Reengineering the NDRindex: A Pythonic Approach to Single-cell RNA-Seq Preprocessing Quality Assessment

Lennert Saerens
Vrije Universiteit Brussel
Brussel, Belgium
lennert.saerens@vub.be

Abstract

This research presents a comprehensive reproduction[1] of the NDRindex, a method to evaluate the outcomes of normalization and dimensionality reduction methods for single-cell RNA-Seq data, using Python, emphasizing structured and maintainable code. Leveraging the principles of computer science, the study not only replicates the original methodology but also offers a more accessible and modular approach, facilitating future adaptations and improvements. The results affirm the accuracy and effectiveness of the Pythonic implementation, underscoring the synergy between bioinformatics and structured software engineering.

Introduction

Bioinformatics, an interdisciplinary field bridging biology and computational science, has emerged as a pivotal domain in modern biological research. By leveraging computational tools and algorithms, bioinformatics deciphers complex biological data, paving the way for groundbreaking discoveries in medicine, genetics, and evolutionary biology.

Single-cell RNA sequencing (scRNA-seq) stands at the forefront of these advancements. Unlike traditional RNA sequencing, which averages gene expression across thousands of cells, scRNA-seq profiles individual cells, offering a granular view of cellular heterogeneity. This granularity unveils the intricacy of cell types and states within tissues, providing unparalleled insights into developmental processes and disease mechanisms such as with COVID-19 (Zou et al. 2020; wu Pan et al. 2020; Lin et al. 2020). The advent of scRNA-seq has revolutionized various biological disciplines. From tracing the lineage of cells during organismal development to unraveling the cellular landscape of tumors (Liu et al. 2023), scRNA-seq has become an indispensable tool in contemporary research.

However, the richness of scRNA-seq data brings forth analytical challenges. The data is often sparse, high-dimensional, and noisy, requiring rigorous preprocessing steps to extract meaningful insights (Feng 2021). Proper preprocessing ensures the accuracy of downstream analyses, from differential expression to pathway enrichment.

Central to scRNA-seq data analysis is the concept of clustering—grouping cells based on their gene expression profiles (Wang, Li, and Nabavi 2021). Clustering identifies subpopulations of cells, revealing novel cell types or states (Kiselev et al. 2016). Yet, the task is non-trivial. The high dimensionality of scRNA-seq data, coupled with technical noise, demands robust algorithms and quality assessment metrics to ensure accurate clustering outcomes.

As the field of bioinformatics continues to grapple with the challenges of scRNA-seq data, the importance of effective preprocessing cannot be overstated. Preprocessing not only refines the raw data but also ensures that subsequent analyses, such as clustering, are based on accurate and meaningful information (Xiao, Lu, and Jin 2019). However, with a plethora of preprocessing methods available, the challenge lies in determining which method or combination of methods is optimal for a given dataset. In this context, the *NDRindex* (Normalization and Dimensionality Reduction index) emerges (Xiao, Lu, and Jin 2019). Introduced as a novel method to evaluate the outcomes of normalization and dimensionality reduction methods, the NDRindex provides a quantitative measure of data quality after preprocessing. By calculating the degree of data aggregation, the NDRindex offers insights into the quality of data before it undergoes clustering, ensuring that researchers can make informed decisions about their preprocessing paths.

While the original NDRindex has proven its efficacy in various studies (Qi et al. 2023; Wang, Li, and Nabavi 2021; Mallick et al. 2023), the dynamic nature of bioinformatics and the ever-evolving landscape of scRNA-seq data necessitate continuous validation and enhancement of such methods. Reproducing foundational research, such as the NDRindex, not only reaffirms its validity but also provides opportunities for refinement and adaptation to newer challenges and datasets.

This paper presents a comprehensive reproduction of the NDRindex using Python, with an emphasis on structured and maintainable code. By leveraging the principles of computer science, we aim to not only replicate the original methodology but also offer a more modular and accessible approach. Through this reproduction, we hope to bridge the gap between bioinformatics and structured software engineering, providing the community with a robust and adaptable tool for scRNA-seq preprocessing quality assessment.

[1] The complete implementation, along with additional scripts and utilities, can be found on GitHub at https://github.com/LennertSaerens/Assignment-NDRindex.

Background

Single-cell RNA sequencing (scRNA-seq) data is inherently complex and high-dimensional. Unlike bulk RNA sequencing, which provides an average gene expression profile across a multitude of cells, scRNA-seq offers a granular view of individual cells. This granularity results in data that is predominantly sparse. In fact, a significant portion of scRNA-seq datasets, often approaching 90%, consists of zero measurements (Xiao, Lu, and Jin 2019). These zero measurements, or dropouts, pose a challenge as they can either represent a genuine lack of expression or arise from technical limitations of the sequencing process.

Given the sparsity and complexity of scRNA-seq data, preprocessing becomes an essential step in the analysis pipeline. Preprocessing encompasses a range of operations, including normalization, dimensionality reduction, and noise filtering. The choice of preprocessing methods is pivotal, as different normalization and size reduction techniques can profoundly influence the outcomes of downstream analyses, such as clustering and cell type enrichment (Xiao, Lu, and Jin 2019; Germain, Sonrel, and Robinson 2020).

