# RgCop-A regularized copula based method for gene selection in single cell rna-seq data

**Snehalika Lall[1,+], Sumanta Ray[2*,+], and Sanghamitra Bandyopadhyay[1]**

[2]Genome Data Science, Bielefeld University, Bielefeld, Germany
[1]Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India.
[*]Sumanta.Ray@cwi.nl
[+]these authors contributed equally to this work

## ABSTRACT

We propose *RgCop*, a novel **reg**ularized **cop**ula based method for gene selection from large single cell RNA-seq data. *RgCop* utilizes copula correlation ($Ccor$), a robust equitable dependence measure that captures multivariate dependency among a set of genes in single cell expression data. We raise an objective function by adding a $l_1$ regularization term with $Ccor$ to penalizes the redundant co-efficient of features/genes, resulting non-redundant effective features/genes set. Results show a significant improvement in the clustering/classification performance of real life scRNA-seq data over the other state-of-the-art. *RgCop* performs extremely well in capturing dependence among the features of noisy data due to the scale invariant property of copula, thereby improving the stability of the method. Moreover, the differentially expressed (DE) genes identified from the clusters of scRNA-seq data are found to provide an accurate annotation of cells. Finally, the features/genes obtained from *RgCop* can able to annotate the unknown cells with high accuracy.
The corresponding software is available in: https://github.com/Snehalikalall/RgCop

## Introduction

With the advancement of single cell RNA-seq (scRNA-seq) technology a wealth of data has been generated allowing researchers to measure and quantify RNA levels on large scales[1,2]. This is important to get valuable insights regarding the complex properties of cell type, which is required for understanding the cell development and disease. A key goal of single cell RNA-seq analysis is to annotate cells within the cell clusters as efficiently as possible. To do this, the basic goal would be to select a few informative genes that can lead to a pure and homogeneous clustering of cells. Generally, the task of selecting the effective genes among all gene panel that can precisely discriminate cell type labels can be regarded as a combinatorially hard problem.

The usual approach for annotating cells is to cluster them into different groups which are further annotated to determine the identity of cells[3,4]. This is considered a popular and unsupervised way of annotating different types of cells present in a large population of scRNA-seq data[5–8]. The general pipeline of downstream analysis of scRNA-seq data typically goes through several steps. Starting from the processing of the raw count matrix, the scRNA-seq data is going through the following steps: i) normalization (and quality control) of the count matrix ii) feature selection, and iii) dimensionality reduction[9,10]. The first step is necessary to adjust discrepancies between samples of individual cells. Several quality measures are also applied to reduces the skewness of the data. The next step identifies the most relevant features/genes from the normalized data. The relevant genes are either selected by identifying the variation (highly variable genes)[11,12] or can be selected by calculating the expression levels across all cells which are higher than the average value (highly expressed genes)[9].

The selection of top genes has a good impact on the cell clustering process in the later stage of downstream analysis. A good clustering can be ensured by the following characteristics of features/genes: the features/genes should have useful information about the biology of the system, while not including features containing any random noise. Thereby the selected features/genes reduce the data size while preserving the useful biological structure, reducing the computational cost of later steps.
The usual approach of gene selection is based on the high variability of the gene expression label of scRAN-seq data. This process is simple and suffers from several disadvantages: i) as the expression variability is dependent on pseudo-count, it can introduce biases in the data, ii) Next, PCA is applied in downstream analysis for dimensionality reduction which is not suitable for sparse and skewed scRNA-seq data.

In this paper, we present a method for finding the most informative features/genes from large scRNA-seq datasets based on a robust and equitable dependence measure called copula correlation (Ccor). Although the major applications of copula can be found in the domain of time series, finance, and economics, but it is now ripe for application in different domain of bioinformatics such as for modeling directional dependence of genes[13], finding differentially co-expressed genes[14] high dimensional clustering for identification of sub-populations of patients selection[15] and many more. However, the application of

copula in single cell domain, particularly on gene selection is less explored.

In this paper, we show that employing a simple $l_1$ regularization term with the proposed objective function, will improve the performance of any clustering/classification model significantly. The objective function is utilized a new robust-equitable copula correlation (Ccor) measure on one hand and a regularization term to control the coefficient on other hand. Thus it is robust due to regularization and not susceptible to noise due to scale-invariant property copula. *RgCop* has major advantages both in clustering/classifying unknown samples and in the identification of meaningful marker detection. The latter point is addressed because novel marker genes for different cell types are identified with the cell clusters. Biologically meaningful marker selection is usually an important step in the downstream analysis of scRNA-seq analysis. This depends on the homogeneity of the cell clusters identified after the gene selection stage. Our proposed method selects the most informative genes that ultimately leads to a homogeneous grouping of cells of the large scRNA-seq data.

Beyond selecting a good informative gene set that leads good clustering/classification of cells, we also demonstrate that our method performs well in completely independent data of the same tissue. We demonstrate this by evaluating the performance of the selected features in completely unknown test samples. We observed that the selected features are equally effective for clustering/classifying the unknown test samples. We further carry out a simulation study on eight Gaussian Mixture datasets to establish the effectiveness of the proposed method. The results show that the proposed method not only select genes with high accuracy, but is also robust and less susceptible to noise when the parameters are appropriately tuned.

### Summary of contributions

The main contributions of the paper are summarized below:

- We provide one of the first regularized copula correlation (Ccor) based robust gene selection approaches from large scRNA-seq expression data. This robustness is a characteristic of our proposed objective function. The method also works equally well for the small sample and large feature scRNA-seq data.

- We derive a new objective function using the copula correlation and regularization term and theoretically prove that the selected feature set is optimal with respect to the minimum redundancy criterion (see Method).

- The objective function of *RgCop* is so designed that it can simultaneously maximize the relevancy criterion and minimize the redundancy criterion among the two sets of features/genes. The regularization term is also added with the objective function so as to control the large coefficient of the relevancy term.

- *RgCop* can able to effectively cluster/classify unseen scRNA-seq expression data. Annotating unknown cells is crucial and is the final goal of scRNA-seq data analysis. We demonstrate the applicability of our framework in this case. We demonstrate that the selected features are effective for clustering completely unseen test data. The annotation of cells can also be done in a supervised way if one can train a classification model with the selected features.

- Our method is less sensitive (robust) against the noises present in the scRNA-seq data. Our objective function uses copula-correlation, a robust-equitable measure which has the advantage of capturing the multivariate dependency between two sets of random variables. The scale invariant property of the copula ensures the robustness of our method.

## Results

In this section, we first describe the workflow of our methodology and then discuss the basic underline theory that supports *RgCop*.

### Workflow

Figure 1 provides a workflow of the whole analysis performed here. Following subsections discussed the important steps:

***A. Preprocessing of raw datasets*** See -A of panel-'*RgCop* framework for gene selection' of figure 1. Raw ScRNA-seq datasets are obtained from public data sources. The counts matrix $M \in \mathscr{R}^{c \times g}$, where $c$ is number of cells and $g$ represents the number of genes, is normalized using a transformation method (Linnorm)[16]. We choose that cells have more than a thousand genes expression values (non zero values) and choose that genes which have the minimum read count greater than 5 in at least 10% among all the cells. Then, the $log_2$ normalization is employed on The transformed matrix by adding one as a pseudo count.

