METHOD ARTICLE

Spatial mapping of single cells in the *Drosophila* embryo from transcriptomic data based on topological consistency [version 1; peer review: 1 approved with reservations]

Maryam Zand, Jianhua Ruan

Computer science, University of Texas at San Antonio, San Antonio, Texas, 78249, USA

---

**Abstract**

The advancement in single-cell RNA sequencing technologies allow us to obtain transcriptome at single cell resolution. However, the original spatial context of cells, a crucial knowledge for understanding cellular and tissue-level functions, is often lost during sequencing. To address this issue, the DREAM Single Cell Transcriptomics Challenge launched a community-wide effort to seek computational solutions for spatial mapping of single cells in tissues using single-cell RNAseq (scRNA-seq) data and a reference atlas obtained from in situ hybridization data. As a top-performing team in this competition, we approach this problem in three steps. The first step involves identifying a set of most informative genes based on the consistency between gene expression similarity and cell proximity. For this step, we propose two different approaches, i.e., an unsupervised approach that does not utilize the gold standard location of the cells provided by the challenge organizers, and a supervised approach that relies on the gold standard locations. In the second step, a Particle Swarm Optimization algorithm is used to optimize the weights of different genes in order to maximize matches between the predicted locations and the gold standard locations. Finally, the information embedded in the cell topology is used to improve the predicted cell-location scores by weighted averaging of scores from neighboring locations. Evaluation results based on DREAM scores show that our method accurately predicts the location of single cells, and the predictions lead to successful recovery of the spatial expression patterns for most of landmark genes. In addition, investigating the selected genes demonstrates that most predictive genes are cluster specific, and stable across our supervised and unsupervised gene selection frameworks. Overall, the promising results obtained by our methods in DREAM challenge demonstrated that topological consistency is a useful concept in identifying marker genes and constructing predictive models for spatial mapping of single cells.
Keywords
Single cell RNA-seq, spatial mapping, feature selection, particle swarm intelligence, nearest neighbor

This article is included in the DREAM Challenges gateway.

Corresponding author: Jianhua Ruan (jianhua.ruan@utsa.edu)

Author roles: Zand M: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Ruan J: Conceptualization, Formal Analysis, Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This research was funded by National Science Foundation [1565076] and National Institutes of Health [US4CA217297].

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2020 Zand M and Ruan J. 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: Zand M and Ruan J. Spatial mapping of single cells in the Drosophila embryo from transcriptomic data based on topological consistency [version 1; peer review: 1 approved with reservations] F1000Research 2020, 9:1014 https://doi.org/10.12688/f1000research.24163.1

First published: 20 Aug 2020, 9:1014 https://doi.org/10.12688/f1000research.24163.1
Introduction
Single cell RNA sequencing (scRNA-seq) is a cost-efficient, high throughput technology that has dramatically enhanced our understanding of developmental biology such as cell type identification, regulatory network inference, and cell trajectories\textsuperscript{7,8}. Despite many breakthroughs in biological sciences made possible by this technology, it yet suffers from the drawback that native cell location in e.g. embryo or complex tissue is often lost, except for in a few experimental methodologies which are either expensive, require highly specialized tools, or are not as widely applicable as standard scRNA-seq protocols\textsuperscript{9–11}. Given the substantial benefit offered through cell location recovery, such as obtaining a basic understanding of tissue function and disease pathology\textsuperscript{12,13}, the cell spatial reconstruction was specifically addressed in recent Single Cell Transcriptome DREAM challenge as a community-wide effort.

Many promising computational approaches dealing with the spatial reconstitution problem are centered around the main idea that an in situ atlas of a set of landmark gene’s expressions is used as a guideline to be combined with scRNA-seq profiles of individually measured cell\textsuperscript{14,15}. For instance, Seurat\textsuperscript{4} first imputes the noisy scRNAseq data then predicts the cell locations by comparing the scRNAseq gene expression pattern to its binary expression level measured by in situ data. This step is done through a mixture model. Finally, original cell location is retrieved by evaluating a posterior probability function constructed for cell-bin pairs. DistMap\textsuperscript{11} was a successful method for spatial reconstruction (of Drosophila embryo) with near single cell resolution, much higher compared to that of Seurat (3039 bins versus 128 bins). It predicts top candidate positions for a given cell by calculating the Mathews Correlation Coefficients (MCC) of binarized landmark gene expressions for every cell-bin combination. While DistMap was to some extent successful in dealing with the cell spatial mapping problem, it was limited to binarized data rather than continuous, utilized simplistic MCC analysis, and more importantly it treats each single cell independently whereas it might be more beneficial to account for collective interrelationships between cells. To more extensively explore the space of better predictive strategies, DREAM challenge aimed to exploit the atlas provided by DistMap with the hope of resolving spatial reconstruction by using incrementally fewer landmark genes (i.e. 60,40,20). Achieving this goal will help with eliminating the need for a priori reference atlas, which is expensive and time-consuming to obtain, in the future transcriptomic studies.

