DOI: 10.15514/ISPRAS–2020–32(5)–8

Application of HDBSCAN Method for Clustering scRNA-seq Data

[1,2] *M.A. Akimenkova, ORCID: 0000-0002-9064-5253 <m.akimenkova@ispras.ru>*
[2] *A. A. Maznina, ORCID: 0000-0002-5780-1330 <aamaznina@gmail.com>*
[1] *A. Y. Naumov, ORCID: 0000-0003-4851-7677 <vandedok@ispras.ru>*
[1,2] *E.A. Karpulevich, ORCID: 0000-0002-6771-2163 <karpulevich@ispras.ru>*

[1] *Ivannikov Institute for System Programming of the Russian Academy of Sciences,*
25, Alexander Solzhenitsyn st., Moscow, 109004, Russia
[2] *Moscow Institute of Physics and Technology,*
9, Institutskiy per., Dolgoprudniy, 141701, Russia

Abstract. One of the main tasks in the analysis of single cell RNA sequencing (scRNA-seq) data is the identification of cell types and subtypes, which is usually based on some method of clustering. There is a number of generally accepted approaches to solving the clustering problem, one of which is implemented in the Seurat package. In addition, the quality of clustering is influenced by the use of preprocessing algorithms, such as imputation, dimensionality reduction, feature selection, etc. In the article, the HDBSCAN hierarchical clustering method is used to cluster scRNA-seq data. For a more complete comparison Experiments and comparisons were made on two labeled datasets: Zeisel (3005 cells) and Romanov (2881 cells). To compare the quality of clustering, two external metrics were used: Adjusted Rand index and V-measure. The experiments demonstrated a higher quality of clustering by the HDBSCAN method on the Zeisel dataset and a poorer quality on the Romanov dataset.

Keywords: HDBSCAN; scRNA-seq clustering; denoising autoencoder

For citation: Akimenkova M.A., Maznina A.A., Naumov A.Y., Karpulevich E.A. Application of HDBSCAN method for clustering scRNA-seq data. Trudy ISP RAN/Proc. ISP RAS, vol. 32, issue 5, 2020, pp. 111-120. DOI: 10.15514/ISPRAS-2020-32(5)-8

Применение метода HDBSCAN для кластеризации данных scRNA-seq

[1,2] *М.А. Акименкова, ORCID: 0000-0002-9064-5253 <m.akimenkova@ispras.ru>*
[2] *А.А. Мазнина, ORCID: 0000-0002-5780-1330 <aamaznina@gmail.com>*
[1] *А.Ю. Наумов, ORCID: 0000-0003-4851-7677 <vandedok@ispras.ru>*
[1,2] *Е.А. Карпулевич, ORCID: 0000-0002-6771-2163 <karpulevich@ispras.ru>*

[1] *Институт системного программирования им. В.П. Иванникова РАН,*
109004, Россия, г. Москва, ул. А. Солженицына, д. 25.
[2] *Московский физико-технический институт,*
Россия, 141701, г. Долгопрудный, Институтский пер., д. 9

Аннотация. Одной из основных задач при анализе данных РНК-секвенирования единичных клеток является идентификация типов и подтипов клеток, которая обычно основана на каком-либо методе кластеризации. Существует ряд общепринятых подходов к решению проблемы кластеризации, один из которых реализован в пакете Seurat. На качество кластеризации, помимо прочего, влияет использование алгоритмов предварительной обработки, таких как импутация, уменьшение размерности, отбор признаков и т. д. В статье для кластеризации данных scRNA-seq используется метод иерархической кластеризации HDBSCAN. Для более полного сравнения эксперименты и сравнения проводились на двух размеченных наборах данных: Zeisel (3005 клеток) и Romanov (2881 клетка). Для сравнения качества кластеризации использовались две внешние метрики: скорректированный индекс Рэнда и V-мера. Эксперименты продемонстрировали более высокое качество кластеризации методом HDBSCAN на наборе данных Zeisel и более низкое качество на наборе данных Romanov.