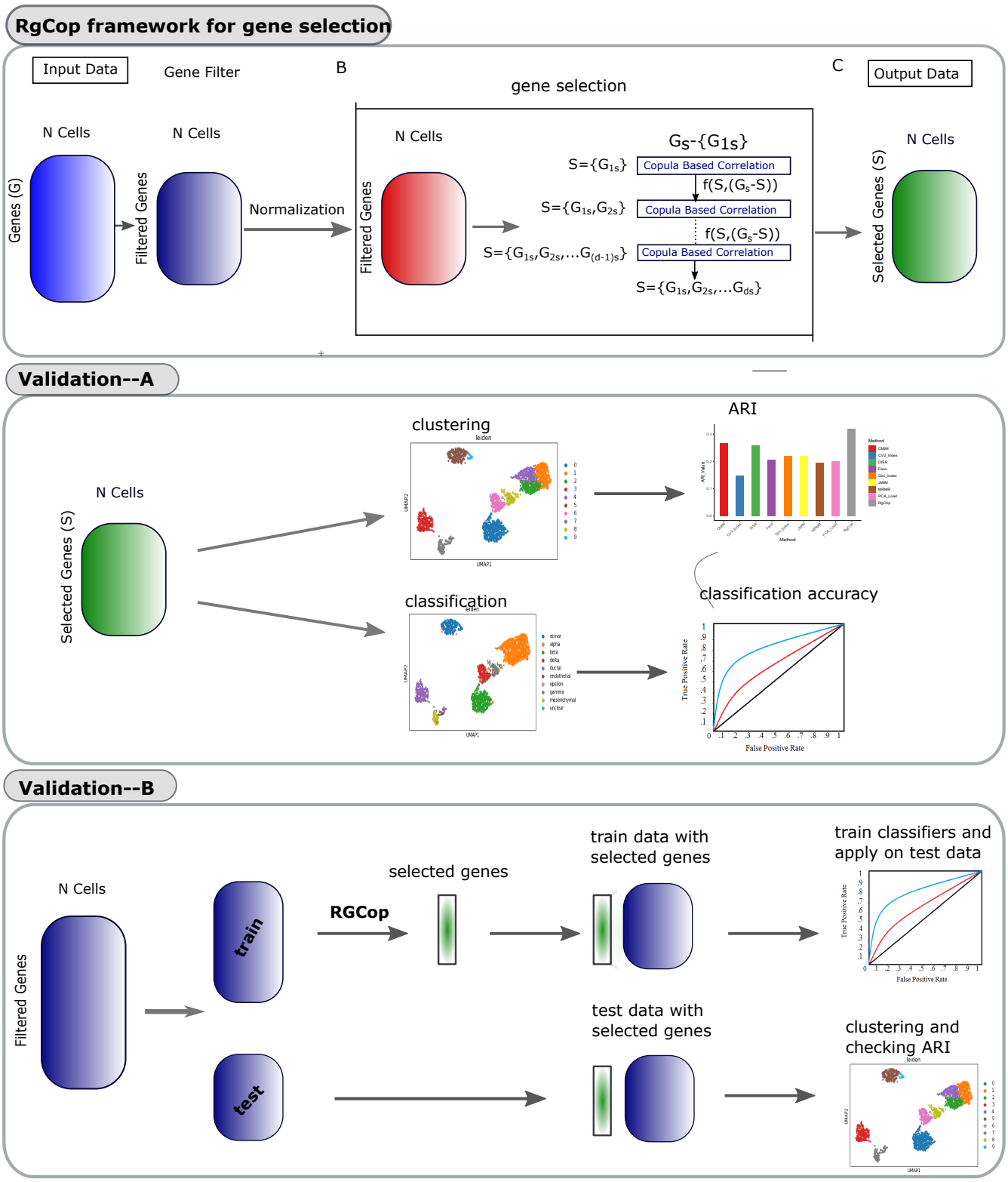


**Figure 1.** The whole workflow of the methodology:*RgCop* framework for gene selection is provided in the top panel. Clustering and classification is performed with the genes obtained from *RgCop* to validate the method (shown in middle panel). *RgCop* is validated for detection of unknown sample by splitting the data into train-test ratio of 7:3 (shown in the bottom panel). The test data is utilized for validation of the selected genes by RgCop.

***B. RgCop framework for feature selection*** See -B of panel-'*RgCop* framework for gene selection' of figure 1. The preprocessed data is used in the proposed copula-correlation (*Ccor*) based feature/gene selection models. First, a feature ranking is performed based on the *Ccor* scores between all features and class labels. We assume the feature having a larger *Ccor* value is the most relevant one and we include it in the selected list. Next, *Ccor* is computed between the selected relevant features and the remaining features. The feature with a minimum score is called the most essential (and not redundant) feature and included in the selected list. The process continued in an iterative way by including the most relevant and minimum redundant features in each step every time in the list. Feature selection in this way ensures the list of genes will be optimal (see proof of correctness). A $l_1$ regularization term is added with the objective function to penalize the large coefficient of relevancy term. The resulting matrix is then normalized (preprocessed) count matrix with selected features (see -C of panel-'*RgCop* framework for gene selection' of figure 1)

***C. Validation through clustering*** See panel-'validation-A' of the figure 1. We adopt the conventional clustering steps of scanpy[17] python package to cluster the resulting matrix obtained from the previous step. First, principal component analysis is performed to reduce the dimension and for creating a neighborhood graph of cells. We employed two clustering techniques (SC3[3], and Leiden clustering[18]) for clustering the neighborhood graph of cells. To validate the clusters we utilize the Adjusted Rand Index (ARI) metric which is usually used as a measure of agreement between two partitions. We compare the ARI score of *RgCop* with different state-of-the-art unsupervised feature selection method.

***D. Validation through classification*** See panel-'validation-A' of the figure 1. We validate the selected features by employing several classifiers to train the resulting matrix obtained from step-B. The features are selected by several supervised feature selection algorithm and the classification accuracy are compared with *RgCop*.

***E. Annotating unknown cells*** See panel-'validation-B' of the figure 1. For cells of the unknown type, *RgCop* can able to accurately cluster/classify the cells using the genes selected in the previous step. Annotating unknown cells are crucial and maybe the final goal of scRNA-seq data analysis. The filtered and preprocessed data is divided into train-test ratio 7:3 and the train set is utilized to obtain the selected features using *RgCop*. Several classifier models are trained on train data with the selected feature set and apply to the test set to know the accuracy. The test data with selected features are also used for clustering purposes. The ARI is computed on the clustering results. This provides the validation of our approach to work in practice.

***F. Marker identification*** We detect highly differentially expressed (DE) genes within each cluster obtained from step-C in the workflow. Here we utilizing Wilcoxon Ranksum test to identify DE genes in each cluster. The top five DE genes are chosen from each cluster according to their p-values.

## RgCop performs well in data with overlapping and non-overlapping classes

For single cell clustering the most common challenge is to discriminating samples between major cell types and its sub-types. Samples of similar cell types tend to overlap within one cluster, discriminating of which required sophisticated method that can extract features from overlapped samples. To explore whether RgCop can address this issue we apply it on data with overlapping and nonoverlapping classes and evaluate clustering performance on it.

**Generating synthetic data with overlapped and non-overlapped classes** We generate four pair of synthetic Gaussian mixture datasets, each pair consisting of overlapping and non-overlapping classes. We create a pair of synthetic data by varying the number of classes as $k = \{2,3,4,5\}$. For each dataset, we generate 50 relevant features by varying the mean ($\mu$) in a known set of range and create 250 irrelevant features by generating white Gaussian noise[19] with mean 0 and standard deviation 1. The covariance matrix ($\Sigma$) is fixed for all data and is generated with the formula $\Sigma = (\rho^{|i-j|})$, where $i, j$ denote the row and column index, and the value of $\rho$ is taken as 0.5. The detail descriptions and parameters distribution for generating data are given in Table 1 and 2. Figure 2, panel-A illustrates a 2D pictorial representation of the synthetic gaussian mixture datasets.

**Exploring performance of RgCop on synthetic data** We selected top 100 features from the generated datasets. $k$-means clustering is employed to cluster these data with the selected features. The clustering performance is evaluated by computing Adjusted Rand Index (ARI), a widely used metric to measure concordance between the predicted clusters and the known groups. Figure 2, panel-B describes the ARI values collected from the clustering results of the data with four and five overlapping and non-overlapped classes. The higher values of ARI suggests RgCop can perform well in both overlapping and non-overlapping classes.

