METHOD ARTICLE

REVISED Unsupervised gene selection for predicting cell spatial positions in the *Drosophila* embryo [version 2; peer review: 1 approved, 1 not approved]

Previously titled: An unsupervised learning method for reconstructing cell spatial organization with application to the DREAM Single Cell Transcriptomics Challenge

Yang Chen ¹, Disheng Mao², Yuping Zhang², Zhengqing Ouyang ¹

¹Department of Biostatistics and Epidemiology, School of Public Health and Health Sciences, University of Massachusetts, Amherst, MA, USA

²Department of Statistics, University of Connecticut, Storrs, CT, USA

V2 First published: 19 Feb 2020, 9:124
<https://doi.org/10.12688/f1000research.20446.1>
 Latest published: 07 Jun 2021, 9:124
<https://doi.org/10.12688/f1000research.20446.2>

Abstract

Analyzing single cell RNA-seq data is important for deciphering the spatial relationships, expression patterns, and developmental processes of cells. Combining in situ hybridization-based gene expression atlas images, some works have successfully recovered spatial locations of cells in zebrafish and *Drosophila* embryos. In this article, we describe a highly ranked method in the DREAM Single Cell Transcriptomics Challenge for predicting cell positions in the *Drosophila* embryo. The method performs unsupervised feature extraction to select a small number of driver genes and then uses them to predict gene expression and spatial position of each individual cell. First, hierarchical clustering is used to select a subset of driver genes. Second, the similarity matrix of single cells in the bins of the reference atlas is computed. Based on the similarity matrix, the spatial positions of cells are then determined by hierarchical clustering. This method is evaluated on the cell positions and gene expressions in the DREAM Single Cell Transcriptomics Challenge. The comparison with the “silver standard” suggests that our method is effective in reconstructing the cell spatial positions and gene expression patterns in tissues.

Keywords

Spatial organization, single cell RNA-seq, *Drosophila* embryo, clustering, DREAM challenge

Open Peer Review

Approval Status

	1	2
version 2 (revision) 07 Jun 2021	 view	
version 1 19 Feb 2020	 view	 view

1. **Xianwen Ren** , Peking University, Beijing, China

2. **Mark S. Cembrowski**, University of British Columbia, Vancouver, Canada

Larissa Kraus, University of British Columbia, Vancouver, Canada

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Zhengqing Ouyang (ouyang@schoolph.umass.edu)

Author roles: **Chen Y:** Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation; **Mao D:** Investigation, Methodology; **Zhang Y:** Funding Acquisition, Investigation, Methodology, Resources, Supervision, Writing – Review & Editing; **Ouyang Z:** Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was partially supported by the Faculty Research Excellence Program Award at UConn (to YZ).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2021 Chen Y *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Chen Y, Mao D, Zhang Y and Ouyang Z. **Unsupervised gene selection for predicting cell spatial positions in the *Drosophila* embryo [version 2; peer review: 1 approved, 1 not approved]** F1000Research 2021, 9:124
<https://doi.org/10.12688/f1000research.20446.2>

First published: 19 Feb 2020, 9:124 <https://doi.org/10.12688/f1000research.20446.1>

REVISED Amendments from Version 1

The title is modified to avoid ambiguity.

The abstract is modified to improve the clarity of method summary.

More citations and description of existing methods are added.

All figures and legends are modified to describe the results more clearly.

More details are added on binarization of scRNA-seq data and prediction of spatial gene expression.

More discussions are added on the interpretation of the results as well as the advantage and limitation of the method.

Any further responses from the reviewers can be found at the end of the article

Introduction

The development of single cell RNA sequencing (scRNA-seq) has provided a powerful solution for building a global transcription landscape of all cell types in tissues, finding new cell types, cell lineage tracing, spatial reconstruction, and combining with other omics^{1–9}. Single cells are originally made from dissociated tissues without spatial information, and spatial gene expression pattern is unknown. *In situ* hybridization (ISH) and its variants can detect the spatial information of mRNA transcripts and produce gene expression reference atlas. Using enough marker genes, users can reconstruct the spatial position of single cell RNA-seq data by combining the ISH reference atlas^{10–15}. Some works have also combined sequential fluorescence *in situ* hybridization (seqFISH) and multiplexed error robust fluorescence *in situ* hybridization (MERFISH) with scRNA-seq data to map cell types to the reference atlas^{16–18}.

