2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-6819-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/BIBM55620.2022.9995176

MicroCellClust 2: a hybrid approach for multivariate rare cell mining in large-scale single-cell data

Alexander Gerniers
ICTEAM / INGI / AIA
UCLouvain
Louvain-la-Neuve, Belgium
alexander.gerniers@uclouvain.be

Pierre Dupont
ICTEAM / INGI / AIA
UCLouvain
Louvain-la-Neuve, Belgium
pierre.dupont@uclouvain.be

***Abstract*—Identifying rare subpopulations in single-cell data is a key aspect when analyzing its heterogeneity. With large datasets now commonly generated, the focus went to scalability when designing rare cell mining methods, often relying on univariate approaches. Yet, MicroCellClust, an approach based on a multivariate optimization problem, has proven effective to jointly identify rare cells and specific genes in small-scale data. The proposed solver had a quadratic complexity, posing a practical limit to analyzing small or middle-scale data.**

Here, we present a new approach that scales MicroCellClust to larger datasets. It first performs a beam search among cells that are identified as rare to find an initial approximation. Then it uses simulated annealing, a classical derivative-free optimization algorithm which efficiently approaches the optimal solution.

MicroCellClust 2 has a linear complexity in terms of the number of cells, which makes it scalable to large data (typically containing over 100 000 cells). Our experiments report the identification of rare megakaryocytes within 68 000 PBMCs, and rare ependymal cells within 160 000 mouse brain cells. These results show that MicroCellClust 2 is more effective at identifying a subpopulation as a whole than typical alternatives, demonstrating the usefulness of jointly selecting cells and genes as opposed to other approaches.

***Index Terms*—single-cell, rare cells, data mining, scRNA-seq**

I. Background

Single-cell expression data, such as scRNA-seq, is useful to analyze the heterogeneity inside cell tissues. Besides classical clustering techniques which group cells into large clusters, dedicated methods were developed to identify rare subpopulations containing only a fraction (e.g. less than 5%) of the cells. In a previous work, we presented MicroCellClust [1], which solves a multivariate optimization problem to jointly identify a rare subpopulation of cells with highly specific genes within single-cell expression data.

Such data can be represented by a matrix $\mathbf{M} \in \mathbb{R}^{|\mathcal{G}| \times |\mathcal{C}|}$ with $\mathcal{G}$ the set of rows associated to the genes (in case of scRNA-seq) and $\mathcal{C}$ the set of columns associated to the cells. The m_{ij} entry of this matrix is here assumed positive whenever gene i is expressed in cell j, and negative otherwise. Such data can be obtained from the (normalized) count data using an appropriate transformation, for instance $\log_{10}(x + 0.1)$.[1] The matrix $\mathbf{M}$ is assumed to be sparse in terms of positive expression values. However, a few genes may be expressed in nearly all cells, and are removed from $\mathcal{G}$ as one searches for gene markers of specific subpopulations rather than generic markers of high expression throughout the cell population.[2]

The goal of MicroCellClust is to select a subset of genes $I \subseteq \mathcal{G}$ and a corresponding subset of cells $J \subseteq \mathcal{C}$, i.e. a bicluster (I, J), representative of a small subpopulation of, by default,[2] highly expressed cells and highly specific genes. This goal is formalized below as a constrained optimization problem for which an optimal solution is denoted by (I^*, J^*):

$$(I^*, J^*) = \underset{\substack{I \subseteq \mathcal{G} \\ J \subseteq \mathcal{C}}}{\operatorname{argmax}} \sum_{i \in I} \left(\sum_{j \in J} m_{ij} - \kappa \sum_{k \in \mathcal{C} \setminus J} \max\{0, m_{ik}\} \right) \tag{1}$$

$$\text{such that } \frac{\left| \left\{ (i,j) \,\middle|\, i \in I, j \in J, m_{ij} < 0 \right\} \right|}{|I| \cdot |J|} \leq \mu \tag{2}$$

The objective function in (1) is composed of two terms. Maximizing the first term, i.e. $\sum_{i \in I} \sum_{j \in J} m_{ij}$, corresponds to the max-sum submatrix problem [2], [3]. One searches for a bicluster for which the sum of all the corresponding expression values (i.e. the submatrix $\mathbf{M}_{IJ}$) is maximal. This global sum criterion allows for variations within the expression values, as they are not compared in a pairwise fashion. This is well suited to scRNA-seq data given the presence of technical and biological noise [4], for instance unusually high expression due to transcriptional bursting or false negative values due to dropouts. Indeed, the sum criterion allows for some genes within the selected bicluster to be lowly, or even negatively,

[1] Using a pseudocount of 0.1 yields negative values (-1) for zero-counts, while (normalized) counts ≥ 0.9 remain positive.

