Scrutinised and Compared: HVG Identification Methods in Terms of Common Metrics

Nadiia Kasianchuk[1,2], Yevhenii Kukuruza[3], Vladyslav Ostash[3], Anastasiia Boshtova[4],
Dmytro Tsvyk[5] and Matvii Mykhailichenko[6]
[1]Faculty of Biology, Adam Mickiewicz University, Uniwersytetu Poznańskiego Str. 6, Poznań, Poland
[2]Faculty of Pharmacy, Bogomolets National Medical University, Taras Shevchenko Str. 13, Kyiv, Ukraine
[3]Faculty of Biotechnology and Biotechnics, Igor Sikorsky Kyiv Polytechnic Institute, Beresteiskyi Avenue 37, Kyiv, Ukraine
[4]Educational and Scientific Centre "Institute of Biology and Medicine', Taras Shevchenko National University of Kyiv, Hlushkova Avenue 2, Kyiv, Ukraine
[5]Educational and Scientific Institute of International Relations, Taras Shevchenko National University of Kyiv, Illyenka Str. 36, Kyiv, Ukraine
[6]Faculty of Biotechnology, University of Wroclaw, Fryderyka Joliot-Curie 14a, Wroclaw, Poland
nadkas2@st.amu.edu.pl, evgenku1508@gmail.com, ostash.vladyslav@lll.kpi.ua, a.boshtova@gmail.com, tsvykdima@gmail.com, 341450@uwr.edu.pl

Keywords: Highly Variable Genes, Single-Cell RNA-Sequencing, Differential Expression, Heterogeneity Analysis, Cellular Heterogeneity, Cellular Diversity.

Abstract: Highly variable gene (HVG) identification plays a critical role in unravelling gene expression patterns and understanding cellular heterogeneity in single-cell RNA-sequencing (scRNA-seq) data. A plethora of software packages have been developed for this purpose; however, their comparative performance is yet to be explored. This study addresses this gap by independently evaluating 22 methods from 9 different packages to provide a comprehensive assessment of the HVG identification methods. For such purpose it was deemed necessary to employ a set of common metrics, namely overlap with highly and lowly expressed genes, runtime, and clustering indices (e.g., Calinski-Harabasz, Davies-Bouldin, and ROGUE). The results reveal substantial disparities not only between different methods but also in the performance of a single method across diverse datasets. That is to say, the dimensionality of the provided data, spike-ins, and background noise are some of the key factors influencing the results. These variations underscore the significant impact of dataset characteristics on analysis outcomes. Therefore, consistent consideration of data nature is imperative. The study emphasises the urgent need for a standardised, data-driven assessment framework to ensure reliable and effective scRNA-seq analyses. This work serves as a valuable resource for both scRNA-seq software developers and experimental researchers seeking optimal methods for their investigations.

1 INTRODUCTION

In the era of data-driven medical research, the ways in which biological systems are screened and evaluated have been redefined. Thus, the challenges implied by complexity, heterogeneity, and multidimensionality of data necessitate innovative computational approaches successfully to collect, preprocess [1] and analyse [2, 3, 4, 5] the records.

Therefore, multiomics techniques have been developed as a pivotal tool in addressing contemporary biomedical tasks and challenges. Among the groundbreaking technologies in this domain, single-cell RNA sequencing (scRNA-seq) is of particular standing, as it characterises gene expression patterns at a single-cell resolution [6]. The identification of highly variable genes (HVGs) is crucial within the abovementioned process for profiling cell subtypes, performing dimensionality reduction techniques and unveiling cellular heterogeneity.

A number of methods have recently been put forward to deal with such identification in the scRNA-seq, with many of them being able to extract valuable insights from the data despite the challenges posed. Concurrently, only a minimal number of independent evaluations have been conducted to underline the most optimal methods of identifying

HVGs [7]. Furthermore, there is a lack of documentation regarding which methods are the most suitable for specific types and dimensions of data. Such thorough assessment is vital, given the methods perform ambivalently across diverse datasets, both real-world and synthetically generated [8, 9, 10].

Hence, the purpose of the present study is to establish an efficient and dependable evaluation of existing tools for HVG identification. The groundwork herein provided consists of 5 diverse scRNA-seq datasets being assessed by commonly used methods to distinguish those with the best performance. The study serves as the foundational element in crafting a comprehensive algorithm for the real-time assessment of the methods utilised in the identification of HVGs on the specific dataset required for the exact research. Thence, the researchers will be able to make well-informed decisions on the methods most suitable for their specific objectives, improving accuracy and reliability in scRNA-seq data analysis.

