A holistic approach for single-cell data trajectory inference using chromosome physical location and ensemble random walk

Jovany Cardoza-Aguilar, Caleb Milbourn, Yifan Zhang, Lei Yang, Sergiu M. Dascalu, and Frederick C. Harris, Jr.

***Abstract*—Single-cell RNA sequencing technology enables the analysis of complex, heterogeneous cell samples. However, errors in data processing, dimension reduction, and clustering can negatively impact subsequent calculations, particularly when inferring cell trajectories using graph methods. We proposed a novel method for single-cell data Trajectory Inference using Chromosome physical location and ensemble Random Walk (scCRW). It utilizes entire chromosomes and their gene identifiers to enhance factor analysis, providing a more comprehensive view of biological processes. For trajectory inference, scCRW employs a random walk, which has been evaluated against other state-of-the-art methods using real single-cell RNA-seq datasets. These datasets include both linear and nonlinear data, showcasing scCRW's capabilities in pseudotime and trajectory inference tasks. The results demonstrate that scCRW consistently achieves top or near-top correlation scores and excels in nonlinear metrics such as F1 branches and milestones. This approach provides accurate trajectory inference that closely aligns with ground truth, highlighting the utility of using chromosomes in factor analysis and random walk techniques for more precise data analysis.**

***Index Terms*—Single-cell, Chromosome, Trajectory Inference, Pseudotime Inference, Clustering, RNA-sequencing, Factor Analysis, Random Walk, Data Processing, Deep Learning.**

I. Introduction

RNA sequencing [1] unveils real-time cellular processes within individual cells at the time of sampling. It yields data highlighting active gene expressions and their levels, enabling diverse biological research on single-cell samples. This data managed via a gene expression matrix, can be computationally modeled using trajectory inference methods. Consequently, it facilitates the creation of graphs depicting cell stage progression and pseudotime, guided by the similarities among single-cell expression patterns.

The datasets obtained through this procedure are frequently extensive and necessitate the use of advanced computational tools and statistical techniques for thorough analysis. A comprehensive workflow involves various components, including data preprocessing, clustering, and creating graphs that illustrate the various developmental stages of cells in a dataset. In this paper, we focus on enhancing the feature selection and graph construction procedure of the trajectory inference. These refinements to the workflow components aim to facilitate the identification of cell stages and the estimation of "pseudotime," a metric indicating a cell's progression through these stages. This enhances our understanding of cell development by providing deeper insights.

Numerous efforts [2] have been made to analyze single-cell RNA sequence data. The minimum spanning tree (MST) has found application in various trajectory inference methods for constructing graphs. An early method that employed MST for calculating pseudotime in cells was Monocle [3], which achieved this by identifying the longest path between each cell. Similarly, Tools for Single Cell Analysis (TSCAN) [4] utilized MST for trajectory inference, but it grouped cells into clusters before applying MST, rather than at the single-cell level. TSCAN assigned cells to the edges of this trajectory and computed pseudo time, significantly reducing computational time. Slingshot [5], on the other hand, amalgamates elements from these methods while introducing some unique modifications. Slingshot initially clusters cells together and then applies MST to the clusters, akin to TSCAN. Subsequently, Slingshot employs a principal curves algorithm to smooth the trajectory, assigning cells to the principal curve and calculating pseudo time.

However, one prominent challenge for inferring trajectories of individual cells lies in their capacity to effectively handle large datasets comprising hundreds of thousands of single cells. The execution time for these methods significantly increases in such scenarios, leading to a detrimental effect on the accuracy of the generated output. Issues related to stability emerge, as these methods should ideally yield similar results when given similar input data across different runs. However, some of these methods exhibit variance in their outputs, which might benefit from potential reduction.