[2] Alternatively, data normalization could be considered instead of plainly ignoring these genes, e.g. replacing their values by their opposite to look for absence of expression within some cells, or subtracting the median value to look for overexpression. Nevertheless, we stick here for clarity to the original interpretation as such genes are marginal in number in our experiments.

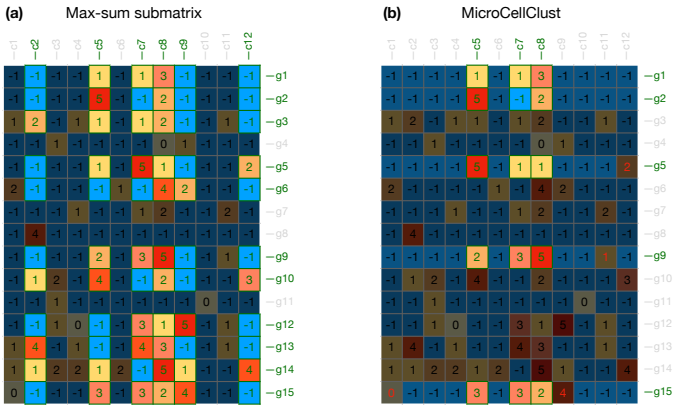

Fig. 1. Toy example of scRNA-seq data. **(a)** The max-sum solution, which forms a large bicluster with a low specificity. It includes, for instance, the cells c2 and c12 as they positively contribute to the objective (the sum of the selected entries in their respective column is positive) despite many negative values and a low similarity with the other selected cells. **(b)** The MicroCellClust solution with $\kappa = 1$ and $\mu = 10\%$. The three cells inside the selected bicluster have a similar expression of their five selected genes. These genes are also specifically expressed in this bicluster with only a few out-of-cluster expressions (red digits in out-of-cluster cells). Other genes, such as g3, are no longer selected since their out-of-cluster positive expression implies a negative contribution to the objective function.

expressed for some cells as long as their inclusion increases the objective value globally. The parameter μ (typically 10%) in the additional constraint (2) controls the proportion of negative values allowed in the solution to ensure the selected genes remain highly expressed across the bicluster as a whole.

As such, the result is not guaranteed to be highly specific, in the sense that the genes selected within the bicluster could also be highly expressed in other cells. The second term of (1) prevents such situation by penalizing positive expression of these genes (hence the $\max\{0, .\}$ operation) in out-of-cluster cells ($k \in \mathcal{C} \setminus J$). The parameter κ controls the relative influence of these two terms within the objective function. The higher κ, the fewer genes tend to be included in the solution, as fewer expression is tolerated outside the bicluster. Our results described in [1] suggest $\kappa = \frac{100}{|\mathcal{C}|}$ and $\mu = 10\%$ as good starting points (which can easily be tuned further without the need for prior knowledge). Fig. 1 shows the solution of MicroCellClust on a toy example.

Using exact approaches to solve this problem, e.g. by adapting CPGC [3], proved inefficient in practice as the max-sum upper bound is not tight. To solve it in reasonable time, we implemented a beam search (see Algorithm 2 when the line with `MCC 1` comment replaces the one with `MCC 2` comment). It explores the search space (composed only by all subsets of cells, as the optimal assignment of genes for each of them can be inferred in linear time) in a breadth-first fashion but keeping only the best solutions at each level to continue the search. It first evaluates all possible pairs of cells, selects the ones with highest objective to generate supersets of 3 cells (by adding each possible cell, one at a time, to these pairs), then evaluates them and selects the best ones to generate candidates with 4 cells, and so forth until no improvement in objective value is found during several successive search levels. Such a beam search algorithm has no guarantee of finding an optimal solution to problem (1)(2). Yet, it proved able to identify biologically relevant solutions on small and middle-scale data [1].

Each level of the beam search has a linear time complexity with respect to the number of cells, except the first where all possible pairs of cells are generated. This quadratic complexity poses a practical limit to solving data typically up to 10 000 cells. In this paper, we present an extension of this solver using new heuristics based on the FiRE or DoRC rareness scores [5], [6] to reduce the complexity of the beam search solver to a linear one, which scales to large data. This heuristic approach is used to quickly find an approximated solution to problem (1)(2) in large-scale data, given its efficiency to identify the region in which the optimum is located. The solution is further refined using simulated annealing [7], a classical derivative-free optimization algorithm that efficiently approximates the global optimum. Consequently, MicroCellClust 2 is able to successfully identify biologically relevant rare subpopulations within data containing over 100 000 cells.

II. The MicroCellClust 2 method

At the beginning of the beam search, the goal is to identify pairs of similar cells, which will consequently yield a higher objective value than a random pair. Let us assume a rareness score which assigns a high value to cells that present similarity with few other cells, i.e. are rare, and a low value to abundant ones. One can reasonably assume that two similar cells (rare or not) should also have relatively similar rareness scores, and therefore restrict the evaluation to pairs of cells that are similar within this score distribution instead of all $\mathcal{O}\left(|\mathcal{C}|^2\right)$ pairs. Finding the best pairs is then done in linear time by pairing each cell to the n closest ones according to the rareness score ($n \ll |\mathcal{C}|$; typically 100).

