A Computational Approach for Cell Characterization Without Prior Isolation: Advances in scRNA-seq

Bruno Sime[1,2], Marco Antônio Zanata Alves[1], Tarcio Teodoro Braga[2]

[1]Department of Informatics, Federal University of Paraná (UFPR), Brazil.
[2]Department of Basic Pathology, Federal University of Paraná (UFPR), Brazil.

*Corresponding authors: brunosime@ufpr.br; tarcio.braga@ufpr.br

Abstract

Single-cell RNA sequencing (scRNA-seq) enables high-resolution analysis of cellular heterogeneity, but traditional cell isolation methods like flow cytometry and laser microdissection often suffer from limitations in efficiency, viability, and bias. To overcome these challenges, computational tissue deconvolution approaches have emerged as effective alternatives. In this work, we introduce a high-performance computational pipeline for scRNA-seq data analysis that identifies and segregates cell populations based on marker gene expression. Our method incorporates advanced preprocessing, normalization, and clustering techniques, optimized for scalability and reproducibility in high-performance computing (HPC) environments. Compared to related tools, our pipeline offers enhanced adaptability across diverse datasets and experimental settings. We validated its performance using zebrafish ventricular tissue post-injury, effectively identifying key regenerative cell types such as immune cells, including macrophages. This approach supports in-depth biological discovery without prior physical cell separation and expands the potential of scRNA-seq applications in regenerative biology, immunology, and single-cell transcriptomics.

1 Introduction

Single-cell RNA sequencing (scRNA-seq) has revolutionized cellular biology by enabling high-resolution characterization of cellular heterogeneity across a wide range of tissues and physiopathological conditions[2, 11, 12]. This technology facilitates the identification of distinct cell populations, their gene expression profiles, and functional states, thereby offering critical insights into complex biological processes such as development, immune response, and tissue regeneration[1, 25].

Traditionally, the separation of specific cell types within a tissue relies on experimental methods such as optimized centrifugation protocols, flow cytometry, or laser microdissection. However, these techniques often present limitations related to efficiency, cell viability, and potential experimental bias[16, 30, 19]. Moreover, the reliance on precharacterized markers for cell identification may hinder the discovery of novel or rare cellular states. In this context, computational approaches for the deconvolution of whole-tissue transcriptomic profiles have emerged as powerful alternatives, enabling the extraction of biologically meaningful information without requiring prior physical cell separation[21, 4, 18, 27].

In this study, we introduce an advanced computational pipeline for scRNA-seq analysis, designed to enable the targeted separation of cell populations based on the expression of known marker genes. Our pipeline integrates robust pre-processing protocols (including SCTransform normalization[8]), advanced batch-correction methods (such as Harmony[12] and Seurat integration[24]), and state-of-the-art clustering algorithms (Leiden and Louvain) to ensure accurate identification and isolation of relevant cell populations from whole-tissue samples. By leveraging previously validated molecular markers, the approach provides a comprehensive characterization of cellular heterogeneity and ensures the precise delineation of target populations for downstream analyses[23, 28].

The pipeline is engineered for high scalability and optimized for execution in high-performance computing (HPC) environments, allowing for efficient processing of large-scale transcriptomic datasets. Its architecture ensures analytical reproducibility, robustness of results, and adaptability across a variety of datasets and experimental contexts[3].

To the best of our knowledge, this is the first study to implement a marker-based, HPC-compatible computational strategy for the targeted separation of cell populations from whole-tissue scRNA-seq data without prior physical sorting. This innovative framework enhances analytical resolution while offering a practical, scalable alternative for contexts in which conventional separation methods are limited or impractical.

To demonstrate the capabilities of our pipeline, we selected publicly available studies that employed scRNA-seq as the primary methodology to investigate cardiac regeneration in adult zebrafish. These studies typically feature time-course experimental designs with sampling at key regenerative stages, including uninjured controls and post-injury time points at 3, 7, and 14–30 days after injury[9, 14, 15, 20]. Common analytical focuses include the characterization of non-myocyte populations—such as fibroblasts, macrophages, endothelial cells, and epicardial cells—with emphasis on their transcriptional dynamics and functional roles during tissue regeneration[5].

