MarcoPolo: a clustering-free approach to the exploration of differentially expressed genes along with group information in single-cell RNA-seq data

Chanwoo Kim[1,*], Hanbin Lee[2,*], Juhee Jeong[3], Keehoon Jung[3,4,5], Buhm Han[3,6]

1. Department of Electrical and Computer Engineering, Seoul National University, Seoul, Republic of Korea
2. Department of Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea
3. Department of Biomedical Sciences, BK21 Plus Biomedical Science Project, Seoul National University College of Medicine, Seoul, Republic of Korea
4. Department of Anatomy and Cell Biology, Seoul National University College of Medicine, Seoul, Republic of Korea
5. Institute of Allergy and Clinical Immunology, Seoul National University Medical Research Center, Seoul, Republic of Korea
6. Interdisciplinary Program in Bioengineering, Seoul National University, Seoul, Republic of Korea

Corresponding Author: Buhm Han (buhm.han.@snu.ac.kr)

*: These authors contributed equally.

ABSTRACT

A common approach to analyzing single-cell RNA-sequencing data is to cluster cells first and then identify differentially expressed genes based on the clustering result. However, clustering has an innate uncertainty and can be imperfect, undermining the reliability of differential expression analysis results. To overcome this challenge, we present MarcoPolo, a clustering-free approach to exploring differentially expressed genes. To find informative genes without clustering, MarcoPolo exploits the bimodality of gene expression to learn the group information of the cells with respect to the expression level directly from given data. Using simulations and real data analyses, we showed that our method puts biologically informative genes at higher ranks more accurately and robustly than other existing methods. As our method provides information on how cells can be grouped for each gene, it can help identify cell types that are not separated well in the standard clustering process. Our method can also be used as a feature selection method to improve the robustness against changes in the number of genes used in clustering.

INTRODUCTION

Single-cell RNA (scRNA) sequencing technology has offered opportunities to study the expressions of individual cells of a biological system. A common approach to analyzing the data in the first place is to perform unsupervised clustering method to group cells[1,2]. Then based on the clustering result, differentially expressed genes (DEGs) between the groups of cells are identified. Since people typically consider only the top-ranked DEGs in the analysis, in case the grouping is wrong, it is nearly impossible to rediscover other missed but informative genes. Thus, the interpretation of data via downstream differential expression analysis largely depends on the predetermined clustering result[3].

However, a clustering result has an innate uncertainty. It is hard to know whether the clustering result appropriately reflects the underlying biological structure of the data. Moreover, it is common that the clustering procedure is done repeatedly to get the seemingly best clustering result. At each trial, one performs clustering with a single parameter setting and verifies the grouping by comparing cluster-level expression profiles with the list of already known marker genes. These two steps can be repeated with many varying parameters until the identity of each group is confidently determined. However, there are numerous options and parameters, such as the methods for highly variable genes (HVGs) selection[4,5], the number of HVGs used, the methods for dimensionality reduction, and the parameter for the clustering resolution. Thus, the process is arduous and confusing in many cases and may often lead to data-specific overfitting because of picking the best observation from a large number of trials. Furthermore, our prior knowledge on marker genes is incomplete, making it hard to conclude that the obtained clustering result is the ground truth.

In this respect, there have been demands for methods that can extract differentially expressed genes from the data without being susceptible to uncertainty rooting from clustering results[3,6]. One possible approach is to manually examine the expression patterns of genes prioritized by HVG selection methods before clustering. However, the ways the HVG methods sort out genes are not adequate for finding DEGs. This is because most HVG methods only exploit the relationship between the variance and the mean, as they were designed to select genes as input for the dimensionality reduction step. In our analysis, we show that they do not assign higher ranks to DEGs that are obviously differentially expressed among cell types with sufficient precision compared to the standard DEG analysis with clustering. Another more advanced approach is singleCellHaystack, a recently developed method that extracts the list of candidates for DEGs by examining non-random expression pattern before clustering[7]. However, although the method may sort out DEGs more accurately than the HVG methods, it does not tell which groups of cells 'differentially' express the identified genes. Accordingly, a concept of fold change in DEG analysis cannot be established. Thus, like HVG methods, the expression patterns of genes should be examined manually. In this sense, strictly speaking, the method is more like finding highly variable genes. Before finding candidates for DEGs, this method uses a fixed threshold to determine whether a gene is either expressed or not in each cell, a concept of hard thresholding; it is known that expression between two subsets of cells may not be binary but bimodal[8–10]. Thus, two subsets of cells can be shown as a single group, if both express a gene but with different expression intensities.

We here propose MarcoPolo, a novel clustering-free approach to identify differentially expressed genes in scRNA data. Our method does not demand prior group information of cells in advance as it learns the tentative grouping of the cells with respect to the expression level directly from the given data. Thus, our approach is robust to uncertainties from clustering or cell type assignment, and it does not require our prior knowledge on marker genes to be complete. The main function of our method is to sort out genes with biologically informative expression patterns in advance of clustering with high precision. Additionally, our framework provides the analysis result in the form of an HTML file, so that researchers can conveniently interpret and make use of the result for various purposes.

MarcoPolo achieves a high precision in finding informative genes by using the following strategies. It first disentangles the bimodality inherent in gene expression and divides cells into two groups by the maximum likelihood estimation under a mixture model. Thus, it utilizes the fact that the difference of expression patterns of a gene between two subsets of cells can be bimodal.[8–10] Then, it goes through additional processes to confirm which genes have a differential expression pattern that is biologically feasible. Specifically, it uses a 'voting' system that it compares how cells are grouped for a gene with how cells are grouped for other genes in order to see if the groupings are similar. This is based on a biological phenomenon that a group of cells in a similar biological status co-expresses a subset of genes dependently[11]. Hence, if the grouping pattern of a gene is repeated, or supported, by many other genes, the gene is considered to be informative. In addition to this strategy, MarcoPolo uses specifically designed statistic to winnow genes of which expressions are bimodal and noteworthy. Using simulations and real data analyses, we show that our method can have utilities in various steps of scRNA analyses.