In this work, we proposed a top-performing method (evaluated based on three distinct scoring criteria defined by DREAM challenge) which allows us to predict the cell location consistently as accurate as DistMap while requiring fewer number of landmark genes. The details of our method and evaluation metrics are provided later in the text.

Methods
Overview of the proposed method
The general overview of our method is such that in the first step we investigate both supervised and unsupervised feature selection methods by defining two biologically rational metrics optimizing the consistency between gene expression similarity and cell proximity. In the unsupervised version we do not use the predicted cell locations given in 15 to obtain the set of most informative genes (e.g. 60,40,20), thus avoiding overfitting. On the other hand, the supervised version uses the cell locations given by DistMap as a reference. In the next, to predict the final cell locations, we use a PSO algorithm to assign proper weights to genes based on fitness functions defined by gene expression patterns. This reflects the intuition that different landmark genes are expected to demonstrate different potential in guiding us toward the proper embryo reconstruction. Finally, we use the information embedded in the cell topology to adjust the associated cell-location score with the hope to improve the predictions.

Datasets and pre-processing steps
To reconstruct Drosophila embryo from single cells, we need reference dataset (in situ), spatial coordinates, and scRNA-seq data, the details of which along with the preprosseing steps are given in the following.

Reference database
The reference database (denoted as W) provides the in situ expression values as a $W_{3039\times3}$ matrix where rows and columns correspond to bin locations and marker genes, respectively. The original data comes from Berkeley Drosophila Transcription Network Project (BDTNP) and in here we used the binarized format as explained in 15.

Spatial coordinates
The spatial coordinate information from one half of Drosophila embryo (denoted as L) is an $L_{1297\times3}$ matrix where the columns are $x$, $y$, and $z$ coordinates of 3039 rows of bins.

Single cell RNA sequencing
The scRNA-seq data (denoted as Y) gives the gene expression values as a $Y_{1297\times9324}$ matrix where rows and columns are single cells and genes, respectively. In here we followed the normalization process as implemented by 15. Briefly, the raw data was first normalized with respect to the total number of unique molecular identifiers (UMI) for each cell, followed by a pseudo count addition and a log transformation. The binarization process was implemented such that the quantile was varied in order to obtain the minimum mean squared root error between the gene correlation matrix of binarized atlas and binarized scRNA-seq.

Finding most informative genes
In this study, our first goal is to identify a subset of genes whose expression patterns are predictive of cell locations. We have proposed two different feature selection methods (supervised and unsupervised) to select informative genes. In the supervised method, our metric was defined based on true cell locations (gold standard). To prevent overfitting we applied a 10 fold cross validation. On the other hand, we designed an unsupervised method based on the intuition that the current locations obtained by matching the normalized and binarized scRNA-seq expression patterns with the in situ expression patterns are not necessarily the true locations of these cells. These two methods are discussed in detail in the following sections.
Unsupervised gene selection

As we believe the current locations obtained by matching the normalized and binarized scRNA-seq expression patterns with the in situ expression patterns are not necessarily the true locations of these cells, we decided to take an unsupervised feature selection approach, which does not depend on the current locations of the cells to be predicted, and therefore avoid overfitting.

The key rationale in our unsupervised feature selection method is that if a set of genes can be used as predictors of cell locations, then the cells showing similar expression patterns of these genes must be geometrically close to each other. Therefore, we defined two complementary metrics to quantitatively measure the proximity of cells with similar expression patterns for different gene subsets, and developed a greedy algorithm to search for a gene subset with the optimal (minimal) score combining the two metrics.

Metrics to measure the power of gene signatures as location predictors. The first metric relies solely on the in situ gene expression patterns in the 3039 location bins, and is calculated as follows: given a set of genes \(G\) as features, the pairwise Pearson Correlation Coefficient (PCC) is computed between the in situ expression data for every pair of the 3039 location bins; the top-10 locations with the highest PCC is then identified for each location bin; the metric \(M^G_1\) is defined as the average Euclidean distance between each location bin and its top-10 most similar location bins:

\[
M^G_1 = \frac{\sum_{i=1}^{n} \sum_{j \in S^G_m(i)} D_{ij}}{k \times n},
\]

where \(S^G_m(i)\) is the set of \(k\) most similar bins for location \(i\) based on the in situ expression pattern of a gene signature \(G\), \(k\) is fixed at 10 in this work, and \(n = 3039\) is the total number of location bins. \(D_{ij}\) is the Euclidean distance between the geometric coordinates of location \(i\) and location \(j\).

