engineering proceedings

MDPI

Proceeding Paper

A Gene Selection Strategy for Enhancing Single-Cell RNA-Seq Data Integration †

Konstantinos Lazaros, Georgios N. Dimitrakopoulos, Panagiotis Vlamos and Aristidis G. Vrahatis *

Bioinformatics and Human Electrophysiology Laboratory, Department of Informatics, Ionian University, 49100 Corfu, Greece; konlazaros@gmail.com (K.L.); geo.dimitrakopoulos@gmail.com (G.N.D.); vlamos@ionio.gr (P.V.)

* Correspondence: aris.vrahatis@ionio.gr

† Presented at the Advances in Biomedical Sciences, Engineering and Technology (ABSET) Conference, Athens, Greece, 10–11 June 2023.

Abstract: Cancer remains a pervasive and formidable disease within modern societies, necessitating the utilization of advanced techniques in both diagnosis and therapy. Molecular biology has emerged as a crucial tool in deciphering the underlying biological mechanisms that contribute to various types of cancer. Notably, single-cell sequencing has garnered significant attention as a state-of-the-art method for profiling gene expression in individual cells, unveiling previously concealed mechanisms and biological phenomena. With the abundance of single-cell datasets available, there is a pressing need to integrate related datasets into larger ones to enhance our understanding of biological processes and augment predictive capabilities. In this study, we investigated the impact of gene selection, achieved through the implementation of feature selection techniques, on the integration of single-cell datasets. By systematically exploring the effects of gene selection, we aim to enhance the integration process, leading to improved biological insights and enhanced predictive power. The proposed method aims to enhance two cutting-edge data integration methodologies for single-cell RNA sequencing (scRNA-seq). The method utilizes a strategy that combines two key components: a statistical approach to isolate the high variability in gene expression across cells or samples and a feature selection strategy based on XgBoost to keep genes that are important for distinguishing among healthy and cancerous cells.

Keywords: cancer; data integration; feature selection; scRNA-seq; machine learning

Citation: Lazaros, K.; Dimitrakopoulos, G.N.; Vlamos, P.; Vrahatis, A.G. A Gene Selection Strategy for Enhancing Single-Cell RNA-Seq Data Integration. *Eng. Proc.* **2023**, *50*, 12. https://doi.org/10.3390/engproc2023050012

Academic Editors: Dimitrios Glotsos, Spiros Kostopoulos, Emmanouil Athanasiadis, Efstratios David, Panagiotis Liaparinos, Ioannis Kakkos

Published: 8 November 2023

1. Introduction

Single-cell RNA sequencing (scRNA-seq) [1] is an advanced next-generation sequencing approach designed for the transcriptomic analysis of individual cells within given cellular populations. This technique affords a more nuanced comprehension of gene expression at the single-cell level, offering insights into the extent of their expression and the differential patterns that exist among cells within the same population. Put differently, scRNA-seq unveils unexpected degrees of heterogeneity in what may initially appear as homogeneous cell populations. Heterogeneity, in this context, denotes variances between cells concerning their function and behavior, and, correspondingly, their gene expression patterns. Leveraging this technology, intricate disease mechanisms, such as those observed in various types of cancer, can be deciphered, paving the way for innovative diagnostic strategies and potentially novel therapeutic interventions [2]. As a case in point, Immucan [3] serves as a digital repository, housing 78 publicly accessible scRNA-seq datasets spanning a spectrum of cancers. Such resources empower researchers and clinicians alike to delve deeper into the genomic intricacies of malignancies, potentially fostering the development of medical decision-support systems to fortify clinical diagnostic prowess.

It should be noted that single-cell datasets typically encompass a vast array of features (genes). This characteristic not only complicates the analytical process but also imposes significant computational demands [4]. This complexity of features can precipitate the 'curse of dimensionality' [5], where an extremely large number of features leads to the worsening performance of machine learning models. Fortunately, this issue can often be mitigated through the application of feature selection methodologies [6]. These techniques facilitate the retention of only the most informative features or genes pertinent to the biological phenomena under investigation, while extraneous or redundant ones are discarded.

