sc-REnF: An Entropy Guided Robust Feature Selection for Single-Cell RNA-seq Data

Snehalika Lall

Indian Statistical Institute

Abhik Ghosh

Indian Statistical Institute

Sumanta Ray (✉ Sumanta.Ray@cwii.nl)

Centrum Wiskunde and Informatica, Amsterdam

Sanghamitra Bandyopadhyay

Indian Statistical Institute

Article

Keywords: sc-REnF, entropy, single-cell, cells, PCA

Posted Date: April 12th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-355014/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Briefings in Bioinformatics on January 1st, 2021. See the published version at <https://doi.org/10.1093/bib/bbab517>.

sc-REnF: An entropy guided robust feature selection for single-cell rna-seq data

Snehalika Lall¹, Abhik Ghosh², Sumanta Ray^{3,*}, and Sanghamitra Bandyopadhyay^{1,*}

¹Machine Intelligence Unit, Indian Statistical Institute, Kolkata, West Bengal 700108, India.

²Interdisciplinary Statistical Research Unit, Indian Statistical Institute, Kolkata, West Bengal 700108, India.

³Department of Computer Science and Engineering, Aliah University, Kolkata, India.

*Corresponding Author: sumanta.ray@cwil.in, sanghami@isical.ac.in

ABSTRACT

Annotation of cells in single-cell clustering requires a homogeneous grouping of cell populations. Since single cell data is susceptible to technical noise, the quality of genes selected prior to clustering is of crucial importance in the preliminary steps of downstream analysis. Therefore, interest in robust gene selection has gained considerable attention in recent years. We introduce sc-REnF, (robust entropy based feature (gene) selection method), aiming to leverage the advantages of Rényi and Tsallis entropies in gene selection for single cell clustering. Experiments demonstrate that with tuned parameter (q), Rényi and Tsallis entropies select genes that improved the clustering results significantly, over the other competing methods. sc-REnF can capture relevancy and redundancy among the features of noisy data extremely well due to its robust objective function. Moreover, the selected features/genes can able to clusters the unknown cells with a high accuracy. Finally, sc-REnF yields good clustering performance in small sample, large feature scRNA-seq data.

Availability: The sc-REnF is available at <https://github.com/Snehalikalall/sc-REnF>

Introduction

In recent times technological advances have made it possible to study RNA-seq data at single cell resolution¹. Single cell RNA sequencing (scRNA-seq) is a powerful tool to capture gene expression snapshots in individual cells. Cell type detection is one of the fundamental steps in downstream analysis of scRNA-seq data². A widely used approach for this is to cluster the cells into different groups, and determine the identity of cells within the individual groups/clusters^{3,4}. This provides an unsupervised method of annotating the different cell types present in the large population of scRNA-seq data⁵⁻⁷. Starting from raw counts, scRNA-seq data analysis typically goes through the following steps before clustering: i) normalization (total-count and log-normalization), ii) feature selection, and iii) dimensionality reduction. While normalization/log-normalization adjusts the differences between the samples of individual cells and reduces the skewness of the data, feature selection seeks to identify the most relevant features (genes) from the large feature space. The top genes are either i) highly variable genes, selected by measuring the coefficient of variation^{8,9} or ii) highly expressed genes having expression levels higher than the average across all cells¹⁰.

The performance of downstream analysis, mainly the clustering process, is heavily dependent on the quality of selected top features/genes. The typical characteristics of good features/genes are: i) it should encode useful information about the biology of the system, ii) should not include features that contain random noise, and iii) preserve the useful biological structure while reducing the size of the data so as to reduce the computational cost of later steps.

The conventional approach of gene selection based on high variability in the very first stage of the downstream analysis is seemingly simple. However there are two major caveats: i) The observed variability of genes depends on the pseudo-count, which is arbitrary and can introduce biases in the data ii) PCA dimensionality reduction implicitly depends on Euclidean geometry which may not be appropriate for highly sparse and skewed scRNA-seq data. To take care of (i) adding a small pseudocount to all normalized counts prior to log normalization is a common practice in this pipeline. It is required because CPM (counts per million) cannot change the dominating zeros in the scRNA-Seq data, and without that log transformation is not possible. For (ii), application of PCA, despite its fast and memory-efficient behavior, is questionable because of the high sparsity, and discrete and skewed nature of the datasets.

In this paper, we address these two challenges, independently and in combination. We first present a method that finds the most informative features/genes from the scRNA-seq datasets based on the widely used Rényi and Tsallis entropy measures. Although entropy-based feature selection is widely-researched, its applications in the single-cell domain is less explored. The handful of works that exist in this domain^{11,12}, are not focused on the informative gene selection from single cell data. While¹¹

aim to use entropy to improve the cluster annotation step,¹² focus on quantifying the differentiation potential in the cells. In this paper, we show that employing *Renyi* and *Tsallis* entropy in the gene filtering process introduces major advantages both in terms of clustering accuracy, and in terms of a biologically meaningful interpretation (a.k.a. marker selection) of the results. The latter point is established because we are able to reveal novel marker genes for different cell types. Note that biological marker selection is usually a crucial step in the downstream analysis and depends on the purity of the cell clusters annotated in the previous stage. Here, we present a method for selecting the most informative genes that ultimately leads to a good clustering and annotation of cells in scRNA-seq data. We have compared the proposed method with six state-of-the-art gene selection selection techniques: two methods are based on Min *Renyi*¹³, and Shannon¹⁴) entropy; and four methods are well known for gene selection in scRNA-seq data (*Gini Clust*¹⁵, *PCA Loading*¹⁶, *CV² Index*¹⁷ and *Fano Factor*¹⁸). We have also utilized three well known scRNA-seq clustering method (seurat⁴, SC3³ and CIDR¹⁹ to validate the features (genes) selected by different competing methods.