The second metric uses information from both the in situ expression data and the scRNA-seq expression data, and is calculated as follows. Given a set of genes \(G\) as features, the pairwise PCC is computed between the scRNA-seq expression pattern of each of the 1297 cells and the in situ expression pattern of each of the 3039 location bins; then for each of the 1297 cells, the top-10 location bins with the highest PCC is identified; the metric \(M^G_2\) is defined as the average Euclidean distance between the geometric coordinates of the location bin most similar to cell \(c\) and the geometric coordinates of the top-10 most similar location bins (including the most similar location):

\[
M^G_2 = \frac{\sum_{c=1}^{n} \sum_{k \in S^G_m(c)} D_{k\bar{c}}}{k \times m},
\]

where \(S^G_m(c)\) is the set of top-\(k\) locations whose in situ expression patterns are most similar to the scRNA-seq expression pattern of cell \(c\) based on gene signature \(G\), \(k\) is fixed at 10 in this work, and \(m = 1297\) is the total number of cells whose locations are to be predicted. \(D_{k\bar{c}}\) is the location bin where the expression pattern of gene signature \(G\) is most similar to cell \(c\).

Note that the currently known most possible location of each cell \(c\), \(\bar{i}_c^G\), which is predicted using all 84 genes with uniform weights, are not used in either \(M_1\) and \(M_2\); therefore, the gene selection process is not biased towards identifying genes to match the original locations predicted by the 84 genes. Rather, the metric provides an intrinsic measurement of the power of any subset of genes as location predictors, independent of the locations predicted with the 84 genes. In fact, the quality of the 84 genes as predictors can also be measured using these two metrics, and compared to any other gene sets; it is possible that a subset of the 84 genes can receive higher scores in these two metrics than the original 84 genes. In contrast, using a supervised feature selection method, where the “true” location is defined using all 84 genes, any subset of genes will necessarily be inferior to the complete set of 84 genes.

Step-wise backward elimination feature selection algorithm. We used a standard backward elimination algorithm to identify a subset of genes \(G\) with the minimal sum of \(M^G_1\) and \(M^G_2\). Briefly, starting with a set of \(q\) genes, we computed \(M^G_1\) and \(M^G_2\) for all possible subsets of \(q - 1\) genes by removing one gene at a time from the set. The subset with the minimal \(M^G_1 + M^G_2\) is then recorded as the best subset of size \(q - 1\). This procedure is then repeated until a desired number of genes is reached. As this algorithm is a greedy approach, it does not guarantee to find the optimal solution. We have also attempted to combine backward elimination with forward selection, which only improved the solution slightly. Due to the excessive running time required, we opted to use the simple algorithm described above while leaving additional improvement as future work.

Supervised gene selection

While in the unsupervised approach metrics \(M_1\) and \(M_2\) were optimized, in the supervised version a single metric \(N\) was defined as explained below. This metric, which relies on both the scRNA-seq gene expression patterns in the 1297 cells and the gold standard location of each cell, is calculated as follows: given a set of genes \(G\) as features, the pairwise PCC is computed between the scRNA-seq expression data for every pair of the 1297 cells; the top-10 cells with the highest PCC is then identified for each cell; the metric \(N^G\) is defined as the average Euclidean distance between the gold standard geometric coordinates of each cell and its top-10 most similar cells:

\[
N^G = \frac{\sum_{c=1}^{n} \sum_{k \in S^G_m(c)} D_{k\bar{c}}}{k \times m},
\]
where $T^G_k$ is the set of top-$k$ cells whose scRNA-seq expression patterns are most similar to the scRNA-seq expression pattern of cell $c$ based on gene signature $G$. $k$ is fixed at 10 in this work, and $m = 1297$ is the total number of cells whose locations are to be predicted. $l^c_*$ is the “gold standard” cell location for cell $c$, which is predicted using all 84 genes.

**Supervised learning to find optimal gene weights**

It is intuitive to assume that the contribution of genes in determining cell locations are not equal. Therefore, we look for a way to learn how to assign proper weight to each selected gene for more accurate prediction of cell locations. To this end, we chose a supervised learning approach, using the cell locations predicted by the highest MCCs with the 84 signature genes as “gold standard” locations. To avoid over-fitting, we performed 10-fold cross-validation: gene weights were determined using the scRNA-seq data of 90% of cells; these weights are then used to predict the locations of the remaining 10% of the cells not used in training. The splitting of the data is saved, for reproducibility of the results.