Ключевые слова: hdbscan; кластеризация данных РНК-секвенирования единичных клеток; шумоподавляющий автокодировщик

Для цитирования: Акименкова М.А., Мазнина А.А., Наумов А.Ю., Карпулевич Е.А. Применение метода HDBSCAN для кластеризации данных РНК-секвенирования единичных клеток. Труды ИСП РАН, том 32, вып. 5, 2020 г., стр. 111-120 (на английском языке). DOI: 10.15514/ISPRAS–2020–32(5)–8

1. Introduction

The cell can be considered the fundamental unit in biology. For centuries, biologists have known that multicellular organisms are characterized by many different types of cells. Cells can be distinguished by their size and shape with a microscope, and attributes based on their appearance have traditionally been the main factor in determining cell type. Advances in microfluidics have made it possible to isolate large numbers of cells, and, along with improvements in methods for isolating and amplifying RNA, it is now possible to profile the transcript of single cells using next generation sequencing technologies.

For researchers to make full use of these rich datasets, efficient computational techniques are needed. Numerous steps are neaded before clustering, such as imputation, feature selection, dimensionality reduction. Moreover, there are also software packages that implement the entire clustering workflow, such as Seurat [1]. In this article, we want to compare the use of the popular HDBSCAN [2] algorithm with the steps leading up to clustering, with a Seurat tool.

2. Methods

Many clustering algorithms can be applied to any type of data that is supplied with a measure of the distance between data points. Due to a large number of genes analysed in scRNA-seq, namely the high dimensionality, the distances between data points (i.e., cells) become similar, which is known as the «curse of dimensionality». Hence, distance differences tend to be small and therefore unreliable for identifying clustering. Applying feature selection and / or dimensionality reduction can reduce noise and speed up computations. Feature selection involves identifying the most informative genes, such as genes with the greatest variance, while decreasing dimensionality projects data into a lower-dimensional space. Many tools use variations of standard methods such as PCA [3], uMap [4], DCA [5].

Usually pipelines for scRNA-seq analysis contain tools for imputation, feature selection, dimensionality reduction, etc. This article compares the 14 pipelines shown in the Table. 1.

Table 1. Pipelines

pipeline	**imputation**	**feature selection**	**dimensionality reduction**	**clustering method**
Seurat	no	yes	PCA	Louvain
Seurat*	yes	yes	PCA	Louvain
3	yes	no	DCA	HDBSCAN
4	yes	yes	DCA	HDBSCAN
5	yes	no	uMAP	HDBSCAN

6	yes	yes	uMAP	HDBSCAN
7	yes	no	PCA	HDBSCAN
8	yes	yes	PCA	HDBSCAN
9	no	no	DCA	HDBSCAN
10	no	yes	DCA	HDBSCAN
11	no	no	uMAP	HDBSCAN
12	no	yes	uMAP	HDBSCAN
13	no	no	PCA	HDBSCAN
14	no	yes	PCA	HDBSCAN

2.1 Dimensionality reduction methods

ScRNA-seq data are always large, which increases the complexity of the analysis to some extent. Therefore, to process the initial data we used dimensionality reduction methods.

1) *Principal Component Analysis*

The most common dimensionality reduction technique is principal component analysis (PCA) [3], which requires no control and aims to find a lower-dimensional representation of the data.

PCA is a widely used method of uncontrolled dimensionality reduction. PCA assumes the data is normally distributed, diagonalizes the covariance matrix of the original matrix, and the resulting covariance matrix is a set of new variables for the diagonal matrix. Orthogonal transformation is used to transform a set of potential linear correlation variables into linear explanatory variables, which means that linear dimensionality reduction is realized. One of the main problems with linear dimensionality reduction algorithms is that, when they concentrate disparate data points in a lower-dimensional area, the data points are far apart.

2) *Deep Count Autoencoder*

The deep counting autoencoder network (DCA) [5] denoises scRNA-seq datasets. DCA accounts for the computed distribution, excess variance, and data sparsity using a negative binomial noise model with or without zero inflation, while capturing nonlinear gene-gene relationships.

One of the main advantages of DCA is that the user only needs to specify the noise model. To provide maximum flexibility, DCA implements a set of scRNA-seq-specific noise models, including negative binomial distribution with (ZINB) and no zero inflation (NB).

