values of each selected offspring, enabling the introduction of a higher variability into the population to prevent a premature convergence to local optima. In particular, the designed mutation strategy allows for randomly changing each value with one of the other values in $\{-G, \dots, -1, 0, 1, \dots, G\}$. For each selected offspring, the mutation is applied to all the components of the offspring with the same mutation rate p_m.

5. Results and discussion

To assess the correctness and accuracy of MAGNETO, we execute several tests starting from the datasets described in Section 3.1. Note

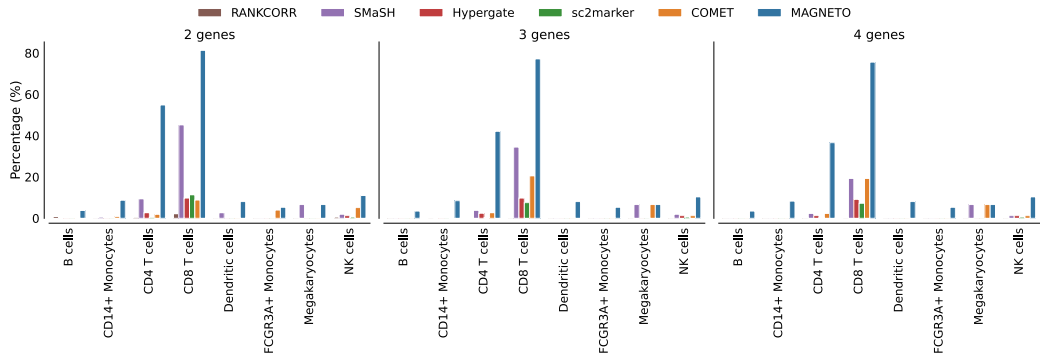
that we only consider the CD molecules (∼400 genes) in all the tests shown hereafter; thus, we filtered out the other genes before applying the tested approaches. Considering the results presented in [16,17], we compared MAGNETO against RANKCORR, SMaSH, Hypergate, sc2marker, and COMET, to thoroughly test its effectiveness in building marker panels.

We first execute the tested methods to identify the optimal marker panels with 2, 3, and 4 genes for the isolation of CD8 *T* cells in the PBMC dataset, alpha cells in the PIC dataset, and B cells in the HFC dataset. The results of this analysis are reported in Fig. 2, where the barplots show the percentage of cells isolated with the genes included in the marker panels built by RANKCORR, SMaSH, Hypergate, sc2marker, COMET, and MAGNETO. We observe that in all cases the percentage of cells of interest isolated with the panel generated by MAGNETO is higher than that obtained with the other approaches. Specifically, in the case of the PBMC dataset, (Fig. 2a) the panels of MAGNETO isolate both CD4 and CD8 cells (up to 80%), while the other approaches remain at very low percentage values, especially using marker panels composed of 4 genes. We argue that this phenomenon can be caused by the fact that the PBMC dataset comprises immune cells, which mainly include *T* cells (∼70%), B cells (∼15%), monocytes (∼5%), dendritic cells (DC) (∼1%), and NK cells (∼10%) [53,54]. In particular, the *T* cell co-receptor can be further divided into two main subtypes: CD4 and CD8 *T* cells [55]; the panel identified by MAGNETO can isolate a high percentage of both subtypes of *T* cells. Considering the marker panels with 4 genes, we also observe that the approaches have different performances both in terms of the percentage of isolated cells and regarding the genes included in the panels (RANKCORR: $ADAM8^{+}LY9^{+}CD5^{+}FCRL3^{+}$; SMaSH: $CD3D^{+}CD8A^{+}CD99^{+}CD8B^{+}$; Hypergate: $CD8A^{+}CD74^{-}CD79A^{-}CCR7^{-}$; sc2marker: $CD8ALAG3^{+}CD8B^{+}CD3D^{+}$; COMET: $CD99^{+}CD8A^{+}CD3D^{+}CD8B^{+}$ MAGNETO: $CD3D^{+}CCR7^{-}CD40LG^{-}TNFRSF4^{-}$). Marker genes listed by RANKCORR are not specific to any particular cell type. For example, RANKCORR identified *CD5* and *FCRL3* as two of the top marker genes for *T* cells. However, *CD5* is a well-known marker for activated B cells [56,57] and *FCRL3* is expressed on B cells, NK cells, and *T* cell subsets [58]. The broad expression patterns of these markers indicate unspecificity, which can make their usage less informative and potentially ambiguous in certain contexts. Although Hypergate identifies *CD8* as a top marker gene for *T* cells, the absence of *CD3D*, a widely recognized *T* cell marker, raises concerns about its ability to identify *T* cells specifically. *CD3D* is crucial for *T* cell receptor signaling and is considered a reliable marker for *T* cells. The absence of comprehensive *T* cell-specific markers may result in incomplete or inaccurate identification of *T* cell subsets. The sc2marker panel includes *CD8ALAG3*, *CD8B*, and *CD3D*. While *CD8ALAG3*, *CD8B*, and *CD3D* are associated with *T* cells, their expression patterns may not be exclusive to *T* cells alone. *CD8B*, for instance, is expressed on both *T* cells and NK cells. This lack of specificity can introduce uncertainty and compromise the accuracy of *T* cell identification. The inclusion of *CD99* in COMET's and SMaSH's marker panel may introduce the risk of non-T cell contaminants influencing the analysis. *CD99* is not *T* cell-specific and can be expressed on various cell types, including NK cells, monocytes, and DCs [59]. MAGNETO's marker combination includes *CD3D* and $CCR7^{-}$ which provide a specific profile for identifying *T* cells. As stated above, *CD3D* is a well-established marker associated with T cell receptor signaling, while $CCR7^{-}$ indicates the absence of *CCR7*, a marker commonly expressed on naive and central memory T cells [60]. *CCR7* is also known to help DCs navigate from peripheral tissues to the lymph nodes, where they then control *T* cell activation [61]. This combination suggests a targeted focus on *T* cells and potentially a specific *T* cell phenotype or subset. MAGNETO excludes *CD40LG* and *TNFRSF4*, which are markers associated with non-T cell populations. By excluding these markers, MAGNETO helps minimize potential interference from non-T cell contaminants, allowing for a more precise and accurate analysis of *T* cell-specific gene expression. T cells have several subsets which lead to an extensive list of genes classed as *T* cell "markers", but it is critical to note that these genes have different immune functions and may not necessarily be useful in identifying the predominant CD4 and CD8 *T* cells.