**Tuning the regularization parameter** To tune the regularization parameter $\gamma$ (see the objective function in equation 13), the clustering process is repeated for nine set of values ranging from 0 to 0.5 ($\gamma = \{0, 0.002, 0.005, 0.009, 0.02, 0.07, 0.09, 0.3, 0.5\}$). The clustering results is reported for the synthetic data with four and five overlapping and non-overlapping classes (figure 2,

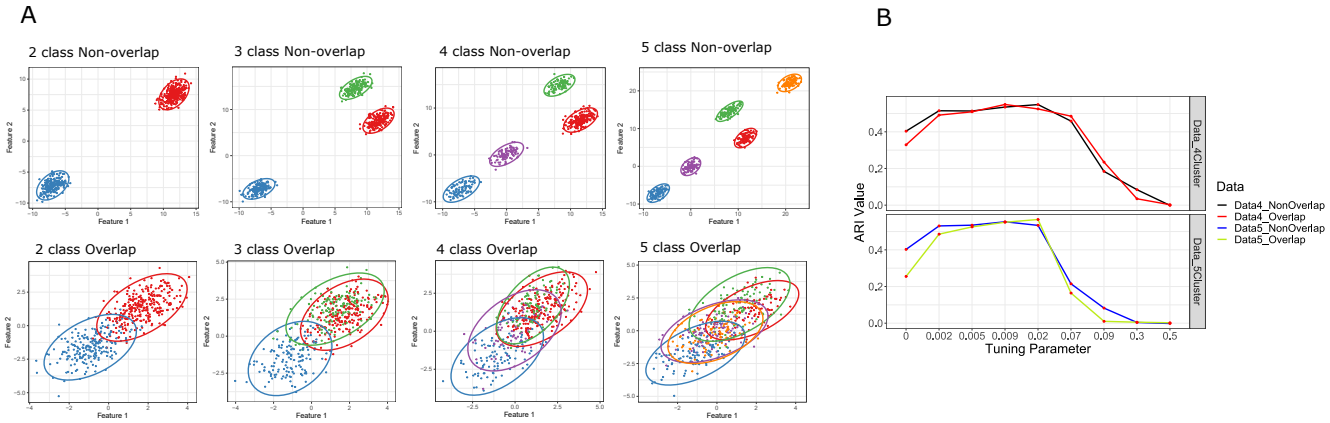


**Figure 2.** Figure shows tSNE visualization of eight synthetic Gaussian Mixture Datasets (Panel-A). The first one represents datasets of non-overlapping classes whereas the latter represents the same for overlapping classes. Panel-B shows the ARI scores of clustering results of synthetic data including four overlapped and non-overlapped clusters. *RgCop* is employed with 9 tuning parameters within the range [0,0.5]

**Table 1.** Description of non-overlapping synthetic Gaussian mixture Data

| # Classes | Mixing Probabilities | # Features | | Range of Means ($\mu$) | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Relevant | Irrelevant | Cluster_1 | Cluster_2 | Cluster_3 | Cluster_4 | Cluster_5 |
| 2 | [0.6,0.4] | 50 | 250 | 50 values from $(7, 17)$ | 50 values from $(-17, -7)$ | - | - | - |
| 3 | [0.4,0.3,0.3] | 50 | 250 | Same as above | Same as above | 25 values from $(7, 17)$ 25 values from $(-17, -7)$ | - | - |
| 4 | [0.4,0.2,0.2,0.2] | 50 | 250 | Same as above | Same as above | Same as above | 25 times 0 25 values from $(7, 17)$ | - |
| 5 | [0.2,0.2,0.2,0.2,0.2] | 50 | 250 | Same as above | Same as above | Same as above | Same as above | 50 values from $(20, 25)$ |

panel-B) It can be noticed from the figure that, *RgCop* produces highest ARI value in the range of $\gamma$ between 0.005 and 0.02. This values are utilized on the real datasets for gene selection.

## Comparison with State-of-the-art

We compared the efficacy of *RgCop* by comparing with four well known techniques for identifying highly dispersed genes in scRNA-seq data[11]. We also compared the performance of *RgCop* with four widely used supervised feature selection techniques.

### *A short description of competing methods and parameter settings*

Four well known gene selection methods in scRNA-seq data are selected for comparisons: *Gini Clust*[20], *PCA Loading*[1], $CV^2$ *Index* and *Fano Factor*[21]. *Gini Clust* uses Gini Index in feature selection which is used in[20] for rare cell detection in scRNA-seq data. *PCA Loading* selects feature with principal component analysis, which is very common and widely used in scRNA-seq data analysis. $CV^2$ *Index* is defined as variance to mean ratio of a variable. Features/genes having higher $CV^2$ Index is selected from scRNA-seq data. *Fano Factor* is a measure of dispersion among the features. It is also defined as a ratio of variance to mean of a variable. In scRNA-seq data the genes having the highest Fano factor is selected.

Four existing mutual information-based supervised feature selection methods are chosen in our work: Conditional Mutual information maximization (CMIM)[22]- It maximizes the mutual information concerning the class while conditioning upon the selected features, Double Input Symmetrical Relevance (DISR)[23]- It selects a set of variables that can return information on the output class higher than the sum of the information's of each variable taken individually, Joint Mutual Information Maximisation (JMIM)[24]- the 'maximum of the minimum' criterion is used, which alleviates the problem of overestimation of the feature significance, and Minimal-Redundancy-Maximal-Relevance criterion (MRMR)[25]- Feature relevance concerning class labels are considered and ensures that redundant features are not present in the final feature subset[26].

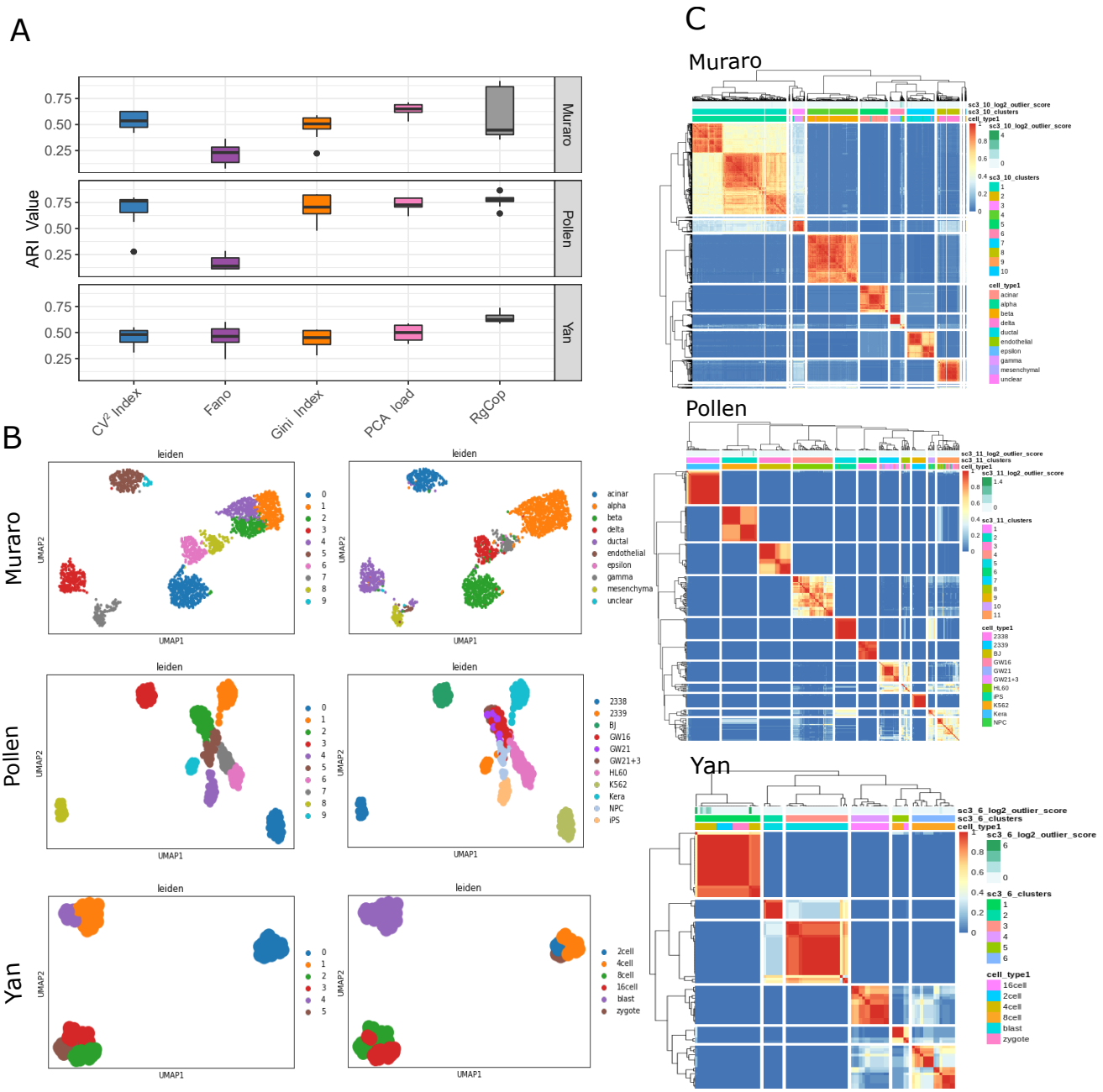


**Figure 3.** Figure shows the comparisons of clustering performance. Panel-A shows the boxplot of ARI values computed from the clustering results of each competing method. Each box represents ten ARI scores of clustering results for selected 10 sets of features ranging from 10 to 100. Panel-B shows the t-sne visualization of clustering results of three datasets for *RgCop*. Panel-C shows the consensus clustering plots of obtained clusters from *RgCop*.