Recent methods have successfully mapped cells from scRNA-seq data to the spatial positions using dozens of landmark genes^{10–15}. Nikos *et al.* developed a DistMap method for mapping the ~1,300 *Drosophila* embryo cells into ~3,000 bins in the spatial position using 84 marker genes by computing the Matthews Correlation Coefficient (MCC) score between *in situ* gene expression and binarized scRNA-seq data¹⁰. Satija *et al.* mapped 851 cells of zebrafish embryo into 64 bins of the spatial embryo with 47 genes. They used a bimodal mixture model to binarize the scRNA-seq data and predict a cell originated from the bin by posterior probability¹¹. Kaia *et al.* computed the correspondence score and mapped 139 cells into a spatial *Platynereis dumerilii* brain using a set of 98 genes by whole-mount *in situ* hybridization (WMISH). They binarized the scRNA-seq data and defined a special score to characterize the relation between every cell-voxel combination. They suggested that 50–100 *in situ* genes were needed to map cells to spatial position with high confidence using simulated spatial reference atlases¹². Andreas *et al.* also reconstructed spatial single enterocytes along the villus axis in 1-D space using 102 bottom and top landmark genes from single-molecule fluorescence *in situ* hybridization (smFISH¹⁹) for 1,383 single cells. They used the summed expression of

landmark genes to infer the locations along the 1-D coordinates and tried 50 and 20 landmark genes for prediction. The reconstruction errors were relatively low for 50 genes, but the prediction errors increased remarkably using 20 landmark genes¹³. Nitzan *et al.* also proposed a new method named novoSpaRc for spatially mapping the scRNA-seq cells into an existing reference atlas. Based on the optimal transport method, it reported the gene expression and spatial position with at most 2 landmark genes for virtual scRNA-seq data and virtual *Drosophila* embryo¹⁴. The performance of novoSpaRc is still not good for real *Drosophila* embryo when the marker gene number is smaller than 30^{14,15}. In the methods mentioned above, the dimension and resolution of the spatial region, as well as the number of marker genes, are key factors that affect the recovery of the spatial position. Although these methods demonstrated good performance, they do not focus on markers genes selection in which biological information were used to determine the spatial gene expression patterns.

To explore new algorithms that include biological information contained in the genes to determine embryonic patterning, the DREAM Single Cell Transcriptomics Challenge asked participating teams to predict the positions in the embryo of 1,297 cells using the expression patterns of 60 (sub-challenge 1), 40 (sub-challenge 2), and 20 (sub-challenge 3) driver genes from the *in situ* hybridization data²⁰. The challenge endeavors to use less marker genes to infer the spatial locations of cells.

In this article, we assume that cells with similar gene expression will have close spatial distances. By clustering the possible spatial bins for a cell, we select the most possible cluster as the predicted spatial position. The details of our method are the following: First, we use hierarchical clustering for gene selection. Then, we map the cells to spatial positions by computing the mapping similarity between the binarized scRNA-seq and *in situ* hybridization data and cluster the bins. Finally, we select the optimal thresholds by computing the maximum MCC score for spatial gene expression prediction. We detail our solution for the DREAM challenge, and validate the results using the “silver standard” derived from DistMap. The paper is organized as follows: in Methods, we briefly describe the solutions for all the three sub-challenges; in Results, we present the results of the three sub-challenges in the DREAM Single Cell Transcriptomics Challenge; finally, we discuss and summarize our work.

Methods

Dataset

The scRNA-seq dataset is from ~1,000 handpicked stage 6 fly embryos using Drop-seq²¹. It contains both raw and normalized UMI counts with 1,297 cells and 8,924 genes per cell. *In situ* hybridization expression patterns of 84 driver genes are from the Berkeley *Drosophila* Transcription Network Project (BDNTP). The BDNTP reference atlas are binarized. The bin number of one half of the symmetric *Drosophila melanogaster* embryo at stage 6-pre-gastrulation (which is used in this study) is 3,039. The spatial coordinates of these bins are also specified. The dataset files can be downloaded

from the DREAM Single Cell Transcriptomics Challenge after registration with Synapse free of charge (<https://www.synapse.org/#!Synapse:syn16782360>).

We directly use the normalized scRNA-seq data, the *in situ* matrix and the geometry of the embryo. The gene names “*E.spl.m5.HLH*” and “*Blimp.1*” are replaced by “*E(spl)m5-HLH*” and “*Blimp-1*”.

The DistMap “silver standard” is the best available cell position reference, which is determined by calculating the maximum similarity for cells and spatial bins using all the 84 *in situ* driver genes in BDNT. It is shown that the 84 driver genes are sufficient to uniquely and individually label most of the 1,297 cells¹⁰.

Gene selection

We use a hierarchical clustering method to select 60, 40, and 20 driver genes from the 84 genes based on the normalized scRNA-seq data (Figure 1). Specifically, based on the belief that the scRNA-seq gene expression pattern is driven by the driver genes’ expression pattern, we propose to select the essential driver genes based on the information provided by scRNA-seq data. If two genes have high correlation in the scRNA-seq data, we choose only one of them without losing too much of the information. To find the correlated genes, we perform hierarchical clustering on the normalized scRNA-seq data to separate all 84 genes into 60 clusters (with the Euclidean distance and the Mcquitty linkage). The Mcquitty linkage gives more weights for objects in small clusters than those in large clusters in calculating the distance between two clusters. Thus, it is suitable for situations with many small clusters. Since the numbers of clusters are fairly large in the sub-challenge 1 and 2, we opt to use the Mcquitty linkage for distance calculation. In sub-challenge 3, since the total number of clusters is shrunk to 20, which is smaller than sub-challenge 1 and 2, we choose to use the ward linkage in the hierarchical clustering part to obtain larger-sized clusters from the data. After this step, the gene selection process remains the same as sub-challenge 1 and 2.