Computational tools such as Seurat[21, 24], Scanpy[28], LIGER[9], and UMAP[17] are consistently used across these studies for dimensionality reduction and clustering, enabling the identification of distinct cell states and transient activation patterns.

Key molecular markers examined include proliferation indicators (e.g., *mKi67*, *PCNA*)[7], extracellular matrix remodeling genes (*col12a1a*, *fn1a*)[2], pro-regenerative signaling molecules (e.g., *nrg1*, *aldh1a2* for retinoic acid synthesis)[9, 15], and inflammatory response mediators[14]. Furthermore, several studies incorporate spatial validation techniques—such as immunohistochemistry and in situ hybridization—to align computational findings with anatomical structures, as well as cell–cell interaction analyses to uncover coordinated cellular behaviors that underlie successful cardiac regeneration[29].

Our results demonstrate the superior capacity of the proposed pipeline to separate and characterize distinct cell populations in zebrafish cardiac tissue following injury. Temporal analysis of regenerative stages revealed dynamic shifts in cell type composition and highlighted the central role of immune cells, particularly macrophages, in orchestrating the regenerative response[14]. Additionally, the ability to isolate specific cell subsets enabled reduction of computational burden and improved analytical efficiency in large-scale datasets[13].

Beyond the zebrafish model, we validated the pipeline's generalizability across species and tissue types, contingent on the availability of prior knowledge regarding marker genes for the target populations. This versatility reinforces the method's potential applicability across a wide range of biological systems and scRNA-seq studies.

In summary, this work introduces a novel and efficient computational tool for scRNA-seq analysis, enabling the exploration of complex tissues without the need for physical cell separation. We anticipate that this pipeline will be broadly applicable across diverse research domains, advancing the understanding of cellular processes and offering new perspectives in regenerative biology, immunology, and single-cell transcriptomics.

2 Materials and Methods

The samples analyzed in this study consist of whole tissue preparations processed in a single scRNA-seq run, encompassing a heterogeneous mix of cell types. Single-cell transcriptional profiling was performed in all major cardiac cell types[5], in association with transgenic animals[18], followed by FACS-sorting cells[20], or only in non-cardiomyocytes followed by cardiomyocytes separation through low-speed centrifugation[15].

The central analytical objective was to focus on a specific cellular population, previously characterized by a well-defined set of gene markers described in the literature.

Our analysis begins with a gene expression matrix derived from scRNA-seq data, regardless of the alignment or quantification algorithm employed (e.g., Cell Ranger[30], STARsolo[10], Salmon[23, 22], or Kallisto|bustools[3]). The first step in the pipeline involves normalization of the data using the Centered Log-Ratio (CLR) method[6, 26]. CLR transforms each gene's expression into a ratio relative to the geometric mean of expression values within the same cell. This method is particularly suited to scRNA-seq data, as it addresses the compositional nature of the dataset, mitigates biases introduced by sequencing depth and capture efficiency, and reduces the dominance of highly expressed cell types over less abundant but biologically relevant populations. By emphasizing relative expression rather than absolute counts, CLR normalization improves the detection of subtle transcriptional variation[26, 19].

$$\mathrm{CLR}_{ij} = \log_2 \left(\frac{x_{ij} + 1}{\left(\prod_{k=1}^{n} (x_{ik} + 1)\right)^{1/n}} \right) \tag{1}$$

Where:

- x_{ij} are the gene counts j on cell i
- n is the total number of genes

Following normalization, cells expressing the canonical markers associated with the target population are selected through a subsetting strategy. This step reduces dataset complexity by removing irrelevant cells from the analysis and concentrating computational effort on biologically pertinent subsets. This also increases sensitivity and statistical power in downstream analyses.

The original, unfiltered dataset is retained to enable optional comparative analysis across cell populations if required. Once the subset of interest is defined, canonical analytical procedures—such as dimensionality reduction, clustering, marker identification, and visualization—are applied, in accordance with widely adopted methodologies in the field.

For validation purposes, reference markers were obtained from the scientific literature and used as a baseline for comparison. Marker tables generated through standard workflows (e.g., Seurat's FindMarkers or FindSubClusters functions) [21] were directly compared to those derived from the proposed filtering approach. Even when applying advanced clustering refinement techniques, the conventional pipeline proved limited in resolving transcriptional states with sufficient granularity, often overlooking low-expression markers or intermediate cellular states.

