ISSN 2616-5791 Vol.5, Issue 1: 21-25, DOI: 10.25236/AJMHS.2024.050104

Single-Cell Automated Annotation Algorithm Based on Reference Expression Profiles

Xiaoqian Huang[1,a], Ruiqi Liu[1,a], Yanmiao Huang[2,a], Xuexia Huang[3,b,*], Xiaozhou Chen[1,a,*]

[1]School of Mathematics and Computer Science, Yunnan Minzu University, Kunming, China
[2]School of Art and Design, Weifang Institute of Technology, Weifang, China
[3]Department of Clinical Pharmacy, Dongchangfu District Maternal and Child Health Hospital of Liaocheng City, Liaocheng, China
[a]h_xiaoq@163.com, [b]18763550646@163.com
[*]Corresponding author

Abstract: *In recent years, the rapid advancement of single-cell RNA sequencing (scRNA-seq) technology has provided a powerful tool for delving into the diversity of cellular populations, offering researchers a unique perspective to explore intracellular heterogeneity. This technology enables us to gain profound insights into the gene expression patterns of individual cells, unveiling latent heterogeneity within cell populations. Accurately predicting single-cell types is a crucial step in understanding the dynamics and functional impacts of cells. This paper aims to introduce the application of hypergeometric testing in gene set enrichment as a foundational tool for predicting single-cell types. In comparison to traditional gene expression analysis methods, scRNA-seq captures individual differences in each cell, presenting unprecedented opportunities for understanding development, diseases, and tissue functionality.*

Keywords: *Single-cell, Auto annotation, Hypergeometric Test, Reference-based*

1. Introduction

The introduction of single-cell RNA sequencing (scRNA-seq) technology has empowered researchers with a potent and precise tool, allowing for a more profound exploration of cellular diversity [1-3]. By accurately identifying and classifying single-cell types, we can delve into the intricacies of biological systems. The continual advancements in this technology have propelled us toward a deeper understanding of intracellular heterogeneity [4-6]. Accurate prediction of single-cell types serves as the foundation for uncovering latent information within scRNA-seq data. Through precise predictions of single-cell types, we can comprehensively identify and understand the existence of various cell subtypes within cell populations, offering indispensable insights for the study of cellular function, developmental processes, and disease progression [7-12].

In this paper, we extensively delve into the application of hypergeometric testing as a robust statistical tool [9, 13, 14]. By evaluating the enrichment of marker gene sets for cell types in the reference dataset and their correlation with single-cell clusters in the test dataset, we establish a reliable foundation for predicting single-cell types. Consequently, accurate predictions of single-cell types hold significant implications for unraveling mysteries at the cellular level. Through sustained efforts in this field, we aim to achieve a more comprehensive and in-depth understanding of the intricate interactions within and beyond cells, propelling continuous advancements in cell biology and medicine [15-18]. This endeavor not only contributes to a holistic comprehension of biological systems but also provides a solid scientific basis for the development of disease treatments and personalized medicine [19-21].

2. Materials and methods

2.1. Data Collection and Preprocessing

During the data collection and preprocessing phase, we extensively utilized the Gene Expression Omnibus (GEO) database [22] and Seurat objects, containing a wealth of single-cell transcriptomic data encompassing crucial information such as cell types and gene expression. To ensure a high degree of correlation and biological relevance within gene sets, we meticulously curated gene sets for each cell

ISSN 2616-5791 Vol.5, Issue 1: 21-25, DOI: 10.25236/AJMHS.2024.050104

type [23]. Through extensive iterative testing, the dataset was partitioned into reference and test datasets at a specific ratio, guided by known biological correlations [24]. This rigorous process was implemented to ensure that the gene sets employed faithfully reflect the distinctive features of each cell type [25, 26].