The FiRE method [5] computes such a score using an approach that is scalable to large datasets, based on the sketching technique [8], [9]. First, FiRE selects 1000 most variable genes (based on their relative dispersion, i.e. $\frac{\text{variance}}{\text{mean}}$). Then, it projects each cell to a low-dimensional bit signature (hash code) by randomly sampling some of the 1000 genes and converting their expression into binary values. The cells are then grouped by their hash code. A high score is given to cells in scarcely populated groups, and a low score to cells that are part of large groups. Several such runs are aggregated to produce a continuous rareness score, which should be high for rare cells. DoRC [6] provides an alternative using a similar approach, where the sketching is replaced by an isolation forest [10], which proved effective to detect anomalies (here rare cells) in data [11]. A tree is grown by splitting the data at each node according to randomly sampled genes (among the 1000 most variable ones). Cells located at a low depth are likely to be rare and are assigned a high score. As for FiRE, a continuous score is obtained by aggregation of several runs.

Pairing cells according to a rareness score reduces the complexity of the beam search from quadratic to linear.[3] Additional heuristics are used to further improve computational speed when considering a large number of cells. Indeed, most of them have a low rareness score, and will unlikely be part of any rare subpopulation. One could preselect cells that have a high rareness score, and consider only those as variables of the optimization problem (i.e. only these cells can be selected in J). Both FiRE and DoRC consider a cell as rare if its score is $\geq q_3 + 1.5 \cdot IQR$, where q_3 denotes the third quartile of the distribution and IQR the interquartile range.[4]

Using such heuristics greatly reduces the search space (typically less than 10% of the cells remain under consideration), but relies on FiRE/DoRC having a high recall, i.e. most interesting cells must be preselected (possibly among others). Results described in section III-B show this recall is not always high enough: some cells might not be included in the MicroCellClust solution as they failed the rareness threshold, even though they express a large part of the genes selected in the bicluster (and are therefore probably part of the same cell subpopulation). Yet, this can easily be detected so as to extend the search to these cells. We therefore propose a hybrid procedure to run MicroCellClust on large-scale data (the pseudocode of which is given in Algorithm 1):

1) Compute the rareness score for each cell using the FiRE (or alternatively DoRC) method, and select the cells passing the rareness threshold.
2) Run the beam search algorithm with only those cells as variables of the optimization problem. The objective is to identify at least part of the subpopulation of interest, with an initial set of marker genes.
3) Create a new search space containing cells that express part of the genes found in step 2 (whether or not they were previously candidates after step 1). Selecting cells expressing at least 50% of these genes seems appropriate according to our experiments to ensure a wide enough selection of candidate cells.
4) Take these cells as variables and search for an optimal solution by performing a local search starting from the solution of step 2.

We implemented a local search based on simulated annealing [7] (see Algorithm 3). The search space is explored by stochastically visiting neighbors of the current solution (obtained by adding or removing a cell from the current assignment). A neighbor is always accepted if it improves the current objective value (i.e. the value of (1) given the current assignment of variables); otherwise it is accepted according to a certain probability to avoid getting stuck in local optima. Multiple restarts are also performed to further diversify the search.

Algorithm 1: `MicroCellClust 2`

Input: $\mathbf{M} \in \mathbb{R}^{|\mathcal{G}| \times |\mathcal{C}|}$ an expression matrix
Input: $\mathbf{r} \in \mathbb{R}^{|\mathcal{C}|}$ a rareness score distribution for the cells
Input: $\kappa \in \mathbb{R}_+,\ \mu \in [0, 1]$ the parameters of (1)(2)
Output: $I^* \subseteq \mathcal{G},\ J^* \subseteq \mathcal{C},\ \Omega^* \in \mathbb{R}$ sets of genes and cells maximizing (1)(2), and corresponding objective value

```
// 1) Preselect rare cells
```
$C_{rare} = \left\{ j \in \mathcal{C} \,\middle|\, r_j \geq q_3(\mathbf{r}) + 1.5 \cdot IQR(\mathbf{r}) \right\}$
```
// 2) Beam search (see Algorithm 2)
```
$I^{\sim}, J^{\sim}, \Omega^{\sim} = \texttt{beam\_search}\left(\mathbf{M}_{\mathcal{G}C_{rare}}, \mathbf{r}, \kappa, \mu\right)$
```
// 3) Cells expressing genes of 2)
```
$C_{cand} = \left\{ j \in \mathcal{C} \,\middle|\, \left|\{i \in I^{\sim} : m_{ij} \geq 0\}\right| \geq 0.5 \cdot |I^{\sim}| \right\}$
```
// 4) Local search (see Algorithm 3)
```
$I^*, J^*, \Omega^* = \texttt{simul\_annealing}\left(\mathbf{M}_{\mathcal{G}C_{cand}}, I^{\sim}, J^{\sim}, \Omega^{\sim}, \kappa, \mu\right)$