For example, using the ZINB noise model, DCA examines the meaning of gene-specific parameters, variance, and dropout probability based on gene expression inputs. The derived average distribution parameter is the denoised reconstruction and DCA output.

In our work, we used a 32-dimensional inner layer.

3) *uMAP*

Uniform Manifold Approximation and Projection (uMAP) [4] is a graph-based dimension reduction method similar to t-SNE, introduced by McInnes et al. in 2018. The algorithm builds a high-dimensional graph representation and then optimizes the low-dimensional graph so that it looks structurally as similar as possible to the original.

The advantages of the algorithm include computational efficiency (compared to t-SNE), preservation of the global structure (also compared to t-SNE). In addition, uMAP has no restrictions on the size of the embedded layer, which allows the use of the algorithm for preprocessing to improve the performance of clustering algorithms.

The disadvantages of the uMAP algorithm are lack of interpretability and false detection of noise, uMAP tends to find a diverse structure in the noise of a dataset. uMAP is more reliable with larger datasets as the amount of structure evident to noise tends to decrease in larger datasets.

2.2 Clustering methods

Density-based methods work well even when the data is noisy and the clusters are oddly shaped.

These methods are not generally used for single cell clustering, but they have their advantages.

Both algorithms have the minimum number of samples parameter, which is the neighbor threshold for a record to become a core point. Both algorithms start by finding the core distance of each point, which is the distance between that point and its farthest neighbor, defined by the minimum samples parameter.

DBSCAN [6] is a density-based clustering algorithm – given a set of points in space, the algorithm groups together points that are closely spaced (points with many close neighbors), with lone points in low-density areas marked as outliers (farthest neighbour).

DBSCAN has the epsilon parameter, which is the radius that those neighbors have to be in for the core to form. This algorithm is well suited for clustering single cell data as it copes well with noisy data.

It also finds clusters of exotic shapes: nested and anomalous clusters, as well as low dimension folds. Additionally, there is no need to specify the number of clusters.

HDBSCAN [2] uses a density-based approach, which makes few implicit assumptions about the clusters. It is a non-parametric method that looks for a cluster hierarchy shaped by the multivariate modes of the underlying distribution. Rather than looking for clusters with a particular shape, it looks for regions of the data that are denser than the surrounding space. In addition to being better for data with varying density, it is also faster than regular DBSCAN. HDBSCAN has a minimum cluster size parameter, which defines how big a cluster needs to be in order to form.

2.3 Available workflows

The steps leading up to clustering can have a significant impact on the outcome, and numerous tools are available for each step. There are software packages that implement the entire clustering workflow, such as Seurat [1].

Satija et al. created Seurat, a single cell data analysis toolkit. The expression matrix includes the number of genes, the number of cells and the number of genes in each cell, as well as the number of cells in which each gene is expressed.

Seurat uses Louvain's graph-based algorithm [7]. The advantage is that most graph-based methods do not require the user to specify the number of clusters for identification, instead, indirect resolution parameters are used. The combination of common nearest neighbor graphs and Louvain community detection was first applied to scRNA-seq data in the PhenoGraph method, and this approach has since been incorporated into Seurat. For dimensionality reduction, PCA is used.

Because of their speed and scalability, the clustering techniques included in Seurat packages are a popular choice for large datasets. However, Louvain clustering has proved to be ineffective for small datasets.

3. Feature selection

Feature selection a collection of statistical approaches that identify and retain only variables that are most relevant to the underlying structure of the data set.

Due to the large number of genes analysed in scRNA-seq, that is, the high dimensionality, the distances between data points (i.e., cells) become similar, which is known as the «curse of dimensionality». Hence, distance differences tend to be small and therefore unreliable for identifying cell groups. Using feature selection can reduce noise and speed up calculations. Feature selection includes identifying the most informative genes, for example, with the highest variance [8].

