UDC 004.9
IRSTI 28.23.25

https://doi.org/10.55452/1998-6688-2024-21-4-45-57

[1]*Kunikeyev A.,
Master of Engineering Sciences, ORCID ID: 0000-0003-3716-0895,
*e-mail: aidyn.daulet@gmail.com
[1,2]Yerimbetova A.,
PhD, Candidate of Technical Sciences, Associate Professor,
ORCID ID: 0000-0002-2013-1513,
e-mail: aigerian8888@gmail.com
[1]Satybaldiyeva R.,
Candidate of Technical Sciences, Professor, ORCID ID: 0000-0002-0678-7583
e-mail: r.satybaldiyeva@satbayev.university

[1]Satbayev University, Almaty, Kazakhstan
[2]Institute of Information and Computational Technologies of the Committee
of Science of the Ministry of Science and Higher Education
of the Republic of Kazakhstan, Almaty, Kazakhstan

A REVIEW OF TOOLS, METHODOLOGIES, AND TECHNIQUES FOR PROCESSING, PRE-PROCESSING, AND CLUSTERING ANALYSIS OF GENETIC DATA

Abstract

Gene expression analysis has become a key component in understanding cellular behavior, disease mechanisms, and drug response. The advent of high-throughput sequencing, particularly single-cell RNA sequencing (scRNA-seq), has expanded our ability to study cellular heterogeneity to an unprecedented level. Clustering algorithms needed to group genes or cells with similar expression profiles have become invaluable for analyzing the massive data sets generated by these technologies. This article reviews various clustering methods applied to gene expression data, particularly single-cell RNA sequencing. The analysis covers traditional methods such as hierarchical clustering and k-means, as well as more advanced approaches such as model-based clustering, machine learning-based methods, and deep learning methods. The primary challenges encompass handling high-dimensional data, mitigating noise, and achieving scalability for large datasets. Moreover, new advancements such as multi-omics data integration, deep learning-based clustering, and federated learning offer potential enhancements in accuracy and biological relevance for clustering applications in gene expression research. The review concludes with a discussion of clustering algorithms in handling increasingly complex gene expression data for more accurate biological insights.

Key words: Clustering methods, Bioinformatics, Machine Learning, Deep learning, single-cell RNA sequencing, Gene expressions.

Introduction

Gene expression analysis has become a cornerstone of modern molecular biology, providing vital information about cellular functions [1], disease mechanisms [2], and drug responses [3]. The advent of high-throughput sequencing technologies, particularly single-cell RNA sequencing (scRNA-seq), has transformed our ability to conduct gene expression research [4] . As we delve into the complex world of cellular heterogeneity and function, clustering algorithms are becoming indispensable tools for making sense of the vast and complex datasets generated by these technologies.

Using clustering, an unsupervised machine learning technique, researchers can categorize genes or cells by similar expression patterns, which helps reveal underlying structures in the data [5].

As the field of genomics continues to evolve, the importance of robust and efficient clustering algorithms cannot be overstated. The aim of this review is to provide a comprehensive overview of the current state of clustering methods in gene expression analysis, with a particular focus on their application to single-cell RNA-seq data.

The Rise of Single-Cell RNA Sequencing. Single-cell RNA sequencing has emerged as a powerful tool in genomics, offering unprecedented insights into cellular heterogeneity and function. Unlike bulk RNA sequencing, which provides an average expression profile across a population of cells, scRNA-seq captures the transcriptomes of individual cells, revealing true diversity in seemingly homogeneous populations [6].

Identifying Complex and Rare Cell Types. One of the main applications of scRNA-seq is the identification and characterization of rare cell populations that may be masked in bulk sequencing approaches. By studying the transcriptomes of individual cells, researchers can detect subtle changes in gene expression that define different cell types or states.

Elucidating Gene Regulatory Networks. The high-resolution data provided by scRNA-seq enable the construction of detailed gene regulatory networks. By studying gene co-expression patterns in individual cells, researchers can infer regulatory relationships and build models of gene interaction networks. This approach has led to the discovery of new regulatory mechanisms and improved our understanding of how gene expression is coordinated at the cellular level [6].

Assessing Developmental Trajectories. scRNA-seq has revolutionized the study of developmental biology by allowing researchers to track the developmental trajectories of individual cell lineages. By analyzing gene expression profiles of cells at different stages of development, scientists can reconstruct the molecular pathways that direct cellular differentiation and maturation [7].

Revealing Cell-to-Cell Variations in Disease. In the context of disease research, scRNA-seq has proven invaluable in identifying cell-to-cell variations in various disease states. The technology has been particularly effective in cancer research, where it has shed light on tumor heterogeneity and the existence of rare cell populations that may contribute to drug resistance or disease progression [8].

Applications in Drug Discovery and Development. scRNA-seq has applications that reach fundamental research, playing a crucial role in drug discovery and development. This technology allows for more precise and effective drug discovery strategies by offering in-depth insights into how individual cells respond to drugs [9].

As we delve into the world of clustering algorithms for gene expression analysis, it is critical to remember the profound impact these computational tools have on our ability to extract meaningful biological insights from the wealth of data generated by single-cell RNA sequencing. The following sections will explore various clustering approaches, their strengths and limitations, and their applications to deciphering the complexities of gene expression data.

Materials and Methods

Clustering is a fundamental technique in gene expression analysis, enabling researchers to group genes or samples based on their expression profiles. This facilitates the identification of patterns and relationships, aiding in the understanding of complex biological processes. Various clustering methods have been developed, each with its strengths and limitations. Traditional methods like hierarchical and k-means clustering offer simplicity and efficiency, while model-based approaches provide flexibility and statistical rigor. This section explores several clustering techniques, highlighting their applications, advantages, and limitations in the context of gene expression analysis.