The recently introduced Single-cell data Trajectory inference method using Ensemble Pseudotime inference (scTEP) [6] offers a approach to RNA sequencing data analysis, which we aim to extend and enhance. scTEP seeks to advance existing methods by incorporating pathway information that groups related genes together and allowing users to specify a starting point for trajectory inference, given that root cells are typically known to users. The scTEP process begins with dataset preprocessing, involving the removal of missing

Y. Zhang, L. Yang, S. Dascalu, and F. Harris, Jr. are with the Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, 89557. E-mail: yfzhang@nevada.unr.edu, {leiy, dascalus, Fred.Harris}@cse.unr.edu

Jovany Cardoza-Aguilar is with the Department of School of Engineering and Technology, University of Washington Tacoma, Tacoma, WA, 98402. E-mail: jca3829@uw.edu

Caleb Milbourn is with the Department of Biology and Biochemistry, University of Houston, Houston, TX, 77004. E-mail: ctmilbourn@gmail.com

Manuscript received October 24, 2023; revised October, 24, 2023, 10:41:35 AM.

values and non-expressed genes. Subsequently, the gene expression matrix undergoes filtration using various gene sets linked to common pathways. Factor analysis is then applied to generate submatrices for each pathway. scTEP utilizes the scDHA [7] package for cell clustering and dimension reduction, marking a significant departure from previous methods by deriving pseudotime from multiple clustering outcomes. To conclude, the graphing stage of scTEP employs an MST and enhances the graph by arranging vertices based on their average pseudotime.

We introduce single-cell data trajectory inference using Chromosome physical location and ensemble Random Walk (scCRW), specifically feature selection and trajectory inference. In the current scTEP method, the focus during data feature selection is rather limited, primarily concentrating on specific pathways. Our proposal introduces an approach that expands the analysis to encompass the entirety of the chromosomes within the datasets. This broader perspective provides a more holistic view of the dataset, granting a deeper understanding of the system and enhancing flexibility when working with diverse biological inputs in the future. The second substantial change we recommend for the scTEP method involves replacing the current employment of the MST with an ensemble random walk algorithm. Utilizing an ensemble random walk algorithm allows for more accurate capture of the inherent complexity and interconnectivity between cells. This is due to the nature of a random walk [8], which is based on the likelihood of transitioning from one cell to another, providing a more precise representation of cellular progression through various stages.

These proposals to the scTEP method aim to address its limitations arising from a restricted focus and deterministic trajectory inference. They offer a more adaptable and extensible approach that future researchers can build upon, thereby overcoming these constraints.

II. Method

In this section, we begin by presenting the comprehensive structure of the proposed pipeline, followed by a detailed discussion of its constituent components. Figure 1 illustrates the complete workflow of the scCRW, which comprises four main parts: (a) Chromosome feature selection, (b) clustering and dimension reduction, (c) pseudotime inference, and (d) ensemble random walk trajectory inference.

A. Chromosome physical location feature selection

Dataset filtering is a common step in many workflows, but it often lacks substantial input from the actual biological processes, potentially resulting in gaps in the analysis. There are methods such as scTEP that seek to mitigate this by applying initial gene filtering through multiple gene sets, each consisting of intersected genes. While this approach incorporates biological insights, it may be criticized for its limited focus on specific pathways. In response, our approach takes a broader view by considering each chromosome, enabling us to reconstruct the dataset from a more comprehensive biological perspective.

There is a connection between histones and gene expression, with the ability to predict gene expression levels of a cell type. Karlic et al. [9] has evaluated and demonstrated the dual functionality of histones, and we take advantage of this relationship by grouping genes according to their chromosomal location. The chromosome data that will be used for the evaluation of scCRW originates from a meticulously curated and extensive dataset, which includes over 50,000 gene identifiers. The individual chromosomes are distinct gene sets in the form of an expression matrix containing tens of thousands of genes. It's essential to emphasize that this study exclusively relies on chromosomal information for its analysis. The foundational data for this compilation was sourced from publicly available files hosted on The Jackson Laboratory's Informatics website [10].