After getting the clusters, we pick the most representative gene of each cluster by calculating the distance between each member gene and the cluster center based on the Euclidean distance and selecting the closest one.

Binarization of scRNA-seq data

We perform binarization on the normalized scRNA-seq data for the selected genes based on the “binarizeSingleCellData()” function in DistMap (<https://github.com/rajewsky-lab/distmap>). The details of binarization is as follows: for each quantiles from 0.15 to 0.5 with 0.01 as interval, we perform binarization on the scRNA-seq data for 60, 40, and 20 driver genes in sub-challenge 1, 2, 3, respectively. If the gene expression value is larger than the quantile gene expression value, it will be set as 1, otherwise it is 0. Then we compute the difference between the correlation matrix of the binarized scRNA-seq data and that of the *in situ* matrix based on the root-mean

square error. Finally, we select the quantile threshold which has the smallest root mean square error.

Next, to predict the spatial gene expression for all the 84 *in situ* genes, we used a similar method to binarize the remaining 24, 44, and 64 *in situ* genes in sub-challenges with the binarized 60, 40, and 20 binarized gene expression. First, we compute the Pearson correlation matrix C_1 between the selected genes (60, 40, and 20) and the remaining genes (24, 44, and 64) using the normalized scRNA-seq data. Second, we binarize the remaining genes based on a series of quantiles from 0.15 to 0.5 with 0.01 as interval. Furthermore, we calculate the Pearson correlation matrix C_2 between the binarized selected genes and the remaining genes. Finally, we select the best quantile for binarizing the remaining genes by minimizing the root mean square error between the Pearson correlation matrices C_1 and C_2 .

Compute the similarity matrix between cells and bins

Given the binarized scRNA-seq data and the *in situ* matrix, we calculate the similarity matrix between the cells and bins based on the selected driver genes. Here, we denote the number of selected driver genes is n_g . The similarity between a cell c_i ($i \in [1, 1,297]$) and a bin b_j ($j \in [1, 3,039]$) is calculated as follows:

$$p_{ij} = \frac{n_s^{ij}}{n_g}, \quad (1)$$

where n_s^{ij} is the number of the same gene expression values (0 or 1) in the two binarized vectors corresponding to cell c_i and bin b_j , and n_g is the *in situ* matrix for cell c_i and bin b_j .

Select candidate the cell positions based on the similarity matrix

Select top bins. The likelihood of a cell mapped to a bin is determined by the gene expression in the bin and the cell. The bins with the higher similarity are possibly the potential cell position. To generate stable results, we select enough bins (see below) based on the similarity values. Then we use clustering to determine a more refined cell position.

To select the potential bins for predicting cell position, we check the number of bins with high similarity values (> 0.8) in the similarity matrix for each cell from sub-challenge 1, 2, and 3 (Figure 2). The average number of bins with mapping similarity > 0.8 increases as the number of driver genes decreases. In sub-challenge 3, there are more bins with high similarity values produced. The small similarity difference among these bins makes it harder to determine real spatial bins where the cell originated. Thus, as the number of driver genes decreases, the accuracy of bins with high similarity values also decreases. Here, we use the third quartile of the similarity values to select the top bins in sub-challenge 1 with 60 driver genes. We use the first quartile of similarity values in sub-challenge 2 with 40 driver genes. And we use all bins in sub-challenge 3 with 20 driver genes. If the number of selected bins is 0, the bin which has the maximum similarity will be the predicted position. If the number of the selected

Figure 1. Workflow of our spatial position prediction method for scRNA-seq with *in situ* reference gene panel. (1) Use hierarchical clustering (HC) to select 60/40/20 driver genes from the given 84 driver genes based on scRNA-seq data with the Euclidean distance and the Mcquitty (for 60 and 40 genes) and ward linkage (for 20 genes). (2) Apply the binarization method for the selected genes by minimizing root-mean-square error between of Pearson correlation matrix between the binarized scRNA-seq and the *in situ* matrix, and use the same criteria to binarize the remaining 24/44/64 predicted genes. (3) Calculate mapping similarity matrix from single cells to spatial bins based on the proportion of matched gene expression between each cell and each bin. (4) For each cell, select top bins based on mapping similarity values by quantile thresholding. Then do HC on the spatial distance of these top bins and select the number of clusters based on the Silhouette score. The top 10 predicted bins are the ones surrounding the largest cluster center. (5) For each gene, we compute the MCC score between predicted gene expression and binarize gene expression given different thresholds. The optimal threshold is corresponding to the maximum MCC. Finally, we apply the optimal threshold for the *computeVISH()* method in DistMap to predict gene expression. The black points in the scatter plot are the predicted MCC. The red points are Pearson correlation between predicted *in situ* gene expression with *in situ* gene expression.

bins is larger than 100, then only the top 100 bins will be kept based on the similarity values.