This hybrid strategy efficiently explores the search space and produces results with high objective values in linear time, allowing the analysis of large-scale datasets. Indeed, the beam search using FiRE/DoRC is particularly effective at identifying a good approximate solution, i.e. it quickly focuses on the zone of interest. On the other hand, simulated annealing thoroughly explores this zone and therefore yields solutions with higher objective value than the beam search.

III. Results

Two scRNA-seq datasets are used to evaluate the ability of MicroCellClust 2 to identify rare cells in large-scale data. We show that MicroCellClust 2 identifies rare megakaryocytes within 68k peripheral blood mononuclear cells (PBMCs) [12]. MicroCellClust 2 also manages to find rare ependymal cells within 160k mouse brain cells [13]. This case study demonstrates that MicroCellClust 2 is more effective at identifying the entire subpopulation than FiRE or DoRC. Finally, subsampling experiments are performed on both datasets to confirm the linear complexity of the solver w.r.t. the number of cells.

For all experiments, read counts are normalized to counts per millions, and $\log_{10}(x+0.1)$ transformed. As in our original publication [1], genes expressed in more than 25% of the cells are left out (respectively 2% and 7% of the genes).

A. Identification of rare megakaryocytes within 68k PBMCs

Our first case study contains 68 579 PBMCs sampled from a healthy donor [12]. FiRE is used to obtain a rareness score for each cell.[5] MicroCellClust 2 executes as follows (see Algorithm 1):

1) 4238 cells are marked as rare using the FiRE threshold.
2) These cells are used as candidate variables for the beam search (Algorithm 2), which returns an initial approximation containing 174 cells and 21 genes.

[3]To produce this pairing, the cells must be sorted by their rareness score, which is $\mathcal{O}(n \log n)$ in general. Yet, linear sorting algorithms exist provided the rareness score is rounded to a fixed precision (which is sufficient for our proposed approach). In practice, this sorting operation is negligible in the overall runtime. Similarly, a naive implementation to choose the 100 best solutions at each level of Algorithm 2 would store all solutions and sort them. We use a more efficient approach using a priority queue to store a fixed number (100) of current best solutions in memory. Inserting a new one is done linearly w.r.t. the size of the queue (i.e. negligible compared to $|\mathcal{C}|$), and the worst one is discarded in constant time to keep a size of 100.

[4]Note that the FiRE or DoRC results form a collection of rare cells, but not *one* rare subpopulation. Since FiRE or DoRC do not contain any clustering mechanism, these cells might come from several different subpopulations.

[5]Using the same protocol as in [5], including a $\log_2(x+1)$ normalization. The FiRE results might differ according to the normalization used, but with limited (if any) impact for the MicroCellClust 2 end result.

Algorithm 2: `beam_search`

Input: $\mathbf{M} \in \mathbb{R}^{|\mathcal{G}|\times|\mathcal{C}|}$ an expression matrix
Input: $\mathbf{r} \in \mathbb{R}^{|\mathcal{C}|}$ a rareness score distribution for the cells (only for MicroCellClust 2)
Input: $\kappa \in \mathbb{R}_+,\ \mu \in [0,1]$ the parameters of (1)(2)
Output: $I^\sim \subseteq \mathcal{G},\ J^\sim \subseteq \mathcal{C},\ \Omega^\sim \in \mathbb{R}$ sets of genes and cells approximating (1)(2), and corresponding objective value

```
I~, J~, Ω~ ← ∅, ∅, 0
// Evaluate pairs of cells (ℓ = 2)
for j ∈ C do
    // for j' ∈ C \ {j} do                          // MCC 1
    for j' ∈ 100_closest(r, j) do                   // MCC 2
        J ← {j, j'}
        I, Ω ← obj(M, J, κ, μ)
        if Ω ≥ Ω~ then I~, J~, Ω~ ← I, J, Ω
// Evaluate next search levels
for ℓ ∈ {3, 4, ...} do
    for J⁰ ∈ 100_best(ℓ − 1) do
        for j ∈ C \ J⁰ do
            J ← J⁰ ∪ {j}
            I, Ω ← obj(M, J, κ, μ)
            if Ω ≥ Ω~ then I~, J~, Ω~ ← I, J, Ω
return I~, J~, Ω~
```

Fig. 2. Beam search algorithm as used by MicroCellClust 2. Only the generation of pairs of cells differs compared to MicroCellClust 1. The latter generates all pairs of cells (see line with `MCC 1` comment) which results in a quadratic complexity. MicroCellClust 2 reduces this complexity to a linear one by considering, for each cell, only a fixed number of cells to form pairs, i.e. the closest ones according to a rareness score (see line with `MCC 2` comment). This number, as well as the number of solutions to keep after each level, have been chosen empirically as their values are not critical to the end result (since one only looks for an approximate solution at this point). A value of 100 produces good results on all (large and small scale) datasets considered.