Normalization methods aim to correct for technical variations and ensure that the gene expression profiles are comparable across cells. On the other hand, dimensionality reduction techniques transform the high-dimensional scRNA-seq data into a lower-dimensional space, making it more tractable for analysis (Melsted et al. 2021). The reduced dimensionality retains the most significant features of the data, enabling more accurate clustering and interpretation. However, the choice of dimensionality reduction method is crucial. An effective method can extract vital information from the intricate raw data, leading to more accurate clustering results and better biological interpretations (Xiao, Lu, and Jin 2019; Imoto et al. 2022; Danda, Vasighizaker, and Rueda 2020).

In essence, the quality assessment of scRNA-seq data preprocessing is not just a technical necessity but a fundamental step to ensure the biological validity of the analyses. The choices made during preprocessing can significantly impact the insights derived from the data, emphasizing the need for robust and effective preprocessing methodologies.

The NDRindex, introduced as a method to evaluate the outcomes of normalization and dimensionality reduction in scRNA-seq data preprocessing, marked a significant step forward in ensuring the quality of processed data. By quantitatively assessing data quality after preprocessing, the NDRindex provided researchers with a tool to make informed decisions about their preprocessing paths, ensuring that subsequent analyses were based on accurate and meaningful information.

The NDRindex has been foundational in assessing the quality of scRNA-seq data preprocessing. However, the dynamic nature of bioinformatics has led to the emergence of various clustering-based methods. A benchmark study conducted in 2019 provided an extensive evaluation of several clustering methods, considering different modes of usage and parameter settings (Krzak et al. 2019). This study applied the methods to both real and simulated datasets, varying in dimensionality, cell populations, and noise levels. Such comparative analyses have been instrumental in understanding the strengths and limitations of various methods in the realm of scRNA-seq data analysis.

Beyond clustering, the realm of scRNA-seq data analysis has seen the development of algorithms focusing on trajectory datasets. One such method, introduced in 2019, ranks the relevance of genes for comparing trajectory datasets (Wang et al. 2019). This algorithm demonstrated its effectiveness on various datasets. While not directly related to the NDRindex, such algorithms underscore the broader landscape of tools and methodologies developed to address the challenges of scRNA-seq data analysis.

The NDRindex, while pivotal, is part of a broader ecosystem of algorithms and methods tailored for scRNA-seq data preprocessing and analysis. The continuous evolution of these tools, driven by the ever-increasing complexity of biological data, underscores the importance of periodic validation, reproduction, and enhancement.

Reproduction

Methods

The NDRindex, standing for *Normalization and Dimensionality Reduction index*, is a specialized method designed to assess the quality of outcomes from normalization and dimensionality reduction techniques applied to single-cell RNA sequencing data (Xiao, Lu, and Jin 2019). The primary objective is to ensure that the preprocessing paths chosen for scRNA-Seq data mining are optimal, extracting the most relevant information from the complex raw data and leading to accurate clustering results (Xiao, Lu, and Jin 2019).

The NDRindex requires a gene expression matrix and can be used to compapare various normalization methods such as TMM (Robinson and Oshlack 2010), Linnorm (Yip et al. 2017), and Seurat (Satija et al. 2015) (Xiao, Lu, and Jin 2019). It also can be used to compare various dimension reduction methods like PCA, tSNE (van der Maaten and Hinton 2008), and sammon (Sammon 1969). These methods transform the high-dimensional data into a more manageable form, making it suitable for further analysis (Xiao, Lu, and Jin 2019).

The essence of the NDRindex method is an algorithm designed to assess data quality. Recognizing that not all data is apt for clustering, the NDRindex evaluates the cluster tendency by calculating the *aggregation degree of data* (Xiao, Lu, and Jin 2019). The higher the degree of clustering, the more points are distributed in a relatively small area, indicating the existence of natural clusters.

The NDRindex algorithm as proposed by the authors consists of several steps, the first one being the calculation of the distance matrix and the *average scale* of the data. Here, the average scale is defined as:

$$M \times \frac{1}{\log_{10} n} \tag{1}$$

where M is the lower quartile distance of all point pairs, and n is the number of samples.

The second step consists of clustering to find the point gathering areas. This involves selecting a point randomly, finding the closest point to the geometric center of the cluster, and determining if it should be added to the existing cluster or form a new cluster based on the previously calculated average scale.

The third and final step is calculating the final index. For each cluster, the average of the distances from all points to the geometric center is defined as the cluster radius. The final index is then defined as:

$$1.0 - \frac{R}{\frac{M}{\log_{10} n}} \tag{2}$$

where R is defined as:

$$\frac{\sum_{i \in set\ of\ all\ clusters} \frac{\sum_{p \in i} distance(p, geometric\ center\ of\ i)}{size\ of\ i}}{K} \tag{3}$$

and K is the total number of clusters (Xiao, Lu, and Jin 2019).