Silhouette score to determine the number of clusters of bins

For bins with high similarity values, we need to perform clustering to select the cluster which has the maximum sum of similarity as the predicted cell position. Here, we use hierarchical clustering on the selected bins based on their three-dimensional distance. The cluster number is determined

by the silhouette score, which measures the average distance of a point to other points in its cluster compared to the smallest average distance to other clusters. The silhouette score ranges from -1 to +1. The higher the silhouette score, the closer the point is closer to its own cluster and the farther it is away from other clusters. We use the average silhouette score across all selected bins to select the optimal clustering number. Here, we use the NbClust package²² to perform hierarchical clustering with the “centroid” method. Based on the silhouette score, we obtain the number of clusters.

Figure 2. Compare the number of high similarity values (> 0.8) in the similarity matrix in each bin for sub-challenge 1, 2, and 3. The numbers of high similarity values in sub-challenge 1 and 2 are much smaller than those in sub-challenge 3.

Predict cell positions based on clustering of bins

We compute the sum of similarities for each cluster and select the cluster with the maximum value. Finally, we assign 10 nearest bins to the selected cluster center as the top 10 most possible cell positions.

Predict spatial gene expression based on the similarity matrix

To predict the final spatial *in situ* gene expression, we use the method `computeVISH()` in DistMap. It multiplies the normalized gene expression matrix (gene X cell) with the mapping similarity matrix (cell X bin) and sets a threshold to binarize final spatial gene expression. For sub-challenge 1, 2, 3, we select an optimal threshold. First, we generate a series of thresholds between 0 and 1. For each threshold, we compute the predicted spatial gene expression using `computeVISH()`, then compute the MCC score between the predicted gene expression and the binarized expression of the driver genes. Finally, we select the best threshold based on the largest MCC score to compute the final spatial gene expression.

To estimate the performance of our method to predict spatial gene expression, we compute the Pearson correlation between the *in situ* gene expression and the predicted gene expression.

Performance evaluation

To evaluate the performance of our method, we use the three performance scores in the DREAM challenge (<https://www.synapse.org/#!/Synapse:syn17091286>, https://github.com/dream-sctc/Scoring/blob/master/dream_scoring_clean.R). The first score (score 1) is the primary score to evaluate the precision of the assignment for the single cells. The second score (score 2) is the average of the relative assignment precision over all the single cells which is used when the first scores are equal for two methods. The third score (score 3) is to compare predictions of gene patterns.

Ambiguous cells: If the predicted top 1 and top 2 positions are the same in the DistMap results, the prediction position will be ambiguous, and the cell will not be computed in the score 1–3. In this challenge, the number of ambiguous cells derived from DistMap are 287.

Results

Select driver genes

We calculated the sums of gene expression values in the *in situ* matrix for each bin of the 3,039 bins. Each bin has at least one gene expressed in the *in situ* matrix. It suggests that our selected driver genes cover all bins as at least one driver gene is expressed in each bin. As the number of expressed driver genes decreases when the number of selected driver genes decreases (from 60 to 40 to 20), the prediction accuracy of cell spatial positions also decreases (Figure 3(a)). We also compared the overlaps when varying the number of the driver genes in Figure 3(b). Among the 40 driver genes of the sub-challenge 2, only one driver gene is not in the selected 60 driver genes of the sub-challenge 1. There are 14 overlapped driver genes between the sub-challenge 1, 2 and 3. It suggests our method is consistent on selecting different number of driver genes.

Evaluate the predicted position and spatial gene expression

We used the score 1–3 to evaluate our method for different number of selected driver genes. Figure 4(a) shows the scores of our submitted results for the sub-challenge 1, 2, 3. The blue bar is the score for the silver standard method using 84 driver genes from DistMap. For score 1, the primary score for cell position prediction, the performance of our method is close to the silver standard when using 60 driver genes. The performance decreases when using 40 and 20 driver genes. For score 2, our method shows high scores when using 60 and 40 driver genes and lower score when using 20 driver genes. As Score 2 is

Figure 3. (a) Compare number of expressed driver genes in each bin in the *in situ* matrix in sub-challenge 1, 2, and 3 under different selected driver genes scenarios. (b) The Venn diagram of the selected 60, 40, and 20 gene in the three sub-challenges 1, 2, 3. There are 14 common genes in three sub-challenges 1, 2 and 3 (labeled as gene60, gene40, and gene20).

the average relative precision for all cells, it suggests that our method is robust for predicting the right position using 60 and 40 driver genes, probably because our algorithm uses the top 10 bins around the largest cluster centers. The score 3 estimates the predicted gene expression patterns. It shows a small difference in our method when using 60 and 40 driver genes. To check the effect of threshold on the prediction results, we test our method under different thresholds. Figure 4(b)–(d) shows the consistency of the score 1–3 over a range of thresholds in the scenarios of different numbers of driver genes. Overall, the results show that our method performs well when using 60 and 40 genes.