An initial data preprocessing step is carried out to enhance the method's performance. This process is depicted in Figure 1(a) of our workflow. The input for the scCRW is chromosomal data that will be in the form of an $m \times n$ matrix, with n representing the genes on m cells. We then normalize the single-cell datasets, while taking measures to be mindful on the scale of genes, as genes with larger scales could become dominant in comparison to other genes. We used a logarithmic transformation (base 2) to rescale the raw expression counts, ensuring that the resulting range of gene expressions is smaller than 100. Additionally, we exclude genes expressed in only a few cells, as their contribution to the data is nearly negligible. Removing such genes not only reduces computation time but also has a limited impact on the overall performance of the method.

The next step of the workflow involves the exclusive filtering of genes situated on the initial chromosome. We select the initial chromosome and its corresponding gene set, and intersect the genes in the expression matrix with each chromosome's gene sets in the process creating an intersected gene expression matrix for each chromosome. There will likely be chromosomes that have little to no genes that are shared in the gene expression matrix of the initial chromosome. We resolve this issue by having a set threshold of 10 genes or less for a chromosome to be marked for removal. This step results in the creation of a gene expression submatrix for each gene set within the chromosome. However, It's important to note that each chromosome may have different ranges of scale having varying numbers of genes associated with it. A larger intersected gene expression submatrix could dominate over a smaller one.

To address this, we generate a latent representation of each chromosome from its gene expression submatrix using the psych [11] package. We use the factor analysis function of the psych package to conduct factor analysis on all the chromosome's gene expression matrices. This process produces a two-dimensional data representation for every chromosome an overall reduction of dimensions, while retaining information. We concatenate the chromosomes into a single whole matrix, with a dimension that will be double the number of remaining chromosomes. The resulting values typically fall within the range of -5 to 5, with very few outliers. We can use this range as a threshold for the safe removal of any results that fall

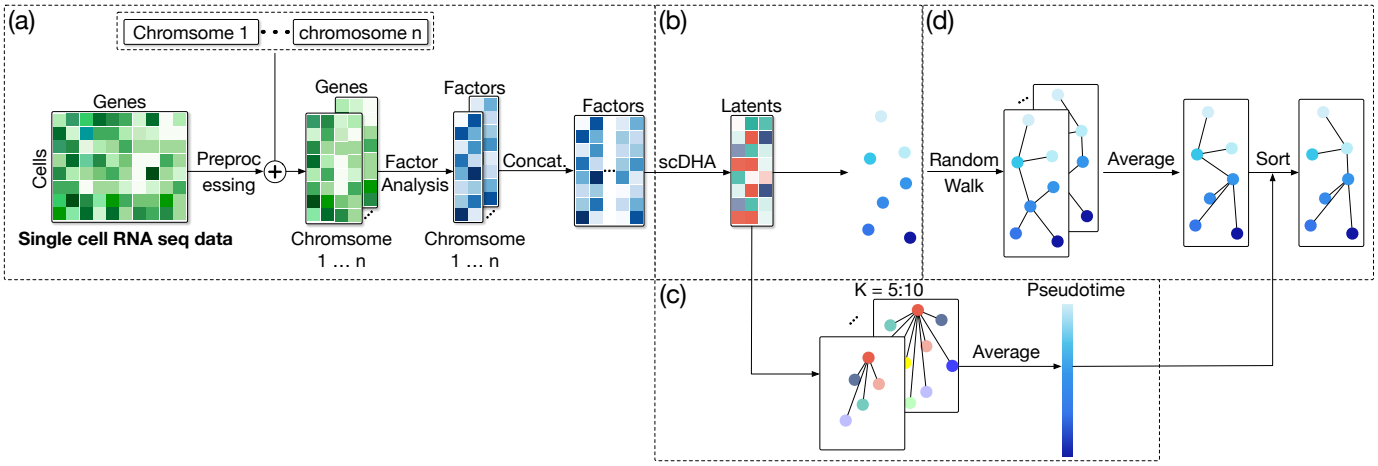

Fig. 1. The workflow of our proposed single-cell data trajectory inference using Chromosome physical location and ensemble Random Walk (scCRW). It consists of four parts: (a) Chromosome feature selection, (b) Dimension reduction and clustering through the use of scDHA, (c) Ensemble pseudotime inference that performs multiple clustering results for a more robust pseudotime of cell clusters, and (d) The construction of a trajectory through the use of an ensemble random walk algorithm generating multiple trajectory graphs, averaging them into a single final graph, and sorted by the average of clusters pseudotime

outside this specified value range.