The basic idea of the PSO algorithm is as follows. We created a set of agents, each of which is initiated with a gene weight vector $w_i$ of size $|G| \times 1$. Each weight vector is evaluated by how closely the weighted gene expression pattern can be used to predict the cell locations when compared to the “gold standard” locations obtained with the 84 genes:

$$M_{3}^{w,G} = \frac{\sum_{i=1}^{m} \sum_{k \in S_{k}^{w,G}} D_{ik}}{k \times m}, \quad (4)$$

where $S_{k}^{w,G}$ is the set of top-$k$ location bins whose in situ expression patterns are most similar to the weighted expression pattern of cell $c$ based on a given gene signature set $G$. The similarity is measured by PCC here. $k$ is fixed at 10, and $m$ is the number of cells in the training set. $l^c_*$ is the “gold standard” cell location for cell $c$, which is predicted using all 84 genes with uniform weights.

During the search, each agent keeps track of a personal best weight vector $P_{best}$, and the global best solution from all agents is denoted $G_{best}$. At each iteration, the weight vector of each agent is updated by the differences between the current weight and the personal best and global best weight vectors:

$$w_i = w_i + \alpha \times r_1 \circ (P_{best} - w_i) + \beta \times r_2 \circ (G_{best} - w_i),$$

where $\alpha$ and $\beta$ are constants to control the granularity of the search and speed of convergence. We choose $\alpha = \beta = 0.2$ with 200 agents and the maximum number of iterations is 40. The operator $\circ$ denotes entry-wise vector multiplication. $r_1$ and $r_2$ are vectors of random numbers uniformly distributed between 0 and 1, generated independently for each agent at each iteration.

**Neighbor-weighted cell location prediction**

The location prediction for each single cell relies on the (weighted) similarity between the expression pattern of selected signature genes in the cell and every location bin. It is important to note that the expression patterns in neighboring cells should be similar in general, and therefore the overall prediction should take the expression of nearby location bins into consideration. Intuitively, if the globally highest scoring location is far away from locations with slightly lower but comparable scores, the confidence score for the highest-scoring location should be reduced; on the other hand, a locally highest-scoring location close to other high-scoring locations should be upweighted. Therefore, to make the final prediction for a given cell, we adjusted the prediction score based on the prediction scores from neighbor locations.

Formally, let $C = (c_j)_{j=1}^n$ be the bin-cell association matrix, where $c_j$ is the PCC between the (weighted) scRNA-seq data and the in situ hybridization data for every pair of cells and locations. $n = 3039$ is the number of candidate location bins, and $m$ is the number of cells in the test set. Let $D = (d_{ij})_{n \times m}$ be the Euclidean distance matrix between the geometric coordinates of every pair of location bins. We define an affinity matrix $A = (a_{ij})_{n \times n}$ such that $a_{ij} = \frac{1}{d_{ij}^\alpha}$ where $d^\alpha$ is a parameter to control how many neighbor locations can impact the final prediction score. A smaller $d^\alpha$ value means fewer neighbor locations to be considered. To have a robust measure of how geometrically close two location bins can be, we first measure the distance between each location and its nearest location, and then computed the median of these shortest distances as $d^\alpha$. As a result, most $a_{ij}$s are much smaller than $e^{-1}$, and only a limited number of neighbor locations with very high scores can impact the final prediction score for each cell.

The final prediction score matrix $P = (p_{ij})_{n \times m}$ is calculated by $P = A \times C$. Since $a_{ij} = 1$ and $a_{ij} \leq 1$ for all $j \neq i$, it is easy to see that

$$P_{ij} = \sum_{k=1}^{n} a_{ik} c_{jk} = c_{ij} + \sum_{k \neq j} a_{ij} c_{jk}.$$ 

Therefore, the final prediction score for a cell $i$ to be at a particular location $j$ is the weighted sum of the similarity scores between the expression pattern of cell $i$ and all locations, where the weight is an exponentially decreasing function of the geometric distance from location $j$.

From the final predicted bin-cell association matrix, we reported the 10 locations with the highest scores for each cell as the most likely positions in embryo.

**How proposed method applied to different subchallenges**

In subchallenge 1 we performed both PSO and neighbor weighting. In subchallenge 2, we did not perform PSO; therefore, the 40 genes obtained from gene selection are utilized with uniform weights. In subchallenge 3, genes were weighted with the PSO procedure, but we chose not to perform neighbor weighting.