In the context of single-cell sequencing, data integration is a term used to describe techniques that are being used in order to merge data from multiple sources into a single cohesive dataset in order to enhance the depth and resolution of the data and gain a more robust identification of trends and patterns [1]. This process, however, is hindered by batch effects. Batch effects is a term used to describe systematic differences between the individual datasets that we want to merge together. They are a result of technical differences that take place during the sequencing process and they mask true biological differences between cells, thus confounding research and leading to either very complex or even false results. Thus, the goal of data integration is to account for these technical variations and eliminate them as much as possible in order to mix batches/datasets in such a way that cells that correspond to the same or identical cell type(s) will be clustered closely together [7].

Here, we present a novel hybrid feature selection (HFS) scheme for single-cell datasets, which synergistically combines a statistical testing approach with a machine learning-based feature importance criterion. We explore our methodology's implications on single-cell data integration techniques to examine its potential modulatory effects on integration outcomes, thereby offering a deeper understanding of its performance and efficacy. By leveraging this unique perspective, we aim to deliniate the wider impacts and potential advancements that such methodologies could usher in for the complex task of single-cell data integration and possibly other related single-cell sequencing analysis pipelines.

2. Related Work

In recent years, a plethora of algorithms and methodologies have been advanced to address data integration and batch effect correction challenges. Notably, while the primary focus of such algorithms is the identification and mitigation of batch effects, many techniques first perform some form of dimensionality reduction. This step aims to enhance the signal-to-noise ratio, subsequently optimizing performance by facilitating batch correction within the latent space [1].

Originating from the realm of bulk transcriptomics, ComBat serves as a global model. This algorithm predicates its function on the presumption that batch effects manifest as uniform (either additive or multiplicative) influences across all cells [8].

Specifically designed for single-cell data, linear embedding models emerged as the most popular category of batch correction techniques. These strategies often employ a modified singular value decomposition (SVD) for data embedding and then identify local cell clusters, termed mutual nearest neighbors, across batches within this embedded space. By doing so, they rectify batch effects through a locally adaptive, non-linear approach. Renowned methods in this category include Scanorama [9], Harmony [10], and Seurat [11], among others.

Unlike other methods, BBKNN [12] utilizes a graph-based methodology for single-cell data integration. It deploys a nearest neighbor graph representing each batch/dataset. Batch effects are rectified by intentionally forcing connections between cells from varying batches, followed by pruning the connections to accommodate cell type variations.

Deep learning (DL) approaches have recently emerged as intricate techniques for batch effect alleviation, typically demanding substantial data for optimal results. These approaches, predominantly rooted in autoencoder networks, either condition their dimensionality reduction on batch variables using conditional variational autoencoders (CVAEs)

or apply a locally linear correction in the embedded domain. For example, SCALEX [13] harnesses a variational autoencoder to discern and amend batch effects. This model is composed of an encoder, which is batch-independent and extracts biologically pertinent latent features, and a decoder, which leverages batch data to reconstruct original data from these latent attributes using a domain-specific batch normalization layer (DSBN). Additionally, SCALEX implements a mini-batch methodology, uniformly sampling from all batches and adjusting for any deviation through a batch normalization layer. Other DL-inspired techniques in this domain encompass scVI [14], scANVI [15], and scGen [16].

While many single-cell analyses employ rudimentary gene selection steps often embedded in computational libraries such as Seurat [11] and Scanpy [17], it is noteworthy that, to our understanding, none capitalize on a refined feature selection pipeline exclusively focused on preserving genes pivotal to the underlying biological study. Our research endeavors to evaluate the potential impact of a refined, hybrid feature selection methodology on single-cell data integration pipelines.

3. Methodology

In light of prior research, which has unveiled that among the thousands of genes found in single-cell sequencing datasets, merely a few hundred significantly influence the specific biological phenomena under analysis [18,19], we introduce a novel Hybrid Feature Selection (HFS) strategy tailored for scRNA-seq datasets. This methodology is underpinned by two pivotal components: a statistical test pinpointing genes exhibiting pronounced variability within the dataset, and a machine learning feature importance criterion rooted in tree-based algorithms, specifically the gain variable importance metric of XgBoost, as shown in Figure 1. The latter quantifies the relative significance of each feature to the model by aggregating the contributions of said feature across all trees. A feature's elevated metric value, relative to another, underscores its paramount importance in generating accurate predictions relevant to the specific biological condition under examination.