Through the use of these techniques to prune the vast dataset that encompasses an excess of 25,000 genes, we have significantly reduced the dimension of the gene expression and its total gene count. This significant reduction in computational requirements greatly benefits the subsequent stages of the workflow.

B. Dimension reduction and clustering

The dimension reduction and clustering technique we've integrated into our method have demonstrated remarkable performance when coupled with the use of single-cell Decomposition using Hierarchical Autoencoder (scDHA). As illustrated in Figure 1(b) within our workflow, scDHA comes into play after the factor analysis. scDHA was developed to address the challenge of efficiently handling the extensive data and noise inherent in single-cell RNA sequencing, yielding representative data from each cell. The scDHA pipeline consists of two core models. The first module is a non-negative kernel autoencoder that eliminates genes with a negligible contribution based on the encoder's weight distribution, thus providing a non-negative data representation. This data is then processed through a stacked Bayesian self-learning network based on the Variational Autoencoder (VAE), which takes and decodes the data produced by the non-negative kernel autoencoder, projecting it into a low-dimensional space, often referred to as the latent space.

The scDHA approach represents data in an informative and compact manner, facilitating excellent performance in the analysis of single-cell data with high accuracy and time efficiency for dimension reduction and clustering. The scDHA package plays a pivotal role in our workflow as it excels in handling dimension reduction and clustering steps, outperforming other potential methods for dimension reduction and clustering. Attempts to replace scDHA with dimension reduction algorithms like PCA and UMAP, or implement the K-means clustering algorithm into scCRW, result in inferior outcomes, leading to an accuracy decline in trajectory inference. The reason for this performance degradation in scCRW is twofold, tied to its integration within the scCRW framework. As demonstrated in Figure 1(c), scDHA is utilized by running it six times with varying values for the parameter 'k,' which represents the cluster number, set from 5 to 10 to cluster all the cells into 'k' clusters. scCRW leverages the results from these multiple clustering runs to create a more robust ensemble pseudotime for cells than would be achieved by running it only once. Figure 1(b) showcases its second application, generating the latent space and clustering results with the help of automatically detected cluster numbers from intersected factors, which are then utilized by scCRW to construct the output graph.

C. Pseudotime inference

Pseudotime inference stands as a pivotal step in our trajectory inference process. As depicted in Figure 1(c), our approach to ensemble pseudotime inference differs significantly from the methods employed by other approaches. The majority of methods typically begin by establishing a trajectory first and then using that trajectory to infer pseudotime. The approach of using a trajectory to infer pseudotime is used in the slingshot method, which initiates by constructing an MST graph based on cell clusters to identify the number of lineages and their branch points. It employs simultaneous principle curves to smooth out the lineages represented by the MST and projects cells onto the principal curves. Subsequently, the slingshot computes the pseudotime of cells by measuring the arc length from the start point to the projected points on the principal curve of cells. While this procedure may offer improved results compared to using an MST alone, it hinges on the generation of an accurate MST graph, and errors in the MST graph can significantly impact the pseudotime. Other methods, like Monocle3, have attempted to mitigate these issues by

learning a principal graph in a low-dimensional space and calculating pseudotime by geodesic distance. These endeavors may enhance the accuracy of MST graph construction, but the reliance on dimension reduction and clustering remains challenging.

In our approach, we aim to circumvent this challenge by inferring the pseudotime of cells first and then incorporating the generated pseudotime into the trajectory inference process. This approach hinges on a key assumption that the closer two cells are to each other on the trajectory, the more similar their gene expression profiles. Our testing reveals that this holds true in a low-dimensional space generated by a dimension reduction algorithm, allowing us to assume that cells belonging to the same developmental state share a similar latent space in a low-dimensional context.