Beyond predicting the cluster annotation of single cells, we also demonstrate that our method provides better clustering and annotation on a completely independent test data. For this, we split the data into a training and test sets and demonstrate that the selected features in the training set are equally effective in the clustering of the test samples. We also carry out a comprehensive simulation study to establish the effectiveness of the proposed method on eight contaminated Gaussian Mixture datasets. The results show that proposed method not only select genes with high accuracy, but is also robust when the parameters are appropriately tuned.

Summary of contributions: Here, we provide the following novelties:

- We explore the first entropy-based gene selection approach for clustering single-cell data. We utilized *Renyi* and *Tsallis* entropy that has major advantages over the Shannon method for their controlling parameter (q), which makes them less sensitive (robust) against different noises present in the data.
- Our approach is the first one to explicitly address how to learn the feature relevancy and redundancy using *Renyi* and *Tsallis* entropies in the scRNA-seq expression data. We raised an objective function that will minimize conditional entropy between the selected features and maximize the conditional entropy between the class label and feature.
- Our framework (with tuned parameter) can able to cluster unseen scRNA-seq expression data with utmost accuracy. Clustering hitherto unclassified data is crucial for the annotation of cell types. We present our framework to be effective in this case. We demonstrate that the selected features from the training set are equally valid for the completely unseen test data of the same tissue.
- Here we derived a new risk factor for *Renyi* (R_q) and *Tsallis* (R_{T_q}) entropies (see Equation 4 and Equation 7 in Method). We theoretically proved that the iterative selection of features eventually minimizes the *Renyi* (R_q) and *Tsallis* (R_{T_q}) risk, which strengthens the robustness of our objective function.
- Our method is less sensitive (robust) against different noises present in the data.
- Our method provides good results for small-sample and large-feature sized single cell data.

Short description of feature selection and related works Selection of relevant features remains a challenging problem in machine learning because of the ever-increasing size of the data. The principle motivation of feature selection is to reduce the dimension of datasets to decrease computational cost and enhance the accuracy of algorithms. Some existing approaches for feature selection in different domains like text categorization, image retrieval, intrusion detection, genomic analysis are described in^{20,21}, respectively. Feature selection algorithms are broadly classified into three categories: *filter*, *wrapper* and *embedded* model (see sec-1 of the supplement).

Several entropy based filter methods was proposed in the literature, among them Shannon's entropy is utilized in most of the cases (see sec-1 of supplement). *Renyi* and *Tsallis* entropies²² are measures which have some interesting characteristics. Recently, their application in the fields of security and privacy have reinvigorated interest in them²³. Their utility in the biological domain is yet to be established. The present article is an attempt in this direction.

Results

In the following, we will describe the workflow of our analysis pipeline.

A

scREnF framework for gene selection

B

Validation of selected genes

cells with selected genes

clustering

computing ARI

Marker selection

C

Validation for predicting unknown cells

Figure 1. A brief framework of our study: Panel-A: scRNA-seq count matrix are downloaded and preprocessed. sc-REnF is applied for gene/feature selection using *Renyi*, *Tsallis* entropy measure. Panel-B: Selected genes are validated by adopting scRNA-seq clustering techniques. Panel-C: validating the selected genes (from a training set) in an unknown test samples using clustering and ARI method.

Workflow

The figure 1 describes the workflow of our analysis pipeline. All important steps are discussed in this subsection.

Preprocessing of raw datasets: Single-cell RNA sequence raw datasets are downloaded from publicly available sources. The RNA counts are organised as a matrix $M_{cl \times ge}$, where cl is the number of cells and ge is the number of genes. Each element $[M]_{ij}$ represents count of the i^{th} cell and the j^{th} gene. If more than a thousand of genes are expressed (non zero values) in one cell, then the cell is termed as good. We assume one gene is expressed if the minimum read count of it exceeds 5 in at least 10% of the good cells. The data matrix M with expressed genes and good cells is normalized using a linear model and normality based normalizing transformation method (Linnorm)²⁴. The resulting matrix ($M_{cl' \times ge'}$) is then $\log_2$ transformed by adding one as a pseudo count.

Feature/gene selection using Renyi and Tsallis entropy See panel-A of the figure 1. The feature/gene selection is driven by an iterative algorithm which select most relevance and non-redundant features using *Renyi* and *Tsallis* entropy in each step. First, relevancy between all features and class labels is computed to select most relevant feature (see Equation 8). Next,

Figure 2. Figure shows the results of applying sc-REnF in synthetic Gaussian mixture data. Panel-A shows a two dimensional representation of the generated data with 4 and 5 overlap and non-overlap classes. Panel-B shows the clustering results of the corresponding data shown in the left side panel. Adjusted Rand Index (ARI) of the clustering results are reported for different q – parameter of Renyi and Tsallis entropy.

non-redundant features are selected by calculating redundancy between remaining features and the relevant one (see Equation 9).