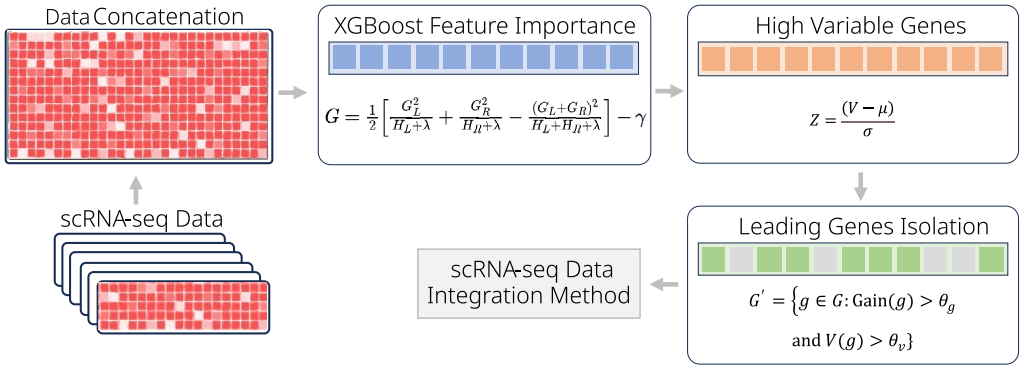

Figure 1. Schematic overview of the hybrid feature selection pipeline. The pipeline initiates with a concatenated single-cell dataset comprising n individual datasets. Feature selection is conducted using XgBoost's gain-based variable importance criterion. The top variably expressed genes are retained to create a refined dataset, which is then subjected to a data integration algorithm for batch effect correction.

For a given concatenated scRNA-seq dataset encompassing n distinct batches, our approach initially leverages XGBoost's gain variable importance metric [20] in order to identify genes that play a pivotal role in distinguishing control from case samples/cells. Through this process, only genes with non-zero importance scores are kept.

In the context of tree-based machine learning algorithms, gain denotes the augmentation in classification accuracy that is attributable to the integration of a specific feature within the branches it impacts. The foundational concept is that, prior to the introduction of a new partition based on feature X, the existing branch contains samples that are inaccurately classified. The introduction of this partition yields two derivative branches,

each of which exhibits an elevated level of classification accuracy. Specifically, one of the emergent branches provides a more accurate condition for categorizing an observation as belonging to a specific class, while the opposing branch provides a counter-condition, thereby enhancing the overall predictive ability of the classifier. XgBoost's gain metric is thus calculated as:

$$G = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma$$

This formula can be decomposed as follows:

- $\frac{G_L^2}{H_L+\lambda}$: score of the new leftmost leaf.
- $\frac{G_R^2}{H_R+\lambda}$: score of the new rightmost leaf.
- $\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}$: score of the original leaf.
- γ: regularization term.

Subsequently, the number of important genes is further reduced by only keeping the genes that exhibit the highest degree of variability among the n batches as determined by the aforementioned statistical test [21]. The mean and a measure of dispersion (variance/mean) are computed for each gene across the entire single-cell dataset. Genes were subsequently categorized into 20 bins according to their average expression levels. Within each bin, the dispersion measures for all genes are z-normalized to isolate those exhibiting high variability, even when compared with genes of comparable average expression. Z-scores are finally employed to identify genes exhibiting significant variability. Z-scores are calculated as follows:

$$Z = \frac{(V-\mu)}{\sigma}$$

This formula can be decomposed as follows:

- Z: Z-score.
- V: observed variance.
- μ: expected variance given the mean expression.
- σ: standard deviation of the variance.

From this process, a streamlined version of the original dataset is obtained, retaining only a select few hundred genes of paramount significance to the biological problem at hand. This refined dataset can be utilized for data integration/batch effect correction using any kind of single-cell data integration algorithm.

4. Experimental Study and Results

In our study, we procured three lung cancer datasets from the ImmuCan database [3]. As an initial step, these individual datasets were merged into a unified entity, retaining only the genes common to all. This fused dataset encompassed a total of 45,228 cells and 8687 genes.