We have confirmed this assumption by conducting tests with true cell types, showing that we can accurately determine pseudotime from true labels. This involves selecting a start group and calculating distances between this group and other cell groups to determine pseudotime. Importantly, we can accurately infer pseudotime solely from the true cell type label. However, when we attempt to use clustering results to accomplish the same task, we observe a significant decrease in pseudotime inference accuracy compared to using true cell types. This discrepancy can be attributed to errors in the clustering method, as cells may be assigned to incorrect clusters due to the limitations of the clustering method. Additionally, the clustering method struggles to infer the number of cell types, further contributing to a decrease in clustering accuracy, which ultimately leads to the construction of an inaccurate graph and a reduction in pseudotime inference accuracy.

To address this issue, we propose employing multiple scDHA clustering results at various resolutions, ranging from coarse scale (e.g., 5 clusters) to finer scales (e.g., 10 clusters) to enhance pseudotime inference accuracy. This method requires one or more cells as starting points to identify the initial cluster. The pseudotime inference algorithm sets scDHA to a clustering result with ‘k’ equal to 5. Using the provided starting points, we identify the starting cluster and assign its cells a pseudotime of 0. We then calculate the Euclidean distance between the center of the starting cluster and other cell clusters, assigning pseudotime to cells based on their respective distances to their corresponding cell cluster. This process is repeated for ‘k’ values ranging from 5 to 10, resulting in six pseudotime values for each cell. Finally, we aggregate these six pseudotime results and divide them by six to generate the final pseudotime value. It’s important to note that increasing the maximum ‘k’ value beyond a certain point offers limited benefits while incurring a higher computational cost.

D. Trajectory inference

The low-dimensional representation of single cells is derived from the single-cell expression data processed through earlier stages in the workflow. The objective of graph construction is to depict the relationships between individual cells and their developmental trajectories, organizing them according to their pseudotime. scTEP endeavors to achieve this by employing an MST that connects the clusters formed during the clustering phase, with edges symbolizing connections between the nodes. While this method does achieve the goal of graph construction, there is room for improvement. An alternative approach that we propose involves the utilization of an ensemble random walk algorithm to trace the developmental trajectory of cells. This approach offers enhanced capability for detecting branching points and is more robust against noise in the datasets.

This stage in our workflow entails leveraging the latent representations generated by scDHA to deduce the cellular trajectories. We initiate this process by computing the centers of the clusters, which are depicted as vertices on a graph representing the centroids of cells corresponding to these clusters. Subsequently, we calculate a distance matrix using the Euclidean distance based on these cluster centers. This distance matrix then serves as an adjacency matrix to establish a fully connected directed graph. The weights of the edges connecting the vertices on the graph are determined by the Euclidean distances between pairs of vertices, with the average pseudotime serving as an attribute of the vertices.

Now, we can initiate our random walk. We commence from a specified root vertex, and the method conducts random walks using the “random_edge_walk()” function, integrated from the igraph [12] package. The algorithm is inclined to move towards nearby cells rather than those farther away. If, during the random walk, the algorithm visits a particular vertex more frequently than what can be expected by chance, that vertex is considered a potential next starting point. Multiple potential starting vertices may emerge from this step, indicating the possibility of branching trajectories. The method then proceeds recursively through the graph based on the next starting vertex, eliminating previously visited vertices and those identified as possible starting points, thus creating a sub-graph for traversal. If there were multiple potential starting vertices, the method traverses sub-graphs created at each selection point. This process repeats until all vertices have been visited. The method eventually returns the trajectory or trajectories, amalgamating into a single graph that may contain multiple edges pointing to the same vertex due to the possibility of being visited multiple times from different potential starting vertices. In such cases, the edge closest to the vertex being pointed to is retained, while others are discarded. This process of running the random walk algorithm to generate a graph is repeated 10 times to obtain ten graphs with their own trajectories based on the random walk algorithm. These ten graphs are then averaged into a single final graph that will be fine-tuned with nodes being sorted based on their pseudotime obtained from the average of all cells in the cluster.