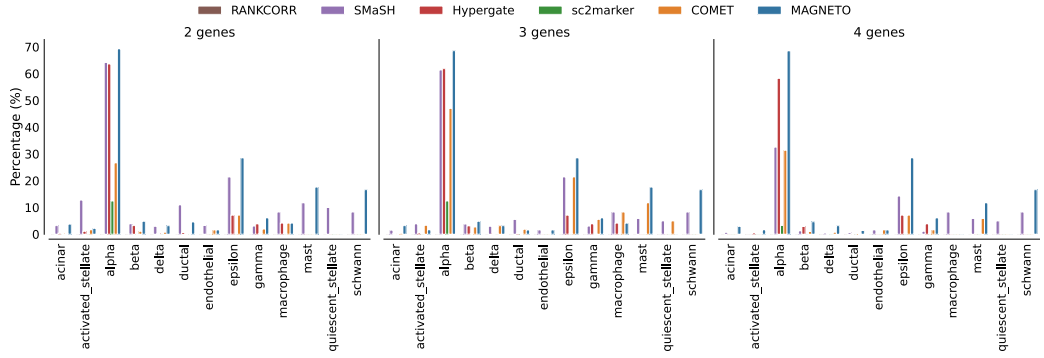
Regarding the PIC dataset (Fig. 2b), all panels isolate other cells besides alpha cells (e.g., epsilon, mast cells) with reduced percentage values. Anyhow, also in this case, the specificity of the panels built with MAGNETO is higher than those identified by the other approaches, as denoted by the very high percentage of alpha cells retrieved, even with only two genes. The majority of PIC cells are made up of beta (65–80%) and alpha (15–20%) cells [62]. Even though alpha cells only comprise ∼15–20% within PIC, the marker genes identified by MAGNETO yield a higher percentage in terms of specificity when compared to COMET consistently and irrespective of the number of genes. This might suggest that MAGNETO has the capability to identify cell populations not present in abundance.

In the case of the HFC dataset (Fig. 2c), MAGNETO allows for isolating around 80% of B cells with all panels. Considering the marker panels with 3 genes, we observe that some of the tested approaches have similar performance (i.e., the percentage of each cell type isolated with the panels is similar) even though the genes included in the panels are different (RANKCORR: $KLRD1^{+}TLR9^{+}CD27^{+}$; SMaSH: $CD79B^{+}CD24^{+}VPREB1^{+}$; Hypergate: $CD24^{+}CD79B^{+}CSF3R^{-}$; sc2marker: $CD24^{+}CD79B^{+}CD79A^{+}$; COMET: $CD74^{+}CD79B^{+}CD24^{+}$ MAGNETO: $CD79B^{+}CD24^{+}ACE^{-}$). *KLRD1* is not a specific marker for B cells but is associated with NK cells [63]. Its presence in RANKCORR suggests that it may detect NK cells instead of B cells. *CD27* is a marker expressed on various immune cells, including B cells, T cells, and NK cells [64]. While it is found on some B cells, its presence in RANKCORR suggests that it may not exclusively target B cells, resulting in potential contamination from other cell types. Hypergate being $CD79B^{+}$ indicates that it relies on *CD79B* expression, which is a reliable marker for B cells. While *CD79A* is a component of the B cell receptor complex, it is not exclusively expressed in B cells. *CD79A* can also be found in other immune cell types, such as plasma cells. The expression level of *CD79A* can vary among different B cell subsets and can be influenced by factors such as activation state and maturation stage. This variability makes it less reliable as a marker for consistent and reliable identification of B cells. Despite including commonly used marker genes such as *CD79B*, *CD24*, and *VPREB1*, the specific combination used in SMaSH does not demonstrate optimal performance or the best results. *CD79B* is a B cell-specific protein whose expression in COMET and MAGNETO reaffirms their ability to isolate B cells accurately. Additionally, *CD24*, a frequently observed marker on B cells, further supports their identification and isolation [65]. The marker genes identified by MAGNETO and COMET are thus standardized B cell markers. While COMET and MAGNETO exhibit similar performance, MAGNETO slightly surpasses COMET in isolating B cells using three genes. MAGNETO demonstrates a slightly reduced expression for various cell types such as LMPs, erythrocytes, and granulocytes, indicating its better ability to specifically target B cells compared to COMET. This suggests that MAGNETO could confidently identify B cells even with a reduced number of genes, consequently implying the reliability and superiority of the combination of genes within its panel. However, it is worth noting that several other cell types were also identified with MAGNETO's panel, albeit in significantly lower percentages.

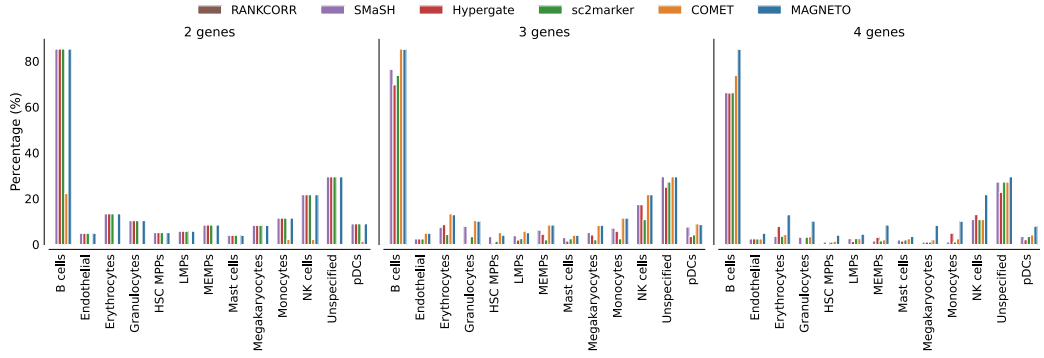
A second test to investigate the performance of MAGNETO regards the identification of marker panels by varying the value of the parameter α, which controls the contribution of the two components of the fitness value related to the first objective of the problem. In particular, we consider the values 0.2, 0.4, 0.6, 0.8 to obtain less (with low α values) or more (with high α values) specific marker panels. In this test, we take into account the same cell types as in the previous analysis. Fig. 3 reports the barplots showing the percentage of cells isolated with the panels identified by RANKCORR, SMaSH, Hypergate, sc2marker, COMET, and MAGNETO using $\alpha = 0.8$. It is evident that MAGNETO



(a) PBMC dataset - marker panels for CD8 T cells



(b) PIC dataset - marker panels for alpha cells

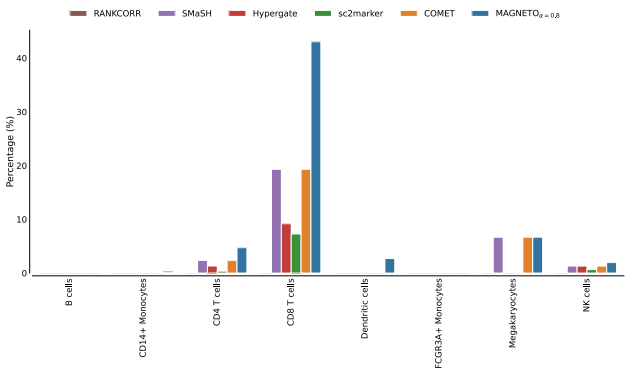


(c) HFC dataset - marker panels for B cells

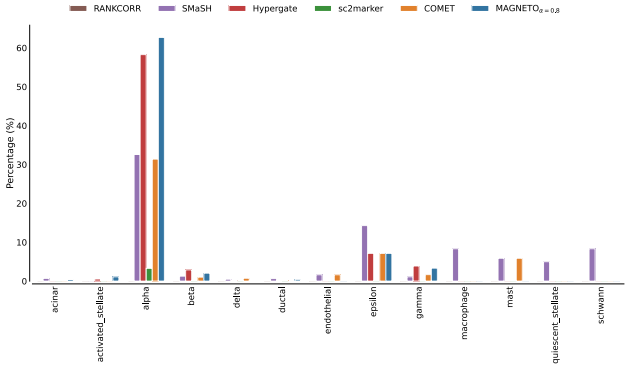
Fig. 2. Barplots showing the percentage of cells isolated by considering the genes included in the best marker panels (with 2, 3, or 4 genes) identified by RANKCORR (brown bars), SMaSH (violet bars), Hypergate (red bars), sc2marker (green bars), COMET (orange bars), and MAGNETO (blue bars) for a specific cell type of interest. The percentage of each cell type is calculated using function g described in Section 4.0.2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