2.2. Hypergeometric Distribution Test

To gain in-depth insights into the gene enrichment patterns of each cell type, we opted for the hypergeometric test as our evaluation tool. In this test, where N represents the total number of genes and M represents the number of marker genes for a specific cell type, the probability (P) of observing i or more marker genes in a randomly selected set of n genes is calculated using the formula:

$$P = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

The hypergeometric test evaluates the statistical significance of gene enrichment for a specific cell type, considering whether this enrichment observed in scRNA-seq data is statistically significant. During the testing, we comprehensively considered the proportion of genes associated with the cell type across the entire dataset, along with the total number of genes related to that cell type. This meticulously designed testing process ensures an accurate and comprehensive assessment of gene enrichment levels.

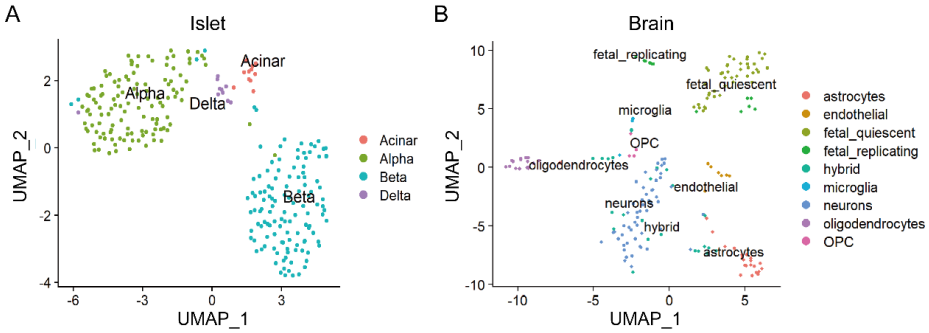
2.3. Statistical Significance

The significance results of the hypergeometric test are presented in the form of p-values, reflecting the probability of observing gene enrichment. Results with $P < 0.05$ are considered significant in the enrichment analysis, thereby further confirming the substantial likelihood of the presence of the identified cell type within the dataset. Through this analysis of statistical significance, we can reliably assess the significant enrichment of cell types, providing robust statistical support for subsequent cell type predictions. This rigorous methodology ensures a profound understanding of cell type features and establishes a credible data foundation for research endeavors.

3. Results

3.1. Cell Type Based Clustering in the Reference Dataset

Through meticulous clustering analysis of the reference dataset annotated with cell type information, we ensured the accurate delineation of cell populations. Each cluster represents a specific cell type, and the annotation of cell types represented by each cluster is illustrated (Figures 1.A, B), constructing a clear and organized foundation of cell types. In this clustering process, our emphasis lies not only in the simple grouping of cells but also in ensuring that each group consistently and significantly represents a specific cell type in terms of gene expression patterns. By adopting this approach, we were able to capture potential cell subtypes and subtle changes in cell states, rendering the final clustering structure more biologically meaningful.



A. Cell type distribution in Islet tissue. B. Cell type distribution in Brain tissue.

Figure 1: Distribution of different cell types in reference datasets.

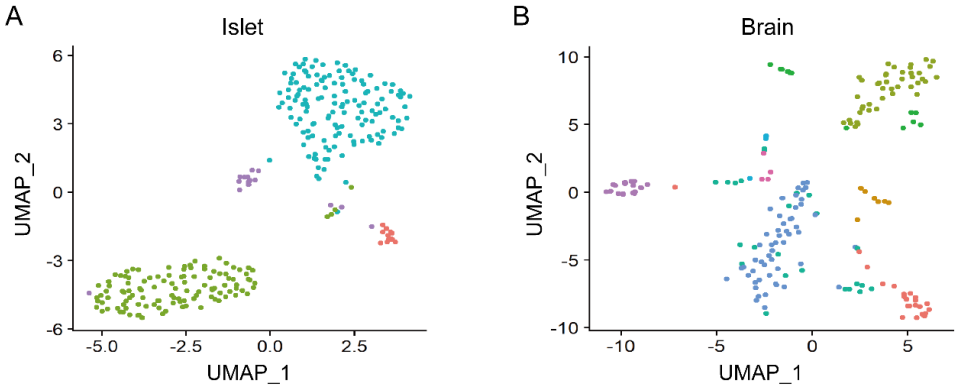
3.2. Clustering of Test Dataset

Subsequently, we performed clustering on the test dataset, leveraging the cell type foundation

ISSN 2616-5791 Vol.5, Issue 1: 21-25, DOI: 10.25236/AJMHS.2024.050104

previously established. By precisely assigning each cell in the test dataset to pre-clustered cell types, we were able to reaffirm the cell type populations established earlier in the test samples, ensuring the consistency and reproducibility of the entire analysis process (Figures 2. A, B).