Algorithm 3: `simul_annealing`

Input: $\mathbf{M} \in \mathbb{R}^{|\mathcal{G}|\times|\mathcal{C}|}$ an expression matrix
Input: $I^0 \subseteq \mathcal{G},\ J^0 \subseteq \mathcal{C},\ \Omega^0 \in \mathbb{R}$ an initial solution to start the local search
Input: $\kappa \in \mathbb{R}_+,\ \mu \in [0,1]$ the parameters of (1)(2)
Output: $I^* \subseteq \mathcal{G},\ J^* \subseteq \mathcal{C},\ \Omega^* \in \mathbb{R}$ sets of genes and cells maximizing (1)(2), and corresponding objective value

```
I*, J*, Ω* ← I⁰, J⁰, Ω⁰
repeat n_r times
    J_curr, Ω_curr ← J⁰, Ω⁰                         // Reset current
    for t ∈ {0, ..., n_i} do
        // Generate a neighbor
        j ← random(C)
        if j ∈ J_curr then J ← J_curr \ {j}
        else J ← J_curr ∪ {j}
        // Evaluate the neighbor
        I, Ω ← obj(M, J, κ, μ)
        // Update current if applicable
        if Ω ≥ Ω_curr or random([0,1]) < e^(−(Ω_curr − Ω)/T(t)) then
            J_curr, Ω_curr ← J, Ω
            if Ω ≥ Ω* then I*, J*, Ω* ← I, J, Ω
return I*, J*, Ω*
```

Fig. 3. Simulated annealing algorithm. n_r is the number of restarts (the more, the higher the probability of approaching the global optimum; a value of 20 is an appropriate trade of w.r.t. time). n_i is the number of iterations within one restart, i.e. neighbors generated ($100 \cdot |\mathcal{C}|$ works well on all datasets). T defines a "temperature" function decreasing with time, e.g. $T(t) = T_0 \cdot \alpha^t$ with $\alpha \in]0,1[$, which ensures a relatively high probability to accept a worse neighbor at the beginning of a restart, so as to diversify the search, which gradually vanishes towards the end. Additionally, cells are ordered at the beginning by their expression over the initial genes. During the generation of a neighbor, the random selection is weighted so as to slightly favor promising cells: the unselected (resp. selected) cells that express these genes most (resp. less) have a higher probability to be added (resp. removed).

3) 226 cells are found to express more than 50% of these genes and are selected as candidates for the local search.
4) The simulated annealing (Algorithm 3) starts from the solution of step 2, and finds a new solution inside the search space defined in step 3 (which improves the objective value by 15%). This final solution contains 202 cells and 22 genes.

Fig. 4 shows the results. The presence of PF4 among the marker genes suggests MicroCellClust 2 identifies a rare subpopulation of megakaryocytes (0.3% of the cells), which is consistent with previous analyses of this data [5], [12].

Alternatively, one could use DoRC as rareness score, in which case 3951 cells are marked as rare (2898 of which are also part of the FiRE result). Interestingly, the exact same solution is returned by the beam search at step 2, and consequently the same subpopulation is identified as end result. This illustrates the stability of MicroCellClust 2 regarding the rareness score chosen as heuristics.

B. Identification of rare ependymal cells within 160k mouse brain cells

Our second case study contains 160 796 mouse brain cells, which were sampled to create a detailed census of cell types in the mouse nervous system [13], which may include rare ones. Fig. 5 shows the result after using MicroCellClust 2 on this data:

1) 1333 cells are marked as rare using the FiRE threshold.
2) The beam search (Algorithm 2) using these cells as candidates identifies a subpopulation of 392 cells and 64 genes. These genes are lowly expressed among the cells that were candidates after step 1 but not selected during the beam search. Interestingly, around 500 cells that were not candidates after step 1 (they failed the FiRE threshold and could therefore not be selected during the beam search) appear to highly express the identified genes.
3) Consequently, a new search space is defined by taking the 1144 cells which express at least 50% of the genes returned at step 2.
4) The local search (Algorithm 3) identifies a much larger subpopulation of 912 cells (0.6% of the data), characterized by 58 genes, which improves the objective value by 155%.

This subpopulation expresses many genes that have been linked to ependymal cells (including among others Rsph1, Dynlrb2, Tmem212, Foxj1, Riiad1, Ccdc153, ...) [13], which further confirms the ability of MicroCellClust 2 to identify biologically coherent rare subpopulations in large-scale data.