Additionally, a technical limitation was identified during post-normalization processing. The use of primitive floating-point data types (such as float32) introduced numerical precision constraints that affected low-abundance transcripts. These values, due to limited decimal representation, were often misclassified as noise and discarded during clustering and differential expression analysis. This phenomenon, referred to here as signal shadowing, is common in datasets with highly diverse cellular compositions and leads to the loss of biologically relevant signals. Furthermore, attempts to mitigate this issue by increasing the number of floating-point decimals (e.g., using float64 or higher precision data types) would, in the best-case scenario, result in a doubling of memory usage, which is not feasible given current computational resource constraints. As such, alternative strategies are required to address the limitations imposed by signal shadowing and ensure that low-abundance transcripts are accurately detected and analyzed.

The proposed approach addresses this issue by introducing a marker-based filtering step prior to downstream processing, thereby minimizing early-stage information loss and improving the resolution and fidelity of the biological inferences drawn from the data.

3 Results

The results obtained were thoroughly validated against established literature on related studies[14, 18, 15, 20, 5]. Following the application of our analytical pipeline, we achieved enhanced resolution and significantly improved recovery of data related to the target cell populations, as exemplified by the distribution shown in Figure 1: the yellow histogram, representing macrophages, displays a more defined and concentrated profile of average log2 fold changes (avg_log2FC), in contrast to the broader and more dispersed distribution observed in blue, which corresponds to the aggregated signal from all cells. This methodological advancement enabled a more granular dissection of cellular heterogeneity, surpassing the resolution typically reported in similar analyses.

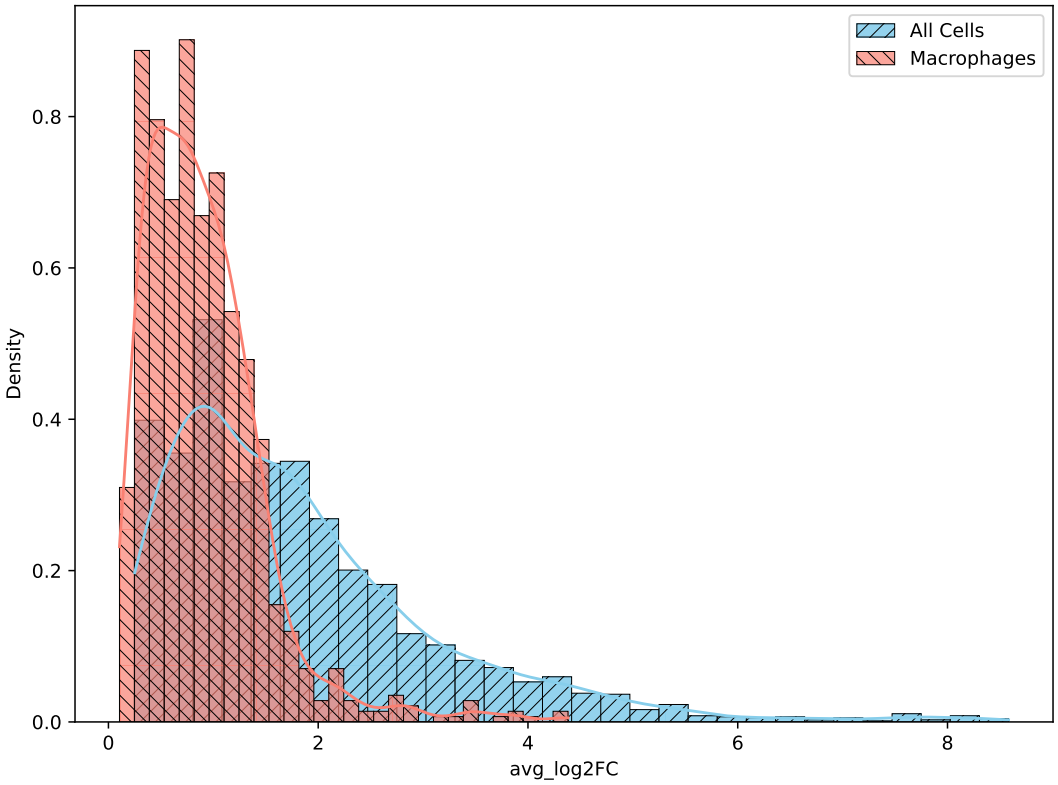

Figure 1: Distribution of average log2 fold change (avg_log2FC) values for genes shared between conditions, comparing all cell populations (blue) versus macrophages only (yellow). The histogram reveals a narrower and more sharply peaked distribution for macrophages, indicating a more consistent and pronounced gene expression response within this specific cell type. In contrast, the broader distribution observed across all cells reflects greater heterogeneity, likely due to the inclusion of multiple cell populations with varying transcriptional profiles.