The presented implementation in Python aligns with this methodology. We have encapsulated the NDRindex algorithm within a class, providing methods to calculate the distance matrix, average scale, perform clustering, and evaluate data quality. The use of libraries such as `numpy` and `scipy` ensures efficient mathematical computations, while the integration with the `rpy2` library allows for seamless interaction with R packages, leveraging their capabilities in the bioinformatics domain.

Data

The experiments conducted in this study utilized four distinct datasets, each sourced from different research papers and imported using various R packages. These datasets were chosen due to their utilization in the original paper's experiments, ensuring consistency and enabling a direct comparison of results.

Yan Dataset This dataset was derived from the paper titled *Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells* and was imported into the code using the `RCSL` R package (Yan et al. 2013). The dataset comprises a matrix of dimensions 20214×90, where columns represent cells and rows signify gene expression values. The Yan dataset focuses on measuring gene expression in individual cells, crucial for understanding the gene regulatory network controlling human embryonic development. The dataset provides a comprehensive framework of the transcriptome landscapes of human early embryos and embryonic stem cells (hESCs) (Yan et al. 2013).

Biase dataset The Biase dataset originates from the paper *Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing* and was imported using the `SparseMDC` R package (Biase, Cao, and Zhong 2014). The dataset dimensions are 49×16514, representing cells and gene expression values, respectively. The Biase dataset delves into the question of when and how the first cell fate decision is made in mammals. It provides insights into the reproducible inter-blastomere differences among mouse embryos and the discovery of point gathering areas (Biase, Cao, and Zhong 2014).

Deng Dataset Sourced from the paper by Deng et al. titled *Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells* this dataset was imported using the `scDatasets` R package (Deng et al. 2014). The dataset has dimensions of 18884×256. The Deng dataset offers a comprehensive analysis of single-cell RNA-seq of mouse embryonic development, from the zygote to the late blastocyst stage. It presents insights into the independent and stochastic allelic transcription that generates random monoallelic expression in mammalian cells (Deng et al. 2014).

Usoskin Dataset This dataset is based on the paper by Usoskin et al. titled *Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing* and was imported via the `scDatasets` R package (Usoskin et al. 2015). The dataset comprises a matrix of dimensions 19252×622. The Usoskin dataset focuses on the RNA-Seq of single cells from the mouse lumbar dorsal root ganglion. It provides a detailed classification of sensory neuron types, revealing the cellular complexity underlying somatic sensation (Usoskin et al. 2015).

Results

The process of scientific validation often hinges on the reproducibility of experimental results. In this section, we delve into a detailed exploration of the experiments conducted in the original NDRindex paper and our attempts to reproduce them. Three primary experiments were undertaken by the original authors, each designed to test and validate the effectiveness and applicability of the NDRindex algorithm in various scenarios. Our objective in this reproduction study is not only to verify the claims made by the original authors but also to provide insights into any discrepancies, challenges, and observations that arose during our replication process. For each experiment, we present the original setup and findings, describe our approach to reproduction, and conclude with a comparative analysis of the results. This systematic approach ensures a comprehensive understanding of the NDRindex's capabilities and potential areas of improvement.

Experiment 1: Evaluating NDRindex on Simulated Datasets The primary objective of the first experiment was to validate the efficacy of the NDRindex in distinguishing datasets based on their suitability for clustering. The authors generated four types of datasets: two-dimensional normal distribution, square, hexagram, and a random shape (Xiao, Lu, and Jin 2019). Each dataset type had three variations, with each subsequent variation having more defined clusters. The results, as depicted in Fig. 3 of the original paper, showed a consistent increase in the NDRindex value as the clusters became more defined, validating the algorithm's ability to recognize and reward more distinguishable clustering patterns.

In the reproduction, due to the absence of the original code, the datasets were recreated in Python through trial and error. The two-dimensional normal distribution and square datasets were successfully replicated. The hexagram dataset was modified to consist of four primary clusters, each containing five sub-clusters arranged in a pentagram shape. This introduced both local and global structures to the dataset. The random shape dataset from the original experiment was replaced with a circular dataset.

A significant challenge faced during the reproduction was the lack of provided code for generating the original datasets. This made the recreation process more challenging and highlighted the importance of transparency and accessibility in academic research.

The results of the reproduced experiments can be seen in Figure 1. The results for the normal distribution and square datasets closely mirrored the findings of the original paper, with the NDRindex value increasing as the data became more aggregate. However, an anomaly was observed in the hexagram dataset, where the third variation had a significantly lower NDRindex value. This could be attributed to the presence of both local and global structures in the dataset, suggesting that the NDRindex might be sensitive to datasets with multiple levels of clustering patterns.

The authors of the original paper concluded that the NDRindex could clearly distinguish between datasets of varying quality. This assertion was mostly supported by the reproduction, with the exception of the hexagram dataset, which provided an interesting insight into the behavior of the NDRindex in the presence of complex clustering structures.