2 MATERIALS AND METHODS

2.1 Data Acquisition and Preprocessing

For the purposes of the present study, publicly available scRNA-seq datasets were collected in such a way that they differ in terms of dimensionality, cell and gene types, ensuring a comprehensive evaluation.

Prior to the analysis, data preprocessing pipeline was implemented to provide quality and comparability of the scRNA-seq data across all the assessed tools. Consistent preprocessing steps were followed for all datasets.

Cells and genes devoid of information regarding expression patterns were excluded from subsequent analyses. Outlier batches were removed only where they significantly impacted detection thresholds. In the datasets with droplet-based sequencing, multiplet droplets and cells with low unique molecular identifier (UMI) counts were omitted (UMI count < 100, false discovery rate threshold 0.001), except when UMI counts had no impact on barcode ranks. Log-normalised counts matrix, size-normalised counts and raw counts normalisation methods were utilised based on the data characteristics within each dataset. Outliers were identified by expression, spike-ins, and mitochondrial gene percentage.

2.2 Selection of HVG Identification Tools

22 methods computed using 9 common packages were chosen for the evaluation. The choice was based on their widespread adoption in the scRNA-seq community, ability to handle large-scale datasets, and capacity to accommodate different data distributions and experimental conditions. Furthermore, the obtained set consisted of both well-studied and not yet assessed packages, so that the comparison could be possible. Most chosen packages were R-based, considering the popularity of this language within biological data science, however Python-written Scanpy was included to cater to the preferences of different researchers and provide a comprehensive evaluation. The methods can be categorised into two groups:

- Variance-based: This group employs metrics related to variance. It selects genes with higher variability under diverse adjustments, assuming their relevance to the dataset's structure. Included methods are M3Drop_Brennecke, Seurat_vst, Seurat_disp, scVEGs, SIEVE_Seurat_vst, SIEVE_Seurat_disp, scLVM_counts, scLVM_log, scLVM_logvar, M3Drop_Brennecke_ERCCs, scLVM_logvar_ERCCs, scLVM_counts_ERCCs, (R-based) and scanpy_seurat, scanpy_cell_ranger (Python-based).
- Distribution-based relies on the dropout rate (prevalence of zeros) or assumption that count data follows a certain distribution. Methods in this category include M3Drop, M3Drop_Basic, ROGUE, ROGUE_n, Seurat_sct, SCHS, scmap, SIEVE_ROGUE, SIEVE_M3Drop (R-based) and scanpy_Pearson (Python-based).

2.3 Assessment Procedure

The process of evaluation involved a systematic test of each tool on the chosen scRNA-seq datasets. To ensure unbiased comparisons, the tools were applied with default parameter settings, and each tool was run on the same computing infrastructure.

2.3.1 Variance-Mean Dependence

Heteroscedasticity, a common issue in single-cell RNA-seq data, can introduce unwanted variability in the analysis. Therefore, to measure the potential influence of expression mean on the identification of HVGs, we quantified the ratio of HVGs that coincided with the top highly and lowly expressing genes, as well as showed Pearson's correlation between the mean expression values and variance.

2.3.2 Clustering Validation Metrics

It is crucial to ensure that selected HVGs follow some distinct expression patterns and are associated with specific clusters of genes. Therefore, several widely used clustering validation metrics were utilised for assessing the performance of the HVG identification methods.

The Calinski-Harabasz (CH) index assesses the compactness and separation between clusters, yielding higher values for better-defined ones. It is calculated by taking the ratio of the between-cluster dispersion to the within-cluster dispersion, where dispersion refers to the sum of squared distances between data points and their cluster centroids. Higher CH index values indicate better-defined and more compact clusters.

The Adjusted Rand Index (ARI) compares the similarity between different partitions of data obtained from the same clustering method, accounting for chance agreements; higher ARI values indicate better internal consistency and separation of clusters.

The Davies-Bouldin (DB) index measures the average similarity between each cluster and its most similar cluster, aiming for a lower value to indicate better-defined parameters.

The ROGUE metric [11] is tailored for scRNA-seq data and utilises an entropy-based method to quantify the purity of cell clusters. Average Silhouette Width quantifies cohesion and separation of data points within clusters, with higher values indicating well-clustered data.