RESULTS

MarcoPolo is a mixture-model-based novel ranking approach to select differentially expressed genes

MarcoPolo fits a two-component Poisson mixture model and ranks the genes using the parameters estimated from the fitted model (**Methods**, **Figure 1a**). For a given marker, cells that are more likely to be part of the low expression component of the mixture distribution are named off-cells while those that are more likely to be part of the high expression component are named on-cells. We developed a novel ranking method called the *voting system* that prioritizes genes that exhibit a common expression pattern with other genes (**Figure 1b**). The intuition behind this ranking method is that a true biological entity will express multiple markers, therefore a gene that reflects this true entity will have other genes with a similar expression pattern. For example, gene 1, gene 3, gene 4, and gene 5 in the figure show the similar on-cell patterns in the top-right cluster. Therefore, the voting system assigns higher ranks to these genes. By contrast, gene 2 and gene 6 have no other genes that share the on-cell patterns, and thus are assigned lower ranks by the voting system. In addition to the voting system, we implement two more complementary ranking methods, *the proximity score* and the *Q-Q ranking system* (**Figure 1c**). We expect that cells within the same cluster will be close in distance in the low-dimensional representation space. Based on this idea, the proximity score computes the variance of the principal component (PC) values of on-cells and assign higher ranks to genes with low variance. Moreover, the two-component mixture model will fit better to genes with larger bimodality. Therefore, we compare the Q-score (the likelihood of the model) of the two-component model versus the one-component model. Genes with a bigger reduction in the Q-score are ranked higher by the Q-Q ranking system. The final ranking of the genes is then determined by combining these three scoring systems. In the end, our framework provides the analysis result in the form of an HTML file so that researchers can conveniently interpret and make use of the result for various purposes (**Figure 1d**). For each gene, the analysis result shows fold change based on the tentative group, a two-dimensional plot of cells, and a histogram of expression with annotated group information. It also contains statistics used by MarcoPolo to winnow genes of which expressions are noteworthy.

Clustering cells can inappropriately hinder the discovery of differentially expressed genes

As the differential expression analysis in downstream depends on the clustering result, if a clustering algorithm fails to properly cluster a group of cells, informative gene of that cluster will be missed. In this sense, the need for a method that can find DEGs independent of clustering was widely discussed[3,6,7]. We demonstrated this situation as a proof-of-concept using realistic simulation datasets generated by Symsim[12], a simulator of single cell RNA-seq experiments. Varying the probability that a gene has a non-zero type-specific expression effect size in the simulator, we generated 40 different simulation datasets (**Methods**). As expected, the more the gene effect sizes were turned off, the less clear the boundaries between the clusters became (**Supplementary Figure 2**). In this setting, we compared MarcoPolo with the standard workflow with clustering, two HVG methods included in Seurat package[13] (VST and DISP), and another clustering-free approach singleCellHaystack. We found that the performance of the standard workflow with clustering was negatively affected by the decrease of this parameter (**Figure 2**). When the probability of non-zero gene effect was set as 1e-2, the medians of standard workflow's area under curve (AUCs) of correct classification were relatively high (0.934~0.978), and the inter-quartile range (IQRs) of them were small (0.048~0.066). However, when the parameter was lowered to 5e-4, the medians went down (0.764~0.847), and the IQRs became large (0.154~0.309). Interestingly, the performance of singleCellHaystack showed a similar trend as the standard workflow. When the parameter was 1e-2, the median was relatively high (0.942), and the IQR was small (0.032); but when the parameter was 5e-4, the median went down (0.788), and the IQR became large (0.378). However, MarcoPolo and HVG methods were more robust to these changes. Their medians did not go below 0.835, and the IQRs remained all below 0.110.

MarcoPolo puts genes of with biologically feasible expression patterns at the top

We applied MarcoPolo to three real datasets. We used the human embryogenic stem cell (hESC) dataset[14], human liver cell dataset[15], and human peripheral blood mononuclear cell (PBMC) dataset[16]. In these datasets, the cells were processed either by fluorescence-activated cell sorting (FACS) or by manual curation based on known markers. Thus, we know the true types of cells and can define genes that are expressed cell-type-specifically, namely the cell-type markers. We defined the marker answer set for each data based on the true cell type labels and checked how well different methods put the markers at the top ranks, if the scRNA data without the true type was given (**Methods**). Overall, in terms of AUC, MarcoPolo showed comparable or even better performance than the standard DEG analysis with clustering (**Figure 3**). In the hESC dataset, different methods showed a relatively large difference in performance. In this dataset, MarcoPolo (AUC: 0.837) showed the best performance followed by HVG (VST) (0.751) and singleCellHaystack (0.738). In the Liver dataset, the performance rank was in the order of MarcoPolo (0.952), the standard DEG analysis (0.941~0.944), singleCellHaystack (0.912). In the PBMC dataset, there were only small differences in performance among the methods, where singleCellHaystack showed the highest AUC (0.994) with a slight difference from the second best method, MarcoPolo (0.988).

The ability of MarcoPolo to exploit bimodality facilitates the identification of cell types

Unlike singleCellHaystack, which uses a fixed threshold to determine whether a gene is either expressed or not in each cell, MarcoPolo adapts flexible thresholds to each gene's expression pattern. Thus, even if two subsets of cells express a gene at the same time, in case they differ in the amounts of expression, MarcoPolo can classify them into two separate groups. **Figure 4a** shows how cells in the hESC dataset were divided by MarcoPolo for each of four exemplary genes. For these genes shown, a certain cell type showed higher level of expression intensity than other cell types. MarcoPolo successfully identified them as a separate group. If MarcoPolo had used a fixed threshold, multiple cell types might have been identified as a single group. The reason why the distribution of on-cells and the distribution of off-cells slightly overlap in **Figure 4a** is that the probabilistic model used by MarcoPolo incorporates the cell size factors (see **Methods**). In other words, the grouping by MarcoPolo is not solely determined by expression but also slightly by cell size factors.

As MarcoPolo provides information on how cells can be grouped for each gene, one can guess how the clustering result would be like even before actually performing a clustering algorithm. We measured how many genes at the top ranks in MarcoPolo give sufficient information to fully segregate all cell types in each dataset (**Methods**). To this end, we used each gene's grouping information like cell-type marker information. We regarded a gene to be the marker of a cell type if the gene was expressed by most (70%) of the cells in the cell type. We started this mapping from the top ranked gene. As we gradually added genes of the next ranks to our set, more cell types started to be distinguished based on how those genes are associated with cell types. We stopped when all true types have a unique marker representation, in which situation we can say that the genes fully segregate all types (**Figure 4b, Supplementary Table 1**). For the hESC, Liver, and PBMC datasets respectively, only 14, 73, and 39 genes were sufficient to recover all labels in each dataset. On the other hand, 19, 117, and 94 genes were needed for singleCellHaystack. Thus, the genes found by MarcoPolo had better segregating power in determining true types compared to those found by singleCellHaystack.

MarcoPolo genes can distinguish cell types that are not distinguished by the standard pipeline

The human embryogenic stem cell (hESC) dataset[14] of Koh et al. is an example showing that the standard clustering approach can often fail to distinguish true cell types. In the 2D t-SNE coordiantes included in the downloaded dataset, the anterior primitive streak (APS) and mid primitive streak (MPS) are mixed together in the 2D t-SNE space (**Supplementary Figure 3**). Since t-SNE reflects the variance in the PC space, this mixture suggests that the clustering algorithm will likely fail to distinguish them as well.