Validation of selected features/genes See figure 1, panel-B. The validation is performed by employing benchmark single cell clustering techniques to group the cells with selected features/genes. The clustering results are evaluated using Adjusted Rand Index (ARI) score which ensures proper and accurate partitioning of cells. We compute differentially expressed (DE) genes from each cluster which can be treated as marker genes of specific groups of cells.

Clustering of unannotated cell samples See figure 1, panel-C. Clustering unknown samples are crucial in scRNA-seq data analysis and can be addressed by a supervised or unsupervised way. For unknown cell types, our method can able to cluster the data with selected genes, yielding a good clustering that is as fine-grained as is justified by the available true labeled data. We split the whole expression data in train:test ratio 70:30, and observed that the selected genes in training data can accurately cluster the cell samples of test data. This provides the key element of our approach to work in practice.

Feature selection on synthetic contaminated data

Data generation To validate sc-REnF on synthetic contaminated data, we first generate four Gaussian mixture datasets with overlapping and non-overlapping classes (k) and then contaminated it by reshuffling the samples among the classes. Figure 2, panel-A shows the pictorial 2-dimensional t-SNE representation of the generated data with $k = 4$ and 5. In each case, we

Table 1. Comparisons with five state-of-the-arts feature selection methods on four scRNA-seq datasets: ARI scores are reported for three clustering methods (SC3, Seurat and CIDR)

Sl No.	Dataset	sc-REnF.Renyi (q=0.7)			sc-REnF.Tsallis (q=0.3)			Shannon			Gini Index			PCA Loading			Fano			CV ² Index		
		SC3	Seurat	CIDR	SC3	Seurat	CIDR	SC3	Seurat	CIDR	SC3	Seurat	CIDR	SC3	Seurat	CIDR	SC3	Seurat	CIDR	SC3	Seurat	CIDR
1	CBMC	0.24	0.21	0.26	0.39	0.36	0.8	0.35	0.32	0.33	0.14	0.22	0.16	0.20	0.19	0.17	0.2	0.18	0.21	0.18	0.2	0.19
2	Darmanis	0.41	0.49	0.15	0.45	0.51	0.1	0.362	0.23	0.04	0.35	0.29	0.07	0.4	0.40	0.18	0.28	0.28	0.06	0.4	0.45	0.11
3	Yan	0.87	0.72	0.825	0.827	0.69	0.63	0.8	0.64	0.60	0.70	0.42	0.48	0.62	0.50	0.41	0.57	0.32	0.49	0.706	0.61	0.39
4	Pollen	0.95	0.79	0.33	0.82	0.75	0.31	0.8	0.56	0.21	0.89	0.55	0.18	0.91	0.71	0.3	0.67	0.30	0.28	0.91	0.73	0.35

generated 500 samples containing 50 relevant and 250 irrelevant features by varying the mean (μ) in the range between -5 to 20 for non-overlapping and in the range between -2 to 3 for overlapping classes. The covariance matrix (Σ) is fixed and is computed using the formula: $\Sigma(i, j) = \rho^{|i-j|}$ with $\rho = 0.5$. For generating the irrelevant feature white Gaussian noise¹¹ is generated ($\mu = 0$, and s.d = 1) and added to the relevant features. Further, to make the constructed data noisy, we reshuffled/mixed the samples by using 8 : 2 mixing ratio among the classes.

Feature selection using different range of q parameter We select features from the four synthetic datasets (see figure 2, panel-A) using sc-REnF for different q -parameter values of *Renyi* and *Tsallis* entropies in the range of [0.1 – 7] (see Method). For each dataset, feature is selected using sc-REnF with *Renyi* and *Tsallis* entropy with a varying q – parameter values. K-means clustering is performed on the datasets to validate the selected features. Figure 2, panel-B shows the Adjusted Rand Index (ARI) values of the k – means clustering results. It can be observed from the figure 2, panel-B that for the *Renyi* entropy ARI achieves a high score in the q – parameter range between 0.3 and 1.5. Similarly, for the *Tsallis* entropy the range is between 0.3 and 1.3 in which ARI achieves high value. The selected range of q -values are utilized in scRNA-seq data for feature selection.

Comparisons with state-of-the-art

Here we compared sc-REnF with five gene selection algorithms (Shannon entropy, *Gini Clust*, *PCA Loading*, *CV² Index*, and *Fano Factor*) widely used in scRNA-seq clustering. For Gini-Clust, we use the R package with the default parameter as provided in the original paper. For PCA loading, we consider the first three PC components as the default parameter. For the other methods, Fano Factor and *CV² index* which are commonly used to define mean normalized variance, are implemented from their original definition.

Three benchmark scRNA-seq clustering algorithm (Seurat⁴, SC3³ and CIDR¹⁹) are employed to validate the selected features obtained from different state-of-the-arts gene selection techniques. We have retained the default parameters as specified in each of the scRNA-seq clustering package.

Top 100 genes are selected for each competing method. The ARI score is computed for each clustering and is reported in table 1. It is observed that sc-REnF (using *Renyi* and *Tsallis* entropy) responds well compare to the other six competing methods. sc-REnF achieves the highest ARI score (for *Renyi* entropy with q – parameter=0.7) for Pollen dataset. Similarly, sc-REnF using *Tsallis* entropy yields the highest ARI score for the darmanis dataset. Considering all the three clustering techniques (Seurat, SC3 and CIDR) and all the datasets, it can be seen that sc-REnF (with *Renyi* and *Tsallis* entropy) provides better results than the other competing methods. The detailed results of clustering analysis and marker analysis on Darmanis and CBMC dataset are provided in supplement (sections 5, 6 and 7).