Experiment 2: Validation of Preprocessing Path Selection by NDRindex The second experiment aimed to validate the efficacy of the NDRindex algorithm in selecting the optimal preprocessing path for scRNA-Seq datasets. The authors utilized four real-life scRNA-Seq datasets: Yan, Biase, Deng, and Usoskin. By applying the NDRindex algorithm, they identified the preprocessing methods that the algorithm deemed optimal for each dataset. Subsequently, they processed each dataset using all possible combinations of preprocessing methods. Each preprocessed dataset was then subjected to four clustering algorithms, and the Adjusted Rand Index (ARI) was computed for each resulting clustering. The ARIs served as a metric to evaluate the quality of the clusterings, with the hypothesis that the preprocessing methods chosen by the NDRindex should yield the highest or near-highest ARI values.

The results from the original paper indicated that the preprocessing methods selected by the NDRindex often achieved the maximum possible ARI across all combinations. Even when they didn't, the selected methods consistently produced ARIs above the average and upper quartile values, validating the algorithm's efficacy in selecting optimal preprocessing paths.

In our reproduction, we closely followed the experimental setup of the original authors. However, due to the unavailability of the Adpclust clustering algorithm, we incorporated the HDBSCAN (Malzer and Baum 2020) clustering algorithm as a substitute. This inclusion served as an additional validation step, ensuring the robustness of the NDRindex algorithm across different clustering methods.

The datasets used in our reproduction were the same as those in the original study: Yan, Biase, Deng, and Usoskin. Each dataset's specifics, including their origins, sizes, and descriptions, are detailed in the Data section of this report.

One significant challenge faced during the reproduction was the absence of direct links to the datasets and clustering algorithms used by the original authors. This omission necessitated additional efforts to locate and utilize the appropriate resources.

Our reproduction yielded results that were strikingly similar to those of the original paper. These can be seen visualized in Figure 2. The preprocessing methods chosen by the NDRindex consistently produced high ARI values, often matching or closely approximating the maximum ARIs obtained across all combinations. This trend was observed even with the HDBSCAN clustering algorithm, which was not part of the original study, further attesting to the NDRindex's robustness.

In conclusion, our reproduction of the second experiment reaffirms the NDRindex algorithm's capability to select optimal preprocessing paths for scRNA-Seq datasets, leading to high-quality clustering results.

Experiment 3: Comparing NDRindex with Other Preprocessing Methods The third experiment in the original paper aimed to benchmark the performance of the NDRindex algorithm against four other prominent preprocessing methods: *SC3* (Kiselev et al. 2017), *pcaReduce* (Žurauskienė and Yau 2016), *SNN-Cliq* (Xu and Su 2015), and *SEURAT* (Satija et al. 2015). The primary objective was to ascertain the relative accuracy and stability of the NDRindex in comparison to these methods. The experimental setup involved preprocessing the four datasets (Yan, Biase, Deng, Usoskin) using each method, followed by clustering using the hierarchical clustering algorithm. The performance of each method was then evaluated using the Adjusted Rand Index (ARI). The results from the original study indicated that the NDRindex algorithm consistently demonstrated high accuracy and stability when juxtaposed with the other methods.

In replicating the experiment, the same datasets and methodologies were employed. The datasets used were Yan, Biase, Deng, and Usoskin, as detailed in the Data section. Each dataset was preprocessed using the NDRindex and the other four methods, followed by hierarchical clustering. The performance was then assessed using the ARI. A notable modification in the reproduction was the visualization of results. Instead of plotting each of the 100 ARI measurements, as done in the original paper, only the average ARI score was plotted for clarity. This change in visualization did not impact the overall findings, which can be seen visualized in Figure 3.

The results from the reproduction closely mirrored those from the original paper. The NDRindex algorithm consistently showcased its high accuracy and stability when compared against SC3, pcaReduce, SNN-Cliq, and SEU-