Traditional Clustering Methods. Traditional clustering methods are widely employed due to their straightforward implementation and interpretability [10]. Hierarchical clustering is particularly popular for its ability to uncover relationships at multiple levels of granularity. By constructing a dendrogram, hierarchical clustering provides a visual representation of nested clusters, making it intuitive for biological interpretation [11]. Its flexibility allows researchers to explore data without specifying the number of clusters in advance. However, it has significant computational demands,

with a time complexity of O(n3)O(n^3)O(n3), and is sensitive to noise and outliers, which can distort the clustering structure.

K-means clustering is another widely used method, known for its computational efficiency and simplicity [12]. It partitions data into kkk clusters by minimizing within-cluster variance, making it well-suited for large datasets. K-means is easy to implement and adapts to various distance metrics, enhancing its flexibility. Nonetheless, it requires the number of clusters to be predefined, which can be a challenging task. Additionally, the algorithm is sensitive to the initial placement of centroids and assumes that clusters are spherical, which may not always align with the actual data structure.

K-medoids clustering addresses some of k-means' limitations by using medoids – actual data points – as cluster centers, which improves robustness to outliers [13]. This method offers better interpretability, as medoids are representative of the data. K-medoids allows for the use of various distance metrics, making it possible to analyze mixed data types; nevertheless, it is more computationally intensive than k-means and still requires the number of clusters to be specified in advance.

Model-Based Clustering Methods. Model-based clustering operates on the assumption that data are produced from a combination of probability distributions, making it especially suitable for gene expression data, which frequently display complex statistical characteristics [14].

Gaussian Mixture Models (GMM), one of the most commonly used model-based approaches, fit data as a mixture of multivariate Gaussian distributions [15]. This allows for flexible cluster shapes and probabilistic assignments, enabling the handling of overlapping clusters. However, GMMs are sensitive to initialization and can overfit in high-dimensional spaces.

Latent Dirichlet Allocation (LDA), originally developed for text analysis, has been adapted for gene expression studies to identify latent functional groups of genes [16]. LDA assigns genes to multiple groups with varying probabilities, offering a nuanced view of gene relationships. While it can reveal biologically meaningful patterns, LDA requires careful interpretation and may not capture all expression data complexities.

Clustering remains a cornerstone of gene expression analysis, with traditional and model-based methods offering complementary strengths. Traditional methods like hierarchical, k-means, and k-medoids clustering are prized for their simplicity and efficiency, though they face challenges with large datasets and noisy data. Model-based approaches, including GMMs, LDA, and DPMMs, provide greater flexibility and adaptability to complex data structures but come at the cost of higher computational complexity. The choice of clustering method depends on the specific goals of the analysis, data characteristics, and available computational resources. Together, these methods provide a robust toolkit for exploring and interpreting the rich information embedded in gene expression data.

Results

RNA-seq Data Analysis Pipeline. The RNA-seq data analysis pipeline shown in Figure 1 is a critical process for extracting meaningful biological information from raw sequencing data [17–19]. This pipeline consists of several key steps, each of which plays an important role in transforming raw reads into interpretable gene expression data. Understanding this pipeline is necessary to appreciate the context in which clustering algorithms are applied.

Preprocessing. Preprocessing is an initial and critical step in the RNA-seq data analysis process. This step aims to ensure the quality and integrity of the raw sequencing data before further analysis. The main goals of preprocessing include:

1. Quality Control: The overall quality of the raw data is assessed using tools such as FastQC [20]. At this stage, various metrics such as per-base sequence quality, GC content, sequence length distribution, and overrepresented sequences are checked.
2. Adapter Trimming: Sequencing adapters, which are artificial sequences added during library preparation, are removed using tools such as Trimmomatic [21] or Cutadapt [22].

Alignment. Alignment, also known as mapping, is the process of determining the genomic origin of each sequence read. This step is fundamental to RNA-seq analysis because it allows one to determine which genes are expressed in a sample. Aligners such as STAR [23], HISAT2 [24], or TopHat2 [25] are used to handle reads that span exon junctions.

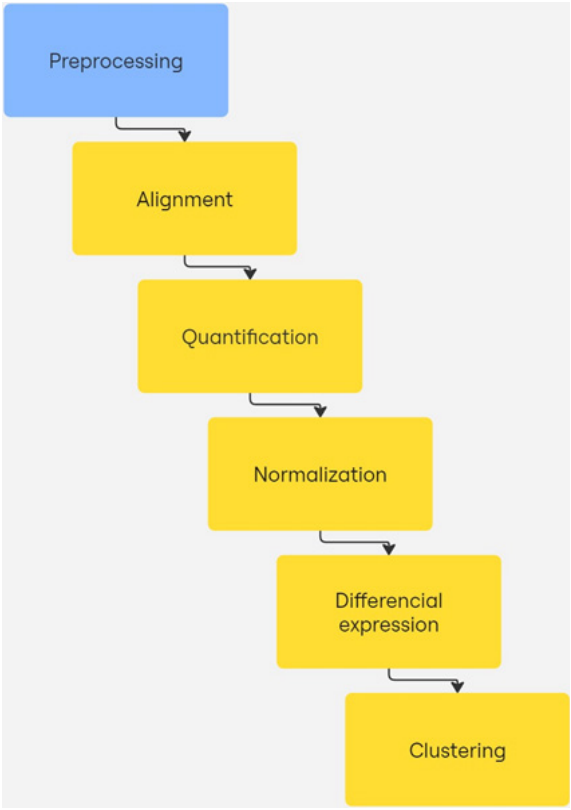

Figure 1 – Gene expression analysis pipeline

Quantification. Quantification is the process of assessing gene or transcript expression levels from aligned RNA-seq reads. This step converts the mapped reads into meaningful expression values that can be used for comparative analysis. Quantification tools such as HTSeq [26] or featureCounts [27] are used to count the number of reads that map to each gene's exons.