built more specific marker panels that allow for better isolating of the desired cells, reducing the percentage of cells of other cell populations. Supplementary Fig. 4 reports the barplots showing the percentage of cells isolated with the panels identified by MAGNETO with increasing values of α. As expected, with $\alpha = 0.2$, a very high percentage of the cell type of interest, as well as the other types, are isolated with the panels. On the contrary, when the value of α is increased, the panels obtained with MAGNETO allow for more specific isolation of the cell type of interest, meaning that the percentage of all other cells is strongly reduced.

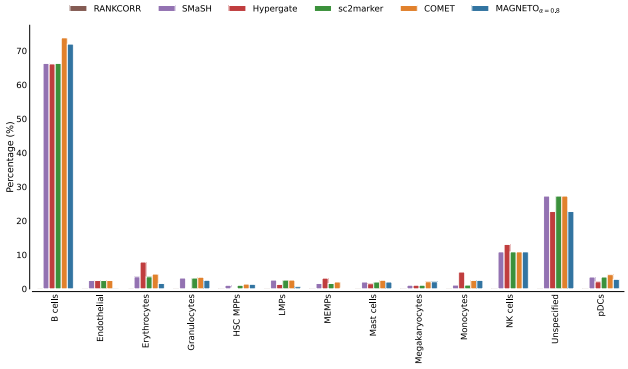
The results obtained by MAGNETO setting $\alpha = 0.8$ permitted for having very specific marker panels. We further analyze and compare this tuning of MAGNETO against the most promising approaches (i.e., SMaSH, Hypergate, and COMET) in an *ad-hoc* test. In particular, we consider the isolation of HSC/MPP cells of the HFC dataset; the marker panel that can be built considering the top 4 genes calculated by SMaSH includes $SELL^+$, $CD34^+$, $CD52^+$, and $PROM1^+$, while Hypergate identified $CD52^+$, $CD74^+$, $CD24^-$, and $CD9^-$ as the best genes for isolating HSC/MPP cells. Despite identifying *CD52* as a marker gene, Hypergate also detects *CD74*, which is not only a significant marker gene for B cells but is also broadly expressed in various other immune cell populations, including T cell subsets, monocytes, and macrophages [66]. This broad expression of *CD74* in different immune cell types potentially hinders Hypergate's ability to isolate HSC/MPP populations specifically. While SMaSH effectively identifies key marker genes for HSC/MPP populations, its performance could be enhanced by incorporating negative markers to eliminate potential contaminating populations that do not strictly represent HSC/MPP cells. The panel built by COMET includes the genes $CSF3R^+$, $CD74^+$, $CD52^+$, and $CD34^+$, while that built by MAGNETO is composed of the genes $PROM1^+$, $CD52^+$, $CSF1R^-$, and $HMMR^-$. MAGNETO identified key marker genes that



(a) PBMC dataset - marker panels for CD8 T cells



(b) PIC dataset - marker panels for alpha cells



(c) HFC dataset - marker panels for B cells

Fig. 3. Barplots showing the percentage of cells isolated by considering the genes included in the best marker panels identified by MAGNETO using $\alpha = 0.8$ and the other state-of-art approaches for a specific cell type of interest. The percentage of each cell type is calculated using function g described in Section 4.0.2.

were previously used in the refined panel by Ranzoni et al. [12] to isolate HSCs/MPPs which include *CD52*$^+$ and *PROM1*$^+$. *CSF1R* and *HMMR* are known marker genes for common myeloid progenitors/common lymphoid progenitors (CMP/CLP) [67] and megakaryocyte progenitors (MKP) [68], respectively. Although CMP/CLP and MKPs are subtypes of HSC/MPPs they are present further down the hematopoietic lineage tree and represent committed progenitors. This possibly explains why we see most of the cells correctly present in the HSC/MPPs cluster and only a small fraction of cells present in other clusters in the UMAP. COMET identified *CD34* and *CD52* that are some of the most common HSC/MPP marker genes [69]. However, it also picked up *CD74* that is a key marker gene of B cells [70] but also broadly expressed in several other immune cell populations as stated above [66]. Although *CSF3R* is known to play a role in hematopoietic stem cell mobilization, it is also found to regulate granulopoiesis and neutrophil function. Together, these genes do play a role in identifying HSC/MPPs but more precisely committed progenitors and this possibly elucidates why we see a significant number of cells also present in the B cell and granulocyte clusters. Fig. 4 (top) shows the UMAP representation of the dataset with the manual annotation where the HSC/MPP cluster is denoted in violet. The manual annotation can be visually compared with the HSC/MPP cells (denoted in green) expressing the best marker panels identified bySMaSH (middle left), Hypergate (middle right), COMET (bottom left), and MAGNETO (bottom right), respectively. We observe that the cluster of interest is better highlighted when considering the genes of the panels identified by SMaSH and MAGNETO, especially for what concerns the cells of other types mistakenly isolated. In particular, the marker panel identified by MAGNETO is expressed by ~62% of HSC/MPP cells and by only ~6% of other cells (see Supplementary Fig. 5). Considering the top 4 genes identified by SMaSH, the marker panel that can be built is expressed by ~63% of HSC/MPP cells and ~10% of other cells. The contribution of positive and negative genes of the marker panel built by MAGNETO is graphically explained in Fig. 5 (top), where the UMAP representations highlight the cells of the HSC/MPP cluster isolated considering only the positive genes (left) that must be all co-expressed by the cells, the cells of any kind that express at least one of the negative genes of the panel (middle), and the cells isolated considering both positive and negative genes (right). The pie-charts reported at the bottom of Fig. 5 denote the percentage of HSC/MPP cells isolated with positive genes (left), negative genes (middle), and with the complete panel (right); thanks to the contribution of all genes included in the panel, the resulting percentage of cells of interest isolated is 81.4%.