The purpose of this step is to validate the applicability of our established cell type classification system on a new test dataset and confirm its generalizability across different datasets. Through precise cell clustering, we can verify the robustness of the model, enabling a more credible application of previously obtained cell type information to new research instances. This ensures that our cell type classification results are widely applicable and reliable.



A. Cellular distribution in islet tissue. B. Cellular distribution in brain tissue.

Figure 2: Distribution of different annotated cell clusters in the test dataset.

3.3. Identification of Marker Genes

Following the confirmation of the basic distribution of cell types and the distribution of different annotated cell clusters, we employed gene expression analysis methods within the Seurat framework to identify marker genes for each cell type (Figure 3). This involved utilizing advanced techniques such as differential expression analysis to ensure that the identified marker genes accurately and representatively reflected the specific characteristics of each cell type. Through this crucial step, we gained a deeper insight into the uniqueness of each cell type at the gene expression level.

By employing sophisticated techniques such as differential expression analysis, we were able to discern genes significantly expressed in specific cell types. These genes not only quantitatively reflected the characteristics of the cell types but also possessed representativeness, providing us with a more comprehensive understanding of the biological properties of cell types. This in-depth exploration of gene expression data enriches the information available for the accurate classification and understanding of cell types, laying a solid foundation for subsequent cell type predictions and biological interpretations.

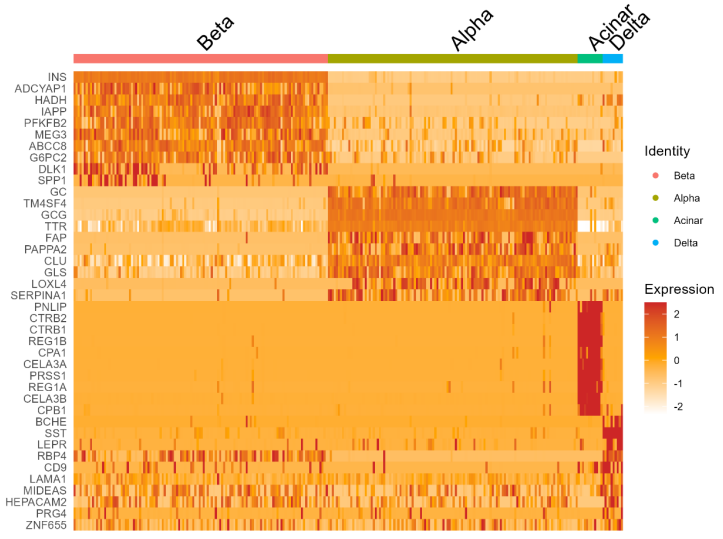


Figure 3: Expression levels of marker genes in different cell types.

ISSN 2616-5791 Vol.5, Issue 1: 21-25, DOI: 10.25236/AJMHS.2024.050104

3.4. Further Evaluation of the Enrichment Levels of Marker Genes in Cell Subpopulations using Hypergeometric Test

To comprehensively assess the enrichment of marker genes in each cell subpopulation, we employed the hypergeometric test as a precise evaluation tool. Through a thorough analysis of the enrichment levels of marker genes in various cell subpopulations, we were able to accurately quantify the close association of these genes with specific cell types. This led to precise annotations of cell clusters in the annotated dataset (Figure 4).