Additionally, we incorporated the Purity of t-SNE k-means Clustering, that is a useful technique utilised for assessing the quality of generated clusters. For such purpose, t-distributed stochastic neighbour embedding with following k-means clustering was performed on the obtained data. Either pre-existing labels or external references were used for assigning the cell subtypes and the cluster subtype was deemed as the most common cell subtype in the cluster. Purity was calculated as the ratio of cells assigned to correct cluster, the calculation was repeated 3 times and the average was taken into the further analyses.

3 RESULTS AND DISCUSSION

The assessment was performed on 5 publicly available scRNA-seq datasets, carefully chosen to represent different types of biological data. Human datasets, such as Mair [12] (peripheral blood mononuclear cells) and Campbell [13] (brain), and mouse datasets, including Richard [14] (T cells) and Buettner [15] (embryonic stem cells), were obtained using the scRNA-seq R package [16] . The HIV [17] dataset (human HIV and CMV-specific CD8+ T-cells) was sourced from the Single Cell Portal database [18]. Data was thoroughly pre-processed, and all missing values and outliers were excluded from the analyses (Table 1).

Table 1: Dimensionality of datasets before and after the Quality Control (QC).

Name of dataset	Genes (Before QC)	Genes (After QC)	Cells (Before QC)	Cells (After QC)
Buettner	38293	11318	288	87
Mair	499	471	29033	6540
HIV	12122	12122	1559	1073
Richard	46603	28402	46603	528
Campbell	26774	26383	21086	17172

3.1 Heteroscedasticity and Variance-Mean Dependence

As proven by a number of papers, a strong correlation (e.g., 0.951, $p < 0.001$ in Mair dataset) exists between mean expression values and variance (Figure 1) causing the heteroscedasticity. Therefore, variance cannot be used as a direct indicator of HVGs and mean–variance relationships should be given attention during analyses.

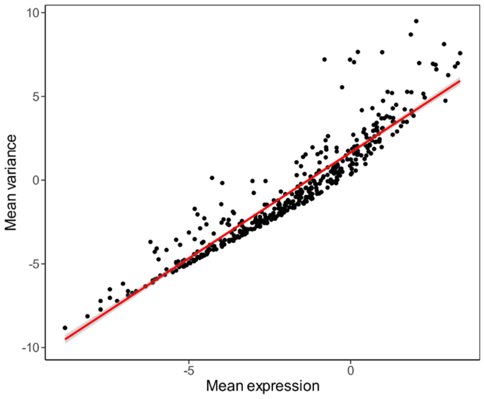


Figure 1: Correlation between the mean expression and variance in the scRNA-seq data assessed on the Mair dataset.

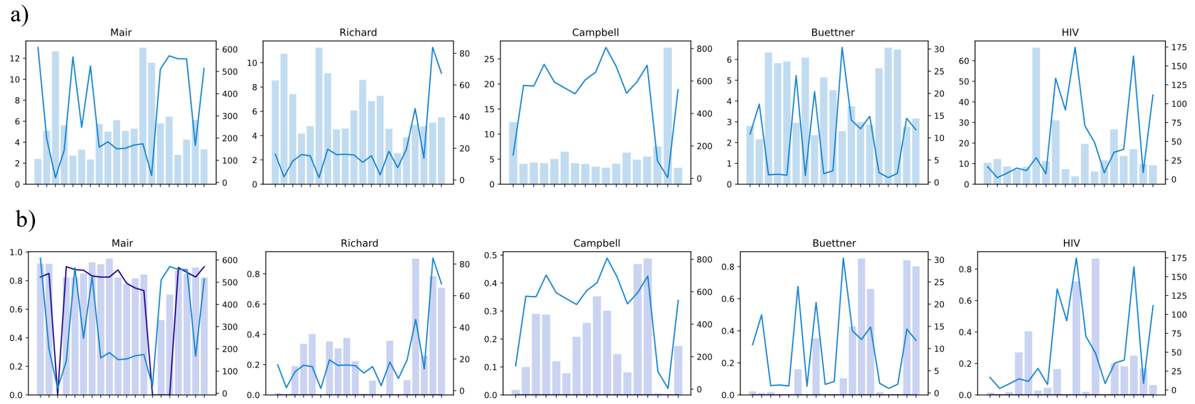


Figure 2: Evaluation of the HVG identification methods based on the quality of clustering and dependence on the heteroscedacity. a) The x-axis displays various tools used for HVG identification. A blue line plot, aligned with the right y-axis, shows the CH index. The histogram, corresponding to the left y-axis, presents the DB results, revealing a general inverse relationship between the CH index and DB values, highlighting differences in clustering efficacy. b) The x-axis again lists the tools. A blue line plot and right y-axis depicts the CH index. A purple line plot, alongside a histogram and left y-axis, shows the percentage overlap of HVGs with HEGs and LEGs. Notably, all datasets exhibited no overlap between LEGs and HVGs, with the exception being the Mair dataset.