We tried running t-SNE algorithm and clustering algorithm on the data from scratch using Seurat. We used the default setting (VST method to select 2,000 genes) to calculate PCs and generated t-SNE coordinates. In our analysis, similarly, APS and MPS were mixed together in the 2D plot (**Figure 5a**). When we applied the widely used clustering pipeline (FindNeighbors and FindClusters function of Seurat with resolution parameter of 2.0), APS and MPS were included in the same cluster (**Figure 5b**). This implies that, if one simply uses the standard pipeline, one would not be able to distinguish these two types and find DEGs between them.

In contrast, we found that the top ranked genes found by MarcoPolo distinguish these two types. For example, PCAT14 gene, ranked 14th by MarcoPolo, is the case. Based on the expression modality of the gene, MarcoPolo successfully identified MPS as one single group (**Figure 5c, d**). Although other cells, including APS, also expressed the gene, the distribution of their expression was shifted to the left. Accordingly, MarcoPolo was able to classify MPS and other cells into separate groups as they differed in the expression intensity.

Using MarcoPolo as a feature selection method to improve the robustness of the clustering process

In the standard scRNA-seq analysis pipeline, a subset of genes is selected to construct low dimensional representation for the clustering step. To obtain a clustering result that well reflects the structure of the underlying biological data structure, it is important to use informative genes as an input. As MarcoPolo and singleCellHaystack are designed to pick genes with informative differential expression, we used them as a feature selection method in the standard pipeline. We wanted to see if the robustness of the clustering procedure improved when we mixed genes identified by our method into the set of genes selected by the standard HVG method. In other words, we examined if the frequency of successful clustering increased when MarcoPolo (or singleCellHaystack) genes were employed. For how we chose the populations and how we determined if each clustering result separates them, see **Methods**. We conducted the analysis using the three real datasets (hESC, Liver, and PBMC datasets). Note that we assumed that the cell-type labels provided by the original publications are true in our analysis.

Briefly, we compared three categories of feature selection methods: only using genes selected by standard HVG method, using the mixture of HVG genes and MarcoPolo genes (HVG with MarcoPolo), and lastly using the mixture of HVG genes and singleCellHaystack genes (namely HVG with Haystack). In case we mix two different criteria such as HVG and MarcoPolo, we extracted the same number of genes from the top-ranked genes in each criterion.

In order to test the robustness of the feature selection method, we ran the same method multiple times using different parameters and settings. Then, we measured how many times a method gave a successful clustering over the wide range of tested parameters. Specifically, we varied the parameters and settings as follows. First, the HVG method is used in all three methods, whether it is used as a standalone or in a mixture. There are different HVG methods available, though. We tried two widely used methods implemented in Seurat (VST and DISP). Second, we varied the number of HVGs from 200 to 1,000 with the interval of 100 (9 numbers). Third, we varied the resolution parameter in Louvain clustering algorithm from 0.6 to 2.0 with the interval of 0.2 (8 parameters). To sum up, for each category of feature selection methods, we repeatedly clustered cells 144 times with different settings (2 HVG methods × 9 HVG numbers × 8 resolution parameters). We then calculated how many of those 144 trials succeeded in isolating populations that we defined as difficult to separate.

Overall, employing MarcoPolo for feature selection improved the separation of cell types (**Figure 6**). For the hESC dataset, when MarcoPolo genes were employed, the frequency of separating the populations improved from 27.8% to 59.7% compared to using HVGs alone. For the Liver dataset, it improved from 66.7% to 78.5%. For the PBMC dataset, it improved from 61.1% to 75.0%. When singleCellHaystack genes were employed, the results for the Liver dataset and the PBMC dataset improved from 66.7% to 69.4% and from 61.1% to 88.9% respectively. However, the result for the hESC dataset was reduced from 27.8% to 7.6%. When it comes to the overall performance of all the datasets, MarcoPolo, singleCellHaystack, and HVGs were 71.1%, 55.3%, and 51.9% respectively. Thus, although there was not a method that performed the best consistently for every dataset, MarcoPolo showed a high and robust performance less dependent on different datasets.

METHODS

Linear Poisson mixture model

To identify the expression modality in scRNA-seq data, we fitted the following Poisson mixture model to each gene's count data.

$$\log \mu_{nt} = \beta_0 + \sum\nolimits_p \beta_p x_{np} + \delta_t + \log s_n$$

Here, $t \in \{0,1\}$ indicates the two groups of bimodality, the on-cells and off-cells. Conditional on that cell n belongs to group t, μ_{nt} is the mean of Poisson distribution followed by the observed read count of cell n, y_{gn}. β_0 is an intercept. β_p are coefficients corresponding to covariate x_{np}. δ_t is group-specific over expression. s_n is the size factor of cell n.

Then, the loss function *Q* of each gene is defined using the log likelihood.

$$Q = -\log\left[\sum_n \left(\sum_t Poisson(y_{gn}|\mu_{nt})\right)\right]$$

We optimized this loss function using the Adam optimizer implemented in PyTorch.

As a result, for each gene, we learn how the cells in the datasets are divided into two groups according to the expression modality. Without loss of generality, we assume that the mean expression of group t=1 is larger than the mean expression of group t=0. We let the indicator variable $I_{gn} \in \{0,1\}$ denote the cluster assignment of a cell n according to gene g.

Sorting out genes of which expression patterns are biologically feasible

To sort out DEGs without taking the group information of the cells as an input, MarcoPolo uses the following multiple strategies.

(1) Voting system

Assuming that a gene of interest is truly related to a biological status, it is likely that there are more genes that are co-expressed in a way similar to the gene. In other words, if a gene reflects the structure of the dataset appropriately, its expression pattern will be replicated by other genes several times. We examine how many times the segregation pattern of a gene is repeated by calculating the voting score for each gene as follows.

$$v_g = \sum_{g'} v_{gg'} \text{ where if } \frac{\sum_n (I_{gn} \cdot I_{g'n})}{\min\left(\sum_n I_{gn},\ \sum_n I_{g'n}\right)} > 0.7 \ \ v_{gg'} = 1 \text{ , or else } v_{gg'} = 0.$$

Thus, the more times a gene is supported by other genes, the higher the gene's voting score becomes. Note that our formula above calculates what proportion of the cells that express a gene with a smaller on-cell count also expresses a gene with a larger on-cell count. We used this formulation because sometimes, one gene can be a marker gene of a group and another gene can be a marker gene of a subtype of that group. In such a case, we wanted to consider them as supportive of each other in our voting system.