As seen in Fig. 5, a majority of the selected cells were not labeled as rare by FiRE or DoRC, despite 26 out of the 58 marker genes being part of the 1000 most variable

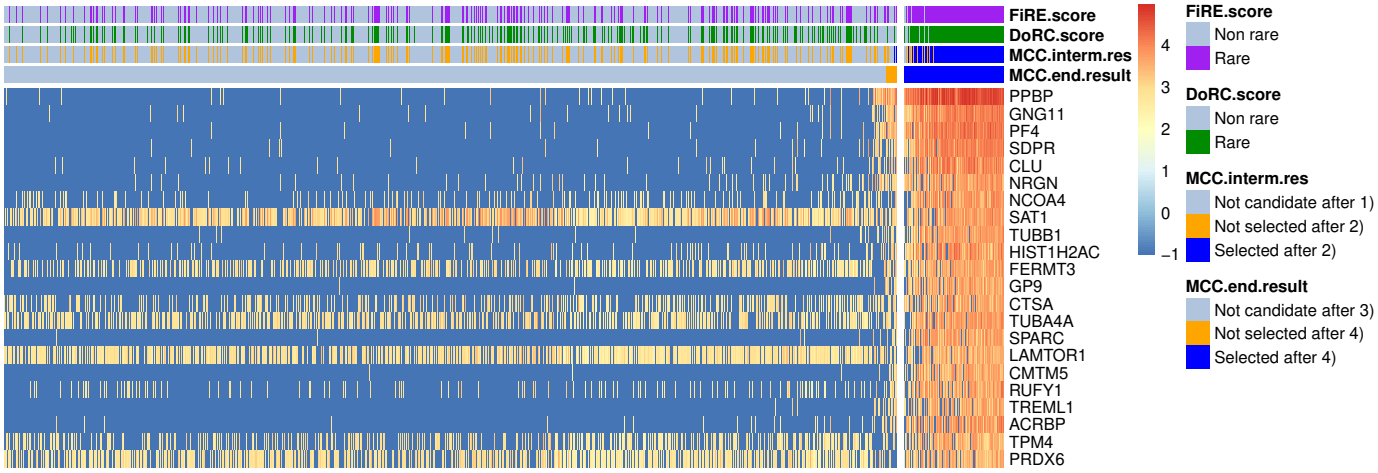

Fig. 4. Megakaryocyte subpopulation identified by MicroCellClust 2 within the 68k PBMCs. For the clarity of the visualization, cells are ordered increasingly by the sum of their expression over the selected genes and only the 2000 cells with highest sum are displayed (3% of the data). The MCC.iterm.res annotation indicates the result of the beam search using only the cells passing the FiRE rareness threshold (indicated by the FiRE.score annotation; the DoRC result is also indicated by DoRC.score). The MCC.end.result annotation indicates the result of the local search, i.e. the end result of MicroCellClust 2. It contains 202 cells that highly express the 22 marker genes, which are lowly expressed in the remaining cells (note that only a small fraction of them are displayed: it may appear that genes such as SAT1 are also highly expressed in the out-of-cluster cells, but this is not the case when considering all 66k of them).