Clustering unannotated cells with selected features

Clustering unannotated cell is crucial for scRNA-seq analysis pipeline. Here we addressed this by performing a cross validation with a train test split of data in the ratio 7:3. First, the training dataset is used in sc-REnF to select informative genes. Top 100 genes are selected for clustering and validation. The performance of sc-REnF is computed on the test data using the top selected genes in the earlier step. We performed clustering on test data with the selected genes and ARI value is reported. The experiment is repeated 20 times with a random split of train-test data (7:3 ratio) in each case. Table 2 shows the median and standard deviation of the ARI score.

Selected genes have reliable overlap with marker genes

Here the aim is to investigate the overlap of the selected selected gene set with the DE genes of specific cell types. We compare the selected gene sets of five competing methods (sc-REnF, Gini Index, PCA Loading, Fano, and *CV² Index*) to know the number of overlap with the marker genes. Figure 3 shows six radarcharts corresponding to six methods, each showing the number of intersection of genes in four datasets. The area of each chart represents the total number of common DE genes among selected top 100 genes. The method which covered larger area have the more number of overlapped DE genes. The figure depicts that sc-REnF (with *Tsallis* entropy) outperforms all the methods. sc-REnF using *Renyi* entropy covered almost

Table 2. Clustering results on unannotated cells. ARI scores are reported for three scRNA-seq clustering methods using the selected features with sc-REnF

SI No.	Dataset	sc-REnF_Renyi (q=0.7)			sc-REnF_Tsallis (q=0.3)		
		SC3	Seurat	CIDR	SC3	Seurat	CIDR
1	CBMC	0.14 ± 0.01	0.12 ± 0.03	0.15 ± 0.13	0.46 ± 0.02	0.39 ± 0.11	0.37 ± 0.04
2	Darmanis	0.41 ± 0.01	0.42 ± 0.07	0.12 ± 0.03	0.35 ± 0.13	0.33 ± 0.02	0.09 ± 0.11
3	Yan	0.71 ± 0.07	0.67 ± 0.04	0.70 ± 0.12	0.86 ± 0.03	0.72 ± 0.05	0.68 ± 0.01
4	Pollen	0.95 ± 0.14	0.78 ± 0.04	0.12 ± 0.03	0.79 ± 0.02	0.65 ± 0.11	0.54 ± 0.02

Figure 3. Figure shows radarplots of six competing methods, each showing the intersection of selected genes with marker genes identified from the whole datasets.

the same area as CV^2 Index. For computing differentially expressed genes from the dataset, we choose Wincoxon Rank Sum test.

Stability of sc-REnF

Here we explore an additional advantage of sc-REnF over the other measures by validating its stability of performance. A non-parametric statistical test Kruskal-Wallis Test²⁵ is utilized to examine the stability of ARI scores resulted from the clustering. We vary the number of features from range 20 to 100 and for each case, we compute the ARI score after clustering. Thus for one method (e.g. *Renyi*) and one dataset, we get nine ARI scores (for #feature=20, 30, 40, 50,60,70,80,90,100) representing the clustering performance with different selected features. To know the variation of ARI scores across all the datasets for a particular method (e.g. *Renyi*), we performed Kruskal-Wallis Test. ARI scores are computed 50 times for each method, and the median of the scores is given to the Kruskal-Wallis test. Although all methods produce a stable results with low p-values, nevertheless the sc-REnF (with *Tsallis* and *Renyi* entropy) show more stable ($p - value=8.01E-05$ and $1.59E-04$) performance among the other methods (see supplementary table-1). These results may be treated as a straightforward implication of the theoretical proof presented in the method section.

Clustering results on CBMC data

Here we provide clustering results on 8000 cord blood mononuclear cells (CBMCs) using the selected features with sc-REnF (with *Tsallis* entropy). Figure 4 panel-A and B shows the t-SNE visualization of cells with original labels and predicted cluster labels, respectively. Most of the clusters such as cluster-2, cluster-4, cluster-14 determine unique cells in the data. For example, cluster-2 captures most of the samples of CD14+ Mono cells, while cluster-4 and cluster-14 represent samples of ‘Nk’ (Natural killer) and erythrocyte cells. Some clusters represent more than one cell, such as cluster-13 includes DCs and pDCs, cluster-11

Figure 4. Clustering results of CBMC data after gene selection. Panel-A and -B represents tSNE-2 visualization of data with original and predicted cluster labels respectively. Panel-C shows a heatmap that represents the percentage of matching samples between 14 identified clusters and 13 different cell types. Panel-D depicts a visualization of samples coming from different immune cells and their corresponding predicted clusters (color-coded)

includes erythrocyte, Mks, and CD34+ cells. Individual mapping of cells to different clusters is depicted in figure 4, panel-D. For example 'NK' cells are captured by two clusters, cluster-4 and cluster-12. Despite being multiple associations of clusters into one particular cell type, in most of the cases one cluster captures major samples of a particular cell type.