(2) Proximity of cells in low dimension

Due to the cell lineage, a hierarchical structure is pervasive in scRNA-seq data. That is, heterogeneity from higher-level grouping determines the global structure, and the cell subtype affects the small signal in the expression. Thus, we assumed that the expression pattern of a gene corresponding to a meaningful biological status tends to align well with the global structure. For each gene g, we calculated the proximity of the on-cells (I_{gn}=1) in a low dimensional representation space. The underlying intuition is that, if a gene can explain the underlying structure, the cells expressing that gene will be clustered or proximal to each other. We first performed principal component analysis (PCA) of the count data with total 50 components. Then we calculated the proximity score as follows.

$$\mathrm{P}_g = \frac{\sum_i^5 \mathrm{Var}(\mathrm{PC}_{i,\mathrm{I}_{gn}=1})}{5}$$

The smaller this score is, we interpret the gene as more informative. Note that although this score tends to capture genes whose on-cells cluster together in the standard clustering approach, our method is not dependent on a specific clustering result. Our method can be interpreted as using the clustering information in PC space in a soft way, so that we can avoid errors induced by fixing the clusters. When calculating each variance of PC among cells, we omitted outliers of which value is above the top 30th percentile or under the bottom 30th percentile.

(3) Discrepancy between two groups

We measured the discrepancy between the high and low expression components of a given gene using several statistics. First, we compared the log-likelihood of the data under the null hypothesis with a single Poisson distribution and the alternative hypothesis with K=2 Poisson distributions with different means. We have developed a unique statistic defined as the ratio of the two log-likelihoods, namely QQ scores, as follows.

$$\mathrm{QQ}_{\mathrm{ratio,g}} = \frac{\mathrm{Q}_{null,g}}{\mathrm{Q}_{alt,g}}$$

This modeling can look unusual because the subtraction of the two log-likelihoods is more common in other statistical areas. We have found that the ratio statistic fits this problem well because the ratio is independent of the absolute read counts if the on-cell and off-cell count distributions are fixed. Because read counts can differ drastically from gene to gene, we found that this statistic performs well for determining the ranks of multiple genes.

Second, we also calculated the traditional subtraction between log-likelihoods under the null hypothesis and the alternative hypothesis. Again, we normalized the value as follows to account for the drastic difference in expression between genes.

$$\mathrm{QQ}_{diff,g} = \frac{\mathrm{Q}_{null,g} - \mathrm{Q}_{alt,g}}{\sum_n y_{gn}/n - \log\left(\sum_n y_{gn}/n\right)}$$

Third, we calculated how much the mean of alternative hypothesis shifts from the mean of the null hypothesis as follows.

$$\mathrm{MS}_g = \log\left(\frac{\sum_n I_{gn} \cdot y_{gn}}{\sum_n y_{gn}}\right)$$

Lastly, we calculated a log fold change between the mean expressions of two groups for each gene as follows.

$$\mathrm{lfc}_g = \log\left(\frac{\sum_n I_{gn} \cdot y_{gn}}{\sum_n (1 - I_{gn}) \cdot y_{gn}}\right)$$

(4) Final step of obtaining MarcoPolo score

Finally, we aggregate the abovementioned statistics to select genes that with biologically feasible expression patterns. We generate the nonparametric rank-based MarcoPolo score for each gene by combining v_g, $\mathrm{QQ}_{\mathrm{ratio,g}}$, $\mathrm{QQ}_{\mathrm{diff,g}}$, and MS_g.

$$\mathrm{MacoPolo}_g = \min\left(rank\left(v_g\right), rank\left(\mathrm{QQ}_{\mathrm{ratio,g}}\right), rank\left(\mathrm{QQ}_{\mathrm{diff,g}}\right), rank\left(\mathrm{MS}_g\right)\right)$$

After calculating this statistic, we removed outlier genes that satisfy one of the following conditions: (1) lfc_g is < 0.6, (2) both v_g and P_g are under the bottom 3rd percentile, or (3) $\sum_n I_{gn}$ is < 10 or under the bottom 30th percentile.

Datasets

Embryonic stem cell scRNA data

The Koh et al.[14] dataset consists of 531 human embryonic stem cells (hESCs) at various stages of differentiation. We extracted the data from the R package DuoClustering2018[17], which can be installed using Bioconductor package manager. The dataset contains 9 cell types. Among them, we used 8 cell types with both scRNA-seq data and bulk RNA-seq data, which are hESC (day 0), anterior primitive streak (day 1), mid primitive streak (day 1), DLL1+ paraxial mesoderm (day 2), lateral mesoderm (day 2), early somite (day 3), sclerotome (day 6), central dermomyotome (day 5). Koh et al. annotated the cell types through fluorescence-activated cell sorting (FACS). We used only the genes of which mean expression (log-normalized count) value across all cells was in the top 30th percentile.

Human liver scRNA data

The MacParland et al.[15] liver dataset consists of 8,444 cells of 11 cell types collected from 5 patients. We extracted the data from the R package HumanLiver, which can be downloaded from https://github.com/BaderLab/HumanLiver. After clustering cells, MacParland et al. determined the identity of each cluster using known gene expression profiles. We mapped 20 discrete cell populations identified by the authors to 11 unique cell types for our analysis (hepatocytes, ab T cells, macrophages, plasma cells, NK cells, gd T cells, LSECs, mature B cells, cholangiocytes, erythroid cells, hepatic stellate cells). We used the same criterion for filtering genes as the hESC dataset.

PBMC 4k scRNA data

We obtained PBMC 4k (peripheral blood mononuclear cell) dataset[16], namely Zhengmix8eq, from the R package DuoClustering2018. This dataset is a mixture of 3,994 FACS purified PBMC cells of 8 cell types, which are B cells, monocytes, naive cytotoxic cells, regulatory T cells, memory T cells, helper T cells, naive T cells, and natural killer cells. We used the same criterion for filtering genes as the hESC dataset.

Simulation data

We generated multiple scRNA-seq simulation datasets using Symsim[12], a simulator of single-cell RNA-seq experiment. Each dataset contained 1,000 cells with 5,000 genes sequenced. We modulated a parameter, the probability that the gene effect size is not zero, in the simulator. We ran simulations ten times for each combination of parameters with different random seeds. In total, 40 datasets were generated. For the rest of the parameters, we used the tree structure of numerous subpopulations (**Supplementary Figure 1**) using the following setting. The number of extrinsic variability factors (EVFs) was 10. The number of different EVFs between subpopulations (Diff-EVFs) was 5. The mean of a normal distribution from which gene effects are sampled was 1. The parameter bimod, modifying the amount of bimodality in the transcript count distribution, was 1.

Visualization

In all three real datasets, meta information of the samples including the t-SNE coordinates was available. Except for the PBMC dataset, we directly adopted the t-SNE coordinates from these metadata to plot **Figure**. As the t-SNE coordinates included in the PBMC dataset did not separate cell types enough, we newly obtained coordinates. For this dataset, we calculated PCs using genes ranked at the top by MarcoPolo. Then, we ran t-SNE again using the PCs.

Marker definition

We defined marker genes for each cell type following the same procedure described in Zhang et al[18]. Briefly, for each gene, we sorted the cell types in ascending order based on the mean expression level. We then calculated log fold change between two consecutive types in this order. We then chose the maximum value among the N-1 log fold change values, given N types. After calculating this maximum value for all genes, we used genes with maximum value in the top 200 as marker genes. Among them, we filtered genes with the maximum log fold change not larger than two-fold for real datasets and four-fold for simulation datasets.