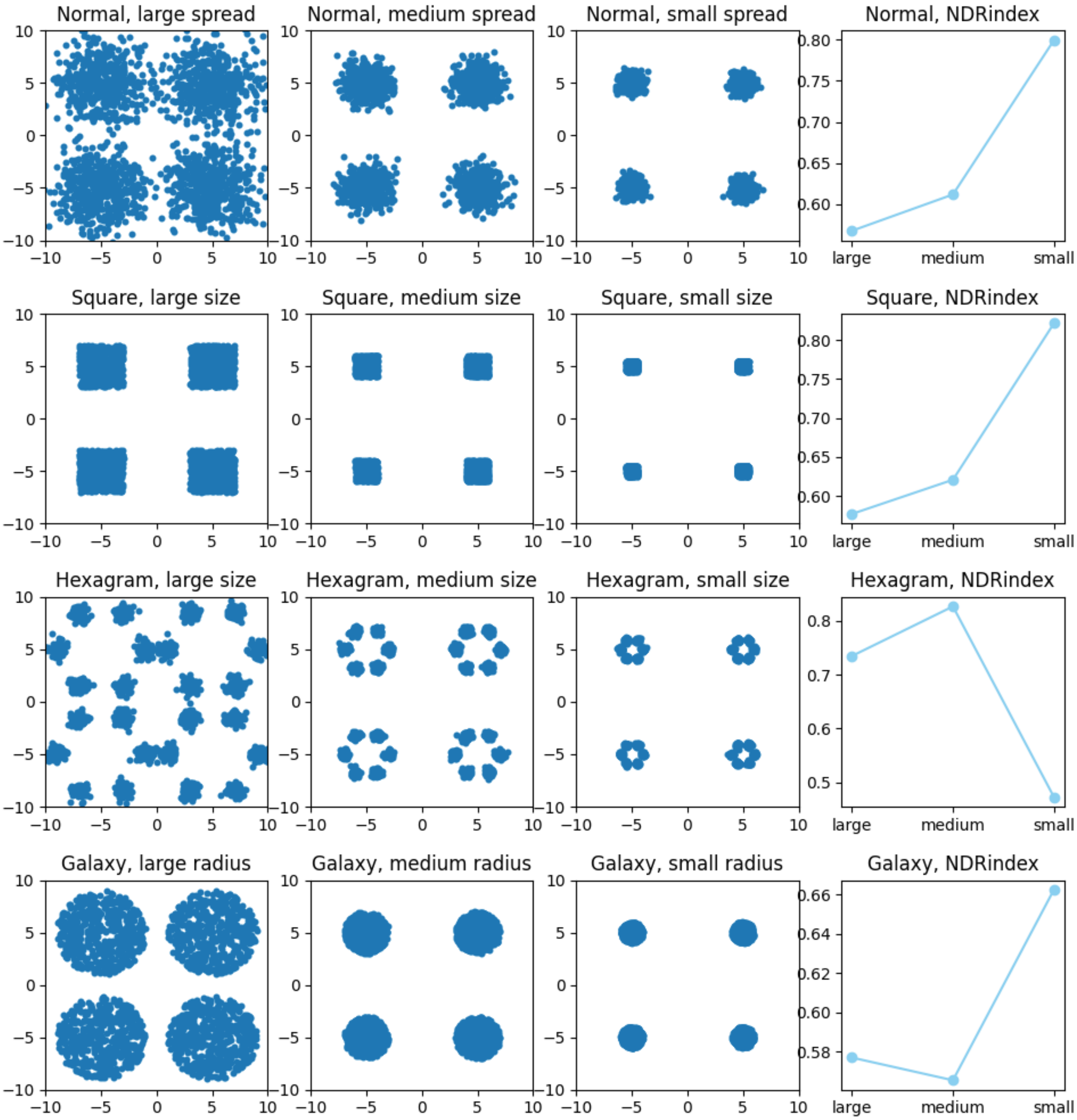

Figure 1: Results of the reproduction of the first experiment.