As shown in Figure 5, we used boxplots to present the spatial gene expression prediction accuracy of the 84 driver genes by computing the Pearson correlation between the predicted and *in situ* gene expression. Consistent with Figure 4(a), the average Pearson correlation of our prediction (60 or 40 driver genes) is higher than that of the silver standard (84 driver genes by DisMap). When the number of driver genes decreases to 20, the average Pearson correlation of our prediction becomes smaller but close to that of the silver standard, which used 84 driver genes. From the results, it suggests that our optimal threshold selection may improve the spatial gene expression prediction. It is feasible to use less genes for spatial gene expression prediction.

Discussion

With high quality *Drosophila* dataset, Karaikos *et al.* mapped single cells to the 3D embryo positions using the DisMap method¹⁰. It assumed that transcriptomes of neighboring cells should show similar patterns in spatial tissues. We further assumed that cells with similar gene expression will have close spatial distance. We applied hierarchical clustering to select a small number of driver genes and group similar cells to be mapped to the 3D embryo. After selecting the largest cluster, we determined the most possible predicted spatial cell position around the cluster center.

We assessed the performance of our method on the selected 60, 40 or 20 driver genes using score 1–3 and compared it with the silver standard (DisMap results). For 60 driver genes, our results show close performance to the silver standard in score 1 and 3, and higher performance in score 2. It suggests that our method applied to 60 driver genes can achieve the same or higher level of performance as the silver standard to predict cell position. Our method obtains better performance for gene expression pattern prediction with 60 and 40 genes. When using 40 driver genes, the score 1 of our method decreases a little but the score 3 is very close to the silver standard. It suggests that our method can still predict gene expression pattern reasonably well using 40 genes. For 20 driver genes, the performance of our method is lower than but still close to the silver standard in both cell position and gene expression prediction.

Overall, our method achieves good performance to predict spatial cell position and gene expression with 60 and 40 driver genes. The primary reason may be that our method selects the most representing genes in the gene clusters for prediction. Also, for potential spatial bins, our method utilizes bin clusters and selects the cluster center as predicted position. The good performance of our method is consistent to our assumption that cells with close spatial position show similar gene expression patterns. The performance of our method decreases when the number of driver genes reduces to 20. This is probably because the similarities between cells and spatial bins have larger variations when measured using smaller number of genes. In this case, it becomes harder to find the most possible bin position based on cell-bin similarities. In Nitzan's work^{14,15}, they successfully reconstructed the virtual *Drosophila* embryo with one marker gene and the structural information between gene expression space and physical space. It suggested that gene-gene correlation and spatial correlation contain useful information for gene expression and spatial cell position prediction. In addition, our method can be extended using an iterative approach by updating the mapping similarity matrix at each step.

Figure 4. (a) Comparing the score 1, 2, 3 for sub-challenge 1 (60 driver genes), 2 (40 driver genes) and 3 (20 driver genes) with the silver standard (84 driver genes). (b-d) Comparing the score 1, 2, 3 using different thresholds (min, 1st quantile, median, mean, 3rd quantile, max value) for (b) sub-challenge 1 with 60 genes; (c) sub challenge 2 with 40 genes; (d) sub-challenge 3 with 20 genes.

Figure 5. Comparing Pearson correlation of predicted spatial gene expression and in situ spatial gene expression for 84 genes between DistMap (84 genes) with our method using (a) 60 genes in subchallenge 1; (b) 40 genes in sub-challenge 2; and (c) 20 genes in sub-challenge 3. For each gene, we calculate the Pearson correlation between the predicted spatial gene expression and the *in situ* spatial gene expression. The average Pearson correlation value of DistMap (84 genes), sub-challenge 1, sub-challenge 2, and sub-challenge 3 are 0.532, 0.622, 0.561, and 0.422, respectively.

Data availability

The dataset associated with the DREAM Single Cell Transcriptomics Challenge is available at <https://www.synapse.org/#!Synapse:syn16782360> and <https://www.synapse.org/#!Synapse:syn18632189>. Limited by the sharing settings of Synapse, users should register in Synapse (free of charge; <https://www.synapse.org/>) using their email address, and agree to the dataset conditions of use (<https://www.synapse.org/#!Synapse:syn15665609/wiki/583240>).