Notably, in the dataset specifically curated for this experiment[14], the data yield was markedly elevated, with improvements reaching approximately 26,89%. This substantial gain underscores the efficacy of our approach in enhancing the sensitivity and specificity of cell population identification.

A critical factor contributing to these results was the strategic curation of the dataset-specifically, the exclusion of non-relevant cells based on predetermined marker profiles. This filtering step not only streamlined the analytical process but also resulted in several computational benefits: increased data capture for the cells of interest, reduced overall computational load, and a significant decrease in the complexity of extracting biologically relevant information from raw FASTQ files.

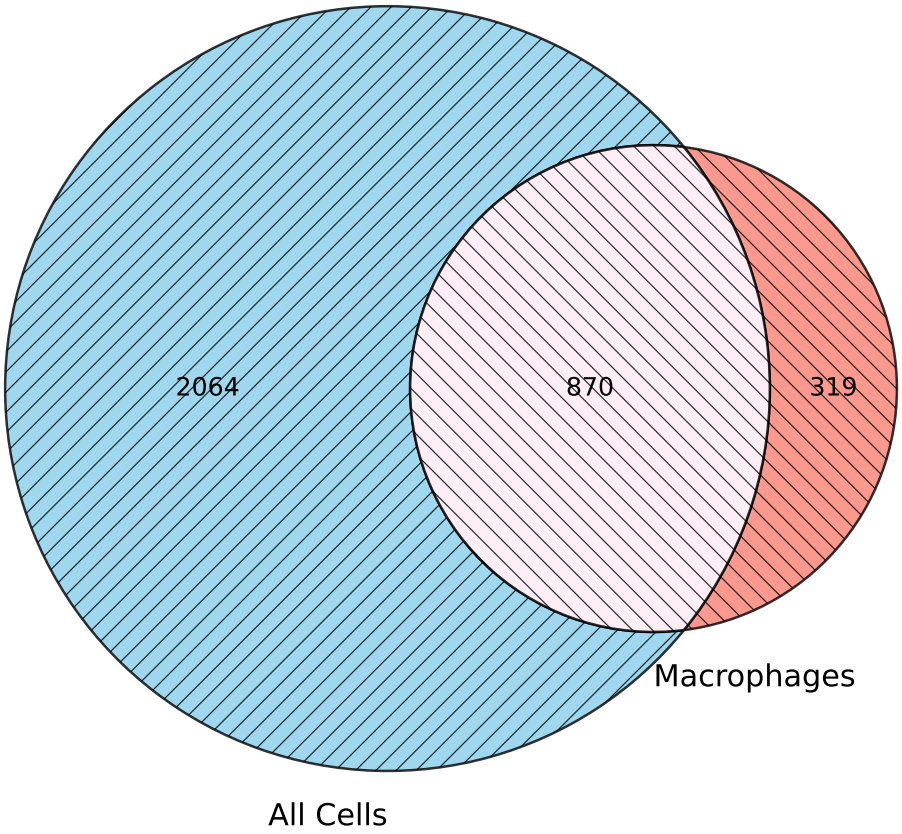

Figure 2: Venn diagram illustrating the overlap of genes detected in the complete dataset comprising all cell populations ("All Cells") versus the macrophage-specific subset ("Macrophages"). A total of 870 genes were commonly detected in both datasets, whereas 2,064 genes were uniquely identified in the global dataset and 319 genes were exclusively recovered within the macrophage-specific subset. This distribution underscores the increased resolution and sensitivity achieved through the targeted, marker-guided curation strategy employed in our analytical pipeline. Notably, the subset-based approach enabled the detection of macrophage-specific transcripts that would likely remain undetected under conventional whole-sample analyses, where signal dilution from heterogeneous cell populations often obscures cell-type-restricted gene expression. These results highlight the methodological advantage of incorporating cell-type-aware filtering steps in single-cell RNA-seq workflows, allowing for the preservation and analysis of transcriptional signatures critical to specialized or transient cell states.