Table 3. A brief summary of the scRNA-seq datasets used in this study

Dataset Name	#Features (Genes)	#Instances (Cells)	#Class
Yan ²⁶	20514	90	7
Darmanis ²⁷	22088	466	9
Pollen ²⁸	23794	299	11
CBMC ²⁹	2000	7895	13

Discussions and conclusions

Clustering of cells in scRNA-seq data is an essential step for cell type discovery from a large population of cells. Owing to the large feature/gene set of scRNA-seq data, the selection of most variable genes are crucial in the preprocessing step, which has immense effect in the later stage of downstream analysis. The proposed method sc-REnF addressed this issue by using an entropy (*Renyi*, *Tsallis*) based feature selection method for identifying possible informative genes in the preprocessing steps. sc-REnF has the advantage over the conventional statistical approach that it can consider the cell-to-cell dependency based on generalized and wide spectrum entropy measures *Renyi* and *Tsallis*. We demonstrated that sc-REnF using *Renyi* and *Tsallis* method introduces major advantages both in terms of clustering accuracy and in terms of marker gene detection in the downstream analysis of scRNA-seq data.

sc-REnF yields a stable feature/gene selection with a controlling parameter (q) for *Renyi* and *Tsallis* entropy. The optimal controlling parameter (q) is determined by applying it in a synthetic contaminated Gaussian mixture dataset. We later demonstrated that the range of selected q – values is applicable in the real life scRNA-seq data clustering task. The four scRNA-seq data where we apply sc-REnF yields accurate clustering results which are validated by the ARI score. The stability of sc-REnF is demonstrated by evaluating the performance of it using KruskalWallis test. While applying sc-REnF multiple times with varying number features, the resulting ARI scores employ a minimum deviation (p -value $\ll 0.05$ for Kruskal-Wallis) for *Renyi* and *Tsallis* entropy.

Although the primary objective of sc-REnF is variable gene selection in the preprocessing step of scRNA-seq data, we extend the process towards the later stage of downstream analysis. We employ the clustering technique to groups the cells using those selected genes. A precise clustering of cells also demonstrates the efficacy of our method for selecting the most variable genes in the first stage. This facilitates the selection of novel marker genes within each cluster. We pinpoint several markers, which shows a high expression level within a particular cluster, among them some are also identified in previously published cell marker database.

Clustering of unknown samples based on the reference data is a crucial problem for the identification of cell types in scRNA-seq cell classification. We addressed the problem by cluster the unknown samples using the selected genes in the reference data. We demonstrate the advantage of using selected genes by sc-REnF in clustering of the unknown test sample. We observed good ARI scores in clustering of test samples, suggesting selected genes from the reference data is also effective to produce a perfect clustering in a completely unknown test sample.

The execution time of sc-REnF is directly proportional to the number of selected features and can be expensive when one needs to select a large number of features. However, this can be easily tackled with ever-increasing computing power in advanced servers. Additionally, the regularization parameter has not been considered in our proposed approach, which may sometimes make the algorithm susceptible to overfitting unless carefully employed.

Taken together, the proposed method sc-REnF not only has good performance on informative gene selection in the preprocessing step but also can explore the classification of unknown cells in the scRNA-seq data. Despite being applied in feature selection of different domains, the application of *Renyi* and *Tsallis* entropy shows good potential in gene selection and cell clustering of scRNA-seq data. Results show that sc-REnF not only leads in the domain of robust feature (gene) selection analyses but accelerate the investigations of cell type definition in large scRNA-seq data as well. We believe that sc-REnF may be an important tool for computational biologists to explore the most informative genes and marker genes in the downstream analysis of scRNA-seq data.

Materials and method

Overview of Datasets

Single Cell RNA sequence dataset Four single-cell RNA-seq datasets are used for the evaluation of the proposed method: Yan, Pollen, Darmanis, and CBMC. Table provides a brief description of the datasets.

- **Yan:**²⁶ This is a human preimplantation embryo and embryonic stem cell dataset. The average total read count in the expression matrix is 25,228,939 reads. There are 7 cell types, including labelled 4-cell, 8-cell, zygote, Late blastocyst and 16-cell.[GEO under accession no. GSE36552].
- **Pollen:**²⁸ The data library was generated from 600 individual cells in parallel. It contains 11 cell types. [GEO under accession no GSM1832359]
- **Darmanis:**²⁷ It contains single cell RNA sequencing on 466 cells to capture the cellular complexity of the adult and fetal human brain at a whole transcriptome level. Healthy adult temporal lobe tissue was obtained from epileptic patients during temporal lobectomy for medically refractory seizures. [GEO under accession no GSE67835].
- **CBMC:**²⁹ Cord blood mononuclear cells (CBMC) were profiled by CITE-seq which is proposed to measure both cellular protein and mRNA expression in one cell, by using oligonucleotide-labeled antibodies. The data set consists of the expression levels of 2000 mRNAs and 13 protein, individually measured in 8,000 cord blood mononuclear cells (CBMCs). The dataset is available in the GEO website under the accession GSE100866.