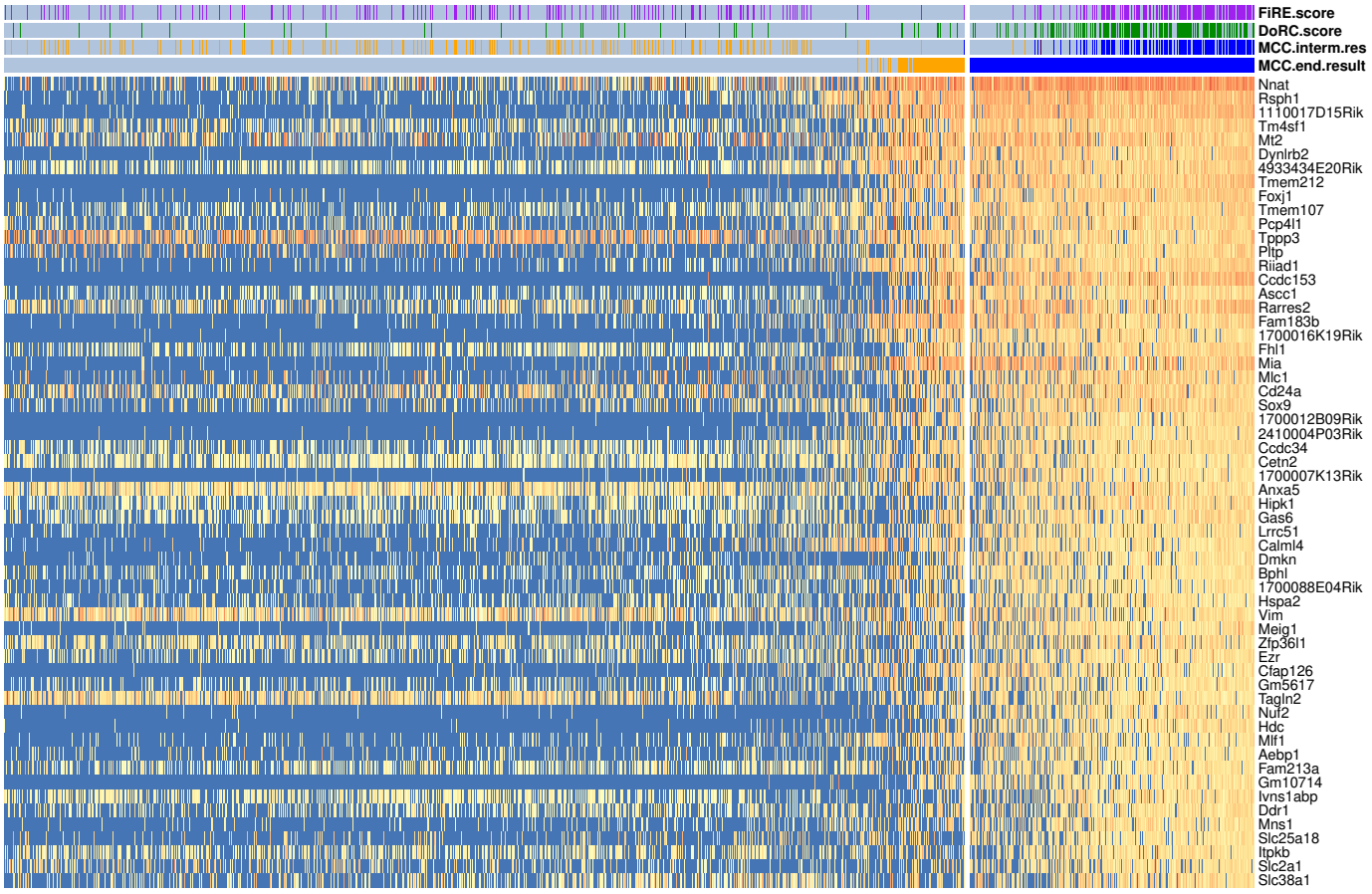

Fig. 5. Ependymal subpopulation identified by MicroCellClust 2 within the 160k mouse brain cells. Only 4000 cells (those with the highest sum of expression over the selected genes) out of the 160k ones are displayed (2.5% of the data). The 58 marker genes are highly expressed among the 912 cells selected by MicroCellClust 2. Both FiRE and DoRC only mark a fraction of these cells (respectively 398 and 447) as rare (see the FiRE.score and DoRC.score annotations), missing the majority of these cells. MicroCellClust 2 is therefore more effective at identifying the subpoulation as a whole. Note that the objective function provides an ordering of the cells (based on their sum of expression values over the selected genes), which allows to identify cells that are closely related even though they do not express all the selected genes (tuning the κ and μ parameters might slightly change which genes and cells are selected [1], but this doesn't change the underlying biological interpretation in any of our experiments).

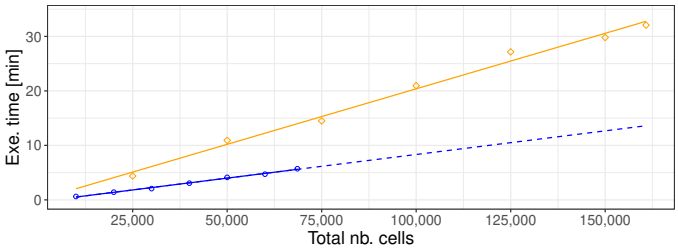

Fig. 6. Evolution of the runtime of MicroCellClust 2 for subsamplings of different size for (○) the 68k PBMCs dataset and (◇) the 160k mouse brain cells (Mac OS 11.6.5; 2.7 GHz Intel Core i7 CPU; 16 GB RAM)

genes used to compute the rareness scores. This illustrates the effectiveness of selecting both the cells and genes at the same time to identify a subpopulation as a whole. Indeed, FiRE and DoRC don't select genes when computing the rareness score, instead relying on an *a posteriori* clustering or differential expression analysis to identify marker genes. Such an approach would in this case only be able to identify half of the identified ependymal subpopulation as only half of them were labeled as rare.

C. Scalability

The use of a beam search algorithm to find a first approximation among rare cells, combined with local search to refine the solution, makes MicroCellClust 2 scalable to large volumes of data. Random subsamplings of both datasets are drawn to evaluate the execution time of the solver in function of the number of cells. Subsamplings of 10k up to 60k cells are drawn from the 68k PBMCs, and sets of 25k to 150k cells are drawn from the 160k mouse brain cells, while ensuring the same relative proportion of megakaryocytes (resp. ependymal cells) is kept. Fig. 6 shows MicroCellClust 2 has a linearly execution time w.r.t. the number of cells. It is also robust regarding variations of the data in the sense that a megakaryocyte (resp. ependymal) subpopulation was identified within each subset of the PBMC (resp. mouse brain) data.