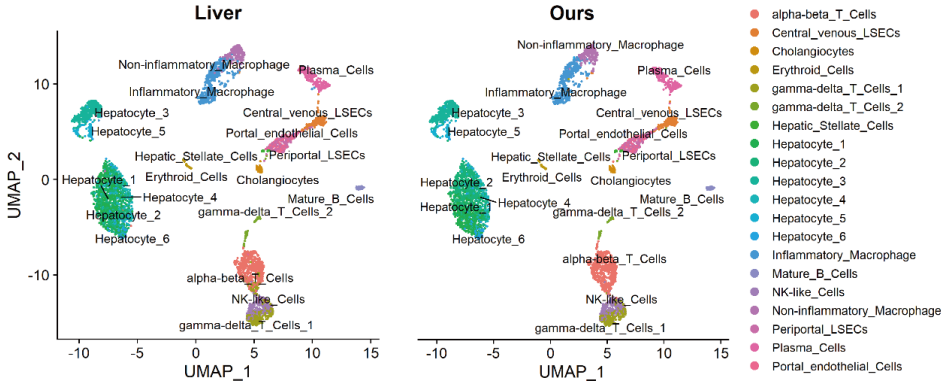


Figure 4: Distribution of raw cell types and hypergeometric test-predicted cell types in the annotated dataset of human liver tissue.

The application of hypergeometric tests enabled precise quantification of enrichment patterns for annotated genes within cell subpopulations. This accurate quantification contributes robust data support for further analysis of cell types. The reliability and accuracy of this step not only strengthen our confidence in the correlation of cell types but also ensure that the resulting classification of cell types possesses clear biological significance. This systematic evaluation approach will lay a solid foundation for subsequent cell type predictions and biological investigations, allowing for a more comprehensive understanding of the biological information within single-cell transcriptomic data.

4. Conclusions

Our study applies rigorous data collection and processing, utilizes hypergeometric testing, and employs precise assessment of statistical significance to apply hypergeometric distribution testing for predicting cell types in scRNA-seq data. The results reveal statistically significant enrichments of multiple cell types (Table 1).

Table 1: Predictive performance evaluation.

Dataset	Number of cell types	Forecast accuracy	p-value
Islet	4	97.12%	1.81e-4
Brain	9	94.60%	2.97e-3
Liver	20	93.91%	1.67e-2

Through the organic integration of the above steps, we have ensured a high-quality analysis of single-cell transcriptomic data, constructing a precise and comprehensive analytical framework. The superiority of this integrated framework is manifested in its ability to ensure the accurate identification and classification of cell types in the test dataset, providing a solid foundation for in-depth exploration of the differences and biological significance between cell types.

By employing a systematic approach, we have been able to establish a reliable classification of cell types, thereby ensuring the intrinsic consistency and reproducibility of our research. The rigor of this methodology contributes to uncovering hidden patterns and associations within cell populations, providing a robust scientific basis for cell type prediction and biological interpretation. Furthermore, it lays a reliable methodological foundation for future exploratory studies and investigations into the biological mechanisms associated with cell types.

ISSN 2616-5791 Vol.5, Issue 1: 21-25, DOI: 10.25236/AJMHS.2024.050104

Acknowledgements

We thank all the authors involved in this study for data collection, preparation, quality control and manuscript writing.