The trajectory resulting from the random walk is subsequently refined by a sorting algorithm that arranges the graph’s vertices based on their pseudotime. The algorithm commences with the vertex designated as the root vertex and proceeds to locate its neighbors along with the vertices on the graph in pursuit of the node with the minimum pseudotime. If the root vertex is not the one with the minimum pseudotime, the two vertices are swapped, making the minimum vertex the new

TABLE I
THE TRAJECTORY INFERENCE CORRELATION RESULTS OF LINEAR DATA SETS

Methods	Proposed	scTEP	Slingshot	TSCAN	SCORPIUS	PAGA	Monocle3
Kowalczyk	**0.84**	0.78	0.79	-0.72	0.78	0.69	0.61
Han	**0.75**	0.73	0.69	-0.76	0.39	0.66	0.19
Manno	**0.82**	0.68	0.77	-0.76	0.81	0.41	0.42
Yuzwa	0.6	**0.66**	0.41	0.62	-0.66	**0.66**	0.35
Mean	**0.696**	0.67	0.574	-0.388	0.274	0.544	0.394

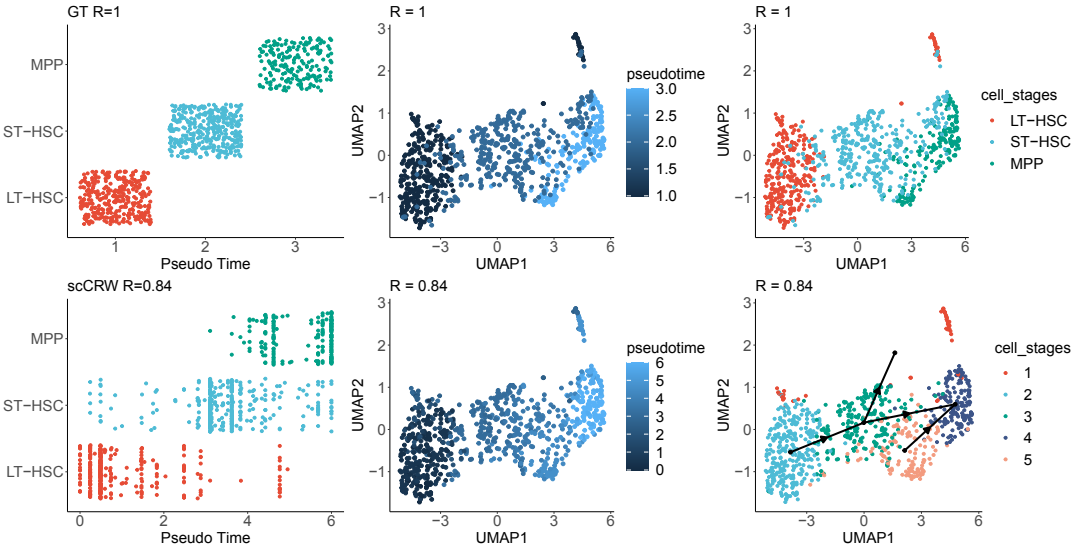

Fig. 2. The visual representation of Kowalczyk's results. The top row displays the ground truth, while the bottom row showcases the scCRW output.

root. This process is then iteratively applied to sub-graphs with the root vertex removed, and is continued until there are no vertices left. These sub-graphs are then combined into a single graph. This process ensures the creation of a polished and precise representation of cell trajectories in the final graph.

III. RESULTS AND DISCUSSION

Analyzing single-cell RNA-seq data presents numerous challenges, with a prominent issue being the substantial level of variability. This variability can stem from various sources, potentially impacting single-cell data, underscoring the significance of employing robust techniques capable of processing and managing any noise inherent in the data.