Though multiple approaches are implemented to overcome heteroscedasticity, identification of lowly expressed genes (LEGs) still remains a daunting task in most of the methods [19, 20]. That is to say, no overlap was witnessed between lowly expressed genes (bottom *x* genes sorted by expression, where *x* is the number of HVGs returned by the respective method) and HVGs in all datasets, except in Mair (Figure 2b). However, such outcome might have been caused by low dimensionality of the dataset in question [17], [21].

The extent of overlap differed significantly across the datasets, underscoring how the methods' effectiveness is greatly influenced by the specific nature and dimensionality of the data. For instance, SCHS showcased substantial highly expressed genes (HEGs) overlap within the Mair and HIV datasets, aligning with the earlier findings [9]. However, within datasets containing a larger number of genes, the observed overlap leaned towards the moderate.

Only scLVM-based methods showed relatively consistent results, usually emerging as leaders in respect of HVG-HEGs overlap. On the other hand, M3Drop-based approaches exhibited relatively low overlap in most cases, possibly suggesting a reduced reliance on the variance (Figure 2b).

3.2 Runtime

In the realm of computational tool evaluation, runtime encapsulates efficiency and performance, shedding light on the methods' scalability, resource utilisation, and overall responsiveness [22].

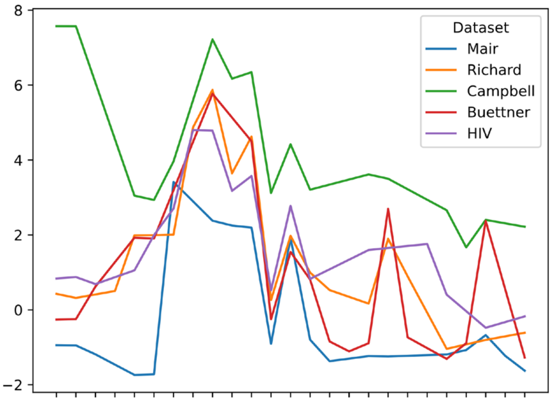
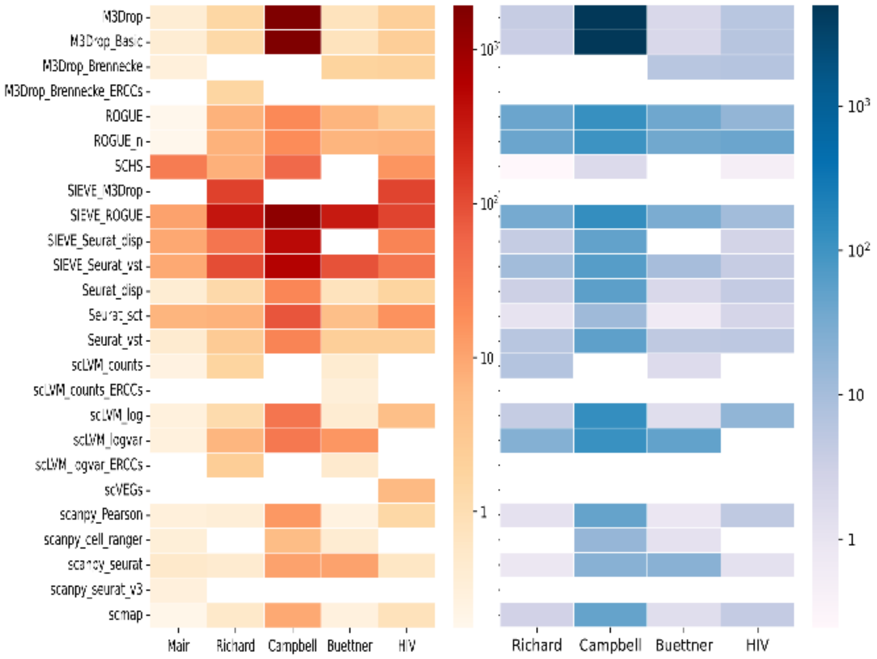


Figure 3: The runtime of the methods as indicated on the x-axis, with the y-axis representing the log-normalized runtime values in seconds.