References

[1] Slovin S, Carissimo A, Panariello F, et al. Single-Cell RNA Sequencing Analysis: A Step-by-Step Overview [J]. Methods Mol Biol. 2021;2284:343-365.
[2] Balzer MS, Ma Z, Zhou J, et al. How to Get Started with Single Cell RNA Sequencing Data Analysis [J]. J Am Soc Nephrol. 2021;32(6):1279-1292.
[3] Rossin EJ, Sobrin L, Kim LA. Single-cell RNA sequencing: An overview for the ophthalmologist[J]. Semin Ophthalmol. 2021;36(4):191-197.
[4] Hickey JW, Becker WR, Nevins SA, et al. Organization of the human intestine at single-cell resolution [J]. Nature. 2023;619(7970):572-584.
[5] Kim D, Chung KB, Kim TG. Application of single-cell RNA sequencing on human skin: Technical evolution and challenges[J]. J Dermatol Sci. 2020;99(2):74-81.
[6] Kolodziejczyk AA, Kim JK, Svensson V, et al. The technology and biology of single-cell RNA sequencing [J]. Mol Cell. 2015;58(4):610-620.
[7] Erfanian N, Heydari AA, Feriz AM, et al. Deep learning applications in single-cell genomics and transcriptomics data analysis[J]. Biomed Pharmacother. 2023;165:115077.
[8] Fan J, Slowikowski K, Zhang F. Single-cell transcriptomics in cancer: computational challenges and opportunities [J]. Exp Mol Med. 2020;52(9):1452-1465.
[9] Ke M, Elshenawy B, Sheldon H, et al. Single cell RNA-sequencing: A powerful yet still challenging technology to study cellular heterogeneity[J]. Bioessays. 2022;44(11):e2200084.
[10] Sullivan KE, Kendrick RM, Cembrowski MS. Elucidating memory in the brain via single-cell transcriptomics[J]. J Neurochem. 2021;157(4):982-992.
[11] Tumminello M, Bertolazzi G, Sottile G, et al. A multivariate statistical test for differential expression analysis [J]. Sci Rep. 2022;12(1):8265.
[12] Ziegenhain C, Vieth B, Parekh S, et al. Comparative Analysis of Single-Cell RNA Sequencing Methods [J]. Mol Cell. 2017;65(4):631-643 e4.
[13] Johannssen A, Chukhrova N, Castagliola P. Efficient algorithms for calculating the probability distribution of the sum of hypergeometric-distributed random variables[J]. MethodsX. 2021;8:101507.
[14] Khozyainova AA, Valyaeva AA, Arbatsky MS, et al. Complex Analysis of Single-Cell RNA Sequencing Data[J]. Biochemistry (Mosc). 2023;88(2):231-252.
[15] Aldridge S, Teichmann SA. Single cell transcriptomics comes of age[J]. Nat Commun. 2020; 11(1):4307.
[16] Bacher R. Normalization for Single-Cell RNA-Seq Data Analysis[J]. Methods Mol Biol. 2019; 1935:11-23.
[17] Bod L, Kye YC, Shi J, et al. B-cell-specific checkpoint molecules that regulate anti-tumour immunity [J]. Nature. 2023;619(7969):348-356.
[18] Brendel M, Su C, Bai Z, et al. Application of Deep Learning on Single-cell RNA Sequencing Data Analysis: A Review[J]. Genomics Proteomics Bioinformatics. 2022;20(5):814-835.
[19] Clarke ZA, Andrews TS, Atif J, et al. Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods[J]. Nat Protoc. 2021;16(6):2749-2764.
[20] Denyer T, Timmermans MCP. Crafting a blueprint for single-cell RNA sequencing[J]. Trends Plant Sci. 2022;27(1):92-103.
[21] Huang X, Liu S, Wu L, et al. High Throughput Single Cell RNA Sequencing, Bioinformatics Analysis and Applications [J]. Adv Exp Med Biol. 2018;1068:33-43.
[22] Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets--update [J]. Nucleic Acids Res. 2013;41(Database issue):D991-995.
[23] Stuart T, Butler A, Hoffman P, et al. Comprehensive Integration of Single-Cell Data[J]. Cell. 2019; 177(7):1888-1902 e21.
[24] Stuart T, Satija R. Integrative single-cell analysis[J]. Nat Rev Genet. 2019;20(5):257-272.
[25] Saura CA, Deprada A, Capilla-Lopez MD, et al. Revealing cell vulnerability in Alzheimer's disease by single-cell transcriptomics[J]. Semin Cell Dev Biol. 2023;139:73-83.
[26] Piovani L, Marletaz F. Single-cell transcriptomics refuels the exploration of spiralian biology[J]. Brief Funct Genomics. 2023;22(6):517-524.