Normalization. Normalization is a critical step to remove systematic bias and ensure comparability between samples. Without proper normalization, differences in sequencing depth, gene length, and other technical factors can obscure true biological differences.

Differential Expression. Differential expression analysis aims to identify genes that show statistically significant differences in expression levels between experimental conditions. This process typically involves tools such as DESeq2 [28], or edgeR [28] use statistical models to test for differential expression.

Clustering. Clustering is a powerful technique used to group genes or samples with similar expression patterns. This step is essential for detecting co-expressed genes [29], uncovering new cell or tissue subtypes, and gaining insights into the overall structure of gene expression data.

Various clustering algorithms can be employed (Table 1). We compared six clustering algorithms using the R programming language, providing plots that demonstrate their applications and interpretability. All analyses were conducted using an open-access dataset from NCBI (PRJNA736095; GEO: GSE176415), following a comprehensive preprocessing workflow up to the extraction of the gene expression set, utilizing the Galaxy Project platform [30], with clustering performed exclusively on the gene expression dataset. In this table, we aimed to highlight the advantages and limitations encountered during our analysis.

Comprehending this pipeline is vital for evaluating the context in which clustering algorithms are utilized, along with the various factors that can affect their performance and interpretation. In the sections that follow, we delve into specific clustering methods used in gene expression analysis, examining their advantages, limitations, and role in interpreting complex gene expression data.

Conclusion

The application of clustering algorithms to gene expression data has transformed our understanding of biological systems, which allow researchers to uncover complex patterns, identify co-expressed genes, and classify cell types with unprecedented accuracy. This review explored a wide range of clustering methods, from traditional approaches to cutting-edge machine learning methods, each offering unique strengths in gene expression data analysis. The challenges and future directions in clustering gene expression data highlight the intricate nature of modern genomic datasets and the increasing need for clustering algorithms that are scalable, interpretable, and biologically meaningful. Addressing the limitations of high dimensionality, data integration, and scalability will be critical to the continued success of clustering in gene expression analysis. Additionally, emerging trends such as deep learning, automated clustering pipelines, and federated learning hold great promise for advancing the field and enabling new biological discoveries. With the shift toward larger, more complex, and multimodal datasets, it will be crucial to develop advanced clustering methods capable of handling these challenges to deepen our understanding of gene regulation, cellular diversity, and disease mechanisms.

The diversity of available clustering methods reflects the complexity and heterogeneity of gene expression data. Although traditional methods like hierarchical clustering and k-means are still widely utilized, advanced machine learning and deep learning techniques are being increasingly adopted to tackle the challenges of high-dimensional and noisy data. As the field of gene expression analysis continues to evolve, integrating these methods with biological knowledge and leveraging the strengths of each approach will be critical to uncovering new insights into gene regulation and cellular function.

To sum up, the field of clustering algorithms for gene expression analysis is advancing swiftly, propelled by breakthroughs in sequencing technologies and computational methods. With the move toward more complex, large-scale, and multimodal datasets, developing sophisticated, scalable, and biologically interpretable clustering algorithms becomes crucial. These advancements hold the potential to enhance our understanding of gene regulation, cellular diversity, and intricate biological systems, ultimately fueling progress in areas from developmental biology to personalized medicine.