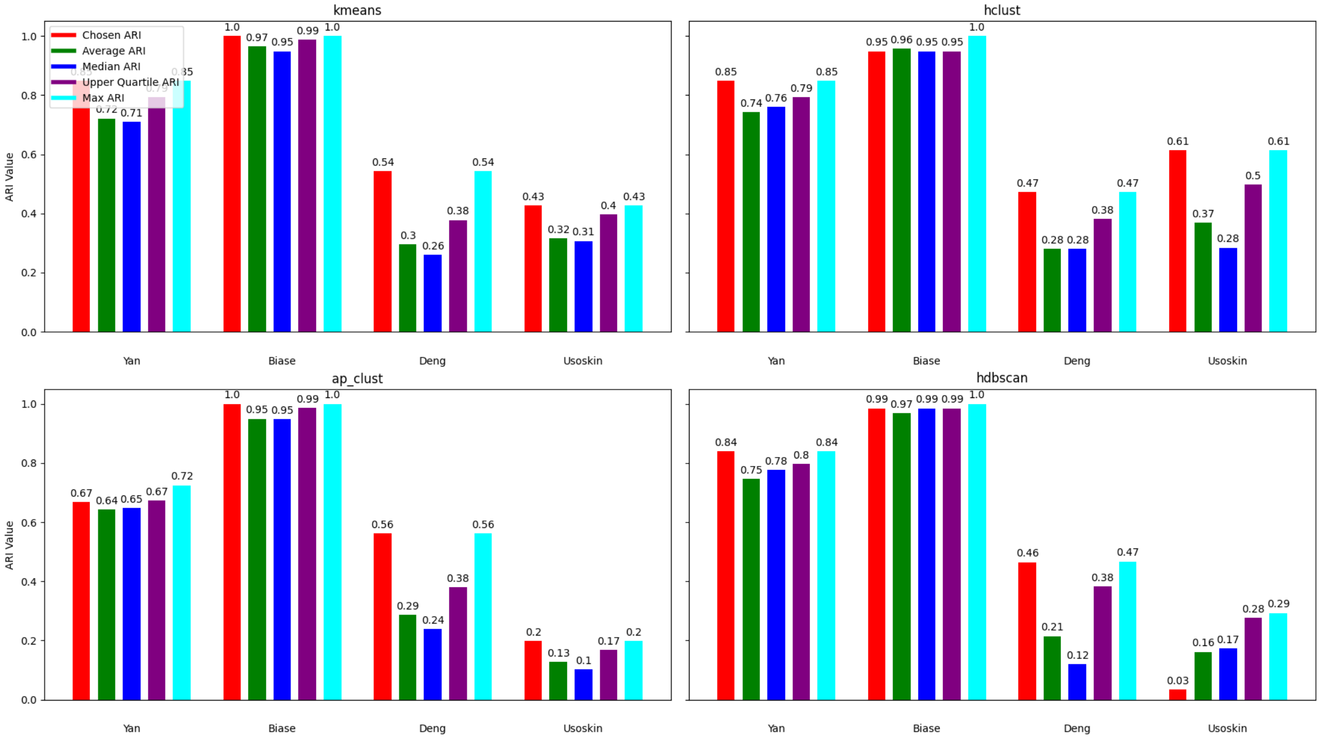

Figure 2: Results of the reproduction of the second experiment.

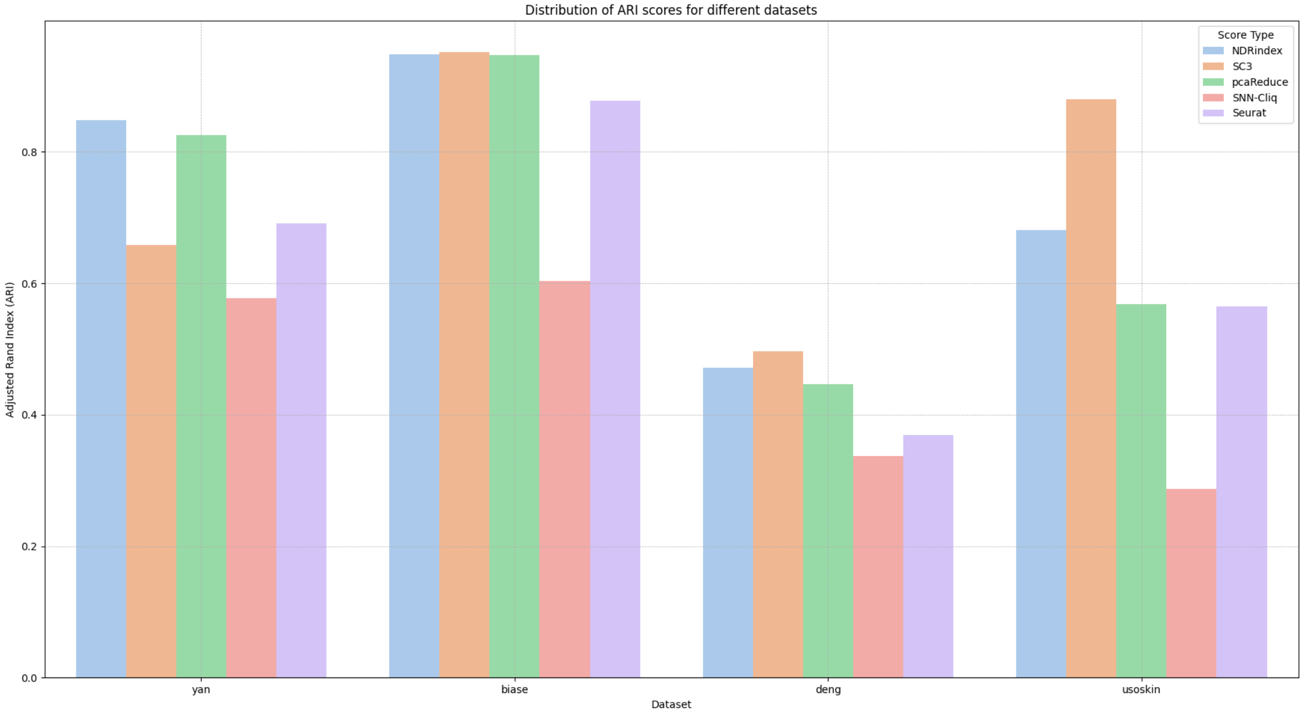

Figure 3: Results of the reproduction of the third experiment.

RAT. Specifically, the average ARI scores obtained from the NDRindex were often comparable or superior to those from the other methods. This congruence between the reproduction and the original findings underscores the robustness and reliability of the NDRindex algorithm in assessing the quality of preprocessing methods for single-cell RNA-Seq data.

Discussion

Reproducing the NDRindex algorithm has been extremely interesting, revealing the robustness of this method in evaluating the quality of preprocessing techniques for single-cell RNA-Seq data. Our findings, which largely align with the original paper, emphasize the NDRindex's potential as a reliable tool for scRNA-seq data researchers.

Our Python-centric approach brings forth several advantages. Python's clear and high-level syntax ensures that our code remains readable and maintainable, a crucial aspect in academic research where reproducibility is paramount. The vast array of libraries and frameworks in Python allows our implementation to adapt and scale efficiently with the growing complexity and size of datasets. The active Python community, with its continuous updates and improvements, further ensures that our work is well-positioned for future enhancements and emerging methodologies.