**Post-challenge phase**

In this phase to evaluate the robustness and soundness of the method, a 10 fold CV scenario was performed to obtain 10
different sets of informative genes using a subset of cells. To compare the similarity of the selected genes, Jaccard similarity was defined as follows:

\[ J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (5) \]

where \( A \) and \( B \) are two sets of informative genes and \( J(A, B) \) measures the ratio of the number of common genes and the total number of genes presented in two sets. In addition, the expected Jaccard similarity was computed as follows:

\[ E(J) = \sum_{k=0}^{m} \binom{n}{k} \binom{n-m}{m-k} \quad (6) \]

where, \( n \) is the total number of genes, here 84, and \( m \) is the number of genes in our selected gene set, 60, 40, and 20 for subchallenge 1, 2, and 3, respectively.

### Validation of model by DREAM consortium

Our method was designated as a top-performing method among 34 participating teams. To evaluate and rank the teams, the challenge organizers had defined three scoring metrics \( s_1, s_2, \) and \( s_3 \), which were not disclosed to participants at the time of submission. The details of each metric are available at [https://github.com/dream-sctc/Scoring](https://github.com/dream-sctc/Scoring). Here, we briefly explain the general intuition and purpose behind each scoring metric. Roughly speaking, \( s_1 \) stands for how the gene expression of the 10 most probable predicted position for all cells closely approximates that of the \textit{in situ} expression pattern of the gold standard locations (as predicted by DistMap). On the other hand, \( s_2 \) only considers how the averaged location prediction of the 10 most probable predictions using 60, 40, and 20 genes is compared to that of the one predicted by using all the 84 genes. As is evident, this metric does not include either of the accuracy of the \textit{in situ} expression profile prediction and the closeness of \textit{in situ} and scRNA-seq data. Meanwhile, \( s_3 \) accounts for how the scRNA-seq expression of 60, 40, and 20 genes of the best predicated locations is closely approximating that of the \textit{in situ} expression patterns.

### Software

The method proposed here is written in Matlab 2018b and the source code is available from [GitHub](https://github.com/dream-sctc/Scoring). It does not utilize or rely on any specific Matlab toolbox. Therefore by following the clear detailed formulation provided in manuscript this method can readily be implemented in any open-access software.

### Results and discussion

#### Performance evaluation

In the two supervised and unsupervised methods discussed so far there exist two important components implemented with the intention to both select a set of most informative genes and locate cells based on the information buried in cell neighborhood network topology. To elucidate the effectiveness of these two components, we designed a set of four experiments in which the first component was replaced either by random gene selection or a simple strategy of high degree gene selection and the second component was inactivated. The subchallenge scores for these total of six methods are given in Table 1. To quantify the effectiveness of the two mentioned components we calculated the average score of the group A, containing supervised and unsupervised, and group B, containing the remaining four methods. Our analysis showed that on average the score of group A were higher by 14%, 20%, 30% in the subchallenges 1, 2, and 3 respectively. The reason for the greater performance in subchallenge 3 was expected given that fewer number of genes were used compare to the other two subchallenges.

The predictions mentioned above did not involve any additional pre-processing steps, e.g. imputation, on the provided

### Table 1. Numerical values of subchallenges scores are given for the ease of comparison with some designed baseline methods.