Overall, Scmap demonstrated the best results in terms of runtime, being within the top three fastest methods across all datasets (Figure 3). Notably, Scmap adeptly handles high-dimensional datasets, exhibiting a 46,6-fold increase in runtime in the Campbell dataset compared to Mair (Figure 4). Scanpy-based methods also exhibited relatively swift execution time among all datasets not relying much on the dimensionality of the provided datasets. Seurat- and scLVM-based approaches also performed with commendable speed. Moreover, the use of External RNA Controls Consortium (ERCCs) significantly enhanced the runtime of the latter (Figure 4).



a) b)

Figure 4: Runtime assessment of the investigated methods: a) Log-normalized analysis time for individual datasets. b) The heatmap visualises the log-normalised augmentation in runtime relative to the Mair dataset, which serves as the baseline with a runtime value of 1.

Intriguingly, SCHS showed satisfactory outcomes across all datasets, other than the smallest Mair, where it performed with a strikingly low speed. M3Drop, on the other hand, distinctively underperformed on the high dimensional datasets. Incorporation of SIEVE substantially slowed down the performance of all methods used for comparison (Figure 3).

3.3 General Clustering Evaluation

Accurately identified HVGs should display distinct expression profiles across various cell populations or conditions. This specificity allows them to serve as valuable signatures for subsequent analyses. Consequently, metrics like cluster purity, heterogeneity and separation play a pivotal role in comparing and assessing different tools in this context.

Considering the high dimensionality and complexity of data discussed, several clustering indices were assessed, assuring non-biased evaluation of the HVG identification methods. Remarkably, no single method exhibited consistent superiority or inferiority across all datasets, nor did any method demonstrate such performance across all indices (Figure 2a).

M3Drop showcased the best results in the low-dimensional Mair dataset, ranking top-3 in CH, DB and ARI metrics, though it struggled while dealing with Campbell data. M3Drop_Brennecke, on the other hand, was among the worst methods on the same Mair dataset. The evaluation of Scanpy_cell_ranger also differed tremendously, as it ended up having one of the lowest scores in Campbell and Buettner, while showing good outcome in Mair. Moreover, ERCCs utilisation heightened both the CH and DB scores in all methods, where it was applied. This intriguing observation warrants deeper investigations to discern the underlying reasons for this phenomenon, as, traditionally, a high CH score signifies improved clustering outcomes, whereas a high DB score indicates the opposite.

Such divergence and somewhat chaos in the scores prove that methods' performance strongly depend on the data used for analyses [10], [23] and, therefore, some standardised method of choosing the optimal tool should be developed.

3.4 Average Score for Clustering Evaluation

As explained in the previous chapter, systematic evaluation of HVG-based clustering is necessary. However, this task is challenging, since performance significantly depends on type, volume, and complexity of the data provided. Consequently, an average score was calculated using the z-normalised results of all clustering-based metrics. Furthermore, in all indices except DB, a higher value indicates better clustering characteristics. To account for this, DB values were inverted.

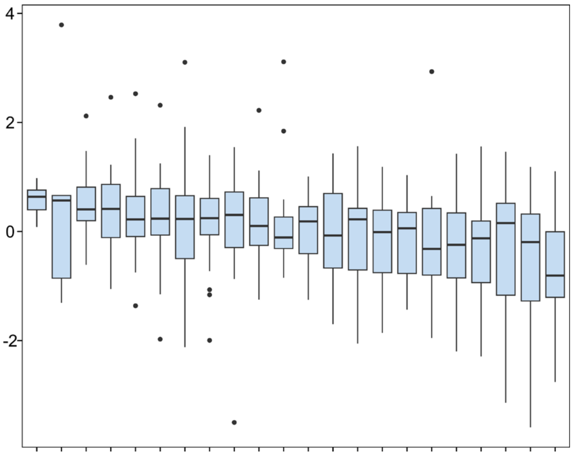


Figure 5: Average score of clustering-based metrics. The x-axis illustrates the various tools, while the y-axis represents the z-normalized average scores of clustering-based metrics.