Software availability

Source code implementation for the method presented in this article and used in the DREAM Single Cell Transcriptomics Challenge is available at: https://github.com/ouyang-lab/SCTC-Challenge-zho_team. The scoring scripts are available at: https://github.com/dream-sctc/Scoring/blob/master/dream_scoring_clean.R.

Archived source code as at time of publication can be accessed at: <https://doi.org/10.5281/zenodo.3592532>

License: GLP 3.0

Author information

All authors actively contributed to the results and discussed the procedures. They have been involved in the test, design and evaluation of various approaches; many of the explored techniques were not used in the final workflow due to strong overfitting. DM: gene selection and coding. YC: data pre-processing, clustering, and writeup. ZO: supervision, realization of workflows, and writeup. YZ: supervision, realization of workflows, and writeup.

Acknowledgements

Data used in this publication were generated by Prof. Dr. Nikolaus Rajewsky, Max Delbrück at the Center for Molecular Medicine, and these results were obtained as part of the DREAM Single Cell Transcriptomics Challenge project through Synapse ID (syn15665609). We thank Pablo Meyer and Jovan Tanevski for providing the documents, code and suggestions for score 1–3.

References

- Farrell JA, Wang Y, Riesenfeld SJ, *et al.*: **Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis.** *Science*. 2018; **360**(6392): eaar3131.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Butler A, Hoffman P, Smibert P, *et al.*: **Integrating single-cell transcriptomic data across different conditions, technologies, and species.** *Nat Biotechnol*. 2018; **36**(5): 411–420.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Colomé-Tatché M, Theis FJ: **Statistical single cell multi-omics integration.** *Curr Opin Syst Biol*. 2018; **7**: 54–59.
[Publisher Full Text](#)
- Gu C, Liu S, Wu Q, *et al.*: **Integrative single-cell analysis of transcriptome, DNA methylome and chromatin accessibility in mouse oocytes.** *Cell Res*. 2019; **29**(2): 110–123.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bian S, Hou Y, Zhou X, *et al.*: **Single-cell multiomics sequencing and analyses of human colorectal cancer.** *Science*. 2018; **362**(6418): 1060–1063.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Kulkarni A, Anderson AG, Merullo Dp, *et al.*: **Beyond bulk: a review of single cell transcriptomics methodologies and applications.** *Curr Opin Biotechnol*. 2019; **58**: 129–136.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Savulescu AF, Jacobs C, Negishi Y, *et al.*: **Pinpointing cell identity in time and space.** *Front Mol Biosci*. 2020; **7**: 209.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Stuart T, Satija R: **Integrative single-cell analysis.** *Nat Rev Genet*. 2019; **20**(5): 257–272.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Mayr U, Serra D, Liberali P: **Exploring single cells in space and time during tissue development, homeostasis and regeneration.** *Development*. 2019; **146**(12): dev176727.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Karaïskos N, Wahle P, Alles J, *et al.*: **The *Drosophila* embryo at single-cell transcriptome resolution.** *Science*. 2017; **358**(6360): 194–199.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Satija R, Farrell JA, Gennert D, *et al.*: **Spatial reconstruction of single-cell gene expression data.** *Nat Biotechnol*. 2015; **33**(5): 495–502.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Achim K, Pettit JB, Saraiva LR, *et al.*: **High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin.** *Nat Biotechnol*. 2015; **33**(5): 503–9.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Moor AE, Harnik Y, Ben-Moshe S, *et al.*: **Spatial reconstruction of single enterocytes uncovers broad zonation along the intestinal villus axis.** *Cell*. 2018; **175**(4): 1156–1167.e15.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Nitzan M, Karaïskos N, Friedman N, *et al.*: **Charting a tissue from single-cell transcriptomes.** *bioRxiv*. 2018; 456350.
[Publisher Full Text](#)
- Nitzan M, Karaïskos N, Friedman N, *et al.*: **Gene expression cartography.** *Nature*. 2019; **576**(7785): 132–137.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Shah S, Lubeck E, Zhou W, *et al.*: **In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus.** *Neuron*. 2016; **92**(2): 342–357.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Zhu Q, Shah S, Dries R, *et al.*: **Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data.** *Nat Biotechnol*. 2018; **36**: 1183–1190.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Moffitt JR, Bambach-Mukku D, Eichhorn SW, *et al.*: **Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region.** *Science*. 2018; **362**(6416): eaau5324.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Halpern KB, Shenav R, Matcovitch-Natan O, *et al.*: **Single-cell spatial reconstruction reveals global division of labour in the mammalian liver.** *Nature*. 2017; **542**(7641): 352–356.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Tanevski J, Nguyen T, Truong B, *et al.*: **Gene selection for optimal prediction of cell position in tissues from single-cell transcriptomics data.** *Life Sci Alliance*. 2020; **3**(11): e202000867.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Macosko EZ, Basu A, Satija R, *et al.*: **Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets.** *Cell*. 2015; **161**(5): 1202–1214.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Charrad M, Ghazzali N, Boiteau V, *et al.*: **NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set.** *J Stat Softw*. 2014; **61**(6): 1–36.
[Publisher Full Text](#)