AUC calculation

We compared the performance of MarcoPolo with those of HVG methods, singleCellHaystack, and Seurat standard pipelines. For HVG methods (VST or DISP), we used the lists of highly variable genes from Seurat. For singleCellHaystack, we used the default setting of haystack function included in the package. For the standard DEG pipeline with clustering, we used the default parameters of the Seurat package. We used 50 PCs as low dimension representation of cells. We used Wilcoxon Rank Sum implemented in FindAllMarkers function to find genes that are differentially expressed. We sorted the genes by their p-values. We used roc_auc_score function in scikit-learn package to calculate the AUC of each method and dataset pair.

Generating MarcoPolo HTML report

To help researchers conveniently interpret and make use of the result for various purposes, we provide the analysis result in the form of an HTML file. The report shows how cells can be grouped for each gene and provides statistics used to prioritize genes. In addition, it contains a biological description of each gene that was adopted from the NCBI Gene database[19]. The gene description information was downloaded from NCBI FTP server (https://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz).

Using the grouping information from MarcoPolo to identify cell types

MarcoPolo provides grouping information according to the bimodality in each gene's expression as a result. Using this information, we measured how many genes at the top in MarcoPolo gave sufficient information to fully segregate all cell types in each dataset. To this end, we regarded a gene to be the marker of a true cell type label if the gene was expressed in more than 70% of cells of the type. Then we reviewed the gene-label linkage in the order of genes appearing in the MarcoPolo report. When there is a unique gene-label linkage sequence that can be used to discriminate the corresponding label from others, we concluded that the label is identified. For singleCellHaystack, we used the same order in its result. However, as it does not estimate grouping information, we used a hard threshold of 1 for expression value to classify cells into two groups.

Using MarcoPolo as a feature selection method to improve the robustness of the clustering process

Since MarcoPolo and singleCellHaystack are designed to pick genes with informative differential expression, we used them as a feature selection method in the standard pipeline. We measured the frequency of successfully clustering populations difficult to separate when MarcoPolo genes or singleCellHaystack genes were employed. We selected the populations as follows. In the case of the hESC dataset and the PBMC dataset, we used populations that mixed together in t-SNE coordinates included in the dataset's meta information. For the hESC dataset, they were anterior primitive streak and mid primitive streak. For the PBMC dataset, they were naive cytotoxic cells, regulatory T cells, memory T cells, naïve T cells, helper T cells. In the case of the Liver dataset, as all populations were separated well in t-SNE coordinates, we used the three rarest populations. They were cholangiocytes, erythroid cells, hepatic stellate cells. We calculated the ARI based on the populations of interest in each trial. If the ARI was larger than the bottom 25th percentile in 432 trials of each dataset, we considered the trial to be separating the populations.

Code availability

MarcoPolo is available at the GitHub repository (https://github.com/ch6845/MarcoPolo). It is coded in Python 3.7 using PyTorch v1.4. The required packages are NumPy, SciPy, scikit-learn, and Pandas.

Data availability

MarcoPolo reports for real datasets used in the analysis are available online.

hESC dataset (https://ch6845.github.io/MarcoPolo/Kohinbulk_filtered/index.html)

Liver dataset (https://ch6845.github.io/MarcoPolo/HumanLiver_filtered/index.html)

PBMC dataset (https://ch6845.github.io/MarcoPolo/Zhengmix8eq_filtered/index.html)

Discussion

It was commonly discussed that the difficulty of clustering cells correctly undermines the reliability of the downstream analysis. Thus, there was a demand for methods to conduct downstream analysis regardless of the clustering result. One important part of the downstream analysis is finding DEGs. In this sense, recently, a clustering-free method to find DEGs by identifying a non-random pattern of expression was presented. However, although they identify genes that have informative expression patterns, this method does not tell about how the identified genes are 'differentially' expressed. It does not tell which groups of cells 'differentially' expressed the identified genes. Accordingly, a concept of log fold change in DEG analysis cannot be established. Before finding candidates for DEGs, the method uses a fixed threshold to determine whether a gene is either expressed or not in each cell, a concept of hard thresholding. Thus, if two groups of cells express a gene but with different expression intensities, they are shown as a single group.

One key missed property of gene expression in identifying DEGs is its bimodality. In other words, two cell types may express a gene at the same time, but they can be distinguished by the difference in expression intensity. Thus, even looking at only one gene's expression, it is possible to divide two cell types into different groups based on their expression modality. We made use of the bimodality by fitting a mixture model. Accordingly, our method was able to classify a cell type of which expression is much higher than others into a single group highly accurately. As a result, by providing grouping information, our method truly showed how a gene is 'differently' expressed among cells. Thus, our key contribution is that we exploited the bimodality of gene expression to learn the group information from the data in the process of finding informative genes without clustering.

Conclusion

We presented MarcoPolo, a clustering-free approach to the exploration of bimodally expressed genes in single-cell RNA-seq data. Our method exploits the bimodality of gene expression to learn the group information from the data in the process of finding informative genes without clustering. Using simulations and real data analyses, we showed that our method has advantages as follows. First, our method puts genes with biologically informative expression patterns at the top ranks accurately and robustly. Second, as our method provides information on how cells can be grouped for each gene, it can help identify cell types that are not separated well in the standard clustering result. Third, our method can also be used as a feature selection method to improve the robustness of the clustering process.

REFERENCES

1. Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet* **20**, 273–282 (2019).

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

3. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol* **21**, 31 (2020).

4. Yip, S. H., Sham, P. C. & Wang, J. Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. *Brief Bioinform* **20**, 1583–1589 (2018).

5. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* **15**, (2019).

6. Zhang, J. M., Kamath, G. M. & Tse, D. N. Valid Post-clustering Differential Analysis for Single-Cell RNA-Seq. *Cell Syst* **9**, 383-392.e6 (2019).

7. Vandenbon, A. & Diez, D. A clustering-independent method for finding differentially expressed genes in single-cell transcriptome data. *Nat Commun* **11**, 4318 (2020).

8. Dobrzyński, M. *et al.* Nonlinear signalling networks and cell-to-cell variability transform external signals into broadly distributed or bimodal responses. *J Roy Soc Interface* **11**, 20140383 (2014).

9. Birtwistle, M. R. *et al.* Emergence of bimodal cell population responses from the interplay between analog single-cell signaling and protein expression noise. *Bmc Syst Biol* **6**, 109 (2012).

10. Korthauer, K. D. *et al.* A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol* **17**, 222 (2016).

11. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* **16**, 278 (2015).

12. Zhang, X., Xu, C. & Yosef, N. Simulating multiple faceted variability in single cell RNA sequencing. *Nat Commun* **10**, 2611 (2019).

13. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

14. Koh, P. W. *et al.* An atlas of transcriptional, chromatin accessibility, and surface marker changes in human mesoderm development. *Sci Data* **3**, 160109 (2016).

15. MacParland, S. A. *et al.* Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat Commun* **9**, 4383 (2018).

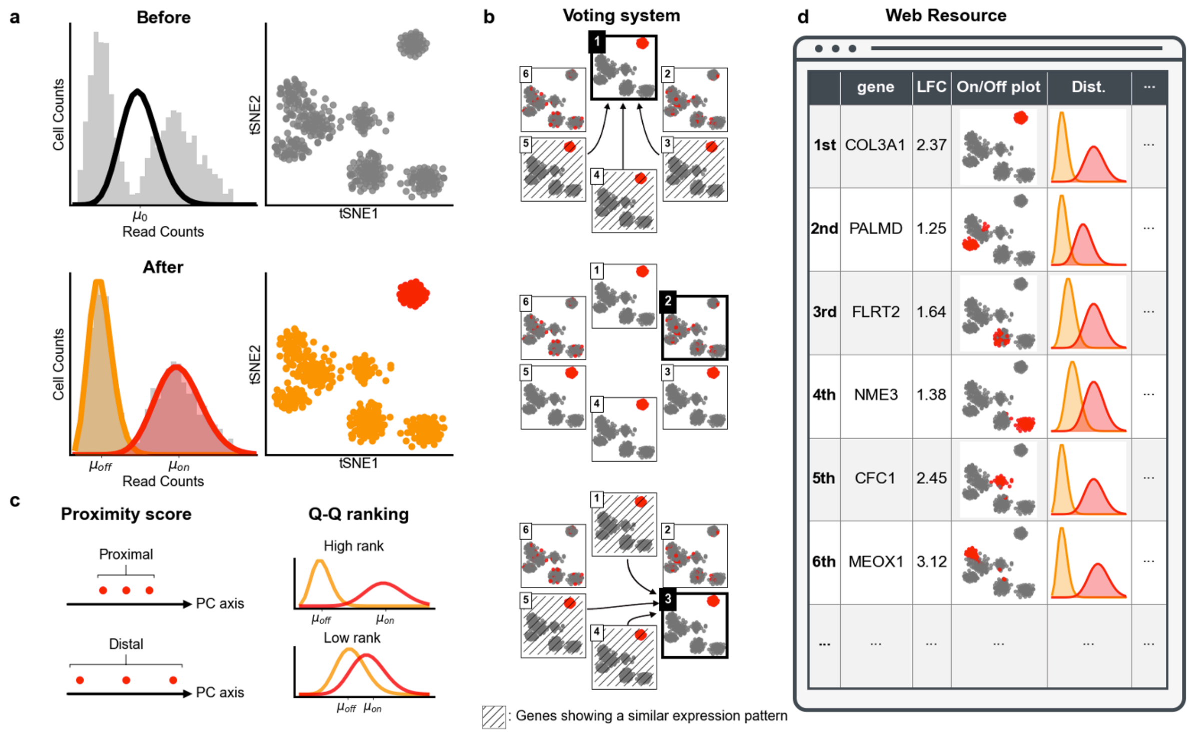
16. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049 (2017).

17. Duò, A., Robinson, M. D. & Soneson, C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000research* **7**, 1141 (2018).

18. Zhang, A. W. *et al.* Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nat Methods* **16**, 1007–1015 (2019).

19. Coordinators, N. R. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **46**, D8–D13 (2017).

FIGURES



	gene	LFC	On/Off plot	Dist.	···
1st	COL3A1	2.37			···
2nd	PALMD	1.25			···
3rd	FLRT2	1.64			···
4th	NME3	1.38			···
5th	CFC1	2.45			···
6th	MEOX1	3.12			···
...	···	···	···	···	···

Figure 1. Overview of MarcoPolo. **(a)** MarcoPolo fits a two-component Poisson mixture model. In t-SNE plot, on-cells, which are more likely to be part of the high expression component of the mixture distribution, are colored red. Off-cells, which are more likely to be part of the high expression component of the mixture distribution, are colored orange. **(b)** In voting system, genes exhibiting a common expression pattern with other genes are prioritized. For each gene, the on-off assignment of the gene is compared with those of other genes. Genes that are supporting the compared one are indicated using arrow. Thus, gene 1 and gene 3 are supported three times. On the other hand, gene 2 in the middle is not supported by any of genes shown. **(c)** In proximity score system, the variance of the principal component (PC) values of on-cells is calculated for each gene. Higher ranks are assigned to genes with low variance. In Q-Q ranking system, Genes with a bigger reduction in the Q-score are ranked higher by the Q-Q ranking system. **(d)** MarcoPolo framework offers the analysis result in the form of an HTML report. For each gene, log fold change, On/Off plot, biological description, and many statistics calculated by MarcoPolo are shown. On/Off plot shows how cells can be grouped for each gene according to the bimodality in its expression.

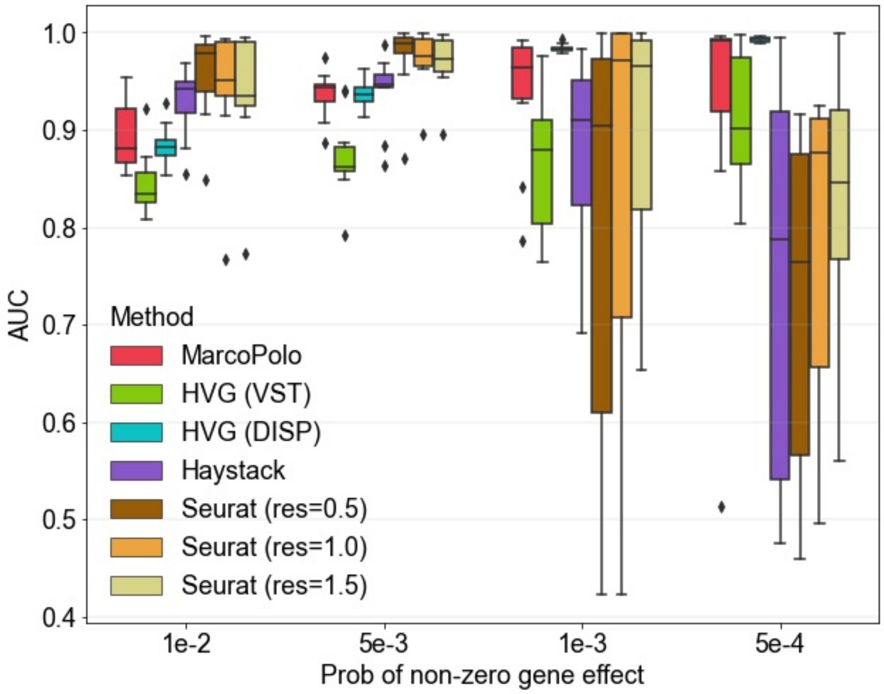


Figure 2. AUC of the genes listed by MarcoPolo, HVG methods, and Seurat pipeline when changing the probability of non-zero gene effects in simulation dataset.