Among the genes identified specifically in the dataset where the method was applied, several features emerged that were not observed using traditional data curation approaches. These genes were validated as essential to macrophage plasticity processes—subtle characteristics that would likely have been missed with less stringent analytical criteria. Such findings highlight transient transcriptional states that are only detectable when cells are undergoing phenotypic transition.

Importantly, we highlight the broader implications of our findings in democratizing access to scRNA-seq data analysis. By enabling a marker-driven, subset-based strategy that can be implemented with familiar tools and frameworks, our pipeline reduces the barrier to entry for researchers without extensive

computational expertise. This approach allows high-resolution, cell-specific analyses without requiring mastery of complex or specialized software environments.

Furthermore, the dimensionality reduction achieved through this method facilitates more intuitive data visualization and interpretation. Researchers can generate plots and visual summaries of their datasets with greater ease, enhancing both exploratory and confirmatory analyses.

Moreover, Seurat remains one of the most widely adopted toolkits for scRNA-Seq analysis and provides a highly modular framework encompassing normalization, scaling, dimensionality reduction, clustering, and differential expression testing [13]. However, it relies heavily on global heuristics - such as minimum feature expression thresholds (often 10–20% of cells) and dispersion-based filtering - to reduce noise and improve statistical power. While effective for broad population-level analyses, these thresholds risk exclude transcriptionally relevant genes that are highly specific to rare cell types or transient activation states. The biological significance of circumventing such thresholds becomes particularly evident when considering cell types characterized by dynamic transcriptional states. Macrophages, for instance, are known for their phenotypic flexibility and context-driven transcriptomic plasticity in response to microenvironmental stimuli [29, 7]. Similarly, dendritic cells, regulatory T lymphocytes, and astrocytes - cells that often play specialized roles in immune surveillance and neural regulation - exhibit context-dependent transcriptional programs that may not reach detection thresholds in full-tissue datasets. The enhanced granularity of the proposed methodology allows for a more faithful representation of these subtle transcriptional shifts, enabling finer dissection of cellular heterogeneity. A compelling illustration of this analytical advantage is provided in Figure 3, which presents a scatterplot comparing the expression frequency (pct.1) of genes shared between the global dataset (encompassing all cell types) versus a macrophage-specific subset. The x-axis represents the proportion of all cells in which a given gene is expressed, while the y-axis indicates the proportion of macrophages expressing the same gene. A pronounced cluster of points emerges in the upper-left quadrant – specifically, genes expressed in over 90–100% of macrophages but in fewer than 20% of all cells.

This region of the plot provides strong evidence that conventional filtering strategies (as implemented in Seurat's standard pipeline) would likely discard these genes due to their low prevalence across the total cell population, despite their clear biological relevance in a specific subpopulation. Since Seurat often applies default cutoffs that exclude genes not detected in a minimum fraction of all cells (e.g., min.pct = 0.1 or min.pct = 0.2 in differential expression tests), this creates a systematic bias against cell-type-restricted genes. The approach presented here overcomes this limitation by evaluating gene prevalence in a cell-type-aware manner, thereby preserving important features that are masked in global analyses.

Finally, our approach enabled the development of streamlined, sample-specific analytical scripts. By tailoring the computational logic to each subset of interest, we avoided the pitfalls of overly generalized pipelines that often lead to data loss or misclassification. As a consequence, we observed a likely reduction in total processing time, encompassing code development, testing, and final data analysis. This reduction is primarily due to the fact that the bioinformatician is no longer required to engineer complex scripts to handle large, heterogeneous datasets—datasets that, under our strategy, are preemptively filtered to retain only biologically relevant signals.

4 Discussion

This study introduces a statistically optimized and computationally efficient methodology for the analysis of single-cell RNA sequencing (scRNA-Seq) data. The principal objective of this approach is to strike a balance between analytical robustness and reduced computational burden, thereby enabling high-resolution transcriptomic analyses even in environments lacking advanced computational infrastructure.