To evaluate the performance of MAGNETO in building larger panels, we execute MAGNETO with the maximum number of genes ρ to be possibly included in the marker panels equal to 10 and with $\alpha = 0.8$, which allows for obtaining selective solutions for the cell type of interest. Fig. 6 reports the barplots identified on the three datasets; in all cases, the marker panels isolate a high percentage of the cell types of interest, while the percentage of the other cell types is below 10%. The only exception regards the panel for the isolation of B cells in the HFC dataset, where the percentage of unspecified cells taken into account is around 30%. We argue that the panel includes the unspecified cells of this dataset, as they are probably some B cell progenitors.

As a final test, we evaluate MAGNETO's running time by varying the number of cells and the values of $\rho \in \{2, 4, 6, 8, 10, 12\}$ (i.e., the maximum number of genes composing the marker panels). To do so, we analyze a subset of the cross-tissue single-cell atlas of developing human immune cells [71]. In particular, we consider only the immune cells (all hematopoietic-derived cells) for a total of 593203 cells. MAGNETO is applied to identify the marker panels for NK cells in three different organs of the immune cell subset: kidney (8444 cells), skin (54122 cells), and spleen (85341 cells). We use a workstation equipped with an Intel Xeon Gold 6208U CPU (clock 2.9 GHz) and 64 GB of RAM, running Ubuntu 22.04 LTS to run MAGNETO, which takes less than 2 minutes to build the marker panels composed of a maximum of 12 genes for NK cells (11763 cells out of 85341 cells) in the spleen (see Supplementary Fig. 6). MAGNETO can be applied to analyze to build the marker panels for a specific cell type from large scRNA-Seq datasets (more than 100000 cells) in a few minutes. MAGNETO also exploits the Python multiprocessing package to build the marker panels for all the clusters of small scRNA-Seq datasets in parallel. For instance, MAGNETO can analyze the 13 clusters of the PIC dataset (~6400 cells) in less than 14 seconds. For the sake of comparison, the approaches tested in [17] (i.e., sc2marker, CombiRock, RANKCORR, and COMET) took several minutes for the analysis of the PBMC dataset, and around one hour for the analysis of the MCA lung data set, which contains 6940 cells characterized in 31 distinct cell types.

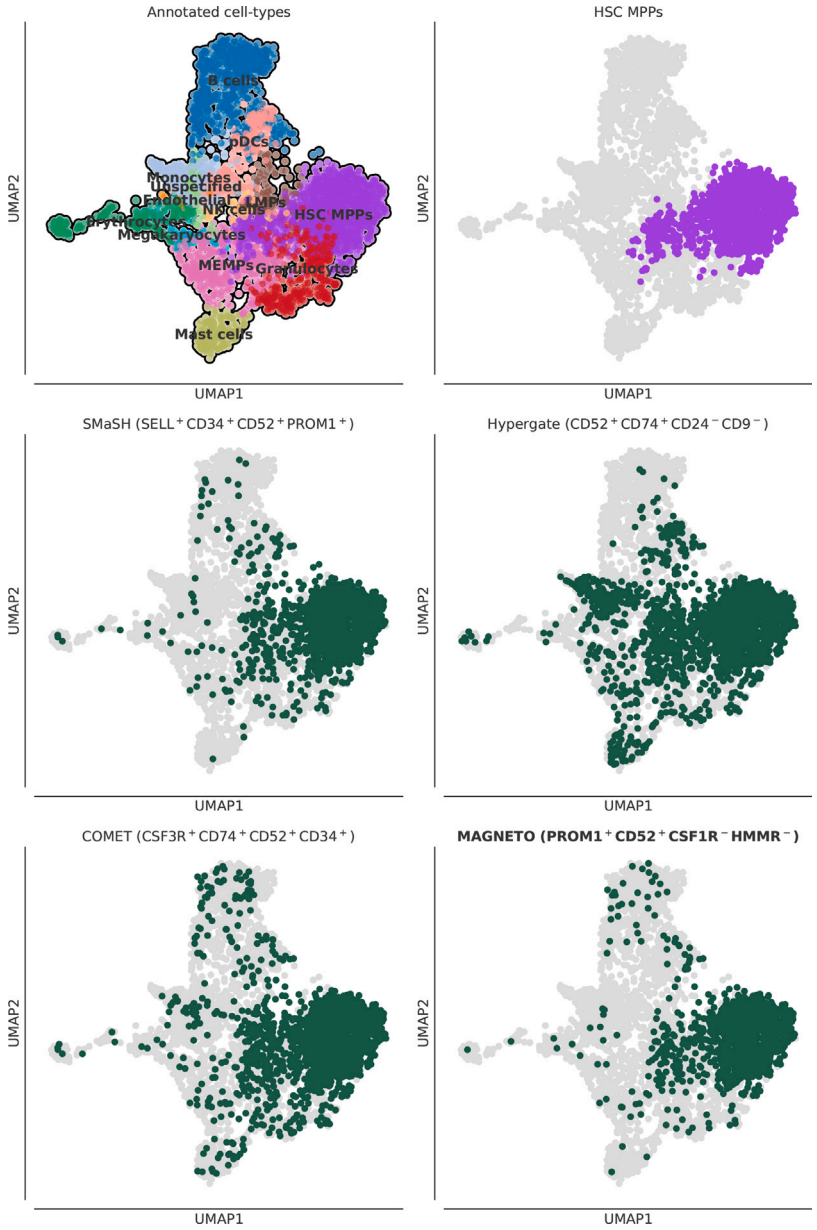


Fig. 4. *Top left*: UMAP representation of the different cell types present in the HFC dataset. *Top right*: UMAP representation highlighting only the HSC/MPP cells in the HFC dataset. UMAP representations highlighting the cells co-expressing all the positive genes and not expressing any negative genes included in the marker panels identified by SMaSH (i.e., $SELL^{+}CD34^{+}CD52^{+}$ $PROM1^{+}$, *middle left*), Hypergate (i.e., $CD52^{+}CD74^{+}$ $CD24^{-}CD9^{-}$, *middle right*), COMET (i.e., $CSF3R^{+}CD74^{+}CD52^{+}CD34^{+}$, *bottom left*), and MAGNETO (i.e., $PROM1^{+}CD52^{+}CSF1R^{-}HMMR^{-}$, *bottom right*) for the HSC/MPP cluster.

6. Conclusions

The fast-increasing number of scRNA-seq experiments is opening a number of computational challenges to be faced as never before. On the one hand, these experiments are allowing us to reveal novel cell types and subtypes in different organisms and tissues, which often play crucial functional roles based on context and the microenvironment where they are. Moving from the high-throughput discovery of these cell populations to targeted functional studies requires the identification of specific marker panels able to isolate them from all the other cell populations specifically. In this work, we presented marker panel generator with multi-Objective optimization, a tool for automatically identifying optimal marker panels for isolating cell types of interest starting from annotated scRNA-Seq data. MAGNETO exploits the AGE-MOEA algorithm to solve the problem of optimizing marker panels, defined as a bi-objective problem. The two conflicting objectives defined in MAGNETO regards the identification of the most appropriate genes to be included in the panel to maximize the number of cells of interest that can be isolated, and the minimization of the number of genes to be included in the panel.