Overall, the SCHS, scVEGs, and scanpy_seurat_v3 methods exhibited the highest performance (Figure 5) with their scores notably surpassing those of other tools (0,57-0,60 compared to 0,30 in the method ranked 4). ROGUE_n (log-normalised counts), along with the SIEVE-incorporated ROGUE, demonstrated commendable performance, securing ranks 4 and 5. However, the same cannot be said for pure ROGUE, as it achieved an average score of -0,18.

Noteworthy, the incorporation of SIEVE improved performance only of the abovementioned ROGUE, as Seurat_vst and Seurat_disp exhibited a fall off in the average score. Scmap ended up with a moderate result of 0,11 and M3Drop underperformed (averaged score < 0) in all cases.

However, it should be noted, that although the averaged normalised score is a convenient way for the basic comparison of methods, it cannot provide insights into the strengths and weaknesses of each exact method. Therefore, next steps are required to enable systematic overview of the HVG identification tools.

4 CONCLUSIONS

HVG identification is a vital component of single cell RNA-sequencing providing a comprehensive overview of gene expression patterns and a deeper comprehension of cellular heterogeneity. While numerous packages have been developed for this purpose, information pertaining to their performance remains limited: many evaluations have been conducted by the developers of these methods, introducing a potential conflict of interests. Hence, an independent assessment may come in handy for both scRNA-seq software developers and wet researchers who look for a suitable method to employ in their research.

Remarkably, the outcomes exhibited significant discrepancies, not merely among methods themselves, but also in performance of a given method when applied to various datasets. However, these variations are in line with findings from other independent research studies in the field [6]. To name a few, M3Drop exhibited moderate to good performance in small datasets, but with the increase of data dimensionality it strongly underperformed both in terms of clustering and runtime. Similar trends were observed with Scanpy_cell_ranger. Conversely, SCHS runtime increased when analysing the Mair dataset, which has the lowest dimensionality. These divergences underscore the substantial influence of the analysed data's nature on the obtained results. Hence, consistently considering this aspect is crucial, and the development of a standardised data-driven assessment system is imperative to ensure effective and reliable analyses in the field.

ACKNOWLEDGMENTS

With heartfelt appreciation, we extend our gratitude to the Bioinformatics for Ukraine initiative, and particularly to Prof. Serghei Mangul, for inspiring us to continue this research project following the East European Bioinformatics and Computational Genomics School. We also express our deep gratitude to everyone who courageously fight for the freedom of Ukraine and the whole civilised world.

REFERENCES

[1] S. Fedushko, M. Gregus, and T. Ustyianovych, 'Medical card data imputation and patient psychological and behavioral profile construction', Procedia Comput Sci, vol. 160, pp. 354-361, 2019, doi: 10.1016/j.procs.2019.11.080.

[2] M. Marczyk et al., 'Treatment Efficacy Score-continuous residual cancer burden-based metric to compare neoadjuvant chemotherapy efficacy between randomized trial arms in breast cancer trials', Annals of Oncology, vol. 33, no. 8, pp. 814-823, Aug. 2022, doi: 10.1016/j.annonc.2022.04.072.

[3] P. Rzymski, N. Kasianchuk, D. Sikora, and B. Poniedziałek, 'COVID-19 vaccinations and rates of infections, hospitalizations, ICU admissions, and deaths in Europe during SARS-CoV-2 Omicron wave in the first quarter of 2022', J Med Virol, vol. 95, no. 1, Jan. 2023, doi: 10.1002/jmv.28131.

[4] N. Kasianchuk, D. Tsvyk, E. Siemens, and H. Falfushynska, 'Random Forest Algorithm in Unravelling Biomarkers of Breast Cancer Progression'. Proceedings of the International Conference on Applied Innovations in IT (ICAIIT), vol. 11, no. 1, pp. 133-141, Mar. 2023, doi: 10.25673/101930.

[5] N. Kasianchuk, D. Tsvyk, E. Siemens, V. Ostash, and H. Falfushynska, "Genomic data machined: The random forest algorithm for discovering breast cancer biomarkers," in Information and Communication Technologies and Sustainable Development, in Lecture notes in networks and systems. Cham: Springer Nature Switzerland, 2023, pp. 428-443. doi: 10.1007/978-3-031-46880-3_25.

[6] D. Deshpande et al., 'RNA-seq data science: From raw data to effective interpretation', Front Genet, vol. 14, Mar. 2023, doi: 10.3389/fgene.2023.997383.

[7] A. Sonrel et al., 'Meta-analysis of (single-cell method) benchmarks reveals the need for extensibility and interoperability', Genome Biol, vol. 24, no. 1, p. 119, May 2023, doi: 10.1186/s13059-023-02962-5.