For Gini-Clust, we use the R package with the default parameter as provided in the original paper[20]. For PCA loading, we consider the first three PC components as the default parameter. We use 'praznik' R package with default parameters for supervised methods (MRMR, DISR, JMIM, and CMIM). For *RgCop*, we use regularization coefficient $\gamma$ as 0.009 (see simulation result on synthetic data). Number of selected features is user defined in our method. In this work, all experiments are performed on top 100 selected features using *RgCop*.

### *Clustering performance on real dataset using unsupervised method*

Here single cell Consensus clustering (SC3) method[3] is employed for clustering expression matrix with selected features. Figure 3, panel-A illustrates the boxplots of ARI Values of the clustering results on Yan, Muraro, and Pollen datasets. We vary

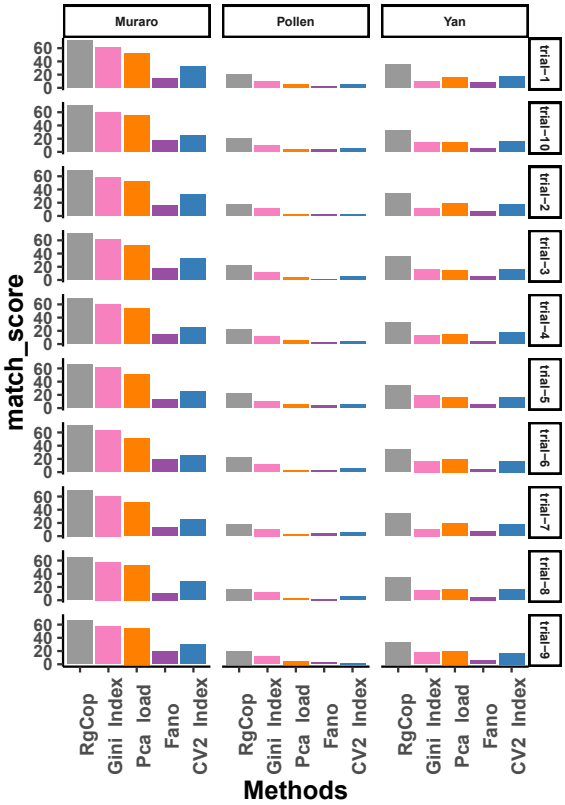


**Figure 4.** Figure shows the median of *match_score* (percentage) of five different competing methods. 10 trials are performed with 100 iteration in each trial to compute the median of *match_score*.

the number of selected features in the range from 20 to 100 and compute the ARI scores for each method. It can be seen from the figure that *RgCop* achieves high ARI values in almost all the three datasets. For the Yan dataset, while the performance of other methods is relatively low, *RgCop* achieves a good ARI value, demonstrating the capability of *RgCop* to perform in small sample data. We also create a visualization of the clustering performance of *RgCop* in Muraro, yan, and Pollen datasets. Figure 3 panel- B, shows two dimensional t-SNE plot of predicted clusters and their original labels. Panel-C of this figure shows heatmaps of cell $\times$ cell consensus matrix representing how often a pair of cells is assigned to the same cluster considering the average of clustering results from all combinations of clustering parameter[3]. Zero score (blue) means two cells are always assigned to different clusters, while score '1' (red) represents two cells are always within the same cluster. The clustering will be perfect if all diagonal blocks are completely red and off-diagonals are blue. A perfect match between the predicted clusters and the original labels can be seen from panel-B and panel-C of figure 3.

### *Classification performance on real dataset using supervised method*

We compare with four well known supervised feature selection methods with *RgCop* and compute the classification accuracy. It is observed that gene selection using the *RgCop* outperforms in any classifier. Four widely used classifiers are considered in our work, Support Vector Machine (*SVM*), Neural Network (*NN*), and Gradient Boosting Machine (*GBM*) for learning the expression matrix with selected features. Table 3 shows the average test accuracy and the corresponding standard errors over 50 runs for each of the competing methods. The caret R package is utilized for classification purpose.

### *Stability performance*

Protocols for preprocessing of scRNA-seq data are complex and often suffer from technical biases that vary across cells. This causes biases in the downstream analysis if the noise is not properly handled. *RgCop* utilizes copula which is well known for its scale invariance property that makes it robust against noise in the data. Moreover, *RgCop* utilizes a regularization technique which also helps to perform well in noisy datasets. To show the performance of *RgCop* in noisy data, white Gaussian noise with a mean ($\mu = 0$) and standard deviation 1 is mixed to each gene/feature of a dataset. The function *Add.Gaussian.noise* of R package RMThreshold is used to generate Gaussian noise. Next, relevant 100 genes/features are chosen from each of the noisy