We tested MAGNETO on three different public datasets (i.e., PBMC, PIC, HFC) and compared its outcome against COMET, RANKCORR, SMaSH, Hypergate, and sc2marker. Our results show that the panels identified by MAGNETO allow for a more specific characterization of the cell types of interest, thanks to the hyper-parameters that can be set by the user (i.e., binarization strategy, the maximum number of genes, specificity of the marker panel). The main assumption behind

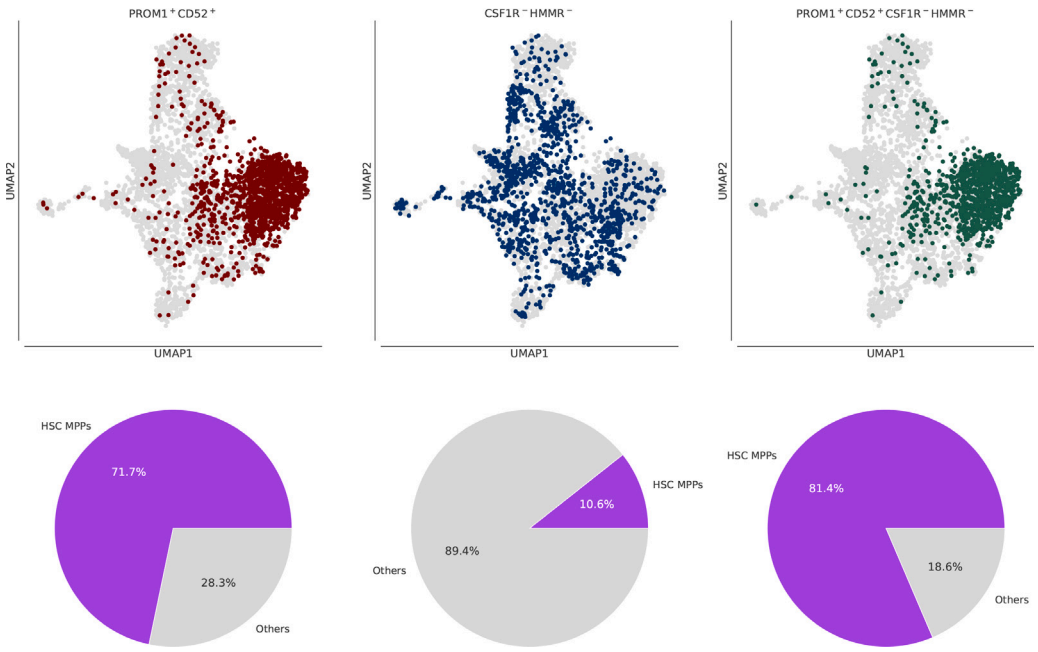


Fig. 5. *Top*: UMAP representations highlighting the cells co-expressing all the positive genes (left), the cells expressing at least a negative gene (middle), and the cells co-expressing all the positive genes and not expressing any negative genes of the marker panel identified by MAGNETO (right) to isolate the HSC/MPP cluster (i.e., *PROM1*$^{+}$*CD52*$^{+}$*CSF1R*$^{-}$*HMMR*$^{-}$). *Bottom*: Pie-charts showing the normalized percentage of HSC/MPP cells (violet) and of other cells (gray) co-expressing all the positive genes (left), expressing at least a negative gene (middle), and co-expressing all the positive genes and not expressing any negative genes of the marker panel identified by MAGNETO (right) to isolate the HSC/MPP cluster. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the idea of building marker panels from scRNA-Seq is that genes correlate well between their transcriptional and protein/cell surface abundance. However, this assumption sometimes is not completely correct due to several biological and technical factors, including mRNA stability, protein stability and transport, and translation efficiency. MAGNETO can be extended by including additional objectives to the problem formulation, such as the cost of the lab experiment or the availability and quality of the different antibodies. Another problem to be taken into account during the identification of the marker panels is the discrepancies between cellular mRNA levels and surface protein detection rates with flow cytometry, which are mainly due to both the antibody quality and specificity. This might be addressed by providing MAGNETO with information about how the protein abundance of each gene correlates with its transcriptional state.

In addition to the validation presented here, we are performing lab experiments to investigate the broader applicability of MAGNETO. To this end, we are planning to experimentally validate using flow cytometry the marker panels identified by MAGNETO from scRNA-Seq experiments, and are performing a series of such experiments in cancer. Cancer is indeed a complex disease where biological heterogeneity poses a clinical challenge [72]. We are designing experiments to isolate cancer cells that are resistant to current clinical therapies, together with other cell populations in the tumor microenvironment that drive or activate resistance via signaling to cancer cells. We anticipate that this work will provide mechanistic insight into therapeutic resistance.

Finally, we are working to extend MAGNETO to be successfully used in CITE-seq experiments.

CRediT authorship contribution statement

Andrea Tangherloni: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Simone G. Riva:** Software, Validation, Investigation, Data curation, Writing – review & editing. **Brynelle Myers:** Validation, Formal analysis, Writing – review & editing. **Francesca M. Buffa:** Resource, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Paolo Cazzaniga:** Validation, Formal analysis, Resources, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

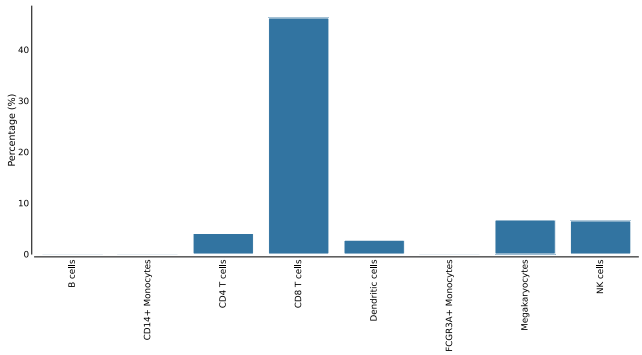
Availability and requirements

MAGNETO's open-source code is available on GitLab: https://gitlab.com/andrea-tango/magneto under the BSD-3 license. MAGNETO can also be easily installed using the Python package installer pip and the conda package manager, which allow the user to use MAGNETO as a Python package that can be integrated into Python scripts and Jupyter Notebooks.

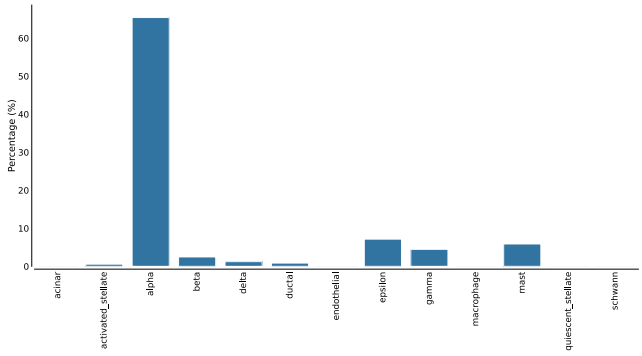
The repository also contains the Jupyter Notebooks used to obtain and analyze the results shown in the paper. We provide a detailed description of MAGNETO parameters and examples so that both novice and expert researchers can use it for analyzing their data.