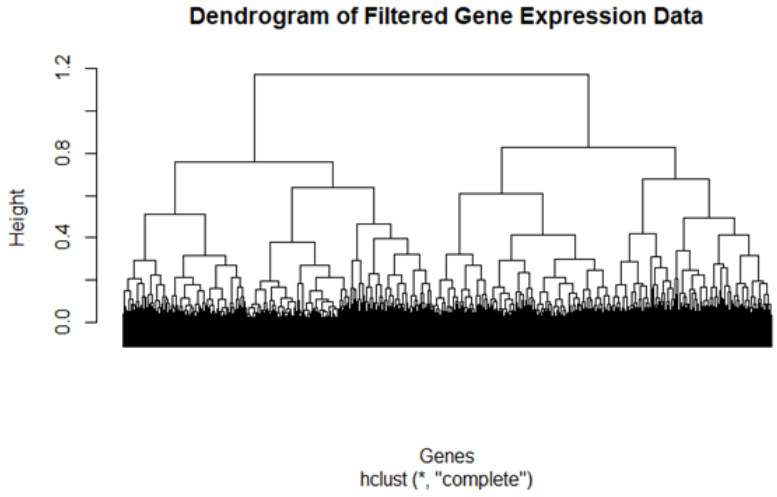

Figure 2 – Dendrogram. Hierarchical clustering of gene expression dataset

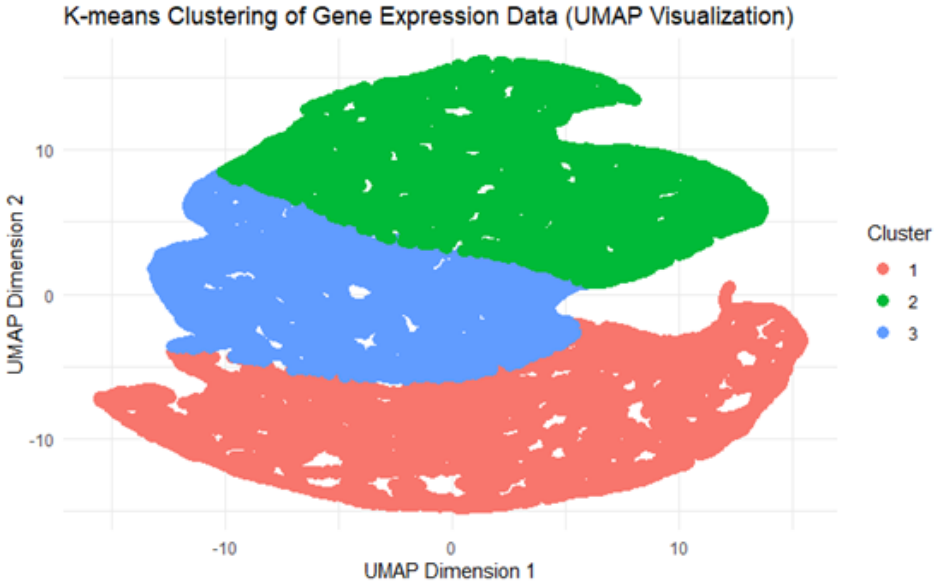

Figure 3 – UMAP. of Hierarchical clustering of gene expression dataset

Table 1 – Comparisonal table of, mostly known clustering algorithms used in searching similar pattern on gene expressions. The given code on R is set of examples to run these clustering algorithms

№	Name	Input	Output, plot and interpretations. Sample.	Advantages and limitations	Reference
1	Hierarchical clustering	Gene expression data	Plot: Dendrogram plot. Output: Dendrogram showing hierarchical structure of clusters. Interpretation: The hierarchy in clusters allows identifying nested subgroups within the data, providing insight into relationships among gene expression patterns and potential biological processes or cell types. Sample: The dendrogram (Figure 2) displays three primary clusters when cut at a height of 0.8, each indicating a broad category of gene expression similarity.	Advantages: ◆ Provides a detailed hierarchical structure, allowing exploration of clusters at different levels. ◆ Does not require pre-specifying the number of clusters, making it flexible for exploratory analysis. ◆ Useful for identifying nested clusters and understanding complex relationships within data. Limitations: ◆ Computationally intensive for large datasets, as in samples.	[11]
2	K-means clustering	Gene expression data	Output: Cluster centroids and assignment of data points to clusters. Plot: Scatter plot with cluster boundaries or bar chart of cluster sizes on UMAP. Interpretation: Each cluster centroid represents a gene expression pattern, helping to identify dominant expression profiles that may correspond to biological functions. Sample: To make this example relevant, we focused on three clusters (Figure 3), where each group may exhibit shared functional characteristics or biological pathways.	Advantages: ◆ Efficient and computationally fast for large datasets, as in samples. Limitations: ◆ Requires specifying the number of clusters in advance, which may not always be known.	[12]

Continuation of table 1

3	K-medoids clustering	Gene expression data	Output: Medoids of clusters with each data point assigned to a medoid. Plot: Similar to K-means, often scatter plots UMAP or heatmaps Interpretation: Medoids provide robust representative profiles of each cluster, which can reduce sensitivity to outliers and reveal distinct gene expression patterns. Sample: The clear separation of clusters (Figure 4) in the plot indicates distinct gene expression profiles, where each group may correspond to different biological functions or pathways. Output is very similar to the output of K-means clustering.	Advantages: ◆ Less sensitive to outliers than K-means, as medoids are more representative of cluster centers. ◆ Does not require spherical clusters, allowing for flexibility in cluster shapes. Limitations: ◆ Slower and more computationally intensive than K-means, especially for large datasets as in samples. ◆ Requires specifying the number of clusters in advance, similar to K-means.	[13]
4	Model-Based Clustering Methods	Gene expression data	Output: Probability of each data point belonging to a specific cluster. Plot: Probability density plots or cluster assignment visualizations on UMAP. Interpretation: This probabilistic approach allows for understanding overlapping clusters and provides statistical confidence in gene expression group assignments. Sample: The UMAP plot (Figure 5) displays nine distinct clusters of gene expression profiles generated using Model-Based Clustering.	Advantages: ◆ Allows for overlapping clusters, capturing complex relationships in the data. ◆ Provides probabilities, adding statistical confidence to cluster assignments. Limitations: ◆ Computationally intensive, especially for large datasets, as in samples. ◆ Requires assumptions about the data distribution, which may not always be accurate.	[14]
5	Gaussian Mixture Models (GMM)	Gene expression data	Output: Mean and covariance of each Gaussian component in the mixture. Plot: Contour plot or ellipses representing cluster densities. Interpretation: GMM reveals the continuous distribution of gene expression clusters, allowing insight into clusters with potential overlap in biological function. Sample: The contour plot (Figure 6) shows the density distribution of gene expression clusters, where contour levels indicate regions of higher probability for gene expression profiles. In the UMAP (Figure 8) visualization, eight clusters are color-coded, revealing complex relationships in gene expression patterns with some degree of overlap. The classification plot (Figure 7) shows clusters with ellipses, indicating the covariance structure of each Gaussian component, highlighting both the central tendency and spread of each cluster. Together, these plots suggest that GMM clustering successfully identifies overlapping clusters, providing insights into genes with mixed expression patterns across biological functions.	Advantages: ◆ Allows for overlapping clusters, which is suitable for complex gene expression data. ◆ Captures the variance within clusters, providing insight into cluster shape and spread. ◆ Provides probabilistic assignment, adding statistical confidence to cluster membership. Limitations: ◆ Computationally intensive for large datasets due to complex calculations.	[15]