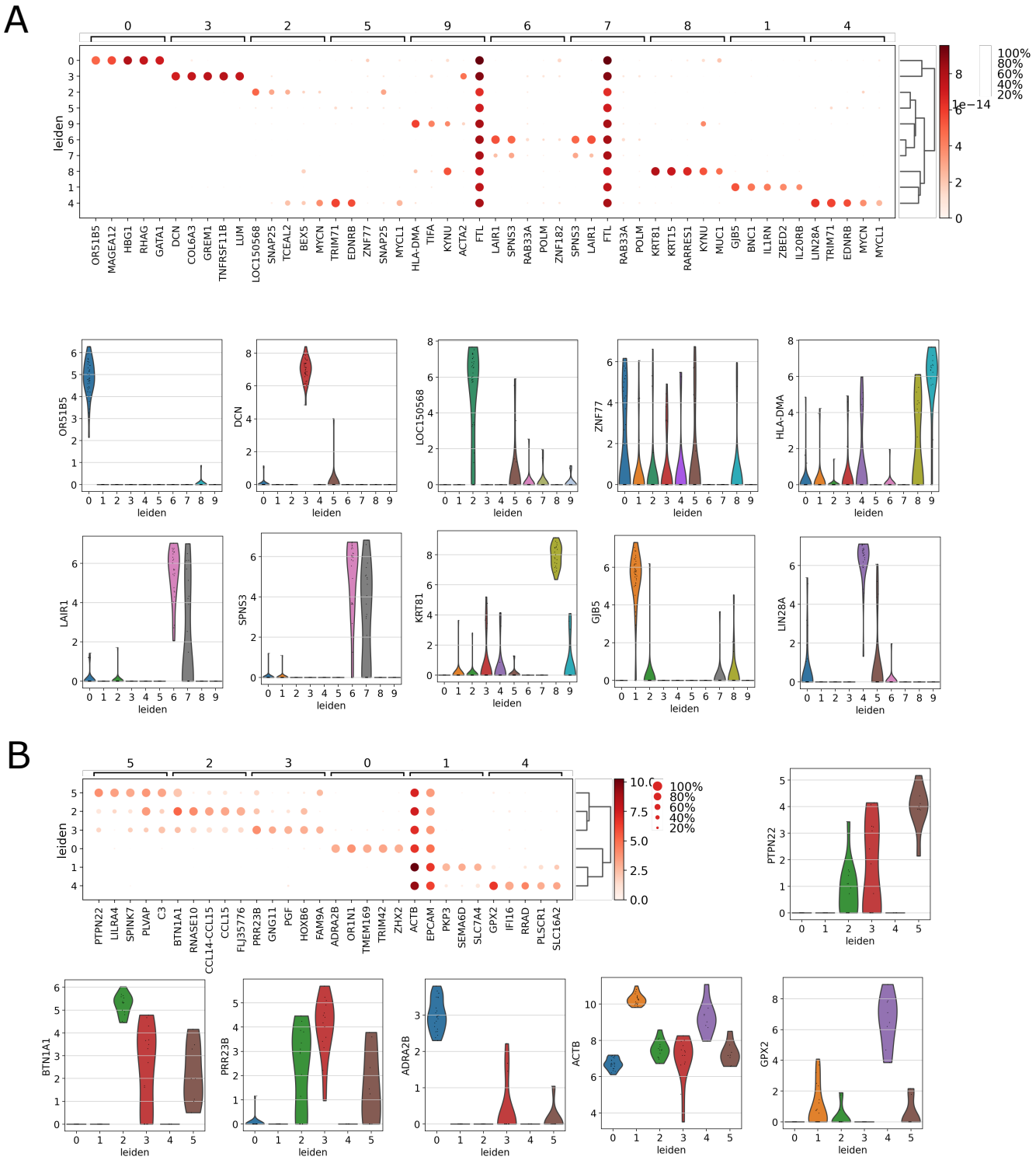


**Figure 5.** Figure shows marker analysis for Pollen dataset (panel-A), and Yan dataset (panel-B). The average expression values of the top five DE genes are shown in heatmap of panel-A, and -B. The violin plots of the expression profiles of those top DE genes within each cluster are shown in panel-A and -B.

**Table 2.** Description of overlapping synthetic Gaussian mixture data

| # Classes | Mixing Probabilities | # Features | | Range of Means ($\mu$) | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Relevant | Irrelevant | Cluster_1 | Cluster_2 | Cluster_3 | Cluster_4 | Cluster_5 |
| 2 | [0.6,0.4] | 50 | 250 | 50 values from $(1,2)$ | 50 values from $(-2,-1)$ | - | - | - |
| 3 | [0.4,0.3,0.3] | 50 | 250 | Same as above | Same as above | 25 values from $(1,2)$ 25 values from $(-2,-1)$ | - | - |
| 4 | [0.4,0.2,0.2,0.2] | 50 | 250 | Same as above | Same as above | Same as above | 25 times 0 25 values from $(1,2)$ | - |
| 5 | [0.2,0.2,0.2,0.2,0.2] | 50 | 250 | Same as above | Same as above | Same as above | Same as above | 25 times 0 25 values from $(-2,-1)$ |

**Table 3.** Classification results on Real Datasets using Supervised Methods

| Classifier | Dataset Name | MRMR | DISR | JMIM | CMIM | *RgCop* ($\gamma = 0.009$) |
|---|---|---|---|---|---|---|
| GBM | Muraro | $0.90 \pm 0.05$ | $0.89 \pm 0.04$ | $0.89 \pm 0.010$ | $0.85 \pm 0.02$ | $0.96 \pm 0.009$ |
| | Pollen | $0.75 \pm 0.03$ | $0.76 \pm 0.01$ | $0.75 \pm 0.05$ | $0.74 \pm 0.02$ | $0.88 \pm 0.002$ |
| | Yan | $1 \pm 0$ | $0.99 \pm 0.02$ | $0.98 \pm 0.02$ | $1 \pm 0$ | $0.97 \pm 0.01$ |
| NNET | Muraro | $0.82 \pm 0.02$ | $0.81 \pm 0.01$ | $0.81 \pm 0.05$ | $0.84 \pm 0.03$ | $0.89 \pm 0.07$ |
| | Pollen | $0.71 \pm 0.01$ | $0.72 \pm 0.003$ | $0.69 \pm 0.02$ | $0.70 \pm 0.01$ | $0.85 \pm 0.02$ |
| | Yan | $0.99 \pm 0.01$ | $0.99 \pm 0.02$ | $0.99 \pm 0.03$ | $0.98 \pm 0.003$ | $0.98 \pm 0.02$ |
| SVM | Muraro | $0.87 \pm 0.04$ | $0.88 \pm 0.01$ | $0.87 \pm 0.011$ | $0.89 \pm 0.01$ | $0.95 \pm 0.01$ |
| | Pollen | $0.74 \pm 0.06$ | $0.75 \pm 0.03$ | $0.70 \pm 0.03$ | $0.71 \pm 0.01$ | $0.87 \pm 0.001$ |
| | Yan | $1 \pm 0$ | $0.99 \pm 0.02$ | $1 \pm 0$ | $0.99 \pm 0.04$ | $0.98 \pm 0.003$ |

datasets, and the percentage of matching is computed with the original genes/feature sets. We define a matching feature score (percentage) as follows

$$match_score = ((N - D)/N) * 100 \tag{1}$$

where $N$ represents the total number of features, and $D$ represents the number of feature discrepancies between the original and noisy dataset.

We perform 10 trials, each contains 100 such experiments. So, for a competing method, each trail contains 100 similarity scores for one dataset. We take the median of these scores and plot this in figure 4. Each row of the figure shows bar plots of the median values for three scRNA-seq datasets over all the competing methods. It can be observed from the figure that the *RgCop* achieves better *match_score* for all the datasets. Particularly for small sample data (Yan dataset) *RgCop* performance is remarkable well than the other methods.

## Validation for annotation of unknown samples

Predicting unknown sample is the final goal of any scRNA-seq analysis pipeline. Here, we addressed this problem by using a cross validation approach. We first split the data in training and test set in the ratio 7:3. Next, top 100 informative genes are chosen using *RgCop* from the training dataset. The clustering performance of *RgCop* is computed on the test data using the selected genes from the trained dataset. We repeat the procedure 50 times with a random split of train-test (7:3 ratio) for each data. The table 4 represents the mean and standard deviation of the ARI scores using *RgCop*. We also train three classifiers SVM, NN, and GBM on the train datasets with the selected feature. The aim is to see whether the trained model can predict the cell type of sample from the test data. Table 4 shows the classification accuracy of three classifiers.

## Marker gene selection

We have chosen marker genes (DE genes) for different cell types from the clustering results. Differentially Expressed (DE) genes are identified from every cluster using Wilcoxon rank-sum test. We use this to directly assess the separation between distributions of expression profiles of genes from different clusters. Figure 5 illustrates the top five DE genes from each cluster of Pollen dataset (panel-A), and Yan dataset (panel-B) (see supplement for the plots of other datasets). The higher expression values of the top five DE genes (shown in the heatmap of panel-A, B) for a particular cluster suggests the presence of marker genes within the selected gene sets. The results are detectable from violin plots of the expression profiles of top DE genes within each cluster (Figure 5 panel-A, and -B). For example, in the Yan dataset, the expression of the gene 'BTN1A1' is very

**Table 4.** Adjusted Rank Index measured on the clustering results on unknown test samples

| Sl No. | Dataset | *RgCop* ($\gamma$=0.009) | | | |
|---|---|---|---|---|---|
| | | ARI | SVM | NN | GBM |
| 1 | PBMC | 0.46 ±0.13 | 0.71 ± 0.08 | 0.69 ± 0.07 | 0.73 ± 0.03 |
| 2 | Muraro | 0.73 ±0.08 | 0.92 ± 0.02 | 0.87 ± 0.01 | 0.93 ± 0.04 |
| 3 | Yan | 0.83 ±0.07 | 0.98 ± 0.04 | 0.97 ± 0.08 | 0.98 ± 0.02 |
| 4 | Pollen | 0.91 ±0.04 | 0.88 ± 0.01 | 0.87 ± 0.04 | 0.88 ± 0.05 |