The expression data of one cell contains a set of missing values and noise data that affects the next step in the analysis. Feature selection with variance has been used to alleviate these issues. Inspired by Prabhakaran et al. [9], we selected groups of genes with the greatest variance in expression. For the Zeisel [10] and Romanov [11], the initial sizes were 19,972 and 24,341 respectively. We took the feature selection data to select genes with high variance. Variance represents the degree of differentiation of gene expression across all cells, and high variance indicates that the gene was important for distinguishing cells. Therefore, we could easily get more biologically significant clusters. Using feature selection, a subset with top 200 genes was generated for Zeisel [9] and Romanov [11] data. We performed the following experiments with three clustering models. We compared all three clustering algorithms (ie HDBSCAN+PCA, HDBSCAN+uMAP) on the subset with the original data (19,972 and 24,341 genes without traits). Selection of the top 200 genes for each of the three algorithms enhanced clustering quality as opposed to the use of the full set of genes (19,972 and 24,341 genes).

In addition, HDBSCAN+DCA algorithm performed best among these clustering algorithms, reaching an accuracy of 0.95 on 200 gene sets. The accuracy was 9.3% higher than the result without gene selection. Meanwhile, using the gene selection method, HDBSCAN+uMAP, HDBSCAN+PCA, it was possible to increase the accuracy by 11.8% and 21.9%, respectively. These results showed that clustering with gene selection gives better performance than methods without it.

4. Imputation

The scRNA-seq data is characterized by excess zero counts, the so-called dropouts due to the low number of mRNAs sequenced within individual cells. In order to reduce the number of dropouts in some experiments, the scRNA-seq scImpute [12] method is used. ScImpute is a statistical method for imputing dropouts. ScImpute automatically detects likely dropouts and imputes only those values without changing the rest of the data. The scImpute algorithm also detects outliers in scRNA-seq data and excludes them from imputation. The effectiveness of scImpute has been demonstrated on both simulated and real human and mouse scRNA-seq data. ScImpute detects and imputes dropouts, thereby enhancing the analysis of differential expression and clustering of cell subpopulations.

In the article scImpute is used to improve the quality of clustering in combination with feature selection and dimensionality reduction methods (PCA [3], DCA [5], uMAP [4]).

5. Evaluation

To evaluate the performance of an array of popular clustering methods, we tested Seurat [1] and HDBSCAN [2] with different methods of dimensionality reduction such as DCA [5], PCA [3], uMAP [4] on 2 published datasets.

Since we have published cell type labels, for the assessment of clustering quality, we used two external quality metrics – Adjusted Rand Index [13] and V-measure [14].

The Adjusted Rand index was chosen as the first metric of clustering quality. This metric is external, i.e. the measures are based on comparing the clustering result with the a priori known division into classes. This metric is robust to the size and number of clusters.

V-measure was used as the second quality metric. The main advantage of this metric is that it is independent of the number of class labels, the number of clusters, the size of the data, and the clustering algorithm used, and is very reliable.

6. Experiments

The operation of the selected algorithms is demonstrated on two scRNA-seq gene expression datasets for house mouse cells. The Zeisel [10] contains scRNA-seq-derived cortical cell expression data from the house mouse and describes 3005 different cells of 9 different types. The data are also presented as a 19,772 x 3005 gene expression matrix. The Romanov [11] contains expression data of hypothalamic cells from the cortex of the house mouse, obtained by scRNA-seq, and describes 2881 different cells of 7 different types. The data are also presented as a 24,341 x 2881 gene expression matrix.

For a sufficient set of statistics, the Zeisel set was divided into 8 non-overlapping sets with a balanced number of cells in each cluster: 7 sets of 19,772 x 353 and one set of 19,772 x 358. The Romanov set was also divided into 8 non-overlapping sets: 7 sets of 24,341 x 346 and one sets of 24,341 x 342.

To investigate the effectiveness of clustering models with and without feature selection, as well as various dimensionality reduction techniques, we directly clustered the original data and feature selection data for 200 genes.

The results are illustrated in the Table 4 for ARI [13] metric and in the Table 5 for V-measure [14] metric.

To test the results for statistical significance, we first used the Friedman test, which the null hypothesis that repeated measurements of the same individuals have the same distribution. If the null hypothesis was rejected, we calculated pairwise comparisons using Conover post hoc test. This test is usually conducted post hoc after significant results of the Friedman test.