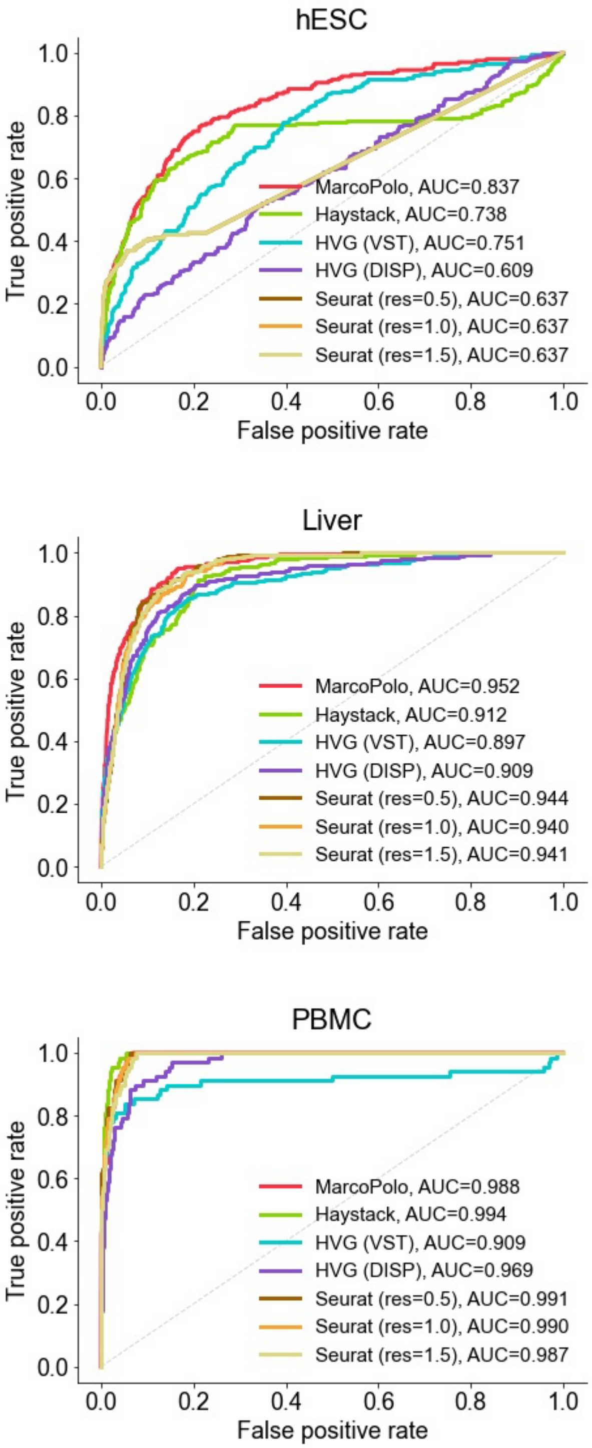


Figure 3. The lists of genes were extracted by MarcoPolo, HVG methods, and Seurat pipeline, respectively. a. ROC curves and their AUCs of called genes in the hESC dataset. b. ROC curves and their AUCs of called genes in the Liver dataset. a. ROC curves and their AUCs of called genes in the PBMC dataset.

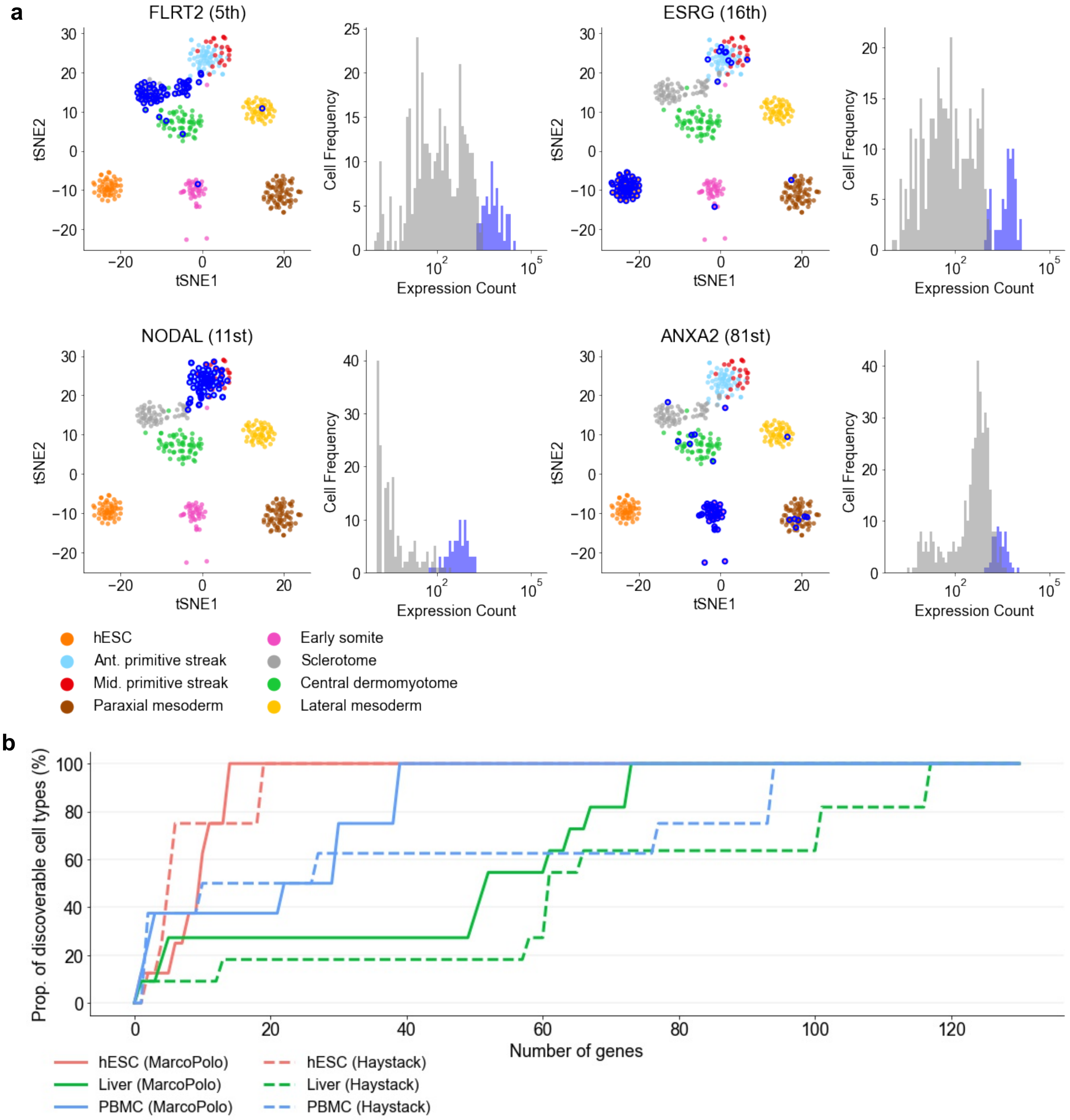


Figure 4. (a) t-SNE plots of the hESC dataset labeled by the true cell-type labels. The number beside the name of each gene is the rank assigned by MarcoPolo. Each histogram shows the expression pattern of a gene in the dataset. The edge of dots that corresponds to on-cells for each gene was colored blue. (b) The proportion of recovered cell types in three real datasets with the increasing number of genes reviewed in MarcoPolo result.