Open Peer Review

Current Peer Review Status:

Version 2

Reviewer Report 17 June 2021

<https://doi.org/10.5256/f1000research.53961.r87000>

© 2021 Ren X. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Xianwen Ren

Biomedical Pioneering Innovation Center, Peking University, Beijing, China

The authors have clarified my concerns.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 23 September 2020

<https://doi.org/10.5256/f1000research.22477.r70705>

© 2020 Cembrowski M et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Mark S. Cembrowski

Dept. of Cellular and Physiological Sciences, University of British Columbia, Vancouver, BC, Canada

Larissa Kraus

Department of Cellular and Physiological Sciences, University of British Columbia, Vancouver, BC, Canada

In this manuscript, the authors describe an approach to infer spatial gene expression using scRNA-seq data. The authors aimed to identify spatial expression of cell clusters using 20, 40 or 60 marker genes, and compared to the “gold standard” Distmap method using 84 marker genes.

The authors begin to explain the rationale of the manuscript by describing the concept of the DREAM challenge, which aims to introduce novel computational techniques to map spatial gene expression using transcriptomic data. Since the DREAM challenge aims to implement novel methods, it is imperative to introduce and carefully compare common methods with the novel proposed approach, which was underdeveloped by the authors. As an example, the “gold standard” method is mentioned several times, though it is not clear what exactly the “gold standard” is or how well-controlled it is. As another key example, a discussion of any kind is completely missing.

The design of figures for a manuscript should be chosen to clearly emphasize the main points of the data. Here, the figures are difficult to interpret because key information and description is missing (e.g., legends are too brief, axis labels are missing, subplot headings are redundant, figure design is inconsistent). In addition, the design and presentation of the figures seem unmotivated and thereby difficult for the reader to see the importance of any particular data or easily draw conclusions from the results. In addition, consistent colour schemes and designs would help the reader to follow the manuscript and understand the results.

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Transcriptomics

We confirm that we have read this submission and believe that we have an appropriate level of expertise to state that we do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 19 Jan 2021

Zhengqing Ouyang, University of Massachusetts, Amherst, USA

Point-by-point response to the comments of Reviewer 2

Reviewer 2

In this manuscript, the authors describe an approach to infer spatial gene expression using scRNA-seq data. The authors aimed to identify spatial expression of cell clusters using 20, 40 or 60 marker genes, and compared to the “gold standard” Distmap method using 84 marker genes.

The authors begin to explain the rationale of the manuscript by describing the concept of the DREAM challenge, which aims to introduce novel computational techniques to map spatial gene expression using transcriptomic data.

Since the DREAM challenge aims to implement novel methods, it is imperative to introduce and carefully compare common methods with the novel proposed approach, which was underdeveloped by the authors. As an example, the “gold standard” method is mentioned several times, though it is not clear what exactly the “gold standard” is or how well-controlled it is. As another key example, a discussion of any kind is completely missing.

We thank the reviewer for the constructive comments. The “gold standard” is generated by DistMap, which is the best available single cell position reference used in the DREAM Challenge. DistMap determines single cell position by searching the maximum similarity for cells and spatial bins using the *in situ* expression of 84 driver genes in the Berkeley Drosophila Transcription Network Project (BDNTP). It is shown that the 84 driver genes are sufficient to uniquely and individually label most of the 1,297 cells (Karaikos, et al. 2017). Thus, it is used in the DREAM Challenge to assess the performance of methods using smaller numbers (60, 40, and 20) of driver genes. In the revised manuscript, we changed “gold standard” to “silver standard”, as the cell positions generated by DistMap are not experimentally validated. We added the description of the DistMap “silver standard” in the Method section.

We also added more introductions to the common methods for the cell position prediction from single cell transcriptome data in the Introduction section. We described the limitation of current methods and explain the motivation to study how the driver gene selection affecting the prediction of cell spatial position.

The comparisons of our methods (Zho_team) and other top performing methods have been described in the DREAM Single Cell Transcriptomics Challenge consortium paper (Gene selection for optimal prediction of cell position in tissues from single-cell transcriptomics data, Life Sci Alliance. 2020 Nov; 3(11): e202000867). Our manuscript is focused on describing the details of our method used in the DREAM Challenge.

We also expanded the Discussion section of our manuscript. In addition to summarizing our method and results, we also pointed that gene-gene correlation and spatial correlation are useful information for spatial cell position and gene expression prediction.

The design of figures for a manuscript should be chosen to clearly emphasize the main points of

the data. Here, the figures are difficult to interpret because key information and description is missing (e.g., legends are too brief, axis labels are missing, subplot headings are redundant, figure design is inconsistent). In addition, the design and presentation of the figures seem unmotivated and thereby difficult for the reader to see the importance of any particular data or easily draw conclusions from the results.