Continuation of table 1

6	Latent Dirichlet Allocation (LDA)	Gene expression data	Output: Topic distribution per gene or sample, with topics representing clusters. Plot: Topic distribution histograms or heatmaps. Interpretation: LDA treats clusters as "topics" of gene expression, revealing latent structures and allowing for thematic categorization of gene functions or pathways. Sample: The document-topic distribution histogram (Figure 10), showing the proportions of each topic (gene expression pattern) across genes. Peaks in blue, red, and green highlight dominant topics, suggesting prevalent expression themes. A term-topic heatmap (Figure 9), illustrates the relationship between specific gene terms and topics, with darker colors indicating higher relevance. This heatmap provides insight into the gene features strongly associated with each topic, useful for understanding gene functions. Together, these graphs demonstrate LDA's ability to capture the latent structure of gene expression patterns, offering a thematic categorization of gene functions and pathways.	Advantages: ◆ Effective for revealing hidden structures within complex data. ◆ Provides a thematic categorization, useful for understanding gene functions and pathways. Limitations: ◆ Requires pre-specifying the number of topics, which may not always match the true structure of the data. ◆ Computationally intensive, especially for large datasets.	[16]

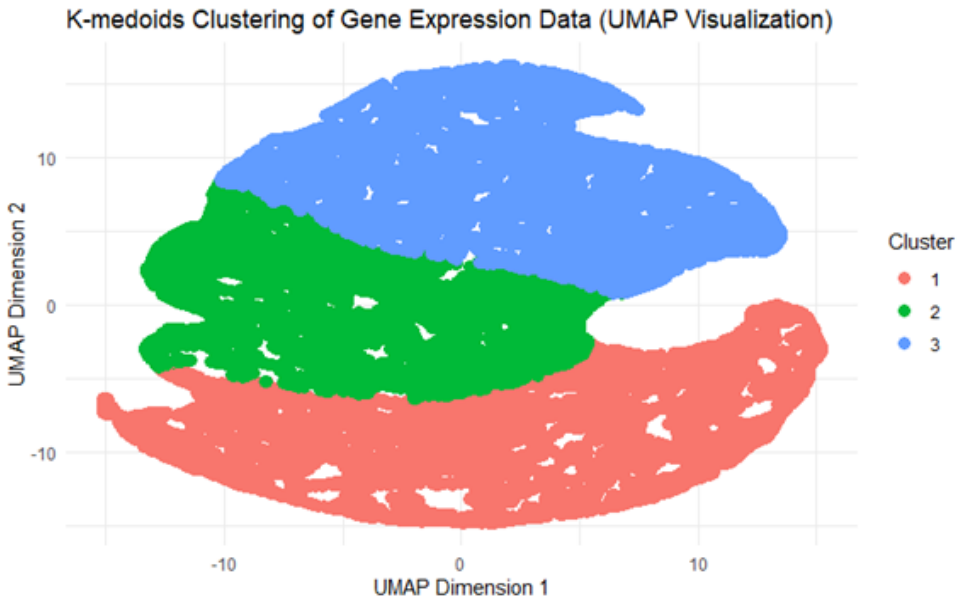

Figure 4 – UMAP. of K-medoids clustering of gene expression dataset

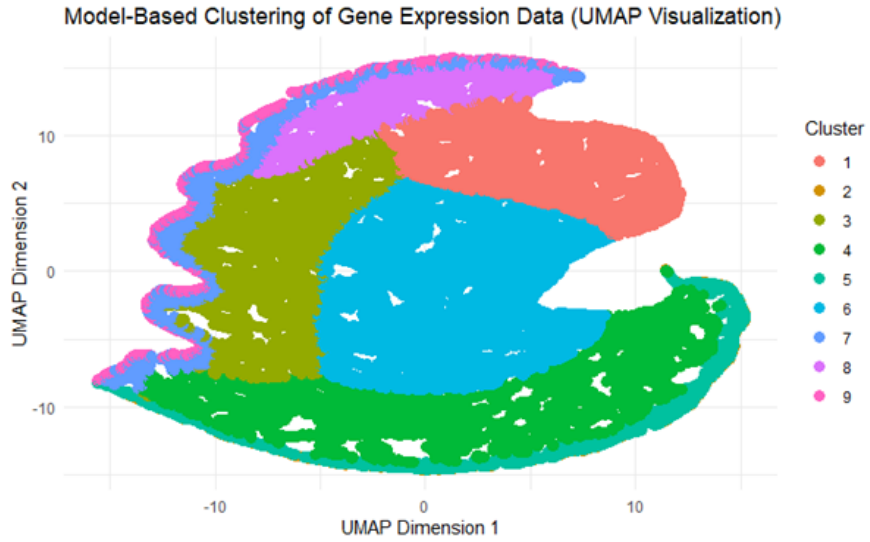

Figure 5 – UMAP. Model-Based Clustering Methods of gene expression dataset

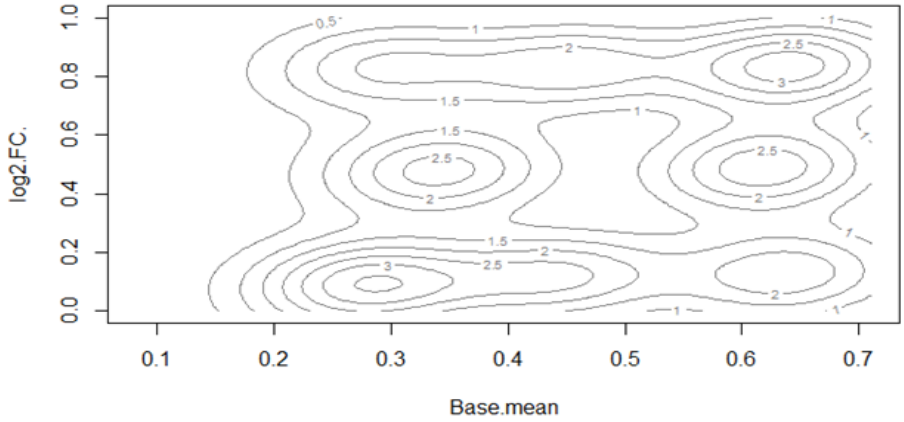

Figure 6 – Probability Density for each Gaussian component in the GMM.
Gaussian Mixture Models (GMM) of gene expression dataset