Initially, the 8687 genes were ranked based on XGBoost's gain feature importance criterion, considering their role in differentiating between healthy and cancerous cells. Through this process, we retained 2024 genes, all of which had non-zero gain values. From these, the top 500 "disease important" genes exhibiting the highest variability across all three batches/datasets were retained through the use of scanpy's **highly_variable_genes** function.

Our hybrid feature selection (HFS) methodology yielded a variant of the dataset comprising 500 genes in total. This refined dataset, along with the original one encompassing 8687 genes, was used to perform data integration/batch effect correction. We utilized two well-established algorithms, Harmony [10] and Scanorama [9], to perform this task, and subsequently compared the results using integration evaluation metrics and UMAP data visualization [22].

In both Figures 2 and 3, the leftmost plot illustrates the original concatenated dataset encompassing all genes (8687 in total). The four rightmost plots showcase the outcome of data integration using Scanorama and Harmony. The two plots at the top showcase the integration results while retaining the complete gene count (8687). The two plots at the bottom demonstrate the integration results derived from the streamlined dataset which was procured through the use of our feature selection pipeline. For clarity, these visualizations are initially colored based on the specific batches or datasets being integrated, and, subsequently, by the distinct cell types present in the data. The ensuing two figures specifically depict outcomes pre and post the application of single-cell integration via Harmony and Scanorama.

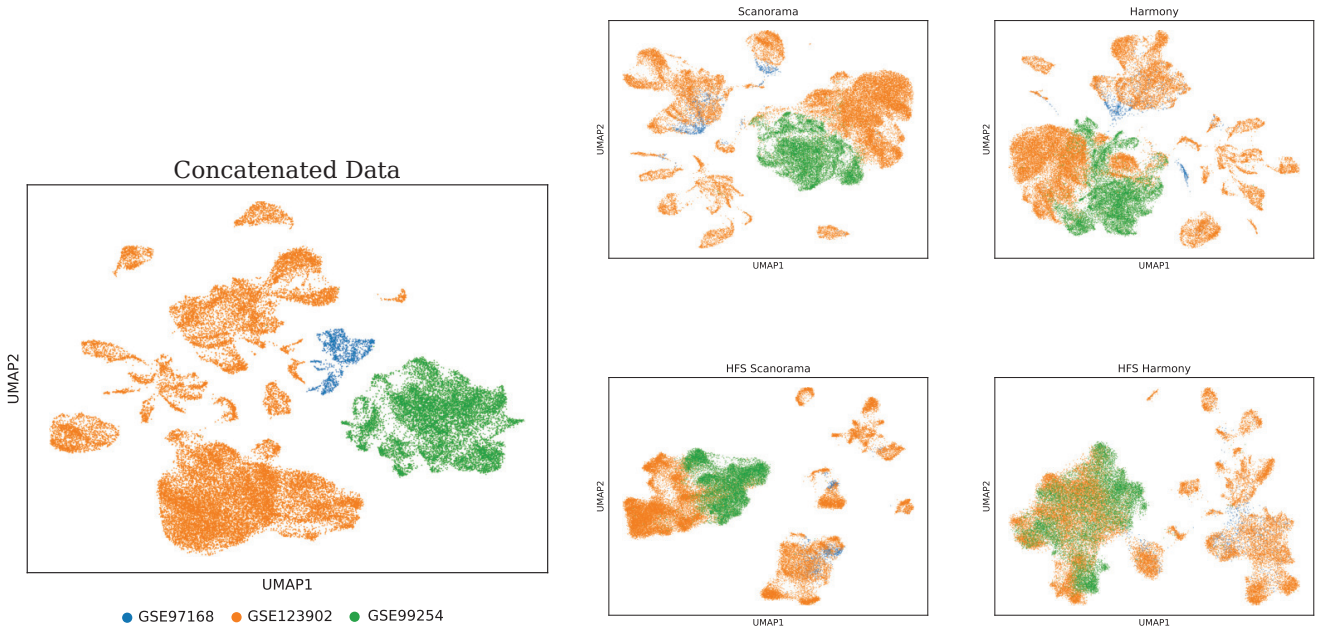

Figure 2. Comparison of 2D UMAP visualization. Points represent cell samples and each color represents a different batch/dataset. We see that, through the use of our HFS scheme, both scRNA-seq integration algorithms (Scanorama and Harmony) managed to mix between batches in a more efficient way.