Deriving *Rényi* and *Tsallis* risk functions

Let us consider three discrete random variables X, Y and Z with supports $\{x_1, \dots, x_d\}$, $\{y_1, \dots, y_p\}$ and $\{z_1, \dots, z_n\}$, respectively. For each $i = 1, 2, \dots, d$, $j = 1, 2, \dots, p$ and $k = 1, 2, \dots, n$, let us denote $p_i = P(X = x_i)$, $p_{ijk} = P(X = x_i, Y = y_j, Z = z_k)$, $p_{i|jk} = P(X = x_i | Y = y_j, Z = z_k)$ and so on. The *Rényi* entropy of the random variable X is defined in terms of a positive real number q , with $q \neq 1$, as

$$H_q(X) = \frac{q}{1-q} \log \left(\sum_i p_i^q \right), \quad q \neq 1. \quad (1)$$

Interestingly, note that, this *Rényi* entropy reduces to the Shannon entropy when $q \rightarrow 1$. It can also be extended for the three random variables X, Y , and Z , so that their joint *Rényi* entropy is given by

$$H_q(X, Y, Z) = \frac{q}{1-q} \log \left(\sum_{i,j,k} p_{ijk}^q \right), \quad q \neq 1. \quad (2)$$

Accordingly, the conditional *Rényi* entropy can be defined as

$$H_q(X|Y, Z) = \frac{q}{1-q} \log \left(\sum_{j,k} \left(\sum_i p_{i|j,k}^q \right)^{1/q} \right), \quad q \neq 1 \quad (3)$$

The *Rényi* entropy is a decreasing function of q . It can be shown that the conditional *Rényi* entropy closely correspond to a risk function, i.e. a form of the expected error when we try to estimate the value of X , given the values of Y and Z . We refer to the associated risk as the *Rényi Risk Function* which is defined as

$$R_q(X|Y, Z) = 1 - \sum_{j,k} p_{j,k} \left(\sum_i p_{i|j,k}^q \right)^{1/q}. \quad (4)$$

Mathematically, the *Tsallis* joint entropy of the three random variables X, Y and Z is defined in terms of a tuning parameter $q > 0$ as

$$H_{T_q}(X, Y, Z) = \frac{1}{q-1} \left[1 - \sum_{i,j,k} p_{i,j,k}^q \right], \quad (5)$$

It also coincides with the Shannon entropy as $q \rightarrow 1$.

The conditional *Tsallis* entropy of random variable X , given the values of Y and Z is defined as

$$H_{T_q}(X|Y, Z) = \frac{1}{q-1} \left[1 - \left(\frac{\sum_{i,j,k} p_{i,j,k}^q}{\sum_{j,k} p_{j,k}^q} \right) \right], \quad q \neq 1. \quad (6)$$

Next, we define a *Tsallis* risk function, i.e. another form of the expected error when we try to estimate the value of X , given the values of Y and Z . This is defined as

$$R_{T_q}(X|Y, Z) = 1 - \sum_{j,k} e p_{j,k}^q \left(\sum_i p_{i|j,k}^q \right), \quad (7)$$

where $e p_{j,k}$ is the joint escort probability distribution³⁰ given by $e p_{j,k} = \frac{p_{j,k}^q}{\sum_{j,k} p_{j,k}^q}$. See section-2 of supplement for a detailed description of the different entropy functions.

Sc-REnF algorithm for feature (gene) selection

Let, any dataset be arranged in a matrix $M_{n \times d}$, where n is number of samples and d is number of features. Let, F be the set of features, $F = \{f_1, f_2, f_3, \dots, f_d\}$, and C be the set of classes. Our algorithm is wrapper based forward selection approach which constructs a monotonically increasing sequence $\{S\}$ of subset of F . At step i , subset $\{f_i\}$ is selected according to the objective function and added to the previous feature subset S_{i-1} . The objective function is defined with a feature relevancy and redundancy criteria driven by entropy measure $\mathcal{E}$. (we utilized *Renyi* and *Tsallis* entropy as the specific choices for $\mathcal{E}$.)

Feature Relevance: Feature f_i is more relevant to the class label C than feature f_j in the context of the already selected subset S , when

$$\mathcal{E}(f_i|C) \geq \mathcal{E}(f_j|C), \quad (8)$$

where $\mathcal{E}(\cdot|\cdot)$ is a (bivariate) conditional entropy function.

Feature Redundance: If feature f_j shares similar information with feature f_i than feature f_{j+1} , then feature f_j is redundant to feature f_i with given information about class label C ; it is characterized as

$$\mathcal{E}(C|f_i, f_j) \geq \mathcal{E}(C|f_i, f_{j+1}), \quad (9)$$

where $\mathcal{E}(\cdot|\cdot, \cdot)$ is an appropriate conditional entropy.

Objective Function: We minimize the conditional entropy function between $f_i \in (F - S)$ and $f_s \in S$ (to reduce the redundancy between them) and maximize the conditional entropy function between class label C and $f_i \in (F - S)$ to select the first feature, where $f_s \in S$ is already selected feature. The selected feature subset, $\{S\}$ and the feature $f_i \in (F - S)$ are inductively defined as

$$\begin{aligned} S &= \emptyset \\ f_1 &= \arg \max_{(f_i \in F)} \mathcal{E}(C|f_i), \\ S &= S \cup \{f_1\}, \\ f_{i+1} &= \arg \min_{(f_i \in (F-S), f_j \in S)} \mathcal{E}(C|f_i, f_j), \\ S &= S \cup \{f_{i+1}\}, \end{aligned} \quad (10)$$

The selected features using the objective function (with *Renyi* and *Tsallis* entropy as a choice of $\mathcal{E}$) are optimal in the sense of minimizing the corresponding risk functions, defined in Equations 4 and 7, respectively, as stated by the proposition:

Theorem: At every step, the selected feature f_{i+1} minimizes the *Renyi* and *Tsallis* risk functions, i.e.,

$$\mathcal{R}(C|f_s, f_{i+1}) \leq \mathcal{R}(C|f_s, f), \forall f \in (F - S), \quad f_s \in S, \quad (11)$$

where $\mathcal{R}$ denotes either R_q or R_{T_q} , defined in Equations 4 and 7, respectively.