| gene selection method | SubCh1 | SubCh2 | SubCh3 |
|------------------------|--------|--------|--------|
|                        | s1     | s2     | s3     | s1     | s2     | s3     | s1     | s2     | s3     |
| Group A                |        |        |        |        |        |        |        |        |        |
| Our Unsupervised method| 0.6610 | 1.4522 | 0.6122 | 0.6552 | 1.3176 | 0.6538 | 0.6620 | 1.0166 | 0.7928 |
| Our Supervised method  | 0.6730 | 1.5463 | 0.5937 | 0.6558 | 1.3719 | 0.6731 | 0.6534 | 1.0994 | 0.7807 |
| Group B                |        |        |        |        |        |        |        |        |        |
| Random gene selection  | 0.6736 | 1.0638 | 0.6289 | 0.6113 | 0.6930 | 0.6240 | 0.5362 | 0.5052 | 0.7283 |
| Random gene selection + neigh based reweighting | 0.6714 | 1.4043 | 0.5762 | 0.6642 | 1.139  | 0.6619 | 0.5734 | 0.6997 | 0.6964 |
| High degree gene selection | 0.6914 | 0.904  | 0.5860 | 0.6061 | 1.0163 | 0.5969 | 0.5653 | 0.6156 | 0.7159 |
| High degree gene selection + neigh based reweighting | 0.6702 | 1.3241 | 0.5706 | 0.6134 | 1.027  | 0.5978 | 0.5593 | 0.7065 | 0.6468 |
| Avg of all metrics in group A for each subchallenge | 0.9231 | 0.8879 | 0.8342 |
| Avg of all metrics in group B for each subchallenge | 0.8137 | 0.7376 | 0.6291 |
input data. We simply used the binarized and normalized *in situ* hybridization and scRNA-seq data. However, for the sake of completeness we also examined the possible role of “imputation” and using raw data instead of the binarized scRNA-seq data. We tried to impute the dropouts in scRNA-seq data using SAVER\(^7\) and netImpute\(^8\), but no significant improvement was gained in terms of enhancing our metric scores. On the other hand, although our analysis indicated that using raw data instead of the binarized data can potentially increase the consistency between gene expression pattern similarity and cell proximity in this challenge (according to \(M_1\) and \(M_2\) metrics), we are limited by the fact that the true locations of the cells to be predicted are unknown, and prediction accuracy is at least partially defined by comparing to the “gold standard” location obtained from binarized data. We speculate that anyone using raw data would probably be disadvantaged. It is noteworthy that our method is applicable if one prefers to use raw data instead of binarized data, and our results (data not shown; available as underlying data) indicate that there is benefit of using raw data instead of binarized data.

Robustness of marker genes

In the post-challenge phase of the competition the data set was divided into train and test subsets using 10-fold cross-validation in order to further investigate to what degree the set of most informative genes are consistent across different subset of cells selected through the 10 fold CV analysis. The results given in Figure 1 show that the Jaccard similarity between different folds are higher than the expected similarity in all three subchallenges indicating that there in fact exists a consistency in the most-informative genes selected across different folds. Moreover, as the number of genes allowed in a subchallenge decreases (from subchallenge 1 to subchallenge 3) the difference between Jaccard similarity of the most-informative genes and its expected value becomes more and more pronounced.

Figure 2 shows the Venn diagram of 20 most informative genes selected from supervised and unsupervised methods. Out of 20 genes selected by each method, there are 11 common genes identified by both methods, which is more than expected (p-value < 0.0005, Fisher’s exact test).

**Figure 1.** Boxplot shows the Jaccard similarity between the genes selected for each of the 10 CV scheme in all 3 subchallenges. Blue stars represent expected Jaccard similarity.

**Figure 2.** The Venn diagram shows 20 genes selected from supervised and unsupervised methods out of which 11 genes are common for both methods. The 12 genes denoted by red color are the scRNA cluster-specific genes reported in DistMap.
Another interesting observation is that cluster-specific genes (denoted by red color) are prevalent in the set of most informative genes obtained from both supervised and unsupervised methods. This finding highlights our method was in fact able to take advantage of those 12 cluster-specific genes which contain cell location information.

**Recovering gene expression pattern**

To virtually reconstruct gene expression patterns, the result of our method (i.e. bin-cell association matrix) was processed based on the methodology of vISH - a tool developed in 15 to derive the expression pattern of each of the 84 genes across the location bins, and compared with the expression patterns obtained by DistMap. Figure 3 shows the distribution of the PCC between DistMap and our results from the three subchallenges. Overall, there is a high correlation among reference patterns (DistMap) and patterns generated by our method. The average correlation in the three sub-challenges are 0.81, 0.76, and 0.68, respectively. In sub-challenge 1, almost all genes have been reliably reconstructed, while for sub-challenge 3, a small number of genes have fairly low reconstruction rate.

Figure 4 shows the reconstructed expression patterns for three genes: *twi, cad, and ftz*, which play key roles in the

---

**Figure 3.** Histogram of correlation between gene patterns predicted by our method and DistMap using 60,40,20 genes.

**Figure 4.** Expression pattern of three sample genes are given for *in situ*, DistMap, and our method using 60,40,20
regulatory network of early *Drosophila* development. Overall, there is good agreement between our predictions and that of DistMap. In case of *twi*, our method and DistMap both very precisely predicted the *in situ* expression pattern. In fact, *twi* is one the 20 genes selected by both the supervised and unsupervised feature selection methods, due to its distinct expression patterns associated with cell spatial arrangement in the embryo. For *cad*, DistMap and our method with as few as 20 genes predicted very similar expression patterns, where there is a higher expression in the posterior domain, consistent with the current knowledge of *cad* in embryo development. On the other hand, the predicted expression patterns seem to be much more diffused than the *in situ* expression pattern, potentially because of the binarization of the *in situ* data, which caused loss of weaker signals. Finally, for *ftz*, while the predicted expression pattern by our method with 60 genes is in general agreement with DistMap and *in situ* data, our method with 40 or 20 genes failed to reconstruct the expression pattern of *ftz* associated with the segmentation of *Drosophila* embryos. While it is possible that more refined parameters such as a smaller number of neighbor cells may improve the prediction of our method, we believe the striped pattern of *ftz* makes it difficult, if not impossible, for any method that aims at a much reduced number of marker genes for spatial mapping.