However, it's essential to recognize the nuances and potential limitations of our approach. The original NDRindex paper highlighted scenarios where the method might not be the best fit, especially for datasets without any sort of clear clusters (Xiao, Lu, and Jin 2019). Our work, while thorough, inherits these intricacies. The challenges we encountered, particularly the absence of the original dataset generation code, underscore the importance of transparency and thoroughness in academic works.

There is much potential for future work. The increasing complexity of data suggests a need for more intuitive visualization techniques. The dynamic field of scRNA-seq data analysis will likely introduce new preprocessing and clustering algorithms. Our foundational work ensures that we can integrate and evaluate these emerging methods seamlessly. While Python offers scalability, there's always room for refinement. Techniques such as parallel processing or the use of efficient data structures can further enhance the NDRindex algorithm's performance.

In conclusion, our efforts to reproduce and validate the NDRindex algorithm have underscored its efficacy and the benefits of a Python-centric approach. As the world of single-cell RNA sequencing continues to evolve, tools like the NDRindex will play a pivotal role in guiding researchers towards optimal preprocessing pathways, ensuring the highest quality of data analysis.

Conclusion

Our comprehensive reproduction of the NDRindex using Python has reaffirmed its robustness and efficacy. By leveraging the principles of computer science, we have not only replicated the original methodology but also introduced a more modular and accessible approach. This synergy between bioinformatics and structured software engineering has the potential to drive further advancements in the field.

Several key insights emerged from our reproduction:

1. Reproducibility and Transparency: The challenges faced during the reproduction process, especially the absence of the original dataset generation code, highlighted the critical importance of transparency and thorough documentation in academic research. Ensuring that methodologies are reproducible is essential for validation and further advancements.
2. Python's Versatility: Our Python-centric approach demonstrated the language's versatility and adaptability. With its vast array of libraries and active community, Python emerges as a suitable platform for bioinformatics research, ensuring scalability, maintainability, and future enhancements.
3. Future Potential: The dynamic nature of scRNA-seq data analysis suggests that new preprocessing and clustering algorithms will continue to emerge. Our foundational work with the NDRindex ensures a platform that can seamlessly integrate and evaluate these emerging methods. Moreover, as datasets grow in complexity, there is potential for further refinement in visualization techniques and performance optimization.

In summary, the NDRindex stands as a testament to the importance of quality assessment in scRNA-seq data preprocessing. Our reproduction has not only validated its efficacy but also paved the way for future enhancements. As bioinformatics continues to push the boundaries of biological research, tools like the NDRindex will remain instrumental in ensuring that researchers are equipped with accurate and high-quality data.

Acknowledgments

This reproduction was conducted within the context of the *Advanced Methods in Bioinformatics* course. This course is lectured at the Vrije Universiteit Brussel by Professor Wim Vranken, who provided me with helpful feedback throughout the process.

References

Biase, F. H.; Cao, X.; and Zhong, S. 2014. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell rna sequencing. *Genome Research* 24(11):1787–1796.

Danda, S.; Vasighizaker, A.; and Rueda, L. 2020. Unsupervised identification of sars-cov-2 target cell groups via non-linear dimensionality reduction on single-cell rna-seq data. In Park, T.; Cho, Y.; Hu, X.; Yoo, I.; Woo, H. G.; Wang, J.; Facelli, J. C.; Nam, S.; and Kang, M., eds., *IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020, Virtual Event, South Korea, December 16-19, 2020*, 2737–2744. IEEE.

Deng, Q.; Ramsköld, D.; Reinius, B.; and Sandberg, R. 2014. Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* 343(6167):193–196.

Feng, J. 2021. Research on tda-effective analytical methods for modern biology. In *2021 3rd International Conference on Intelligent Medicine and Image Processing*, IMIP '21, 109–115. New York, NY, USA: Association for Computing Machinery.

Germain, P.-L.; Sonrel, A.; and Robinson, M. D. 2020. pipecomp, a general framework for the evaluation of computational pipelines, reveals performant single-cell rna-seq preprocessing tools. *bioRxiv*.

Imoto, Y.; Nakamura, T.; Escolar, E. G.; Yoshiwaki, M.; Kojima, Y.; Yabuta, Y.; Katou, Y.; Yamamoto, T.; Hiraoka, Y.; and Saitou, M. 2022. Resolution of the curse of dimensionality in single-cell rna sequencing data analysis. *Life Science Alliance* 5(12).

Kiselev, V. Y.; Kirschner, K.; Schaub, M. T.; Andrews, T.; Chandra, T.; Natarajan, K. N.; Reik, W.; Barahona, M.; Green, A. R.; and Hemberg, M. 2016. Sc3 – consensus clustering of single-cell rna-seq data. *bioRxiv*.

Kiselev, V. Y.; Kirschner, K.; Schaub, M. T.; Andrews, T.; Yiu, A.; Chandra, T.; Natarajan, K. N.; Reik, W.; Barahona, M.; Green, A. R.; and Hemberg, M. 2017. Sc3: consensus clustering of single-cell rna-seq data. *Nature Methods* 14(5):483–486.