Proof. In order to proof the theorem, We will start from the objective function in Equation 10. According to our objective function

$$\mathcal{E}(C|f_s, f_{i+1}) \leq \mathcal{E}(C|f_s, f), \forall f \in (F - S), \quad f_s \in S. \quad (12)$$

Let, u, v', v represent generic value tuples and values of $f_s \in S$ (Selected feature), f_{i+1} (To be selected feature at $(i+1)^{th}$ step), and $f \in (F - S)$ (Non selected features) respectively. After putting the generic representation, Equation 12 for the *Renyi*

entropy can be written as

$$\begin{aligned}
\mathcal{E}(C|u, v') &\leq \mathcal{E}(C|u, v) \\
&\Rightarrow \frac{q}{1-q} \log \sum_{u, v'} \left(\sum_c p_{c, u, v'}^q \right)^{1/q} \\
&\leq \frac{q}{1-q} \log \sum_{u, v} \left(\sum_c p_{c, u, v}^q \right)^{1/q}, \quad (\text{see Equation 3.}) \\
&\Rightarrow \text{If } q > 1, \text{ it leads to} \\
&\sum_{u, v'} \left(\sum_c p_{c, u, v'}^q \right)^{1/q} \geq \sum_{u, v} \left(\sum_c p_{c, u, v}^q \right)^{1/q} \Rightarrow \sum_{u, v'} p_{u, v'} \left(\sum_c p_{c|u, v'}^q \right)^{1/q} \geq \sum_{u, v} p_{u, v} \left(\sum_c p_{c|u, v}^q \right)^{1/q}. \quad (13)
\end{aligned}$$

On the other hand, for $q < 1$, we get

$$\sum_{u, v'} \left(\sum_c p_{c, u, v'}^q \right)^{1/q} \leq \sum_{u, v} \left(\sum_c p_{c, u, v}^q \right)^{1/q} \Rightarrow \sum_{u, v'} p_{u, v'} \left(\sum_c p_{c|u, v'}^q \right)^{1/q} \leq \sum_{u, v} p_{u, v} \left(\sum_c p_{c|u, v}^q \right)^{1/q}. \quad (14)$$

Now, multiplying a constant $k = \frac{q}{1-q}$ in Equations 13 and 14 we get, for any $q \neq 1$, that

$$k \sum_{u, v'} p_{u, v'} \left(\sum_c p_{c|u, v'}^q \right)^{1/q} \geq k \sum_{u, v} p_{u, v} \left(\sum_c p_{c|u, v}^q \right)^{1/q}. \quad (15)$$

Then, by definition of *Renyi* risk function in Equation 4, we get

$$R_q(C|f_s, f_{i+1}) \leq R_q(C|f_s, f). \quad (16)$$

In case of minimization of *Tsallis* risk function, after putting the generic representation, the Equation 12 for the *Tsallis* entropy can be written as

$$\begin{aligned}
\mathcal{E}(C|u, v') &\leq \mathcal{E}(C|u, v) \\
&\Rightarrow \frac{1}{q-1} \left[1 - \left(\sum_{c, u, v'} p_{c, u, v'}^q / \sum_{u, v'} p_{u, v'}^q \right) \right] \leq \frac{1}{q-1} \left[1 - \left(\sum_{c, u, v} p_{c, u, v}^q / \sum_{u, v} p_{u, v}^q \right) \right], \quad (\text{see Equation 6.}) \quad (17)
\end{aligned}$$

Now, for $q > 1$, we have

$$\begin{aligned}
-\left(\sum_{c, u, v'} p_{c, u, v'}^q / \sum_{u, v'} p_{u, v'}^q \right) &\leq -\left(\sum_{c, u, v} p_{c, u, v}^q / \sum_{u, v} p_{u, v}^q \right) \Rightarrow \left(\sum_{c, u, v'} p_{c, u, v'}^q / \sum_{u, v'} p_{u, v'}^q \right) \geq \left(\sum_{c, u, v} p_{c, u, v}^q / \sum_{u, v} p_{u, v}^q \right), \\
&\Rightarrow \left(\sum_{u, v'} p_{u, v'}^q \sum_c p_{c|u, v'}^q \right) / \sum_{u, v'} p_{u, v'}^q \geq \left(\sum_{u, v} p_{u, v}^q \sum_c p_{c|u, v}^q \right) / \sum_{u, v} p_{u, v}^q \Rightarrow \left(\sum_{u, v'} e p_{j, k}^q \sum_c p_{c|u, v'}^q \right) \geq \left(\sum_{u, v} e p_{j, k}^q \sum_c p_{c|u, v}^q \right). \quad (18)
\end{aligned}$$

Also, for $q < 1$, Equation 17 can be written as

$$\left(\sum_{u, v'} e p_{j, k}^q \sum_c p_{c|u, v'}^q \right) \leq \left(\sum_{u, v} e p_{j, k}^q \sum_c p_{c|u, v}^q \right). \quad (19)$$

Now, Multiplying a constant $k = \frac{1}{q-1}$ in Equation 19, we get

$$k \left(\sum_{u, v'} e p_{j, k}^q \sum_c p_{c|u, v'}^q \right) \geq k \left(\sum_{u, v} e p_{j, k}^q \sum_c p_{c|u, v}^q \right). \quad (20)$$

Then, by definition of *Tsallis* risk function, described in Equation 7, we have

$$R_{T_q}(C|f_s, f_{i+1}) \leq R_{T_q}(C|f_s, f). \quad (21)$$

This completes the proof. $\square$

Competing interests

The authors declare that they have no competing interests.