Acknowledgments

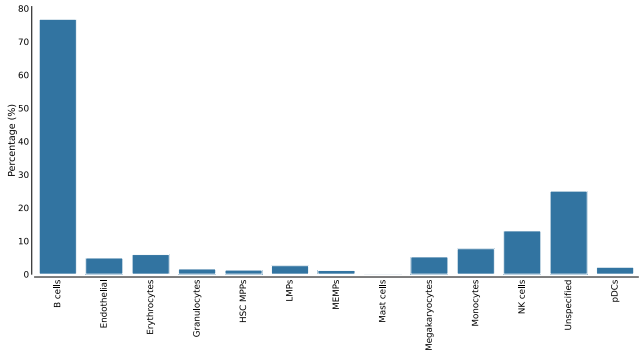
This work was supported in part by the European Research Council Project 772970 – MicroC (A.T. and F.M.B.), and by the National Plan for NRRP Complementary Investments (PNC, established with the decree-law 6 May 2021, n. 59, converted by law n. 101 of 2021) in the call for the funding of research initiatives for technologies and innovative trajectories in the health and care sectors (Directorial Decree n. 931



(a) PBMC dataset - marker panel for CD8 T cells



(b) PIC dataset - marker panel for alpha cells



(c) HFC dataset - marker panel for B cells

Fig. 6. Barplots showing the percentage of cells expressing the genes included in the best marker panels identified by MAGNETO using $\rho = 10$ (i.e., marker panels composed of up to 10 different genes) and $\alpha = 0.8$ for the isolation of a specific cell type of interest. The percentage of each cell type is calculated using function g described in Section 4.0.2.

of 06-06-2022) - project n. PNC0000003 - AdvaNced Technologies for Human-centrEd Medicine (project acronym: ANTHEM). This work reflects only the authors' views and opinions, neither the Ministry for University and Research nor the European Commission can be considered responsible for them (P.C.).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jbi.2023.104510.

References

[1] F. Paul, Y. Arkin, A. Giladi, D.A. Jaitin, E. Kenigsberg, H. Keren-Shaul, D. Winter, D. Lara-Astiaso, M. Gury, A. Weiner, et al., Transcriptional heterogeneity and lineage commitment in myeloid progenitors, Cell 163 (7) (2015) 1663–1677.

[2] R. Satija, J.A. Farrell, D. Gennert, A.F. Schier, A. Regev, Spatial reconstruction of single-cell gene expression data, Nat. Biotechnol. 33 (5) (2015) 495–502.

[3] M. Baron, A. Veres, S.L. Wolock, A.L. Faust, R. Gaujoux, A. Vetere, J.H. Ryu, B.K. Wagner, S.S. Shen-Orr, A.M. Klein, et al., A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure, Cell Syst. 3 (4) (2016) 346–360.

[4] K. Shekhar, S.W. Lapan, I.E. Whitney, N.M. Tran, E.Z. Macosko, M. Kowalczyk, X. Adiconis, J.Z. Levin, J. Nemesh, M. Goldman, et al., Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics, Cell 166 (5) (2016) 1308–1323.

[5] A.-C. Villani, R. Satija, G. Reynolds, S. Sarkizova, K. Shekhar, J. Fletcher, M. Griesbeck, A. Butler, S. Zheng, S. Lazo, et al., Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors, Science 356 (6335) (2017).

[6] G. Jia, J. Preussner, X. Chen, S. Guenther, X. Yuan, M. Yekelchyk, C. Kuenne, M. Looso, Y. Zhou, S. Teichmann, et al., Single cell RNA-seq and ATAC-seq analysis of cardiac progenitor cell transition states and lineage settlement, Nature Commun. 9 (1) (2018) 1–17.

[7] E.M. Kernfeld, R.M. Genga, K. Neherin, M.E. Magaletta, P. Xu, R. Maehr, A single-cell transcriptomic atlas of thymus organogenesis resolves cell types and developmental maturation, Immunity 48 (6) (2018) 1258–1270.

[8] R. Vento-Tormo, M. Efremova, R.A. Botting, M.Y. Turco, M. Vento-Tormo, K.B. Meyer, J.-E. Park, E. Stephenson, K. Polański, A. Goncalves, et al., Single-cell reconstruction of the early maternal–fetal interface in humans, Nature 563 (7731) (2018) 347–353.

[9] M. Efremova, S.A. Teichmann, Computational methods for single-cell omics across modalities, Nature Methods 17 (1) (2020) 14–17.

[10] M.M. Gladka, B. Molenaar, H. De Ruiter, S. Van Der Elst, H. Tsui, D. Versteeg, G.P. Lacraz, M.M. Huibers, A. Van Oudenaarden, E. Van Rooij, Single-cell sequencing of the healthy and diseased heart reveals cytoskeleton-associated protein 4 as a new modulator of fibroblasts activation, Circulation 138 (2) (2018) 166–180.

[11] H. Keren-Shaul, A. Spinrad, A. Weiner, O. Matcovitch-Natan, R. Dvir-Szternfeld, T.K. Ulland, E. David, K. Baruch, D. Lara-Astaiso, B. Toth, et al., A unique microglia type associated with restricting development of Alzheimer's disease, Cell 169 (7) (2017) 1276–1290.

[12] A.M. Ranzoni, A. Tangherloni, I. Berest, S.G. Riva, B. Myers, P.M. Strzelecka, J. Xu, E. Panada, I. Mohorianu, J.B. Zaugg, et al., Integrative single-cell RNA-seq and ATAC-seq analysis of human developmental hematopoiesis, Cell Stem Cell (2020).

[13] D. Lähnemann, J. Köster, E. Szczurek, D.J. McCarthy, S.C. Hicks, M.D. Robinson, C.A. Vallejos, K.R. Campbell, N. Beerenwinkel, A. Mahfouz, et al., Eleven grand challenges in single-cell data science, Genome Biol. 21 (1) (2020) 1–35.

[14] M.D. Luecken, F.J. Theis, Current best practices in single-cell RNA-seq analysis: a tutorial, Mol. Syst. Biol. 15 (6) (2019) e8746.

[15] C. Delaney, A. Schnell, L.V. Cammarata, A. Yao-Smith, A. Regev, V.K. Kuchroo, M. Singer, Combinatorial prediction of marker panels from single-cell transcriptomic data, Mol. Syst. Biol. 15 (10) (2019) e9005.

[16] A. Tangherloni, S.G. Riva, B. Myers, P. Cazzaniga, Multi-objective optimization for marker panel identification in single-cell data, in: Proc. Conference on Computational Intelligence in Bioinformatics and Computational Biology, IEEE, 2022, pp. 1–8.

[17] R. Li, B. Banjanin, R.K. Schneider, I.G. Costa, Detection of cell markers from single cell RNA-seq with sc2marker, BMC Bioinform. 23 (1) (2022) 276.