[8] S. H. Yip, P. C. Sham, and J. Wang, 'Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data', Brief Bioinform, vol. 20, no. 4, pp. 1583-1589, Jul. 2019, doi: 10.1093/bib/bby011.

[9] Y. Zhang, X. Xie, P. Wu, and P. Zhu, 'SIEVE: identifying robust single cell variable genes for single-cell RNA sequencing data', Blood Science, vol. 3, no. 2, pp. 35-39, Apr. 2021, doi: 10.1097/BS9.0000000000000072.

[10] T. S. Andrews and M. Hemberg, 'M3Drop: dropout-based feature selection for scRNASeq', Bioinformatics, vol. 35, no. 16, pp. 2865–2867, Aug. 2019, doi: 10.1093/bioinformatics/bty1044.

[11] B. Liu, C. Li, Z. Li, D. Wang, X. Ren, and Z. Zhang, 'An entropy-based metric for assessing the purity of single cell populations', Nat Commun, vol. 11, no. 1, p. 3155, Jun. 2020, doi: 10.1038/s41467-020-16904-3.

[12] F. Mair et al., 'A Targeted Multi-omic Analysis Approach Measures Protein Expression and Low-Abundance Transcripts on the Single-Cell Level', Cell Rep, vol. 31, no. 1, p. 107499, Apr. 2020, doi: 10.1016/j.celrep.2020.03.063.

[13] J. N. Campbell et al., 'A molecular census of arcuate hypothalamus and median eminence cell types', Nat Neurosci, vol. 20, no. 3, pp. 484-496, Mar. 2017, doi: 10.1038/nn.4495.

[14] A. C. Richard, A. T. L. Lun, W. W. Y. Lau, B. Göttgens, J. C. Marioni, and G. M. Griffiths, 'T cell cytolytic capacity is independent of initial stimulation strength', Nat Immunol, vol. 19, no. 8, pp. 849-858, Aug. 2018, doi: 10.1038/s41590-018-0160-9.

[15] F. Buettner et al., 'Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells', Nat Biotechnol, vol. 33, no. 2, pp. 155-160, Feb. 2015, doi: 10.1038/nbt.3102.

[16] D. Risso et al., 'scRNAseq: Collection of Public Single-Cell RNA-Seq Datasets', R package version 2.14.0, 2023, doi: 10.18129/B9.bioc.scRNAseq.

[17] R. Fardoos et al., 'HIV specific CD8+ TRM-like cells in tonsils express exhaustive signatures in the absence of natural HIV control', Front Immunol, vol. 13, Oct. 2022, doi: 10.3389/fimmu.2022.912038.

[18] 'Single Cell Portal'. Accessed: Jul. 22, 2023. [Online]. Available: https://singlecell.broadinstitute.org/single_cell.

[19] H.-I. H. Chen, Y. Jin, Y. Huang, and Y. Chen, 'Detection of high variability in gene expression from single-cell RNA-seq profiling', BMC Genomics, vol. 17, no. S7, p. 508, Aug. 2016, doi: 10.1186/s12864-016-2897-6.

[20] C. A. Vallejos, J. C. Marioni, and S. Richardson, 'BASiCS: Bayesian Analysis of Single-Cell Sequencing Data', PLoS Comput Biol, vol. 11, no. 6, p. e1004333, Jun. 2015, doi: 10.1371/journal.pcbi.1004333.

[21] F. A. Wolf, P. Angerer, and F. J. Theis, 'SCANPY: large-scale single-cell gene expression data analysis', Genome Biol, vol. 19, no. 1, p. 15, Dec. 2018, doi: 10.1186/s13059-017-1382-0.

[22] A. Tyryshkina, N. Coraor, and A. Nekrutenko, 'Predicting runtimes of bioinformatics tools based on historical data: five years of Galaxy usage', Bioinformatics, vol. 35, no. 18, pp. 3453-3460, Sep. 2019, doi: 10.1093/bioinformatics/btz054.

[23] P. Brennecke et al., 'Accounting for technical noise in single-cell RNA-seq experiments', Nat Methods, vol. 10, no. 11, pp. 1093-1095, Nov. 2013, doi: 10.1038/nmeth.2645.

APPENDIX

For access to the code, supplementary data, and results pertaining to this research, please visit our GitHub repository at https://github.com/RI3NO/HiVaGe.