In addition, consistent colour schemes and designs would help the reader to follow the manuscript and understand the results.

We thank the reviewer for the constructive comments. In the revised manuscript, we improved all figures and legends. We enlarged the font sizes for all figures. In Figure 1, we clarified the whole workflow to display the motivation of our method. In Figure 2, we changed it to display the number of bins with high similarity values (> 0.8) in the similarity matrix for each cell when the number of driver genes is 60, 40, or 20. In Figure 3, we enlarged the font size and added more information in the figure labels. In Figure 4, we enlarged the font size. In figure 5, we changed it to the comparison of the Pearson correlation of predicted gene expression from DistMap (using 84 driver genes) and our method (using 60, 40, or 20 genes).

Competing Interests: No competing interests were disclosed.

Reviewer Report 02 September 2020

<https://doi.org/10.5256/f1000research.22477.r69520>

© 2020 Ren X. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Xianwen Ren

Biomedical Pioneering Innovation Center, Peking University, Beijing, China

Reconstructing the spatial information of single cells from single-cell RNA-seq data is a pivotal question to further release the revolutionary power of the scRNA-seq technology. Here the authors propose a computational method to infer the spatial positions of single cells of *Drosophila* embryos based on scRNA-seq data and a reference spatial map based on in situ hybridization of 84 driver genes. While the method is finely tuned for this specific scientific question, some concerns exist.

First, the title is misleading. The authors claimed "an unsupervised learning method". This is not valid because this method uses the reference spatial map of *Drosophila* embryos. Unless the method only uses the scRNA-seq data, it cannot be claimed "an unsupervised learning method".

Second, the current method has only been evaluated on one dataset. More independent evaluations should be added to demonstrate the generalizing capacity of this method.

Overall, this manuscript addresses an important scientific question. But the method needs more justifications and evidence to demonstrate its novelty and power.

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

No source data required

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 19 Jan 2021

Zhengqing Ouyang, University of Massachusetts, Amherst, USA

Point-by-point response to the comments of Reviewer 1

Reviewer 1

*Reconstructing the spatial information of single cells from single-cell RNA-seq data is a pivotal question to further release the revolutionary power of the scRNA-seq technology. Here the authors propose a computational method to infer the spatial positions of single cells of *Drosophila* embryos based on scRNA-seq data and a reference spatial map based on in situ hybridization of 84 driver genes. While the method is finely tuned for this specific scientific question, some concerns exist.*

*1. First, the title is misleading. The authors claimed "an unsupervised learning method". This is not valid because this method uses the reference spatial map of *Drosophila* embryos. Unless the*

method only uses the scRNA-seq data, it cannot be claimed "an unsupervised learning method".

We thank the reviewer for the comments. To clarify, we used unsupervised hierarchical clustering for gene selection. Based on the selected genes, we predicted the spatial position of each individual cell. Unsupervised and supervised gene selection approaches are used by the methods in the DREAM Single Cell Transcriptomics Challenge for predicting cell spatial position. Table S2 of the consortium paper (Gene selection for optimal prediction of cell position in tissues from single-cell transcriptomics data, Life Sci Alliance. 2020 Nov; 3(11): e202000867) listed the top performing methods (included ours) using either unsupervised or supervised gene selection.

2.Second, the current method has only been evaluated on one dataset. More independent evaluations should be added to demonstrate the generalizing capacity of this method.

We thank the reviewer for the constructive comments. The results of our method (Zho_team) and other top performing methods have been systematically and independently evaluated on the Drosophila embryo dataset in the DREAM Single Cell Transcriptomics Challenge (<https://www.synapse.org/#!Synapse:syn15665609/wiki/583234>) and in the consortium paper aforementioned.

Our manuscript, collected in the F1000Research DREAM Challenges Gateway (<https://f1000research.com/gateways/dreamchallenges/singlecelltranscriptomics>), is focused on describing the details of our method used in the DREAM Challenge. To avoid ambiguity, in the revised manuscript, we changed the title of our manuscript to "Unsupervised gene selection for predicting cell spatial positions in the Drosophila embryo".

Overall, this manuscript addresses an important scientific question. But the method needs more justifications and evidence to demonstrate its novelty and power.

We thank the reviewer for the comments. Our method is the only one that uses unsupervised hierarchical clustering for gene selection among the top performing methods as described in Table S2 of the DREAM Challenge consortium paper mentioned above. The power has been systematically demonstrated in our manuscript as well as the consortium paper using independent measurements and comprehensive comparisons.

Competing Interests: No competing interests were disclosed.

Reviewer Response 15 Jun 2021

Xianwen Ren, Peking University, Beijing, China

The authors have clarified my concerns.

Competing Interests: NA

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research