[18] Y. Saeys, S. Van Gassen, B.N. Lambrecht, Computational flow cytometry: helping to make sense of high-dimensional immunology data, Nat. Rev. Immunol. 16 (7) (2016) 449–462.

[19] M. Uhlén, L. Fagerberg, B.M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, et al., Tissue-based map of the human proteome, Science 347 (6220) (2015) 1260419.

[20] X. Zhang, Y. Lan, J. Xu, F. Quan, E. Zhao, C. Deng, T. Luo, L. Xu, G. Liao, M. Yan, et al., CellMarker: a manually curated resource of cell markers in human and mouse, Nucleic Acids Res. 47 (D1) (2019) D721–D728.

[21] S. Mazzara, R.L. Rossi, R. Grifantini, S. Donizetti, S. Abrignani, M. Bombaci, CombiROC: an interactive web tool for selecting accurate marker combinations of omics data, Sci. Rep. 7 (1) (2017) 45477.

[22] E. Becht, Y. Simoni, E. Coustan-Smith, M. Evrard, Y. Cheng, L.G. Ng, D. Campana, E.W. Newell, Reverse-engineering flow-cytometry gating strategies for phenotypic labelling and high-performance cell sorting, Bioinformatics 35 (2) (2019) 301–308.

[23] A.H. Vargo, A.C. Gilbert, A rank-based marker selection method for high throughput scRNA-seq data, BMC Bioinform. 21 (1) (2020) 1–51.

[24] B. Dumitrascu, S. Villar, D.G. Mixon, B.E. Engelhardt, Optimal marker gene selection for cell type discrimination in single cell analyses, Nature Commun. 12 (1) (2021) 1186.

[25] A. Tangherloni, S.G. Riva, S. Spolaor, D. Besozzi, M.S. Nobile, P. Cazzaniga, The impact of representation on the optimization of marker panels for single-cell RNA data, in: Proc. Congress on Evolutionary Computation, IEEE, 2021, pp. 1423–1430.

[26] A. Panichella, An adaptive evolutionary algorithm based on non-euclidean geometry for many-objective optimization, in: Proc. Genetic and Evolutionary Computation Conference, 2019, pp. 595–603.

[27] M.E. Nelson, S.G. Riva, A. Cvejic, SMaSH: a scalable, general marker gene identification framework for single-cell RNA-sequencing, BMC Bioinform. 23 (1) (2022) 328.

[28] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species, Nat. Biotechnol. 36 (5) (2018) 411–420.

[29] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W.M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, R. Satija, Comprehensive integration of single-cell data, Cell (2019) http://dx.doi.org/10.1016/j.cell.2019.05.031.

[30] Y. Hao, S. Hao, E. Andersen-Nissen, W.M. Mauck, S. Zheng, A. Butler, M.J. Lee, A.J. Wilk, C. Darby, M. Zager, et al., Integrated analysis of multimodal single-cell data, Cell 184 (13) (2021) 3573–3587.

[31] F.A. Wolf, P. Angerer, F.J. Theis, SCANPY: large-scale single-cell gene expression data analysis, Genome Biol. 19 (1) (2018) 15.

[32] G. Finak, A. McDavid, M. Yajima, J. Deng, V. Gersuk, A.K. Shalek, C.K. Slichter, H.W. Miller, M.J. McElrath, M. Prlic, et al., MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data, Genome Biol. 16 (1) (2015) 1–13.

[33] M. Bombaci, R.L. Rossi, Computation and selection of optimal biomarker combinations by integrative ROC analysis using combiROC, Proteom. Biomark. Discov. Methods Prot. (2019) 247–259.

[34] L. Breiman, et al., Classification and Regression Trees, CRC Press, Boca Raton, 1984.

[35] L.S. Shapley, Notes on the *n*-Person Game –– II: The Value of an *n*-Person, RAND Corporation, Santa Monica, Calif., 1951, 195. https://www.rand.org/content/dam/rand/pubs/research_memoranda/2008/RM670.pdf.

[36] L. Rundo, A. Tangherloni, P. Cazzaniga, M.S. Nobile, G. Russo, M.C. Gilardi, S. Vitabile, G. Mauri, D. Besozzi, C. Militello, A novel framework for MR image segmentation and quantification by using MedGA, Comput. Methods Programs Biomed. 176 (2019) 159–172.

[37] A. Tangherloni, S. Spolaor, P. Cazzaniga, D. Besozzi, L. Rundo, G. Mauri, M.S. Nobile, Biochemical parameter estimation vs. benchmark functions: A comparative study of optimization performance and representation design, Appl. Soft Comput. 81 (2019) 105494.

[38] A. Tangherloni, S. Spolaor, L. Rundo, M.S. Nobile, P. Cazzaniga, G. Mauri, P. Liò, I. Merelli, D. Besozzi, GenHap: a novel computational method based on genetic algorithms for haplotype assembly, BMC Bioinform. 20 (4) (2019) 172.

[39] D. Grün, M.J. Muraro, J.-C. Boisset, K. Wiebrands, A. Lyubimova, G. Dharmadhikari, M. van den Born, J. van Es, E. Jansen, H. Clevers, et al., De novo prediction of stem cell identity using single-cell transcriptome data, Cell Stem Cell 19 (2) (2016) 266–277, http://dx.doi.org/10.1016/j.stem.2016.05.010.

[40] M.J. Muraro, G. Dharmadhikari, D. Grün, N. Groen, T. Dielen, E. Jansen, L. van Gurp, M.A. Engelse, F. Carlotti, E.J. de Koning, et al., A single-cell transcriptome atlas of the human pancreas, Cell Syst. 3 (4) (2016) 385–394, http://dx.doi.org/10.1016/j.cels.2016.09.002.

[41] N. Lawlor, J. George, M. Bolisetty, R. Kursawe, L. Sun, V. Sivakamasundari, I. Kycia, P. Robson, M.L. Stitzel, Single-cell transcriptomes identify human islet cell signatures and reveal cell-type–specific expression changes in type 2 diabetes, Genome Res. 27 (2) (2017) 208–222, http://dx.doi.org/10.1101/gr.212720.116.

[42] Å. Segerstolpe, A. Palasantza, P. Eliasson, E.-M. Andersson, A.-C. Andréasson, X. Sun, S. Picelli, A. Sabirsh, M. Clausen, M.K. Bjursell, et al., Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes, Cell Metab. 24 (4) (2016) 593–607, http://dx.doi.org/10.1016/j.cmet.2016.08.020.

[43] S. Picelli, O.R. Faridani, Å.K. Björklund, G. Winberg, S. Sagasser, R. Sandberg, Full-length RNA-seq from single cells using smart-seq2, Nat. Protoc. 9 (1) (2014) 171.

[44] A. Tangherloni, F. Ricciuti, D. Besozzi, P. Liò, A. Cvejic, Analysis of single-cell RNA sequencing data based on autoencoders, BMC Bioinform. 22 (1) (2021) 1–27.