We assess the capability and resilience of scCRW by utilizing datasets encompassing both linear and nonlinear single-cell data. We compare the outcomes achieved by scCRW to those of other state-of-the-art methods, highlighting the advancements made by scCRW over previous approaches. The methods selected for this comparison include scTEP, Slingshot, TSCAN, SCORPIUS [13], PAGA [14], and Monocle 3 [15].

A. Linear data sets

Table I presents the outcomes for the linear datasets, utilizing a correlation metric to quantify the similarity between each method's cell geodesic distances from the starting point within the milestone network and the ground truth. Notably, scCRW outperforms other methods with a mean correlation of 0.696, securing the top position. scTEP follows closely with a mean correlation of 0.67. Given that scCRW builds upon scTEP, a similar mean score is expected. Slingshot and PAGA yield similar results, occupying the third and fourth positions with mean scores of 0.574 and 0.544, respectively. Both methods achieve top results in individual data samples. Monocle3 and SCORPIUS exhibit more significant drops in correlation values, with scores of 0.394 and 0.274, respectively. Lastly, TSCAN records a mean correlation value that is negative (-0.388) and displays negative correlation values in all data samples except one, where it secures the second-highest correlation value. The results demonstrate a clear improvement for scCRW, as it consistently achieves top correlation values over previous methods.

Figure 2 provides a visual representation of the results for the Kowalczyk datasets. The visualization can be separated into three columns. The left column depicts the pseudotime of cells in three groups "MPP", "ST-HSC", and "LT-HSC" with scCRW having pseudotime for cells that are often consistent to the ground truth, with cells aligning together towards the center of the ground truth pseudotime. Although there are a few instances where some cells appear to be in reverse order, relative to the ground truth. The middle column shows a UMAP visualization with cells pseudotime belonging to three main pseudotime groups with lighter blue cells having higher pseudotime. The pseudotime generated by scCRW having

TABLE II
THE TRAJECTORY INFERENCE CORRELATION RESULTS OF NONLINEAR DATA SETS

Methods	Metrics	Proposed	scTEP	Slingshot	TSCAN	SCORPIUS	PAGA	Monocle3
Macrophage	F1 Branches	0.43	0.40	0.40	0.40	0.40	**0.5**	0.40
	F1 Milestone	0.49	0.41	**0.57**	0.37	0.48	0.46	0.46
	Correlation	**0.38**	0.20	0.03	0.008	0.03	0	0.16
NKT	F1 Branches	0.59	0.45	0.50	0.38	0.38	**0.71**	0.32
	F1 Milestone	0.44	0.48	**0.66**	0.40	0.44	0.37	0.28
	Correlation	**0.55**	0.46	0.52	0.27	0.35	0.50	0.22

similar cell clustering pseudotime to the ground truth. The right column illustrates the trajectory inferred by scCRW, and the cells are classified into five groups. The resulting trajectory correctly identifies and follows the main lineage, but there are two additional branches identified by scCRW, one that branches off from cluster 3 and one that originates from cluster 5 and towards cluster 4.

B. Nonlinear data sets

Table II presents the results for the nonlinear datasets, where we utilize a correlation metric, along with F1 branch and F1 milestone metrics. scCRW consistently produces top correlation values for the nonlinear datasets, and its F1 branch and F1 milestone values are often near the top of the rankings. A notable example is the macrophage dataset, where scCRW excels in the correlation metric with a value of 0.38, while the second-highest correlation is 0.20 from scTEP. In terms of the F1 branch and F1 milestone metrics, scCRW records the second-highest top result with values of 0.43 and 0.49, which are in close proximity to the top results of 0.5 for the branch metric by PAGA and 0.57 for the milestone metric by Slingshot. These results underscore the robustness of scCRW in effectively handling various data types.