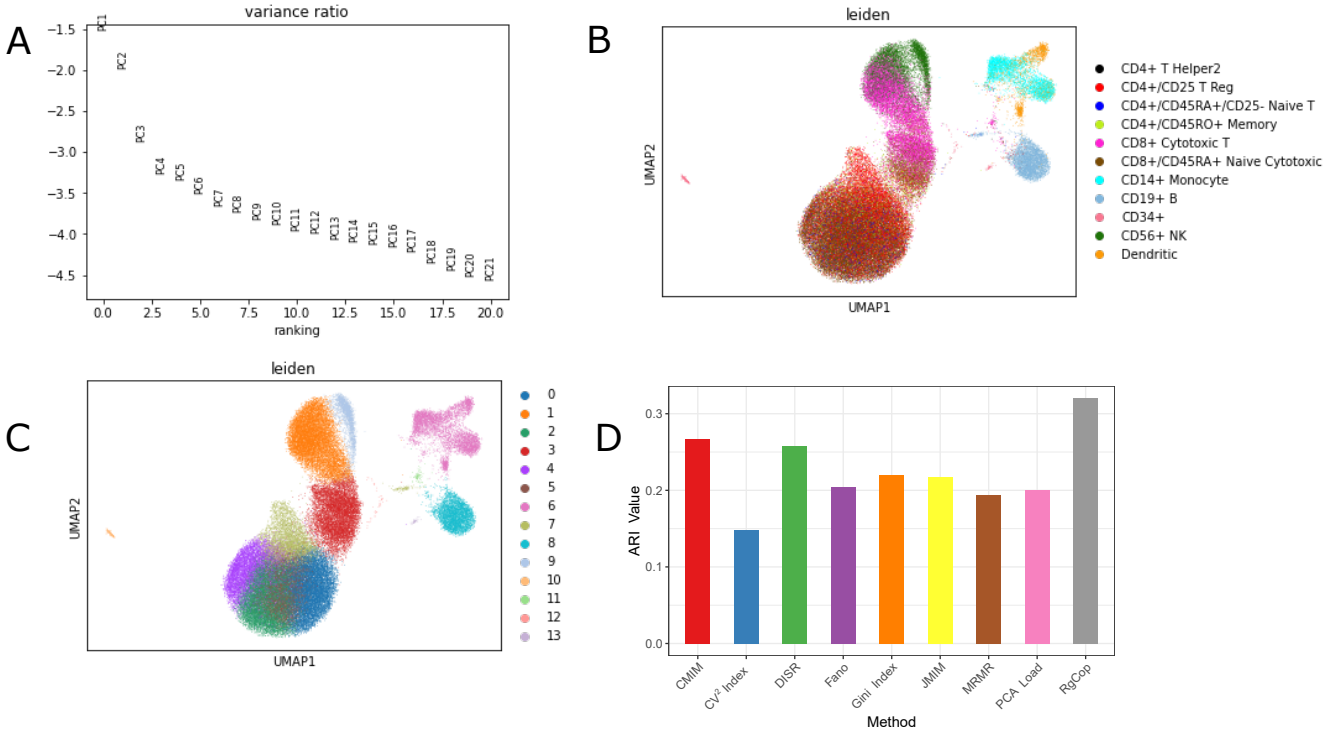


**Figure 6.** Figure shows the clustering results on large PBMC68k datasets. Panel-A, -B and -C represents the t-sne visualization the clusters obtained from the selected feature sets of *RgCop*. Panel-D shows a comparisons of ARI scores among all the competing methods (all supervised and unsupervised methods).

high in cluster-2. Similarly 'ADRA2Ab' and 'GPX2' show high expression in cluster-0 and cluster-4, respectively suggesting a potential marker for these clusters.

### A case study on ultra large scRNA-seq data

*RgCop* is applied on large PBMC68k dataset with sample size 68000. After preprocessing the data, *RgCop* selects 100 genes which are utilized for further processing. We adopt the conventional way for dimensional reduction, neighborhood (k-NN) graph construction, and clustering used in Scanpy[17] python package. The neighborhood graph of PBMC cells is clustered by Leiden graph-clustering method[18] to group the similar cells. Figure 6, panel-A, -B and -C show the t-SNE visualization of the clustering results for *RgCop*. For comparison purposes, we performed the gene selection step of all other competing methods in PBMC68k data and follow similar steps for clustering. Figure 6, panel-D shows the resulting ARI of all the competing methods. It is evident from the figure that *RgCop* outperform the other state-of-the-arts, suggesting its capability of selecting an informative feature from ultra large scRNA-seq dataset.

## Discussions

The selection of informative genes in scRNA-seq data is crucial and an essential step for the downstream analysis. Because of the large feature/gene set of scRNA-seq data, selecting important genes is a challenging task which has an immense effect in

**Table 5.** A brief summary of the real scRNA sequence Dataset

| Dataset Name | Features | Instances | Class |
|---|---|---|---|
| Yan | 20214 | 90 | 7 |
| Muraro | 19127 | 2126 | 10 |
| Pollen | 23794 | 299 | 11 |
| PBMC | 32738 | 68793 | 11 |

clustering and annotation results in the later stage of downstream analysis. The proposed method *RgCop* addressed this task by employing a robust and equitable dependence measure called copula-correlation (*Ccor*). It can accurately measure relevancy and redundancy simultaneously between two sets of gene. *RgCop* also add simple $l_1$ regularization technique with its objective function to control the large coefficients of relevancy terms. Realistic simulations confirm the utmost accuracy of the *RgCop* in overlapped and non-overlapped Gaussian mixture data. We also demonstrated that *RgCop* results high accuracy both in clustering and classification performance with the selected genes from real-life scRNA-seq datasets. The identified marker genes can also be able to dissect cell clusters, suggesting the inclusions of marker genes within the selected sets.

*RgCop* introduces a stable feature/gene selection which is evaluated by applying it in noisy data. By virtue of the important *scale invariant property* of copula, the selected features are invariant under any transformation of data due to the most common technical noise present in the scRNA-seq experiment. The range of tuning parameter (regularization coefficient ($\lambda$)) is determined using *RgCop* on synthetic Gaussian mixture dataset. *RgCop* also produces accurate clustering/classification results on four sc-RNA seq datasets. The results are validated using ARI score/classification accuracy. The stability of *RgCop* is evaluated by applying it in noisy data and matching the resulting feature set with the original one. This was performed multiple times with varying number of selected features. The resulting ARI scores utilize a minimum deviation suggesting a robust and stable approach for feature selection.

It can be noted that although *RgCop* primarily detect variable genes from scRNA-seq data, we extended the process by employing a clustering/classification technique with it to annotate unknown cells. The efficacy of *RgCop* is demonstrated by applying it to cluster/group unknown cells with the selected genes/features. A precise annotation of cell clusters also illustrates the applicability of *RgCop* to select the most variable genes in the early stage. The most general classifiers trained with the selected features can accurately predict the cell types of an unknown sample. Several genes are highlighted having a high expression level within clusters, which are acting as markers.

Taken together, the proposed method *RgCop* not only outperforms in informative gene selection but also can able to annotate unknown cells/cell-clusters in scRNA-seq data. It can be observed from the results that *RgCop* leads both in the domain of robust gene/feature selection and type annotation of unknown cell in large scRNA-seq. The results prove that *RgCop* may be treated as an important tool for computational biologist to investigate the primary steps of downstream analysis of scRNA-seq data.

# Materials and Methods

## Datasets description

### *Single cell RNA sequence Datasets*

The study used The following single-cell RNA sequence datasets are used in our work.

- Yan: The dataset consists of a transcriptome of 124 individual cells from a human preimplantation embryo and embryonic stem cell. The 7 unique cell types accommodates labelled 4-cell, 8-cell, zygote, Late blastocyst, and 16-cell.[GEO under accession no. GSE36552;[27]].

- Pollen: Single cell RNA seq pair-end 100 reads from single cell cDNA libraries were quality trimmed using Trim Galore with the flags. It contains 11 cell types. [GEO accession no GSM1832359;[28]]

- Muraro: Single-cell transcriptomics was carried out on live cells from a mixture using an automated version of CEL-seq2 on live, FACS sorted cells. It contains 2126 number of cells. It is a human pancreas cell tissue with 10 cell types. The dataset was downloaded from GEO under accession no GSE85241[29].

- PBMC68k: The dataset[2], is downloaded from https://support.10xgenomics.com/single-cell-gene expression/datasets. The data is sequenced on Illumina NextSeq 500 high output with 20,000 reads per cell.

A brief description about the dataset is given in Table 5.