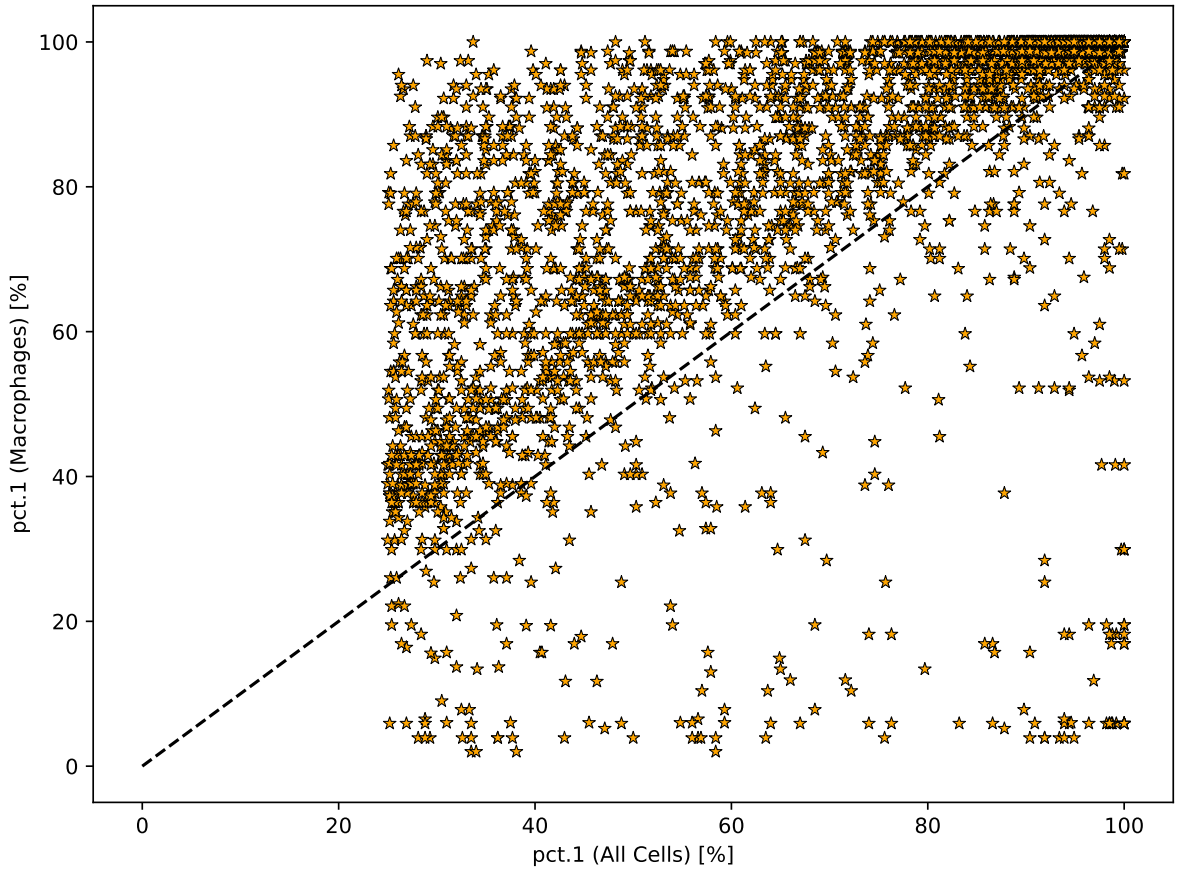

Figure 3: Scatterplot comparing gene expression percentages (*pct.1*) between all cells and macrophages for genes shared across both datasets. The dashed diagonal line represents parity in expression levels across conditions. Genes above the line are more highly expressed in macrophages.

By minimizing the reliance on complex processing pipelines or high-performance computing clusters, the method seeks to democratize access to state-of-the-art single-cell data interpretation across diverse research settings.

When applied to the selected dataset, the proposed methodology demonstrated a consistent and substantial improvement in the identification of marker genes, yielding an average increase of approximately 26.89% compared to traditional workflows implemented in Seurat [21, 1]. This enhancement is primarily attributed to the algorithm's improved sensitivity in detecting subtle yet biologically meaningful transcriptional signatures, particularly within rare or transcriptionally plastic cell populations that are often underrepresented or masked by global thresholds in canonical pipelines– such as macrophage-specific genes.