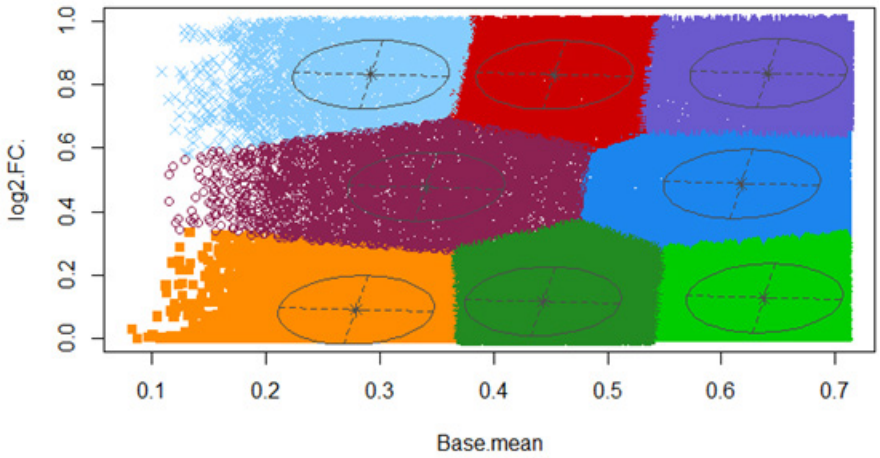

Figure 7 – Classification (Cluster Assignment) with Ellipses representing Gaussian Components.
Gaussian Mixture Models (GMM) of gene expression dataset

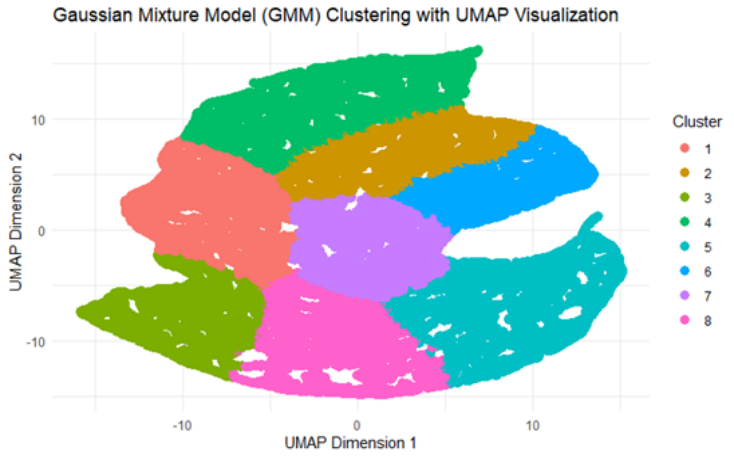

Figure 8 – UMAP on the clustering data for dimensionality reduction. Gaussian Mixture Models (GMM) of gene expression dataset

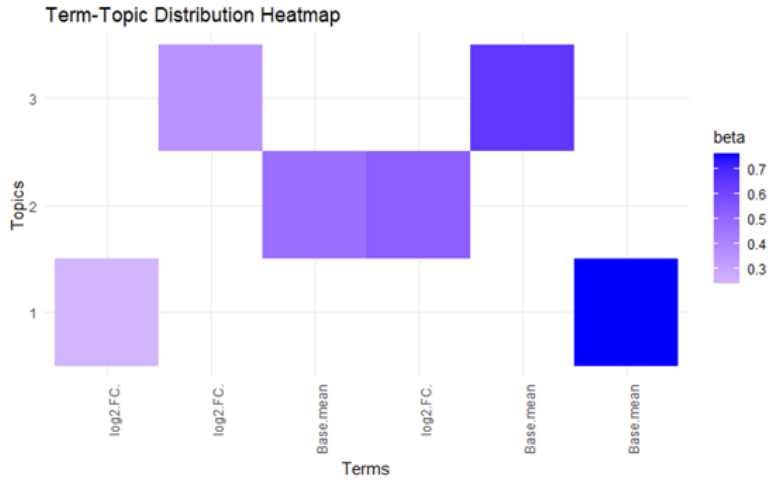

Figure 9 – Heatmap of the term-topic distribution for the top 10 terms in each topic. Latent Dirichlet Allocation (LDA) of gene expression dataset

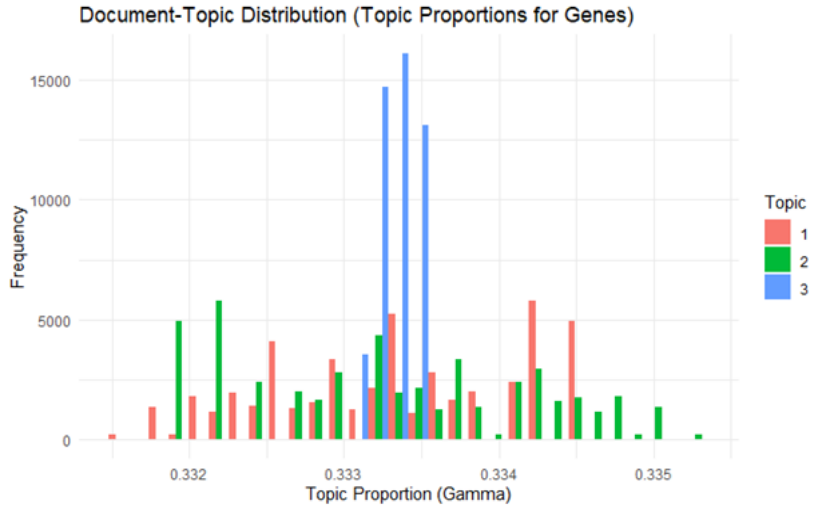

Figure 10 – Histogram Document-Topic Distribution (Topic Proportions for Genes). Latent Dirichlet Allocation (LDA) of gene expression dataset

Funding information. This research was funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. AP22686112 Study of somatic mutations from single-cell RNA data using machine learning methods in patients with peripheral artery disease).