## Underline theory that supports *RgCop*

**Short description on Copula** The 'Copula' term[30] is originated from a Latin word *copulare*, which joins multivariate distributions to its one dimensional distribution function. The copula is considerably employed in high dimensional datasets to obtain joint distributions using uniform marginal distributions and vice versa. According to the famous statistician *Sklar's*, copula function can be defined[31] as:

***Copula:*** A Copula $C$ is an $n$ dimensional function, $C : [0,1]^n \to [0,1]$, which satisfies the following properties:

1. $C(u_1, \cdots, u_{i-1}, 0, u_{i+1}, \cdots, u_n) = 0$, i.e., the copula is 0 if one of any variable is 0.
2. $C(1, \cdots, 1, u, 1, \cdots, 1) = u$, i.e., the copula function is just $u$ if one of the variable is u with all others being 1.
3. $C(u_1, \cdots, u_n)$ is a non-decreasing function. This implies that, for any hyper rectangle, $R = \{(x_i, y_i) : i$ is an integer and $1 \leq i \leq n\}$, the $C$ volume of $R$ must be non negative, where $(x_i, y_i) \in [0,1]$.

$$V_R(C) = \sum_{i_1=1}^{2} \cdots \sum_{i_n=1}^{2} (-1)^{i_1+\cdots+i_n} C(u_{1,i_1}, \cdots, u_{n,i_n}) \geq 0 \tag{2}$$

where, $u_{j,1} = x_j$ and $u_{j,2} = y_j$, $j \in (1, \cdots, n)$.

The definition can be properly described by the famous $Sklar's$ theorem.

***Sklar's Theorem:*** Let $X_1, X_2, \cdots, X_n$ be the random vectors whose uniform marginal distributions are $F_1(x_1), F_2(x_2), \cdots, F_n(x_n)$. So, for any joint cumulative distribution function $H$, their exists a copula function $C$ of its univariate marginal distributions such that,

$$H(x_1, x_2, \cdots, x_n) = C(F_1(x_1), F_2(x_2), \cdots, F_n(x_n)). \tag{3}$$

There are many examples of copulas[32].

One of the extensively used, non-parametric copula family is empirical copula. It is employed to find joint distribution where marginal distributions are unknown. It is defined as follows.

***Empirical Copula:*** Let $X_1, X_2, \cdots, X_n$ be the random variables with marginals cumulative distribution function $(F_1(x_1), F_2(x_2), \cdots, F_n(x_n))$ respectively.
The empirical estimate of $(F_i, i = 1, \cdots, n)$, based on a sample, $\{x_{i1}, x_{i2}, \cdots, x_{im}\}$ of size $m$ is given by

$$\widehat{F_i}(x) = \frac{1}{m} \sum_{j=1}^{m} 1_{\{X_{ij} \leq x\}}, [i = 1, \cdots, n] \tag{4}$$

The *Empirical Copula* of $X_1, X_2, \cdots, X_n$ is then defined as

$$\begin{aligned} &\widehat{C}(u_1, u_2, \cdots, u_n) \\ &= \frac{1}{m} \sum_{j=1}^{m} 1\{\widehat{F_1}(x_{1,j}) \leq u_1, \widehat{F_2}(x_{2,j}) \leq u_2, \cdots, \widehat{F_n}(x_{n,j}) \leq u_n\}, \end{aligned} \tag{5}$$

for $u_i \in [0,1], [i = 1, \cdots, n]$.
The *Empirical Copula* is employed to model our feature selection method.

***Copula correlation measure*** Let, $Y = \{y_1, y_2\}$ and $Z = \{z_1, z_2\}$ are two bivariate random variables. Their joint distribution function and marginal distributions are
$H_{YZ}$, $F_Y(y)$ and, $F_Z(z)$ respectively, here, $C$ is their copula distribution function,

$$H_{YZ}(y, z) = C(F_Y(y), F_Z(z)). \tag{6}$$

Kendall tau($\tau$), the measure of association,[33] can be expressed in terms of concordance and discordance between random variables. Kendall tau is the difference between probability of concordance and discordance of $(y_1, y_2)$ and $(z_1, z_2)$. It can be described as

$$\tau_{YZ} = [P(y_1 - y_2)(z_1 - z_2) \geq 0] - [P(y_1 - y_2)(z_1 - z_2) \leq 0] \tag{7}$$

The relation of Kendall tau measure with copula can be mathematically expressed as:

$$\tau(C_{Y,Z}) = \tau_{YZ} = 4 \iint_{0}^{+1} C(u, v)\, dC(u, v) - 1 \tag{8}$$

Where, $u \in F_Y(y)$ and $v \in F_Z(z)$.
Thus Kendall tau correlation can be described by copula function and is termed as copula-correlation ($Ccor$) in our study.

## A note on regularization

Regularization is a type of regression that penalizes the coefficient of redundant feature towards zero[34]. Thus, it removes outliers (here non informative feature) from datasets. The simplest regularization is $l_1$ norm or Lasso Regression, which adds "absolute value of magnitude" of coefficient as penalty term to the loss function. Another widely used regularization is $l_2$ norm or Ridge Regression, which adds "squared magnitude" of coefficient as penalty term to the loss function. The key difference between these two is that Lasso shrinks the less important feature's coefficient to zero and thus, removes some features as well. So, this will be applicable where we would have huge number of features. On the contrary, $l_1$ norm regularization produces sparse solutions by making higher coefficients of the loss function to zero. $l_1$ norm or Lasso Regression is used in our model to handle the scRNA-seq data with the large number of features. For any vector $A \in \mathscr{R}^m$, the $l_1$ norm is

$$||A||_1 = \gamma \sum_{i=1}^{m} |A_i| \tag{9}$$

where $\gamma$ is a tuning parameter, controls penalization. There is no regularization effect when $\gamma = 0$. As $\gamma$ increases, it penalizes the coefficients to zero. But after a certain value of $\gamma$, the model starts losing important properties, increasing bias in the model and thus causes under-fitting. We tuned $\gamma$ using eight synthetic Gaussian mixture dataset in our study.

## Gene selection using *RgCop* algorithm

**Max relevancy:** A gene $g_i$ is more relevant to class labels $C_D$ than another gene $g_j$, if $g_i$ has higher *Ccor* score with $C_D$ than $g_j$. This is called *Relevancy* test for the genes[26] and is used to select most relevant gene from a gene set. Formally it can be described as: $g_i \prec g_j$ if the following is true,

$$\tau(C_{g_i,C_D}) > \tau(C_{g_j,C_D}), \tag{10}$$

Where $\tau[C(x,y)]$ represents copula correlation between two random variable $X$ and $Y$. For estimating the copula density we have used empirical copula. The maximum-relevancy method choose the gene (feature) subset among gene set $G$ as

$$G_{max-relevancy} = \arg\max_{g_i \in G} \frac{1}{|G|} \sum_{g_i \in G} \tau(C_{g_i,C_D}). \tag{11}$$

$G_{max-relevancy}$ may contains genes that are mutually dependent. This is because we only consider the *Ccor* between a gene and class labels, overlooking the mutual dependency among the selected and non-selected genes. This results spurious genes in the selected list.

**Min Redundancy:** Redundancy is a measure that computes the mutual dependence among set random variables. Here we used the same definition of *Ccor* to compute multivariate dependency between selected gene ($g_i$) and non selected gene sets ($g_s$). Formally it can be expressed as

$$G_{min-redundancy} = \arg\min_{s:g_s \in S} \tau(C_{g_i;g_s}) \tag{12}$$

**Objective function:** *RgCop* utilizes a forward selection wrapper approach to select gene iteratively from a gene set. It uses multivariate copula-based dependency instead of the classical information measure. The objective function integrates the relevancy and the redundancy terms defined using the *Ccor*. Mathematically, it can be expressed as follows.

Let us assume genes $(g_1, \ldots, g_i)$ are in the selected list $G_s$. The next gene $g_{i+1} \in (G - G_s)$ in at $(i+1)$ iteration is using the objective function

$$f = \arg\max_{g_i \in (G-G_s)} [(\tau(C_{g_i;C_D}) - \tau(C_{g_i;g_1;g_2;\cdots;g_s})) + \gamma||\tau(C_{g_i;C_D}) * Var(g_i)||_1] \tag{13}$$

Where, $\tau(C_{g_i;g_j}) = [4 \iint_0^{+1} C(g_i;g_j)\, dC(g_i;g_j) - 1]$ is Kendall tau dependency score of Empirical copula between two genes $g_i$ and $g_j$. Here, $\gamma$ represents the regularization coefficient. An overview of the *RgCop* algorithm is given in algorithm **1**.