IV. Conclusion

This paper describes a new approach to solve the optimization problem of MicroCellClust which scales to large single-cell dataset. MicroCellClust 2 first uses a rareness score, such as FiRE [5] or DoRC [6], as heuristics to efficiently find a good approximated solution in a limited search space using a beam search. It then uses a local search algorithm to enlarge the search space in order to approach the optimal solution. This method has a linear time complexity in terms of the number of cells, which is confirmed by the reported experiments on scRNA-seq data. Consequently, MicroCellClust 2 easily scales to datasets containing hundreds of thousands of cells to identify rare subpopulations with highly specific genes.

These experiments show the ability of MicroCellClust 2 to extract relevant subpopulations from large-scale data, together with specific marker genes. The joint selection of cells and genes ensures the globality of a subpopulation is highlighted, whereas methods that rely on *a posteriori* gene selection tend to miss part of the relevant cells. This confirms the relevance of the multivariate approach used by MicroCellClust 2.

MicroCellClust 2 takes as input a single data matrix representing expression values for a population of cells coming from a single biological sample, or from a few samples (e.g. different donors) simply put together. A natural extension would consider the sample (or donor) identity as a third dimension, beyond genes and cells. It would indeed be interesting to find specific subpopulations of cells and their associated marker genes that would also be common across different conditions or patients. Our future research work will adapt the optimization problem of MicroCellClust to address this generalized objective.

Software availability

`github.com/agerniers/MicroCellClust/`

Acknowledgment

The authors thank Prof. Siegfried Nijssen, Vincent Branders and Victor Hamer from UCLouvain for their useful insights.

References

[1] A. Gerniers, O. Bricard, and P. Dupont, "MicroCellClust: mining rare and highly specific subpopulations from single-cell expression data," *Bioinformatics*, vol. 37, no. 19, pp. 3220–3227, 2021.

[2] V. Branders, P. Schaus, and P. Dupont, "Identifying gene-specific subgroups: an alternative to biclustering," *BMC Bioinformatics*, vol. 20, p. 625, 2019.

[3] V. Branders, "Finding submatrices of maximal sum: applications to the analysis of gene expression data," Ph.D. dissertation, UCLouvain, 2021.

[4] H. Todorov and Y. Saeys, "Computational approaches for high-throughput single-cell data analysis," *FEBS journal*, vol. 286, no. 8, pp. 1451–1467, 2019.

[5] A. Jindal, P. Gupta, Jayadeva, and D. Sengupta, "Discovery of rare cells from voluminous single cell expression data," *Nature communications*, vol. 9, p. 4719, 2018.

[6] X. Chen, F.-X. Wu, J. Chen, and M. Li, "DoRC: Discovery of rare cells from ultra-large scrna-seq data," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2019, pp. 111–116.

[7] P. Van Laarhoven and E. Aarts, *Simulated annealing: Theory and applications*. Dordrecht: Springer, 1987.

[8] Z. Wang, W. Dong, W. Josephson, Q. Lv, M. Charikar, and K. Li, "Sizing sketches: A rank-based analysis for similarity search," in *Proceedings of the 2007 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, ser. SIGMETRICS '07. New York, NY, USA: Association for Computing Machinery, 2007, pp. 157—-168.

[9] Q. Lv, W. Josephson, Z. Wang, M. Charikar, and K. Li, "Ferret: A toolkit for content-based similarity search of feature-rich data," in *Proceedings of the 1st ACM SIGOPS/EuroSys European Conference on Computer Systems 2006*, ser. EuroSys '06. New York, NY, USA: Association for Computing Machinery, 2006, pp. 317—-330.

[10] S. Hariri and M. C. Kind, "Batch and online anomaly detection for scientific applications in a kubernetes environment," in *Proceedings of the 9th Workshop on Scientific Cloud Computing*, ser. ScienceCloud'18. New York, NY, USA: Association for Computing Machinery, 2018.

[11] G. A. Susto, A. Beghi, and S. McLoone, "Anomaly detection through on-line isolation forest: An application to plasma etching," in *2017 28th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, 2017, pp. 89–94.

[12] G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent *et al.*, "Massively parallel digital transcriptional profiling of single cells," *Nature communications*, vol. 8, p. 14049, 2017.

[13] A. Zeisel, H. Hochgerner, P. Lönnerberg, A. Johnsson, F. Memic *et al.*, "Molecular architecture of the mouse nervous system," *Cell*, vol. 174, no. 4, pp. 999–1014.e22, 2018.