Datasets available in NCBI (PRJNA736095; GEO: GSE176415): https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE176415

Sample code on R, available on github: https://github.com/dauletulyaidyn/scrna_clusterin_samples/upload

REFERENCES

1 Casadei R. et al. Identification of housekeeping genes suitable for gene expression analysis in the zebrafish. Gene Expression Patterns, 2011, vol. 11, no. 3–4, pp. 271–276.

2 Seo D., Ginsburg G.S., Goldschmidt-Clermont P.J. Gene Expression Analysis of Cardiovascular Diseases. J Am Coll Cardiol, 2006, vol. 48, no. 2, pp. 227–235.

3 Predicting drug response based on gene expression. Crit Rev Oncol Hematol, 2004, vol. 51, no. 3, pp. 205–227.

4 Huang X. et al. High Throughput Single Cell RNA Sequencing, Bioinformatics Analysis and Applications, 2018, pp. 33–43.

5 Perera M.A.I., Wijesinghe C.R., Weerasinghe A.R. Analysis of Expression Data Using Unsupervised Techniques. 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer). IEEE, 2020, pp. 119–124.

6 Li X., Wang C.-Y. From bulk, single-cell to spatial RNA sequencing. Int J Oral Sci, 2021, vol. 13, no. 1, p. 36.

7 Nathans J.F. et al. Genetic Tools for Cell Lineage Tracing and Profiling Developmental Trajectories in the Skin. Journal of Investigative Dermatology, 2024, vol. 144, no. 5, pp. 936–949.

8 Yao D.W. et al. Quantifying genetic effects on disease mediated by assayed gene expression levels. Nat Genet, 2020, vol. 52, no. 6, pp. 626–633.

9 Huang C.-T. et al. Perturbational Gene-Expression Signatures for Combinatorial Drug Discovery. iScience, 2019, vol. 15, pp. 291–306.

10 Qi R. et al. Clustering and classification methods for single-cell RNA-sequencing data. Brief Bioinform, 2020, vol. 21, no. 4, pp. 1196–1208.

11 Badsha Md.B. et al. Robust complementary hierarchical clustering for gene expression data analysis by β-divergence. J Biosci Bioeng, 2013, vol. 116, no. 3, pp. 397–407.

12 Chen L. et al. Deep soft K-means clustering with self-training for single-cell RNA sequence data. NAR Genom Bioinform, 2020, vol. 2, no. 2.

13 Li L. et al. Selecting Representative Samples From Complex Biological Datasets Using K-Medoids Clustering. Front Genet., 2022, vol. 13.

14 Gormley I.C., Murphy T.B., Raftery A.E. Model-Based Clustering. Annu Rev Stat Appl., 2023, vol. 10, no. 1, pp. 573–595.

15 Yu B. et al. scGMAI: a Gaussian mixture model for clustering single-cell RNA-Seq data based on deep autoencoder. Brief Bioinform, 2021, vol. 22, no.4.

16 Wu X., Wu H., Wu Z. Penalized Latent Dirichlet Allocation Model in Single-Cell RNA Sequencing. Stat Biosci., 2021, vol. 13, no. 3, pp. 543–562.

17 Arora S. et al. Variability in estimated gene expression among commonly used RNA-seq pipelines. Sci Rep., 2020, vol. 10, no.1, p. 2734.

18. Lataretu M., Hölzer M. RNAflow: An Effective and Simple RNA-Seq Differential Gene Expression Pipeline Using Nextflow. Genes (Basel), 2020, vol. 11, no. 12, p. 1487.

19. Rosati D. et al. Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A review. Comput Struct Biotechnol J., 2024, vol. 23, pp. 1154–1168.

20. Lo C.-C., Chain P.S.G. Rapid evaluation and quality control of next generation sequencing data with FaQCs. BMC Bioinformatics, 2014, vol. 15, no. 1, p. 366.

21 Bolger A.M., Lohse M., Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 2014, vol. 30, no. 15, pp. 2114–2120.
22 Sun K. Ktrim: an extra-fast and accurate adapter- and quality-trimmer for sequencing data. Bioinformatics, 2020, vol. 36, no. 11, pp. 3561–3562.
23 Dobin A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 2013, vol. 29, no. 1, pp. 15–21.
24 Kim D. et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol., 2019, vol. 37, no. 8, pp. 907–915.
25 Kim D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol., 2013, vol. 14, no.4, p. R36.
26 Anders S., Pyl P.T., Huber W. HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics, 2015, vol. 31, no. 2, pp. 166–169.
27 Kim T. et al. Impact of similarity metrics on single-cell RNA-seq data clustering, Brief Bioinform., 2019, vol. 20, no. 6, pp. 2316–2326.
28 Liu S. et al. Three Differential Expression Analysis Methods for RNA Sequencing: limma, EdgeR, DESeq2. Journal of Visualized Experiments, 2021, no. 175.
29 Abu-Jamous B., Kelly S. Clust: automatic extraction of optimal co-expressed gene clusters from gene expression data. Genome Biol., 2018, vol. 19, no. 1, p. 172.
30 Abueg L.A.L. et al. The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. Nucleic Acids Res., 2024, vol. 52, no. W1, pp. W83–W94.