**Proof of correctness:** Let, a subset of genes $G_s = \{g_1, g_2, \ldots, g_i, g_{i+1}, \ldots g_d\}$ are obtained from a gene set $G$ using *RgCop*. Here $g_i$ represents selected gene at iteration $i$. We claim that the set $G_s$ is optimal.

**Proof** Let us prove this by the method of contradiction. If we assume the claim is not true, then there should exist some another optimal gene set $G'_s$. Without loss of generality, let us assume $G'_s$ has a maximum number of initial genes ($i$ number genes) common with $G_s$.
Now $G'_s$ can be written as $G'_s = \{g_1, g_2, \ldots, g_i, g_k, \ldots, g_d\}$. So, $G'_s$ contains $\{g_1, g_2, \ldots, g_i\}$ from $G_s$, but not $g_{i+1}$. Then by our assumption $g_{i+1}$ cannot be included in any of the optimal gene list, because of our assumption ($G'_s$ has maximum $i$ number of initial genes overlapped with $G_s$).

Now we claim that $k > (i+1)$. This is because $k$ cannot be equal to $i+1$, otherwise $G'_s$ would have $(i+1)$ genes overlapped with $G_s$. Similarly $k \nleq i$, because otherwise $G'_s$ will contains redundant genes (same gene multiple times).

Now by the definition of our objective function ($f$) we can write: $f(g_k) < f(g_{i+1})$. So we can substitute $g_k$ with $g_{i+1}$ in the $G'_s$ list, and the list will be still optimal. This contradicts our assumption that $g_{i+1}$ cannot be included in any optimal list. This proves our claim.

**Algorithm 1** $l_1$ Regularized Copula Based Feature Selection (*RgCop*)

**Input:** Preprocessed Data Matrix $\mathbf{D}$, Cell Type $\mathbf{C_D}$, Number of Selected Features, $\mathbf{d}$.
**Output:** Optimal Feature subset,($\mathbf{G_s}$).
  **Initialisation**:
  $\mathbf{G_s} = \emptyset$, {$\mathbf{T}$ will hold sub-set feature indices.}
  $\mathbf{g_{1s}} \leftarrow \arg\max_{g_i} \tau(C_{g_i,C_D})$, {Maximum Relevancy}
  **for all** $i = 0$ to $(\mathbf{d}-\mathbf{1})$ **do**
    $\mathbf{E} = \emptyset$
    $\mathbf{R} \leftarrow \tau(C_{g_i;C_D})$, {Relevancy Criterion}
    $\mathbf{S} \leftarrow \tau(C_{g_i;g_{1s};\cdots;g_{ds}})$, {Redundancy Criterion}
    $\mathbf{E} \leftarrow (R-S)$
    $\mathbf{G_s} \leftarrow \{\mathbf{G_s} \bigcup \arg\max(\lim_{g_i}\{E\})$
    $\mathbf{G} \leftarrow G - \{g_i\}$
  **end for**
  **return** $\mathbf{G_s}$

## 1 Acknowledgment

We would like to acknowledge support from JC Bose Fellowship Grant No. SB/SJ/JCB-033/201.6 dated 01/02 12017 of DST, Govt. of India.; SyMeC Project grant [BT/Med-II/NIBMG/SyMeC/2014/Vol. II] given to the Indian Statistical Institute by the Department of Biotechnology (DBT), Govt. of India.

## Competing interests

The authors declare that they have no competing interests.

## Author's contributions

SL and SR equally contributed to this work

## References

1. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
2. Zheng, G. X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8**, 14049 (2017).
3. Kiselev, V. Y. *et al.* Sc3: consensus clustering of single-cell rna-seq data. *Nature methods* **14**, 483 (2017).
4. Ji, Z. & Ji, H. Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. *Nucleic acids research* **44**, e117–e117 (2016).
5. Plass, M. *et al.* Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* **360** (2018).
6. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).

7. Fincher, C. T., Wurtzel, O., de Hoog, T., Kravarik, K. M. & Reddien, P. W. Cell type transcriptome atlas for the planarian schmidtea mediterranea. *Science* **360** (2018).

8. Han, X. *et al.* Mapping human pluripotent stem cell differentiation pathways using high throughput single-cell rna-sequencing. *Genome biology* **19**, 1–19 (2018).

9. Duò, A., Robinson, M. D. & Soneson, C. A systematic performance evaluation of clustering methods for single-cell rna-seq data. *F1000Research* **7** (2018).

10. Luecken, M. D. & Theis, F. J. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology* **15**, e8746 (2019).

11. Brennecke, P. *et al.* Accounting for technical noise in single-cell rna-seq experiments. *Nature methods* **10**, 1093 (2013).

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

13. Kim, J.-M. *et al.* A copula method for modeling directional dependence of genes. *BMC bioinformatics* **9**, 225 (2008).

14. Ray, S., Lall, S. & Bandyopadhyay, S. Codc: a copula-based model to identify differential coexpression. *NPJ systems biology applications* **6**, 1–13 (2020).

15. Kasa, S. R., Bhattacharya, S. & Rajan, V. Gaussian mixture copulas for high-dimensional clustering and dependency-based subtyping. *Bioinformatics* **36**, 621–628 (2020).

16. Yip, S. H., Wang, P., Kocher, J.-P. A., Sham, P. C. & Wang, J. Linnorm: improved statistical analysis for single cell rna-seq expression data. *Nucleic acids research* **45**, e179–e179 (2017).

17. Wolf, F. A., Angerer, P. & Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology* **19**, 15 (2018).

18. Traag, V. A., Waltman, L. & van Eck, N. J. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports* **9**, 1–12 (2019).

19. Liu, W. & Lin, W. Additive white gaussian noise level estimation in svd domain for images. *IEEE Transactions on Image processing* **22**, 872–883 (2012).

20. Jiang, L., Chen, H., Pinello, L. & Yuan, G.-C. Giniclust: detecting rare cell types from single-cell gene expression data with gini index. *Genome biology* **17**, 144 (2016).

21. Grün, D., Kester, L. & Van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nature methods* **11**, 637 (2014).

22. Fleuret, F. Fast binary feature selection with conditional mutual information. *Journal Machine Learning Research* **5**, 1531–1555 (2004).

23. Meyer, P. E. & Bontempi, G. On the use of variable complementarity for feature selection in cancer classification. In *Workshops on Applications of Evolutionary Computation*, 91–102 (Springer, 2006).

24. Bennasar, M., Hicks, Y. & Setchi, R. Feature selection using joint mutual information maximisation. *Expert Systems with Applications* **42**, 8520–8532 (2015).

25. Peng, H., Long, F. & Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis machine intelligence* **27**, 1226–1238 (2005).

26. Brown, G., Pocock, A., Zhao, M.-J. & Luján, M. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *Journal machine learning research* **13**, 27–66 (2012).

27. Yan, L. *et al.* Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells. *Nature structural & molecular biology* **20**, 1131 (2013).

28. Pollen, A. A. *et al.* Low-coverage single-cell mrna sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature biotechnology* **32**, 1053 (2014).

29. Muraro, M. J. *et al.* A single-cell transcriptome atlas of the human pancreas. *Cell systems* **3**, 385–394 (2016).

30. Nelsen, R. B. *An introduction to copulas* (Springer Science & Business Media, 2007).

31. Jaworski, P., Durante, F., Hardle, W. K. & Rychlik, T. *Copula theory and its applications*, vol. 198 (Springer, 2010).

32. Durante, F. & Sempi, C. Copula theory: an introduction. In *Copula theory and its applications*, 3–31 (Springer, 2010).

33. Kruskal, W. H. Ordinal measures of association. *Journal American Statistical Association* **53**, 814–861 (1958).

**34.** Xing, E., Jordan, M., Karp, I. & Richard, M. Feature selection for high-dimensional genomic microarray data. In *ICML*, vol. 1, 601–608 (2001).