From the UMAP visualizations, it becomes clear that prior to integration distinct batches are separated in such a manner that cells of analogous cell types are not congruently clustered together. When employing Harmony on the "full gene" (8687 genes) dataset, there is a modest improvement in batch mixing; however, results remain suboptimal. Remarkably, upon using Harmony with the streamlined dataset comprising 500 genes, there is a marked enhancement in batch integration, leading to more cohesive clusters of cells sharing identical or biologically related cell types. Comparable trends in results were observed when integration was executed using Scanorama.

To bolster the validity of our findings, we incorporated two salient data integration evaluation metrics from the SCIB (Single-Cell Integration Benchmark) framework: namely, kBET [23] and silhouette batch [7].

kBET measures the bias associated with a batch variable within the kNN graph. More precisely, kBET is expressed as the average rejection rate arising from Chi-squared tests, comparing local and global batch label distributions. Notably, a lower kBET value signifies enhanced batch mixing. For clarity, it is pertinent to mention that the default scaling of the kBET score ranges between 0 and 1, with higher scores representing superior batch mixing.

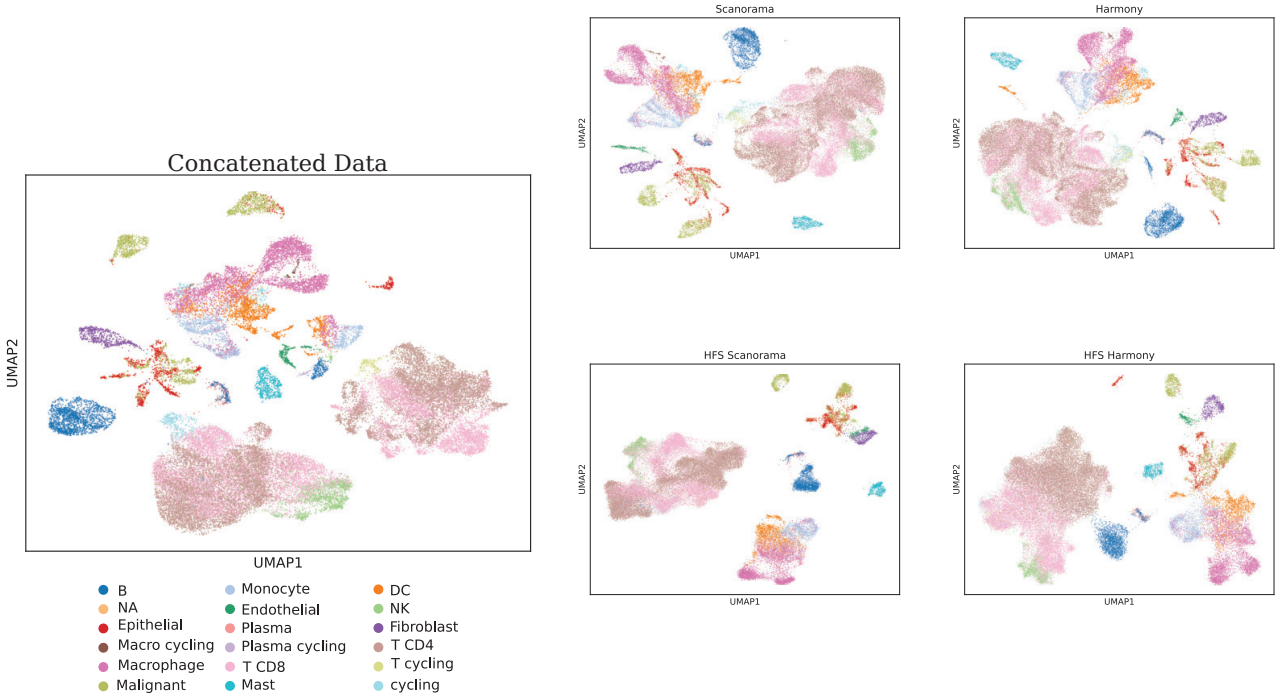

Figure 3. Comparison of 2D UMAP visualization. Points represent cell samples and each color represents a cell type. We see that, through the use of our HFS scheme, both scRNA-seq integration algorithms (Scanorama and Harmony) managed to account for batch effects in a more efficient way and cluster cells of the same or closely corresponding cell types closer together.