[45] S.G. Riva, P. Cazzaniga, A. Tangherloni, Integration of multiple scRNA-seq datasets on the autoencoder latent space, in: Proc. International Conference on Bioinformatics and Biomedicine, IEEE, 2021, pp. 1–8.

[46] A. Zhou, B.-Y. Qu, H. Li, S.-Z. Zhao, P.N. Suganthan, Q. Zhang, Multiobjective evolutionary algorithms: A survey of the state of the art, Swarm Evol. Comput. 1 (1) (2011) 32–49.

[47] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGA-II, IEEE Trans. Evol. Comput. 6 (2) (2002) 182–197.

[48] N. Beume, B. Naujoks, M. Emmerich, SMS-EMOA: Multiobjective selection based on dominated hypervolume, European J. Oper. Res. 181 (3) (2007) 1653–1669.

[49] A. Panichella, An improved Pareto front modeling algorithm for large-scale many-objective optimization, in: Proc. Genetic and Evolutionary Computation Conference, Association for Computer Machinery, 2022.

[50] The pandas development team, pandas-dev/pandas: Pandas, Zenodo, 2020, http://dx.doi.org/10.5281/zenodo.3509134.

[51] W. McKinney, Data structures for statistical computing in python, in: S. van der Walt, J. Millman (Eds.), Proc. Python in Science Conference, 2010, pp. 56–61, http://dx.doi.org/10.25080/Majora-92bf1922-00a.

[52] A.J. Collier, S.P. Panula, J.P. Schell, P. Chovanec, A.P. Reyes, S. Petropoulos, A.E. Corcoran, R. Walker, I. Douagi, F. Lanner, et al., Comprehensive cell surface protein profiling identifies specific markers of human naive and primed pluripotent states, Cell Stem cell 20 (6) (2017) 874–890.

[53] P. Autissier, C. Soulas, T.H. Burdo, K.C. Williams, Evaluation of a 12-color flow cytometry panel to study lymphocyte, monocyte, and dendritic cell subsets in humans, Cytom. Part A J. Int. Soc. Adv. Cytom. 77 (5) (2010) 410–419.

[54] C.R. Kleiveland, Peripheral blood mononuclear cells, in: The Impact of Food Bioactives on Health: In Vitro and Ex Vivo Models, Springer International Publishing, 2015, pp. 161–167.

[55] R.S. Sauls, C. McCausland, B.N. Taylor, Histology, T-cell lymphocyte, 2018.

[56] C. Werner-Favre, T.L. Vischer, D. Wohlwend, R.H. Zubler, Cell surface antigen CD5 is a marker for activated human B cells, Eur. J. Immunol. 19 (7) (1989) 1209–1213.

[57] H. Gary-Gouy, J. Harriague, G. Bismuth, C. Platzer, C. Schmitt, A.H. Dalloul, Human CD5 promotes B-cell survival through stimulation of autocrine IL-10 production, Blood 100 (13) (2002) 4537–4543.

[58] F. Li, W. Won, E. Becker, J. Easlick, E. Tabengwa, R. Li, M. Shakhmatov, K. Honjo, P. Burrows, R. Davis, Emerging roles for the FCRL family members in lymphocyte biology and disease, Fc Recept. (2014) 29–50.

[59] N. Takheaw, S. Pata, W. Laopajon, S. Roytrakul, W. Kasinrerk, The presence of membrane bound CD99 ligands on leukocyte surface, BMC Res. Notes 13 (1) (2020) 1–6.

[60] J.J. Campbell, K.E. Murphy, E.J. Kunkel, C.E. Brightling, D. Soler, Z. Shen, J. Boisvert, H.B. Greenberg, M.A. Vierra, S.B. Goodman, et al., CCR7 expression and memory T cell diversity in humans, J. Immunol. 166 (2) (2001) 877–884.

[61] J.L. Rodríguez-Fernández, O. Criado-García, The chemokine receptor CCR7 uses distinct signaling modules with biased functionality to regulate dendritic cells, Front. Immunol. 11 (2020) 528.

[62] H. Chen, B. Martin, H. Cai, J.L. Fiori, J.M. Egan, S. Siddiqui, S. Maudsley, Pancreas++: automated quantification of pancreatic islet cells in microscopy images, Front. Physiol. 3 (2013) 482.

[63] C. Yang, J.R. Siebert, R. Burns, Z.J. Gerbec, B. Bonacci, A. Rymaszewski, M. Rau, M.J. Riese, S. Rao, K.-S. Carlson, et al., Heterogeneity of human bone marrow and blood natural killer cells defined by single-cell transcriptome, Nature Commun. 10 (1) (2019) 3931.

[64] B. Fu, F. Wang, R. Sun, B. Ling, Z. Tian, H. Wei, CD11b and CD27 reflect distinct population and functional specialization in human natural killer cells, Immunology 133 (3) (2011) 350–359.

[65] F.F. Mensah, C.W. Armstrong, V. Reddy, A.S. Bansal, S. Berkovitz, M.J. Leandro, G. Cambridge, CD24 expression and b cell maturation shows a novel link with energy metabolism: potential implications for patients with myalgic encephalomyelitis/chronic fatigue syndrome, Front. Immunol. 9 (2018) 2421.

[66] S. Zhao, A. Molina, A. Yu, J. Hanson, H. Cheung, X. Li, Y. Natkunam, High frequency of CD74 expression in lymphomas: implications for targeted therapy using a novel anti-CD74-drug conjugate, J. Pathol.: Clin. Res. 5 (1) (2019) 12–24.

[67] E.R. Stanley, V. Chitu, CSF-1 receptor signaling in myeloid cells, Cold Spring Harb. Perspect. Biol. 6 (6) (2014) a021857.

[68] M. Lawrence, A. Shahsavari, S. Bornelöv, T. Moreau, R. McDonald, T.M. Vallance, K. Kania, M. Paramor, J. Baye, M. Perrin, et al., Mapping the biogenesis of forward programmed megakaryocytes from induced pluripotent stem cells, Sci. Adv. 8 (7) (2022) eabj8618.

[69] L.E. Sidney, M.J. Branch, S.E. Dunphy, H.S. Dua, A. Hopkinson, Concise review: evidence for CD34 as a common marker for diverse progenitors, Stem Cells 32 (6) (2014) 1380–1389.

[70] S. Lapter, H. Ben-David, A. Sharabi, H. Zinger, A. Telerman, M. Gordin, L. Leng, R. Bucala, I. Shachar, E. Mozes, A role for the B-cell CD74/macrophage migration inhibitory factor pathway in the immunomodulation of systemic lupus erythematosus by a therapeutic tolerogenic peptide, Immunology 132 (1) (2011) 87–95.

[71] C. Suo, E. Dann, I. Goh, L. Jardine, V. Kleshchevnikov, J.-E. Park, R.A. Botting, E. Stephenson, J. Engelbert, Z.K. Tuong, et al., Mapping the developing human immune system across organs, Science 376 (6597) (2022) eabo0510.

[72] N. Vasan, J. Baselga, D.M. Hyman, A view on drug resistance in cancer, Nature 575 (7782) (2019) 299–309.