We discovered that HDBSCAN+DCA (algorithms 9 and 10) clustering achieved the best results on the original and feature-selected data. On the original data, HDBSCAN+DCA reached an accuracy of 0.87, which was 16.3% and 21.3% higher than those of HDBSCAN+uMAP and HDBSCAN+PCA, respectively. For 200 genes, HDBSCAN+DCA+fs achieved an accuracy of 0.95, which was 4.7%, 17.6% and 25.2% higher than Seurat, HDBSCAN+uMAP+fs and HDBSCAN+PCA+fs, respectively. P-value on Friedman test for HDBSCAN+DCA+fs, HDBSCAN+uMAP+fs, HDBSCAN+PCA+fs and Seurat pipelines is 9.8×10^{-5} and we calculated pairwise comparisons using Conover post hoc test. From the Table 2, HDBSCAN+DCA+fs demonstrated statistically significant result for all three other pipelines.

Table 2. Non-imputed algorithms posthoc p-values

	seurat	**fs+dca+hdbscan**	**fs+umap+hdbscan**	**fs+pca+hdbscan**
seurat	1	0.0417	0.0146	0.0198
fs+dca+hdbscan	0.0417	1	0.038	0.0039
fs+umap+hdbsca	0.0146	0.038	1	0.5329
fs+pca+hdbscan	0.0198	0.0039	0.5329	1

For the imputed data, P-value on Friedman test is 0.0002 we calculated pairwise comparisons using Conover post hoc test. From the Table 3, scImpute+HDBSCAN+DCA+fs demonstrated statistically significant result for all three other pipelines.

Table. 3. Imputed algorithms posthoc p-values

	seurat	**fs+dca+hdbscan**	**fs+umap+hdbscan**	**fs+pca+hdbscan**
seurat	1	0.0475	0.0024	0.0078
fs+dca+hdbscan	0.0475	1	0.038	0.0029

fs+umap+hdbscan	0.0024	0.038	1	0.2079
fs+pca+hdbscan	0.0078	0.0029	0.2079	1

Table 4. Adjusted Rand Index for different experiments

Dataset	Seurat	Impute Seurat	Impute DCA hdbscan	Impute fs DCA hdbscan	Impute uMAP hdbscan	Impute fs uMAP hdbscan	Impute PCA hdbscan
zeisel1	0.754	0.81	0.55	0.81	0.518	0.55	0.65
zeisel2	0.83	0.829	0.861	**0.851**	0.651	0.651	0.47
zeisel3	0.748	0.794	0.738	**0.803**	0.608	0.607	0.17
zeisel4	0.793	0.798	0.604	**0.858**	0.549	0.633	0.45
zeisel5	0.781	0.798	0.747	0.785	0.676	0.734	0.5
zeisel6	0.827	0.777	0.628	**0.865**	0.555	0.621	0.15
zeisel7	0.76	0.705	0.828	0.763	**0.893**	0.819	0.17
zeisel8	0.827	0.869	0.861	0.907	0.57	0.657	0.508
romanov1	**0.809**	0.68	0.605	0.604	0.644	0.626	0.364
romanov2	**0.772**	0.625	0.608	0.667	0.599	0.696	0.325
romanov3	0.65	0.643	0.546	0.587	0.476	0.643	0.327
romanov4	**0.696**	0.537	0.553	0.621	0.533	0.675	0.256
romanov5	**0.801**	0.535	0.516	0.614	0.568	0.654	0.077
romanov6	0.565	0.472	0.518	0.55	0.541	0.482	0.071
romanov7	0.561	0.672	0.647	0.66	0.642	0.645	0.671
romanov8	**0.704**	0.641	0.66	0.691	0.637	0.656	0.291

Table 4 (cont.)