IV. CONCLUSION

We have introduced a novel framework, scCRW, for pseudotime and trajectory inference of single-cell data RNA sequencing data. scCRW utilization of chromosomes in its factor analysis and random walk for trajectory inference has allowed for more accuracy and robustness in its output. scCRW effectiveness is demonstrated with results produced from the linear and nonlinear datasets. scCRW due to filtering by chromosome, we capture broader gene interactions that might be missed when focusing on individual genes or narrower gene sets. This approach also leaves room for incorporating other biological data in the future, adding flexibility to the analysis workflow.

ACKNOWLEDGMENTS

This material is based in part upon work supported by the National Science Foundation under grant number(s) CNS-1950485, DUE-2142360, OIA-2019609, and OIA-2148788. Any opinions, findings conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] G. Chen, B. Ning, and T. Shi, "Single-cell rna-seq technologies and related computational data analysis," *Frontiers in genetics*, vol. 10, p. 317, 2019.

[2] W. Saelens, R. Cannoodt, H. Todorov, and Y. Saeys, "A comparison of single-cell trajectory inference methods," *Nature biotechnology*, vol. 37, no. 5, pp. 547–554, 2019.

[3] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn, "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells," *Nature biotechnology*, vol. 32, no. 4, pp. 381–386, 2014.

[4] Z. Ji and H. Ji, "Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis," *Nucleic acids research*, vol. 44, no. 13, pp. e117–e117, 2016.

[5] K. Street, D. Risso, R. B. Fletcher, D. Das, J. Ngai, N. Yosef, E. Purdom, and S. Dudoit, "Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics," *BMC genomics*, vol. 19, pp. 1–16, 2018.

[6] Y. Zhang, D. Tran, T. Nguyen, S. M. Dascalu, and F. C. Harris, "A robust and accurate single-cell data trajectory inference method using ensemble pseudotime," *BMC bioinformatics*, vol. 24, no. 1, pp. 1–21, 2023.

[7] D. Tran, H. Nguyen, B. Tran, C. La Vecchia, H. N. Luu, and T. Nguyen, "Fast and precise single-cell data analysis using a hierarchical autoencoder," *Nature communications*, vol. 12, no. 1, pp. 1–10, 2021.

[8] S. V. Stassen, G. G. Yip, K. K. Wong, J. W. Ho, and K. K. Tsia, "Generalized and scalable trajectory inference in single-cell omics data with via," *Nature communications*, vol. 12, no. 1, p. 5528, 2021.

[9] R. Karlić, H.-R. Chung, J. Lasserre, K. Vlahoviček, and M. Vingron, "Histone modification levels are predictive for gene expression," *Proceedings of the National Academy of Sciences*, vol. 107, no. 7, pp. 2926–2931, 2010.

[10] "Mgi data and statistical reports," Jackson Laboratory Informatics, 2023. [Online]. Available: https://www.informatics.jax.org/downloads/reports/index.html

[11] W. Revelle, *psych: Procedures for Psychological, Psychometric, and Personality Research*, Northwestern University, Evanston, Illinois, 2021, r package version 2.1.6. [Online]. Available: https://CRAN.R-project.org/package=psych

[12] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal*, vol. Complex Systems, p. 1695, 2006. [Online]. Available: https://igraph.org

[13] R. Cannoodt, W. Saelens, D. Sichien, S. Tavernier, S. Janssens, M. Guilliams, B. Lambrecht, K. D. Preter, and Y. Saeys, "Scorpius improves trajectory inference and identifies novel modules in dendritic cell development," *Biorxiv*, p. 079509, 2016.

[14] F. A. Wolf, F. K. Hamey, M. Plass, J. Solana, J. S. Dahlin, B. Göttgens, N. Rajewsky, L. Simon, and F. J. Theis, "Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells," *Genome biology*, vol. 20, pp. 1–9, 2019.

[15] J. Cao, M. Spielmann, X. Qiu, X. Huang, D. M. Ibrahim, A. J. Hill, F. Zhang, S. Mundlos, L. Christiansen, F. J. Steemers *et al.*, "The single-cell transcriptional landscape of mammalian organogenesis," *Nature*, vol. 566, no. 7745, pp. 496–502, 2019.