Beyond its analytical strengths, the methodology also offers significant practical and economic benefits. By allowing for the extraction of comprehensive transcriptomic profiles from a single sequencing run, regardless of cell-type abundance, it eliminates the necessity for multiple targeted sequencing efforts. This is particularly advantageous in the context of large-scale studies or longitudinal projects constrained by financial or logistical limitations. The approach thus enhances the scalability of scRNA-Seq experiments while simultaneously promoting greater reproducibility and data comparability across experiments.

The results obtained in this study not only validate the methodological framework but also underscore its potential for further refinement and expansion. Future directions may include algorithmic parallelization to facilitate more efficient handling of ultra-large datasets, as well as the integration of GPU-based acceleration to optimize runtime performance.

Furthermore, the methodology supports the intelligent pre-filtering of cell populations not directly relevant to the primary research question. By enabling the exclusion of biologically irrelevant subsets prior to full-scale analysis, the method enhances both computational efficiency and interpretative clarity. This targeted data reduction strategy is particularly beneficial in studies with specific cell-type foci, allowing researchers to allocate resources more effectively and draw more precise biological inferences.

In sum, the methodology presented here represents a significant advancement in the field of single-cell transcriptomics, offering a powerful combination of analytical sensitivity, computational efficiency, and practical applicability. When used in conjunction or as a complement to established platforms like Seurat, it has the potential to elevate the resolution, scalability, and biological interpretability of scRNA-Seq studies.

5 Conclusion

This study proposed and validated an alternative approach for the analysis of scRNA-seq data, aiming to reduce computational complexity, enhance analytical sensitivity, and preserve the biological integrity of the dataset. The methodology is centered on CLR normalization, followed by a marker-driven subsetting process that isolates specific cellular populations prior to the application of canonical analysis pipelines.

The results demonstrated that this strategy enables the identification of a significantly higher number of marker genes compared to conventional workflows, while also capturing intermediate transcriptional states that are often overlooked. The observed issue of signal shadowing—stemming from limitations in floating-point precision after normalization—highlights the importance of early intervention in the analytical pipeline to prevent critical information loss.

Additionally, the proposed workflow offers advantages in terms of cost-effectiveness and operational feasibility, allowing the analysis of entire tissues from a single sequencing run. This reduces the need for high-performance computational infrastructure and lowers experimental costs, making the approach especially suitable for resource-constrained environments or large-scale studies.

In conclusion, the presented strategy represents a viable, scalable, and scientifically robust alternative for the investigation of dynamic cellular populations, particularly in contexts characterized by high cellular heterogeneity. Future developments may focus on optimizing the algorithm for parallel processing or GPU acceleration, further extending its applicability to large and complex datasets.

6 Declarations

This work was carried out with the support of the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES) – Funding Code 001, whose contribution was essential for the development of the research activities herein reported. Additionally, we acknowledge the assistance of artificial intelligence tools in refining the composition and coherence of this text, which helped to enhance its overall clarity and flow. Specifically, these tools were used exclusively for correcting grammatical, syntactical, and orthographical errors, as well as improving the text's coherence and cohesion. We emphasize that no knowledge generation or content creation was performed by AI, and all research findings and conclusions presented in this work are the result of human investigation and analysis.

References

[1] Asif Adil, Vijay Kumar, Arif Tasleem Jan, and Mohammed Asger. Single-cell transcriptomics: current methods and challenges in data acquisition and analysis. *Front. Neurosci.*, 15:591122, 2021.

[2] Constantin Ahlmann-Eltze and Wolfgang Huber. Comparison of transformations for single-cell rna-seq data. *Nat. Methods*, 20(5):665–672, 2023.

[3] Nicolas L. Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic rna-seq quantification. *Nat. Biotechnol.*, 34(5):525–527, 2016.

[4] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, 36(5):411–420, 2018.

[5] Clayton M. Carey, Hailey L. Hollins, Alexis V. Schmid, and James A. Gagnon. Distinct features of the regenerating heart uncovered through comparative single-cell profiling. *bioRxiv*, 2023.

[6] Ardelio Galletti and Antonio Maratea. Numerical stability analysis of the centered log-ratio transformation. In *Proc. Int. Conf. Signal-Image Technol. Internet-Based Syst. (SITIS)*, pages 713–716, 2016.