[1]*Куникеев А.,
техника ғылымдарының магистрі, ORCID ID: 0000-0003-3716-0895,
*e-mail: a.kunikeyev@satbayev.university
[1,2]Еримбетова А.
PhD докторы, техникалық ғылымдар кандидаты, доцент,
ORCID ID: 0000-0002-2013-1513,
e-mail: aigerian8888@gmail.com
[1]Сатыбалдиева Р.,
техникалық ғылымдар кандидаты, профессор, ORCID ID: 0000-0002-0678-7583,
e-mail: r.satybaldiyeva@satbayev.university

[1]Сәтбаев университеті, Алматы қ., Қазақстан
[2]Қазақстан Республикасы Ғылым және жоғары білім министрлігі Ғылым комитетінің Ақпараттық және есептеуіш технологиялар институты, Алматы қ., Қазақстан

ГЕНЕТИКАЛЫҚ ДЕРЕКТЕРДІ ӨҢДЕУГЕ, АЛДЫН АЛА ӨҢДЕУ МЕН КЛАСТЕРЛІК ТАЛДАУҒА АРНАЛҒАН ҚҰРАЛДАРҒА, ӘДІСТЕМЕЛЕР МЕН ӘДІСТЕРГЕ ШОЛУ

Аңдатпа

Ген экспрессиясын талдау – жасушалардың әрекеттерін, ауру механизмдерін және дәрілік реакцияны түсінудің негізгі құрамдас бөлігі. Жоғары өнімді секвенирлеудің, әсіресе бір жасушалы РНҚ секвенирлеуінің (scRNA-seq) пайда болуы жасушалық гетерогенділікті бұрын-соңды болмаған деңгейге дейін зерттеу мүмкіндігін кеңейтті. Ұқсас экспрессиялық профильдері бар гендер немесе жасушаларды топтастыру үшін қолданылатын кластерлеу алгоритмдері осы технологиялар арқылы алынған үлкен деректер жиынын талдау барысында баға жетпес құралға айналды. Бұл мақалада гендік экспрессия деректерін талдауда, әсіресе бір жасушалы РНҚ секвенциясына негізделген зерттеулерде қолданылатын әртүрлі кластерлеу әдістері қарастырылды. Талдау иерархиялық кластерлеу мен k-means сияқты дәстүрлі әдістерді, сондай-ақ үлгіге негізделген кластерлеу, машиналық оқыту және терең оқыту тәсілдері сияқты жетілдірілген әдістерді қамтиды. Негізгі міндеттерге жоғары өлшемді деректерді өңдеу, шуды азайту және үлкен деректер жиынын тиімді масштабтау жатады. Сонымен қатар мульти-омикалық деректерді біріктіру, терең оқытуға

негізделген кластерлеу және федеративті оқыту сияқты жаңа жетістіктер гендік экспрессияны зерттеудегі кластерлеу қосымшаларының дәлдігі мен биологиялық маңыздылығын арттыруға мүмкіндік береді. Мақала кластерлеу алгоритмдерінің күрделі гендік экспрессия деректерін өңдеудегі болашақ бағыттарын талқылап, биологиялық түсініктерді жақсарту жолдарын ұсынады.

Тірек сөздер: кластерлеу әдістері, биоинформатика, машиналық оқыту, терең оқыту, бір жасушалы РНҚ секвенциясы, гендік экспрессиялар.

[1]*Куникеев А.,
магистр технических наук, ORCID ID: 0000-0003-3716-0895,
*e-mail: a.kunikeyev@satbayev.university
[1,2]Еримбетова А.
доктор Ph.D., канд. техн. наук, ассоц. профессор, ORCID ID: 0000-0002-2013-151,
e-mail: aigerian8888@gmail.com
[1]Сатыбалдиева Р.
канд. техн. наук, профессор, ORCID ID: 0000-0002-0678-758,
e-mail: r.satybaldiyeva@satbayev.university

[1]Сатбаев Университет, г. Алматы, Казахстан
[2]Институт информационных и вычислительных технологий Комитета науки Министерства науки и высшего образования Республики Казахстан, Алматы, Казахстан

ОБЗОР ИНСТРУМЕНТОВ, МЕТОДОЛОГИЙ И МЕТОДОВ ОБРАБОТКИ, ПРЕДВАРИТЕЛЬНОЙ ОБРАБОТКИ И КЛАСТЕРНОГО АНАЛИЗА ГЕНЕТИЧЕСКИХ ДАННЫХ

Аннотация

Анализ экспрессии генов стал ключевым компонентом в понимании поведения клеток, механизмов заболеваний и реакции на лекарства. Появление высокопроизводительного секвенирования, в частности секвенирования РНК отдельных клеток (scRNA-seq), расширило наши возможности изучения клеточной гетерогенности до беспрецедентного уровня. Алгоритмы кластеризации, необходимые для группировки генов или клеток со схожими профилями экспрессии, стали бесценными для анализа огромных наборов данных, генерируемых этими технологиями. В этой статье рассматриваются различные методы кластеризации, применяемые к данным об экспрессии генов, в частности секвенирования РНК отдельных клеток. Анализ охватывает традиционные методы, такие как иерархическая кластеризация и k-means, а также более продвинутые подходы, такие как кластеризация на основе моделей, методы на основе машинного обучения и глубокого обучения. Основные проблемы включают обработку многомерных данных, снижение шума и достижение масштабируемости для больших наборов данных. Более того, новые достижения, такие как интеграция данных мультиомики, кластеризация на основе глубокого обучения и федеративное обучение, предлагают потенциальные улучшения точности и биологической значимости для приложений кластеризации в исследовании экспрессии генов. Обзор завершается обсуждением будущих направлений развития алгоритмов кластеризации для обработки все более сложных данных об экспрессии генов для получения более точных биологических пониманий.

Ключевые слова: методы кластеризации, биоинформатика, машинное обучение, лубокое обучение, секвенирование РНК отдельных клеток, экспрессия генов.

Article submission date: 12.11.2024