Conversely, the silhouette batch metric evaluates the silhouette width of a specified batch. The metric assumes that a silhouette width nearing 0 typifies an optimal overlap between batches. As such, the absolute value of the silhouette width serves as a measure for batch mixing efficacy. In its scaled version, which is the default option, the absolute Average Silhouette Width (ASW) for each group is subtracted by one prior to averaging. This ensures that a score of 0 corresponds to a suboptimal label depiction, whereas a score of 1 represents optimal label representation. The results of these two integration evaluation metrics are summarized in Table 1.

Table 1. Integration evaluation metrics results

Dataset	kbet	Silhouette Batch
Concatenated data	0.128	0.62
Scanorama	0.193	0.78
HFS Scanorama	**0.258**	**0.86**
Harmony	0.334	0.84
HFS Harmony	**0.423**	**0.90**

It is apparent, once more, that both single-cell integration algorithms yielded superior performance on the streamlined dataset, a product of our feature selection pipeline. When combining these quantitative results with the aforementioned visual interpretations, there emerges a compelling indication that feature selection has a beneficial impact on data integration and batch effect rectification within single-cell sequencing datasets.

5. Conclusions

In our study, we devised a hybrid feature selection methodology tailored for single-cell RNA-sequencing data. This approach combined a statistical measure—capturing high variability in gene expression—and the variable importance criterion from a machine

learning model, specifically the gain metric from an XGBoost classifier. When evaluating the impact of this refined feature selection approach on data integration, we observed a marked enhancement in the subsequent batch effect correction using two renowned algorithms, Scanorama and Harmony. Our findings not only underscore the efficacy of sophisticated feature selection strategies in data integration but also emphasize their potential significance in analyzing datasets linked to intricate and widely studied biological conditions, such as cancer. As the biomedical research community continues its relentless pursuit of precision medicine, such methodologies might prove instrumental in unearthing nuanced insights from intricate datasets, potentially paving the way for groundbreaking discoveries.