Krzak, M.; Raykov, Y.; Boukouvalas, A.; Cutillo, L.; and Angelini, C. 2019. Benchmark and parameter sensitivity analysis of single-cell rna sequencing clustering methods. *Frontiers in Genetics* 10.

Lin, W.; Hu, L.; Zhang, Y.; Ooi, J. D.; Meng, T.; Jin, P.; Ding, X.; Peng, L.; Song, L.; Xiao, Z.; Ao, X.; Xiao, X.; Zhou, Q.; Xiao, P.; Fan, J.; and Zhong, Y. 2020. Single-cell analysis of ace2 expression in human kidneys and bladders reveals a potential route of 2019-ncov infection. *bioRxiv*.

Liu, K.; Zhang, Y.; Martin, C.; Ma, X.; and Shen, B. 2023. Translational bioinformatics for human reproductive biology research: Examples, opportunities and challenges for a future reproductive medicine. *International Journal of Molecular Sciences* 24(1).

Mallick, K.; Chakraborty, S.; Mallik, S.; and Bandyopadhyay, S. 2023. A scalable unsupervised learning of scrnaseq data detects rare cells through integration of structure-preserving embedding, clustering and outlier detection. *Briefings Bioinform.* 24(3).

Malzer, C., and Baum, M. 2020. A hybrid approach to hierarchical density-based cluster selection. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE.

Melsted, P.; Booeshaghi, A. S.; Liu, L.; Gao, F.; Lu, L.; Min, K. H. J.; da Veiga Beltrame, E.; Hjörleifsson, K. E.; Gehring, J.; and Pachter, L. 2021. Modular, efficient and constant-memory single-cell rna-seq preprocessing. *Nature Biotechnology* 39(7):813–818.

Qi, J.; Lin, J.; Wu, C.; He, H.; Yao, J.; Xu, Y.; Yang, Y.; Wei, Y.; Huang, D.; and Mao, Y. 2023. Combined scrnaseq and bulk rnaseq analysis to reveal the dual roles of oxidative stress-related genes in acute myeloid leukemia. *Oxidative Medicine and Cellular Longevity* 2023:1–20.

Robinson, M. D., and Oshlack, A. 2010. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biology* 11(3):R25.

Sammon, J. W. 1969. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers* C-18:401–409.

Satija, R.; Farrell, J. A.; Gennert, D.; Schier, A. F.; and Regev, A. 2015. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* 33(5):495–502.

Usoskin, D.; Furlan, A.; Islam, S.; Abdo, H.; Lönnerberg, P.; Lou, D.; Hjerling-Leffler, J.; Haeggström, J.; Kharchenko, O.; Kharchenko, P. V.; Linnarsson, S.; and Ernfors, P. 2015. Unbiased classification of sensory neuron types by large-scale single-cell rna sequencing. *Nature neuroscience* 18(1):145—153.

van der Maaten, L., and Hinton, G. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research* 9(86):2579–2605.

Wang, Y.; Thong, T.; Saligrama, V.; Colacino, J.; Balzano, L.; and Scott, C. 2019. A gene filter for comparative analysis of single-cell rna-sequencing trajectory datasets. *bioRxiv*.

Wang, T.; Li, B.; and Nabavi, S. 2021. Single-cell rna sequencing data clustering using graph convolutional networks. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2163–2170.

wu Pan, X.; Xu, D.; Zhang, H.; Zhou, W.; hui Wang, L.; and gang Cui, X. 2020. Identification of a potential mechanism of acute kidney injury during the covid-19 outbreak: a study based on single-cell transcriptome analysis. *Intensive Care Medicine* 46:1114 – 1116.

Xiao, R.; Lu, G.; and Jin, S. 2019. Ndrindex: A method for the quality assessment of single-cell rna-seq preprocessing data. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1792–1800.

Xu, C., and Su, Z. 2015. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* 31(12):1974–1980.

Yan, L.; Yang, M.; Guo, H.; Yang, L.; Wu, J.; Li, R.; Liu, P.; Lian, Y.; Zheng, X.; Yan, J.; Huang, J.; Li, M.; Wu, X.; Wen, L.; Lao, K.; Li, R.; Qiao, J.; and Tang, F. 2013. Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells. *Nature Structural & Molecular Biology* 20(9):1131–1139.

Yip, S. H.; Wang, P.; Kocher, J.-P. A.; Sham, P. C.; and Wang, J. 2017. Linnorm: improved statistical analysis for single cell rna-seq expression data. *Nucleic Acids Research* 45(22):e179.

Zou, X.; Chen, K.; Zou, J.; Han, P.; Hao, J.; and Han, Z. 2020. Single-cell rna-seq data analysis on the receptor ace2 expression reveals the potential risk of different human organs vulnerable to 2019-ncov infection. *Frontiers of medicine* 14(2):185—192.

Žurauskienė, J., and Yau, C. 2016. pcareduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinformatics* 17(1):140.