Dataset	Impute fs PCA hdbscan	DCA hdbscan	fs DCA hdbscan	uMAP hdbscan	fs uMAP hdbscan	PCA hdbscan	fs PCA hdbscan
zeisel1	0.253	0.819	**0.827**	0.476	0.644	0.23	0.278
zeisel2	0.387	0.607	0.838	0.572	0.671	0.55	0.565
zeisel3	0.214	0.727	0.731	0.563	0.644	0.17	0.123
zeisel4	0.306	0.815	0.844	0.548	0.634	0.55	0.281
zeisel5	0.453	0.772	**0.907**	0.52	0.630	0.41	0.436
zeisel6	0.146	0.698	0.802	0.739	0.626	0.19	0.32
zeisel7	0.197	0.795	0.824	0.545	0.619	0.34	0.332
zeisel8	0.549	0.869	**0.95**	0.57	0.609	0.579	0.61
romanov1	0.495	0.754	0.768	0.747	0.668	0.184	0.356
romanov2	0.048	0.791	0.756	0.722	0.731	0.117	0.041
romanov3	0.195	0.641	**0.702**	0.629	0.689	0.694	0.372
romanov4	0.288	0.675	0.69	0.694	0.459	0.152	0.078
romanov5	0.064	0.795	0.797	0.771	0.671	0.152	0.218
romanov6	0.083	0.598	**0.624**	0.472	0.626	0.121	0.153
romanov7	0.447	**0.704**	**0.704**	0.64	0.613	0.799	0.655
romanov8	0.382	0.677	0.686	0.512	0.656	0.293	0.266

Table 5. V-measure for different experiments

Dataset	Seurat	Impute Seurat	Impute DCA hdbscan	Impute fs DCA hdbscan	Impute uMAP hdbscan	Impute fs uMAP hdbscan	Impute PCA hdbscan
zeisel1	0.741	0.802	0.696	**0.835**	0.695	0.71	0.412
zeisel2	0.793	0.798	0.824	**0.894**	0.709	0.785	0.424
zeisel3	0.762	0.786	0.834	**0.885**	0.702	0.705	0.411
zeisel4	0.778	0.813	0.736	**0.846**	0.677	0.702	0.402
zeisel5	0.768	0.8	0.733	0.777	0.695	0.699	0.41
zeisel6	**0.792**	0.754	0.678	0.759	0.643	0.651	0.395
zeisel7	0.779	0.745	0.829	0.851	**0.875**	0.75	0.394
zeisel8	0.817	0.869	0.822	0.873	0.71	0.794	0.42
romanov1	**0.773**	0.629	0.597	0.596	0.591	0.522	0.315
romanov2	0.773	0.58	0.597	0.665	0.512	0.584	0.325
romanov3	**0.661**	0.624	0.531	0.587	0.459	0.521	0.328
romanov4	**0.712**	0.591	0.631	0.684	0.587	0.606	0.348
romanov5	**0.76**	0.59	0.565	0.57	0.515	0.543	0.309
romanov6	0.587	0.533	0.527	0.522	0.547	0.489	0.318
romanov7	0.636	0.636	0.623	0.638	0.545	0.553	0.322
romanov8	**0.702**	0.611	0.638	0.649	0.575	0.558	0.331

Table 5 (cont.)

Dataset	Impute fs PCA hdbscan	DCA hdbscan	fs DCA hdbscan	uMAP hdbscan	fs uMAP hdbscan	PCA hdbscan	fs PCA hdbscan
zeisel1	0.413	0.774	0.826	0.664	0.69	0.415	0.4
zeisel2	0.418	0.638	0.79	0.672	0.685	0.428	0.422
zeisel3	0.407	0.765	0.775	0.718	0.709	0.409	0.402
zeisel4	0.401	0.789	0.809	0.682	0.698	0.414	0.404
zeisel5	0.394	0.765	**0.863**	0.696	0.676	0.405	0.399
zeisel6	0.396	0.713	0.778	0.744	0.635	0.402	0.405
zeisel7	0.386	0.778	0.812	0.709	0.65	0.413	0.404
zeisel8	0.406	0.819	**0.929**	0.715	0.703	0.416	0.401
romanov1	0.316	0.704	0.72	0.687	0.585	0.347	0.345
romanov2	0.317	0.733	**0.791**	0.669	0.651	0.34	0.333
romanov3	0.325	0.616	0.631	0.598	0.592	0.332	0.329
romanov4	0.339	0.669	0.677	0.669	0.513	0.347	0.327
romanov5	0.305	0.739	0.745	0.722	0.617	0.349	0.345
romanov6	0.308	0.611	**0.685**	0.48	0.569	0.343	0.332
romanov7	0.314	0.679	**0.712**	0.64	0.595	0.333	0.343
romanov8	0.321	0.648	0.65	0.54	0.588	0.345	0.329

7. Conclusion

Dimensional reduction and clustering are important when analysing scRNA-seq data. A comparative framework is pro-posed that combines three dimensionality reduction methods and feature selection with HDBSCAN [2] clustering with an en-tire Seurat [1] pipeline. Fourteen experiments were performed on two large scRNA-seq datasets using these combinations.