Author Contributions: Conceptualization, K.L., G.N.D., P.V. and A.G.V.; methodology, K.L., G.N.D., P.V. and A.G.V.; software, K.L.; validation, K.L.; resources, K.L., G.N.D., P.V. and A.G.V.; writing—original draft preparation, K.L., G.N.D., P.V. and A.G.V.; writing—review and editing, A.G.V. and P.V.; visualization, K.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was approved by the Institutional Review Board of Ionian University.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The lung cancer data utilized for this research are publicly available at https://immucanscdb.vital-it.ch/ (immucan-vue database (accessed on 27 May 2023)). The three datasets utilized for this research are: https://immucanscdb.vital-it.ch/LUAD_MYE_MRS_GSE97168 (GSE97168 (accessed on 27 May 2023)), https://immucanscdb.vital-it.ch/LUAD_UNB_10X_GSE123902 (GSE123902 (accessed on 27 May 2023)), and https://immucanscdb.vital-it.ch/NSCLC_T_SS2_GSE99254 (GSE99254 (accessed on 27 May 2023)).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Luecken, M.D.; Theis, F.J. Current best practices in single-cell RNA-seq analysis: A tutorial. *Mol. Syst. Biol.* **2019**, *15*, e8746. [CrossRef] [PubMed]
2. Haque, A.; Engel, J.; Teichmann, S.A.; Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* **2017**, *9*,75. [CrossRef] [PubMed]
3. Camps, J.; Noël, F.; Liechti, R.; Massenet-Regad, L.; Rigade, S.; Götz, L.; Hoffmann, C.; Amblard, E.; Saichi, M.; Ibrahim, M.M.; et al. Meta-Analysis of Human Cancer Single-Cell RNA-Seq Datasets Using the IMMUcan Database. *Cancer Res.* **2023**, *83*, 363–373. [CrossRef] [PubMed]
4. Vrahatis, A.G.; Tasoulis, S.K.; Dimitrakopoulos, G.N.; Plagianakos, V.P. Visualizing high-dimensional single-cell RNA-seq data via random projections and geodesic distances. In Proceedings of the 2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Siena, Italy, 9–11 July 2019; pp. 1–6.
5. Kharchenko, P.V. The triumphs and limitations of computational methods for scRNA-seq. *Nat. Methods* **2021**, *18*, 723–732. [CrossRef] [PubMed]
6. Yang, P.; Huang, H.; Liu, C. Feature selection revisited in the single-cell era. *Genome Biol.* **2021**, *22*, 321. [CrossRef] [PubMed]
7. Luecken, M.D.; Büttner, M.; Chaichoompu, K.; Danese, A.; Interlandi, M.; Müller, M.F.; Strobl, D.C.; Zappia, L.; Dugas, M.; Colomé-Tatché, M.; et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **2022**, *19*, 41–50. [CrossRef] [PubMed]
8. Johnson, W.E.; Li, C.; Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **2007**, *8*, 118–127. [CrossRef] [PubMed]
9. Hie, B.; Bryson, B.; Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **2019**, *37*, 685–691. [CrossRef] [PubMed]
10. Korsunsky, I.; Millard, N.; Fan, J.; Slowikowski, K.; Zhang, F.; Wei, K.; Baglaenko, Y.; Brenner, M.; Loh, P.r.; Raychaudhuri, S. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **2019**, *16*, 1289–1296. [CrossRef] [PubMed]
11. Stuart, T.; Butler, A.; Hoffman, P.; Hafemeister, C.; Papalexi, E.; Mauck, W.M.; Hao, Y.; Stoeckius, M.; Smibert, P.; Satija, R. Comprehensive integration of single-cell data. *Cell* **2019**, *177*, 1888–1902. [CrossRef] [PubMed]
12. Polański, K.; Young, M.D.; Miao, Z.; Meyer, K.B.; Teichmann, S.A.; Park, J.E. BBKNN: Fast batch alignment of single cell transcriptomes. *Bioinformatics* **2020**, *36*, 964–965. [CrossRef] [PubMed]
13. Xiong, L.; Tian, K.; Li, Y.; Ning, W.; Gao, X.; Zhang, Q.C. Online single-cell data integration through projecting heterogeneous datasets into a common cell-embedding space. *Nat. Commun.* **2022**, *13*, 6118. [CrossRef] [PubMed]

14. Lopez, R.; Regier, J.; Cole, M.B.; Jordan, M.I.; Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **2018**, *15*, 1053–1058. [CrossRef] [PubMed]
15. Xu, C.; Lopez, R.; Mehlman, E.; Regier, J.; Jordan, M.I.; Yosef, N. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* **2021**, *17*, e9620. [CrossRef] [PubMed]
16. Lotfollahi, M.; Wolf, F.A.; Theis, F.J. scGen predicts single-cell perturbation responses. *Nat. Methods* **2019**, *16*, 715–721. [CrossRef] [PubMed]
17. Wolf, F.A.; Angerer, P.; Theis, F.J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **2018**, *19*,15. [CrossRef] [PubMed]
18. Chatzilygeroudis, K.I.; Vrahatis, A.G.; Tasoulis, S.K.; Vrahatis, M.N. Feature Selection in single-cell RNA-seq data via a Genetic Algorithm. In Proceedings of the Learning and Intelligent Optimization: 15th International Conference, LION 15, Athens, Greece, 20–25 June 2021; Revised Selected Papers 15; Springer: Berlin/Heidelberg, Germany 2021; pp. 66–79.
19. Lazaros, K.; Tasoulis, S.; Vrahatis, A.; Plagianakos, V. Feature Selection For High Dimensional Data Using Supervised Machine Learning Techniques. In Proceedings of the 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 17–20 December 2022; pp. 3891–3894. [CrossRef]
20. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
21. Satija, R.; Farrell, J.A.; Gennert, D.; Schier, A.F.; Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **2015**, *33*, 495–502. [CrossRef] [PubMed]
22. McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv* **2018**, arXiv:1802.03426.
23. Büttner, M.; Miao, Z.; Wolf, F.A.; Teichmann, S.A.; Theis, F.J. A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* **2019**, *16*, 43–49. [CrossRef] [PubMed]