[7] Emmanuel L. Gautier, Tan Shay, Jason Miller, et al. Gene-expression profiles and transcriptional regulatory pathways that underlie the identity and diversity of mouse tissue macrophages. *Nat. Immunol.*, 13(11):1118–1128, 2012.

[8] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome Biol.*, 20(1):296, 2019.

[9] Bo Hu, Sara Lelek, Bastiaan Spanjaard, et al. Origin and function of activated fibroblast states during zebrafish heart regeneration. *Nat. Genet.*, 54(8):1227–1237, 2022.

[10] Benjamin Kaminow, Dinar Yunusov, and Alexander Dobin. Starsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus rna-seq data. *bioRxiv*, page 2021.05.05.442755, 2021.

[11] Peter V. Kharchenko. The triumphs and limitations of computational methods for scrna-seq. *Nat. Methods*, 18(7):723–732, 2021.

[12] Ilya Korsunsky, Nathan Millard, Jin Fan, et al. Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. Methods*, 16(12):1289–1296, 2019.

[13] Lei Li, Meina Lu, Lidong Guo, Xuejiao Zhang, Qun Liu, Meiling Zhang, Junying Gao, Mengyang Xu, Yijian Lu, Fang Zhang, et al. An organ-wide spatiotemporal transcriptomic and cellular atlas of the regenerating zebrafish heart. *Nat. Commun.*, 16(1):3716, 2025.

[14] Rebeca Bosso dos Santos Luz, André Guilherme Portela Paula, et al. Macrophages and cardiac lesion in zebrafish: what can single-cell rna sequencing reveal? *Front. Cardiovasc. Med.*, 12:1570582, 2025.

[15] Hong Ma, Ziqing Liu, Yuchen Yang, et al. Functional coordination of non-myocytes plays a key role in adult zebrafish heart regeneration. *EMBO Rep.*, 22(11):e52901, 2021.

[16] Evan Z. Macosko, Anindita Basu, Rahul Satija, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[17] Leland McInnes, John Healy, and James Melville. Umap: uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[18] Aaron M. Newman, Christopher B. Steen, Claire L. Liu, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.*, 37(7):773–782, 2019.

[19] Felix Raimundo, Celine Vallot, and Jean-Philippe Vert. Tuning parameters of dimensionality reduction methods for single-cell rna-seq analysis. *Genome Biol.*, 21(1):212, 2020.

[20] Laura Rolland, Alenca Harrington, Adèle Faucherre, et al. The regenerative response of cardiac interstitial cells. *J. Mol. Cell Biol.*, 14(10):mjac059, 2022.

[21] Rahul Satija, Jeffrey A. Farrell, David Gennert, Alexander F. Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, 33(5):495–502, 2015.

[22] Avi Srivastava, Laraib Malik, Hirak Sarkar, and Rob Patro. A bayesian framework for inter-cellular information sharing improves dscrna-seq quantification. *Bioinformatics*, 36(Suppl 1):i292–i299, 2020.

[23] Avi Srivastava, Laraib Malik, Tom Smith, Ian Sudbery, and Rob Patro. Alevin efficiently estimates accurate gene abundances from dscrna-seq data. *Genome Biol.*, 20(1):65, 2019.

[24] Tim Stuart, Andrew Butler, Paul Hoffman, et al. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21, 2019.

[25] Cole Trapnell, Davide Cacchiarelli, Joseph Grimsby, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, 32(4):381–386, 2014.

[26] Beibei Wang, Fengzhu Sun, and Yihui Luan. Comparison of the effectiveness of different normalization methods for metagenomic cross-study phenotype prediction under heterogeneity. *Sci. Rep.*, 14(1):7024, 2024.

[27] Xiuming Wang, James Park, Katalin Susztak, Nan Zhang, et al. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.*, 10:380, 2019.

[28] Florian A. Wolf, Philipp Angerer, and Fabian J. Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biol.*, 19(1):15, 2018.

[29] Jing Xue, Susanne V. Schmidt, Johanna Sander, et al. Transcriptional programming of human macrophages by distinct environmental stimuli. *Cell*, 157(3):659–673, 2014.

[30] Grace X. Zheng, Jason M. Terry, Phillip Belgrader, et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, 8:14049, 2017.