Two conclusions can be drawn from the results. Thus, feature selection and dimensionality reduction with DCA [5] are critical to achieving better clustering results. If the result is unsatisfactory, imputation methods maybe introduced. HDBSCAN clustering can give satisfactory results in most cases.

Contributions

M.A., A.M. and E.K. wrote the article. A.N. designed experiments and directed research. M.A. performed research and analysed data. E.K. supervised the project. A.N. provided feedback on the text.

References

[1] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature biotechnology, vol. 36, no. 5, 2018, pp. 411–420.

[2] L. McInnes, J. Healy, and S. Astels. hdbscan: Hierarchical density based clustering. Journal of Open Source Software, vol. 2, no. 11, 2017, article no. 205.

[3] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. Chemometrics and intelligent laboratory systems, vol. 2, no. 1-3, 1987, pp. 37–52.

[4] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426, 2018.

[5] G. Eraslan, L. M. Simon, M. Mircea, N. S. Mueller, and F. J. Theis. Single-cell rna-seq denoising using a deep count autoencoder. Nature communications, vol. 10, no. 1, 2019, pp. 1–14.

[6] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu. Dbscan revisited, revisited: why and how you should (still) use dbscan. ACM Transactions on Database Systems, vol. 42, no. 3, 2017, pp. 1–21.

[7] H. Lu, M. Halappanavar, and A. Kalyanaraman. Parallel heuristics for scalable community detection. Parallel Computing, vol. 47, 2015, pp. 19–37.

[8] P. Brennecke, S. Anders, J. Kim et al. Accounting for technical noise in single-cell rna-seq experiments. Nature Methods, vol. 10, 2013, pp. 1093–1095.

[9] S. Prabhakaran, E. Azizi, A. Carr, D. Pe'er. Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. In Proc. of the 33rd International Conference on Machine Learning, 2016, pp 1070-1079.

[10] A. Zeisel, A.B. Mũnoz-Manchado et al. Cell types in the mousecortex and hippocampus revealed by single-cell rna-seq. Science, vol. 347, no. 6226, 2015, pp.1138–1142.

[11] R.A. Romanov, A. Zeisel et al. Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. Nature neuroscience, vol. 20, no. 2, 2017, pp. 176–188.

[12] W. Li and J.J. Li. An accurate and robust imputation method scimpute for single-cell rna-seq data. Nature communications, vol. 9, no. 1, 2018, article no. 997.

[13] D. Steinley. Properties of the hubert-arable adjusted rand index. Psychological methods, vol. 9, no. 3, 2004, pp. 386–396.

[14] A. Rosenberg and J. Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), 2007, pp. 410–420.

Информация об авторах / Information about authors

Maria Andreevna AKIMENKOVA – laboratory assistant, Information Systems Department. Research interests: analysis of biomedical data, analysis of single cell sequencing data.

Мария Андреевна АКИМЕНКОВА – лаборант отдела «Информационные системы». Сфера научных интересов: анализ биомедицинских данных, анализ данных секвенирования единичных клеток.

Анна Анатольевна МАЗНИНА – лаборант лаборатории геномной инженерии. Сфера научных интересов: РНК-секвенирование, онкогенетика.

Anna Anatolyevna MAZNINA – laboratory assistant at the Genomic Engineering Laboratory. Research interests: RNA sequencing, oncogenetics.

Anton Yurievich NAUMOV – research assistant, Information Systems Department. Research interests: machine learning, convolutional neural networks.

Антон Юрьевич НАУМОВ – стажер-исследователь отдела «Информационные системы». Сфера научных интересов: машинное обучение, сверточные нейронные сети.

Evgeny Andreevich KARPULEVICH – researcher, Information Systems Department. Research interests: analysis of biomedical data, information systems.

Евгений Андреевич КАРПУЛЕВИЧ – научный сотрудник отдела «Информационные системы». Сфера научных интересов: анализ биомедицинских данных, информационные системы.