Data and Code Availability

All the data and codes are available in the github page <https://github.com/Snehalikalall/sc-REnF>

Author's contributions

SL and SR equally contributed to this work

Acknowledgment

We would like to acknowledge support from J.C. Bose Fellowship [SB/S1/JCB- 033/2016 to S.B.] by the 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. Govt. of India; Inspire DST Project.

References

1. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell rna-seq in the past decade. *Nat. protocols* **13**, 599–604 (2018).
2. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome biology* **21**, 1–35 (2020).
3. Kiselev, V. Y. *et al.* Sc3: consensus clustering of single-cell rna-seq data. *Nat. methods* **14**, 483–486 (2017).
4. Gribov, A. *et al.* Seurat: visual analytics for the integrated analysis of microarray data. *BMC medical genomics* **3**, 21 (2010).
5. Plass, M. *et al.* Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* **360** (2018).
6. 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).
7. Ray, S. & Schonhuth, A. Markercapsule: Explainable single cell typing using capsule networks. *bioRxiv* (2020).
8. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. biotechnology* **36**, 411–420 (2018).
9. Lall, S., Ray, S. & Bandyopadhyay, S. Rgcop-a regularized copula based method for gene selection in single cell rna-seq data. *bioRxiv* (2020).
10. Duò, A., Robinson, M. D. & Soneson, C. A systematic performance evaluation of clustering methods for single-cell rna-seq data. *F1000Research* **7** (2018).
11. Liu, W. & Lin, W. Additive white gaussian noise level estimation in svd domain for images. *IEEE Transactions on Image processing* **22**, 872–883 (2012).
12. Enver, T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nat. communications* **8**, 1–15 (2017).
13. Palamidessi, C. Feature selection with rényi min-entropy. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, 226–239 (Springer, 2018).
14. Lin, J. Divergence measures based on the shannon entropy. *IEEE Transactions on Inf. theory* **37**, 145–151 (1991).
15. 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).
16. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
17. Brennecke, P. *et al.* Accounting for technical noise in single-cell rna-seq experiments. *Nat. methods* **10**, 1093–1095 (2013).
18. Grün, D., Kester, L. & Van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat. methods* **11**, 637 (2014).

19. Lin, P., Troup, M. & Ho, J. W. Cidr: Ultrafast and accurate clustering through imputation for single-cell rna-seq data. *Genome biology* **18**, 1–11 (2017).
20. Genkin, A., Lewis, D. D. & Madigan, D. Large-scale bayesian logistic regression for text categorization. *Technometrics* **49**, 291–304 (2007).
21. Xing, E. P., Jordan, M. I., Karp, R. M. *et al.* Feature selection for high-dimensional genomic microarray data. In *ICML*, vol. 1, 601–608 (2001).
22. Gajowniczek, K., Zabkowski, T. & Orłowski, A. Comparison of decision trees with rényi and tsallis entropy applied for imbalanced churn dataset. In *2015 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 39–44 (IEEE, 2015).
23. Rajagopal, A., Nayak, A. S., Devi, A. U. *et al.* From the quantum relative tsallis entropy to its conditional form: separability criterion beyond local and global spectra. *Phys. Rev. A* **89**, 012331 (2014).
24. 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).
25. Couch, S., Kazan, Z., Shi, K., Bray, A. & Groce, A. Differentially private nonparametric hypothesis testing. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 737–751 (2019).
26. Yan, L. *et al.* Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. structural & molecular biology* **20**, 1131 (2013).
27. Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci.* **112**, 7285–7290 (2015).
28. Pollen, A. A. *et al.* Low-coverage single-cell mrna sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. biotechnology* **32**, 1053 (2014).
29. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. methods* **14**, 865 (2017).
30. Abe, S. Geometry of escort distributions. *Phys. Rev. E* **68**, 031101 (2003).

Figures

Figure 1

A brief framework of our study: Panel-A: scRNA-seq count matrix are downloaded and preprocessed. scREnF is applied for gene/feature selection using Renyi, Tsalli entropy measure. Panel-B: Selected genes are validated by adopting scRNA-seq clustering techniques. Panel-C: validating the selected genes (from a training set) in an unknown test samples using clustering and ARI method.

Figure 2

Figure shows the results of applying sc-REnF in synthetic Gaussian mixture data. Panel-A shows a two dimensional representation of the generated data with 4 and 5 overlap and non-overlap classes. Panel-B shows the clustering results of the corresponding data shown in the left side panel. Adjusted Rand Index (ARI) of the clustering results are reported for different q-parameter of Renyi and Tsallis entropy.

Figure 3

Figure shows radarplots of six competing methods, each showing the intersection of selected genes with marker genes identified from the whole datasets.

A

B

C

D

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplescrenffinal.pdf](#)