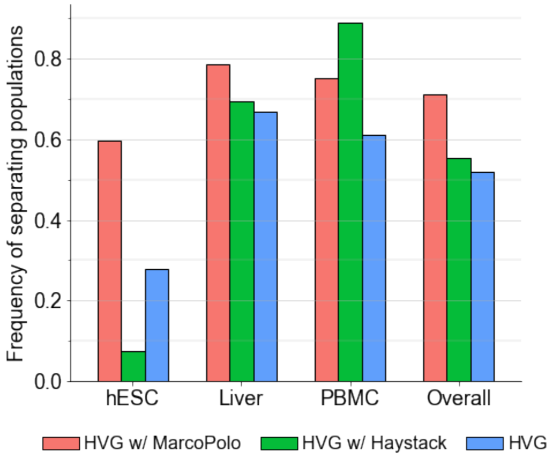


Figure 5. Frequency of successfully separating populations of interest over multiple trials with varying parameters. Using each feature selection method, we selected genes that are used by the downstream clustering analysis. We tried 288 different settings and parameters for the clustering analysis and measured the frequency that the clustering was successful. See Methods for how we defined the success of clustering.

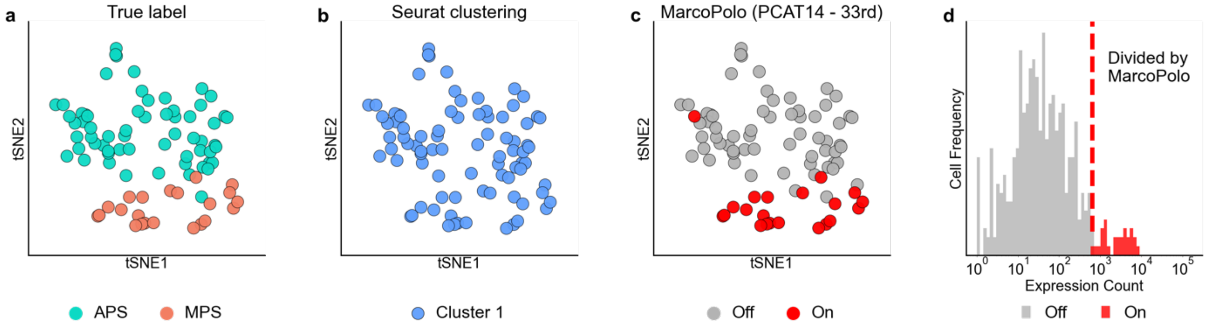
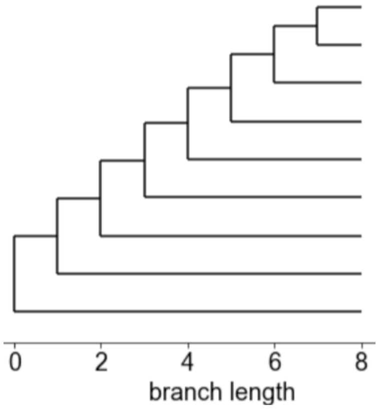
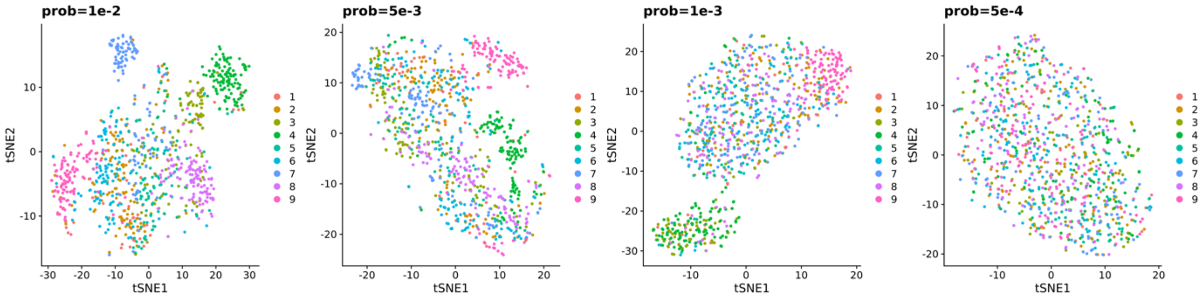


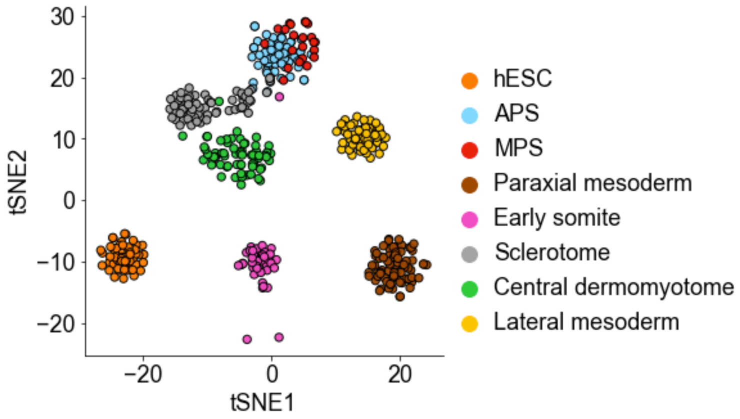
Figure 6. (a)-(c) t-SNE plot of the hESC dataset generated by the default pipeline of Seurat. We only showed anterior primitive streak (APS) and mid primitive streak (MPS). (a) We colored cells by the true cell-type label. (b) We colored cells by Seurat clustering result. (c) For each gene, MarcoPolo learns how the cells in the datasets are divided into two groups according to the expression modality. We colored cells based on how MarcoPolo separated cells for the PCAT14 gene. (d) We showed the expression pattern of the PCAT14 gene for all cells. The on-cells identified by MarcoPolo are colored red.



Supplementary Figure 1. Tree structure of subpopulations used to generate simulation datasets.



Supplementary Figure 2. t-SNE plot of simulation datasets when changing probability of non-zero gene effects.



Supplementary Figure 3. t-SNE plot included in the hESC dataset. The dataset was preprocessed by Duò et al[17].