**Conclusion**

In this work, we proposed a method to identify gene markers for RNAseq-based reconstruction of cell spatial information that were lost during single-cell transcriptomics sequencing of *Drosophila* embryo. The main hypothesis of this study is that the topology of the marker gene expression based cell-cell similarity graph should be consistent with the topology of the cell-cell geometric location map. To test the hypothesis, several metrics were defined based on this biological rationale to capture the consistency between gene expression similarity and cell proximity. A greedy step-wise backward elimination feature selection algorithm was implemented to find a set of most informative genes to optimize these metrics. Next, a Particle Swarm Optimization algorithm was developed to obtain optimal gene weights to construct the cell-location association matrix. Finally, the prediction score of a cell’s location was further improved by considering the expression similarity between neighboring locations. It was shown that our method can successfully identify markers genes capable of predicting cell locations with high accuracy. In addition, it was also demonstrated that our method can recover the spatial expression patterns of most embryo marker genes. Even though the method proposed here was custom designed for this *Drosophila* embryo problem, it has the potential to be readily applied to other organisms as well.

**Data availability**

**Underlying data**

The challenge datasets can be accessed at https://www.synapse.org/#!Synapse:syn16782375

Challenge documentation, including the detailed description of the Challenge design, overall results, scoring scripts, and the clinical trials data dictionary can be found at: https://www.synapse.org/#!Synapse:syn15665609/wiki/582909

**Software availability**

Source code is available from: https://github.com/mary77/scSpatialMapping.git

Archived source code at time of publication: https://doi.org/10.5281/zenodo.3877571

License: MIT

**References**

1. Kolodziejczyk AA, Kim JK, Svensson V, et al.: The Technology and Biology of Single-Cell RNA Sequencing. Mol Cell. 2015; 58(4): 610–620. [PubMed Abstract](https://www.ncbi.nlm.nih.gov/pubmed/26225533) [Publisher Full Text](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4424074/)

2. Aitschuler SJ, Wu LF: Cellular Heterogeneity: Do Differences Make a Difference? Cell. 2010; 141(4): 559–563. [PubMed Abstract](https://www.ncbi.nlm.nih.gov/pubmed/20639019) [Publisher Full Text](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2887383/)

3. Tang F, Barbaciao C, Wang Y, et al.: Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature. 2014; 510(7505): 363–369. [PubMed Abstract](https://www.ncbi.nlm.nih.gov/pubmed/24697797) [Publisher Full Text](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103055/)

4. Wang Z, Genstein M, Snyder M: RNA-Seq: a revolutionary tool for measuring genome-wide expression at high spatial resolution. Science. 2019; 363(6434): 1463–1467. [PubMed Abstract](https://www.ncbi.nlm.nih.gov/pubmed/30829412) [Publisher Full Text](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6893300/)

5. Lovatt D, Ruble BK, Lee J, et al.: Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature. 2014; 510(7505): 363–369. [PubMed Abstract](https://www.ncbi.nlm.nih.gov/pubmed/24697797) [Publisher Full Text](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103055/)

6. Shalek AK, Satija R, Shuga J, et al.: Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature. 2014; 510(7505): 363–369. [PubMed Abstract](https://www.ncbi.nlm.nih.gov/pubmed/24697797) [Publisher Full Text](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103055/)

7. Wagner A, Regev A, Yosef N: Revealing the vectors of cellular identity with single-cell genomics. Nat Biotechnol. 2016; 34(11): 1165–1160. [PubMed Abstract](https://www.ncbi.nlm.nih.gov/pubmed/27460041) [Publisher Full Text](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4968542/)

8. Casasent AK, Schalck A, Gao R, et al.: Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing. Cell. 2018; 172(1–2): 205–217.e12. [PubMed Abstract](https://www.ncbi.nlm.nih.gov/pubmed/29524842) [Publisher Full Text](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5916556/)

9. Stähl PL, Salmén F, Viskovic S, et al.: Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016; 353(6294): 78–82. [PubMed Abstract](https://www.ncbi.nlm.nih.gov/pubmed/27053497) [Publisher Full Text](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5171074/)

10. Lovatt D, Ruble BK, Lee J, et al.: Transcriptome in vivo analysis (TIVA) of spatially defined single cells in live tissue. Nat Methods. 2014; 11(2): 190–196. [PubMed Abstract](https://www.ncbi.nlm.nih.gov/pubmed/24532616) [Publisher Full Text](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3955667/)

11. Rodrigues SG, Stickels RR, Goeva A, et al.: Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science. 2019; 363(6434): 1463–1467. [PubMed Abstract](https://www.ncbi.nlm.nih.gov/pubmed/30829412) [Publisher Full Text](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6893300/)

12. Wang Z, Genstein M, Snyder M: RNA-Seq: a revolutionary tool for measuring genome-wide expression at high spatial resolution. Science. 2019; 363(6434): 1463–1467. [PubMed Abstract](https://www.ncbi.nlm.nih.gov/pubmed/30829412) [Publisher Full Text](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6893300/)

13. Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, et al.: An anatomically comprehensive atlas of the adult human brain transcriptome. Nature. 2012; 489(7416): 391–399. [PubMed Abstract](https://www.ncbi.nlm.nih.gov/pubmed/23340475) [Publisher Full Text](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3506365/)

14. Satija R, Farnel JA, Gennert D, et al.: Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015; 33(5): 495–502. [PubMed Abstract](https://www.ncbi.nlm.nih.gov/pubmed/25993845) [Publisher Full Text](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4475213/)
15. Karaiskos 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

16. Zand M, Ruan J: mary77/scspatialmapping: First release of scspatialmapping. Zenodo. 2020. http://www.doi.org/10.5281/zenodo.387757

17. Huang M, Wang J, Torre E, et al.: SAVER: gene expression recovery for single-cell RNA sequencing. Nat Methods. 2018; 15(7): 539–542. PubMed Abstract | Publisher Full Text | Free Full Text

18. Zand M, Ruan J: Network-based single-cell rna-seq data imputation enhances cell type identification. Genes (Basel). 2020; 11(4): 377. PubMed Abstract | Publisher Full Text | Free Full Text

19. Stauber M, Lernke S, Schmidt-Ott U: Expression and regulation of caudal in the lower cyclorrhaphan fly megaselia. Dev Genes Evol. 2008; 218(2): 81–87. PubMed Abstract | Publisher Full Text | Free Full Text

20. Lim B, Fukaya T, Heist T, et al.: Temporal dynamics of pair-rule stripes in living drosophila embryos. Proc Natl Acad Sci U S A. 2018; 115(33): 8376–8381. PubMed Abstract | Publisher Full Text | Free Full Text
Open Peer Review

Current Peer Review Status: ?

Version 1

Reviewer Report 07 September 2020

https://doi.org/10.5256/f1000research.26653.r69945

© 2020 Kuang R 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.

Tianci Song
Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA

Rui Kuang
Department of Computer Science and Engineering, University of Minnesota, Twin Cities, Minneapolis, MN, USA

This manuscript describes the methodology the authors employed to perform well in the DREAM Single cell Transcriptomics Challenge. The theme of the challenge is to select a subset of 84 known marker genes to characterize the spatial gene expression patterns in Drosophila embryo. The team developed methods for unsupervised gene selection, supervised gene selection, gene weighting and neighbor-smoothing based prediction of cell location. Overall, the manuscript provides sufficient detail to reproduce the results and some interpretations. It appears some further improvements can be done as follows:

1. Where it is understood that the team has optimized the hyper-parameter including k, alpha and beta with some fixed values for this dataset. For others to use the same methods on other similar studies, there is no clue how to tune or select the hyper-parameters. Discussion or more experimental results should be provided in this regard.

2. Similarly, it is also a mystery why different strategies of combining the methods are necessary for the three sub-challenges. There is no discussion of how the combinations are selected or optimized.

3. The results in Table 1 are very poorly explained. How to calculate the 14%, 20%, 30% improvement over the three measures is unexplained. Possibly, it is also helpful to explain the metric used to rank different methods in the competition.

4. In Figure 4, why not also plot the original in-situ hybridization and scRNAseq expressions of the three genes for comparison? The use of vISH seems to be unnecessary.

5. In the section "Datasets and pre-processing steps", a subsection explaining the "gold standard" prediction by DistMap should be added. It is confusing in the description of
supervised vs unsupervised gene selection, when the DistMap prediction and the nearest locations are used in the measures in equation (2) and (3).

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

**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?**
Yes

**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:** Computational Biology, Machine Learning, Biological Network Analysis, Single-cell Genomics, Spatial Genomics

*We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.*
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