METHOD ARTICLE

REVISED Prediction of cell position using single-cell transcriptomic data: an iterative procedure [version 2; peer review: 2 approved]

Andrés M. Alonso [1,2], Alejandra Carrea[1], Luis Diambra [1]

[1]CREG-CONICET, Universidad Nacional de La Plata, La Plata, Buenos Aires, 1900, Argentina
[2]INTech-CONICET, Universidad Nacional de San Martin, Chascomus, Buenos Aires, Argentina

v2 **First published:** 18 Oct 2019, **8**:1775 (https://doi.org/10.12688/f1000research.20715.1)

Latest published: 09 Apr 2020, **8**:1775 (https://doi.org/10.12688/f1000research.20715.2)

Abstract

Single-cell sequencing reveals cellular heterogeneity but not cell localization. However, by combining single-cell transcriptomic data with a reference atlas of a small set of genes, it would be possible to predict the position of individual cells and reconstruct the spatial expression profile of thousands of genes reported in the single-cell study. With the purpose of developing new algorithms, the Dialogue for Reverse Engineering Assessments and Methods (DREAM) consortium organized a crowd-sourced competition known as DREAM Single Cell Transcriptomics Challenge (SCTC). Within this context, we describe here our proposed procedures for adequate reference genes selection, and an iterative procedure to predict spatial expression profile of other genes.

Keywords

Single-Cell RNA sequencing, Drosophila Embryo, Gene expression Patterns, DREAM Challenge

This article is included in the DREAM Challenges gateway.

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers 1	2
version 2 (revision) 09 Apr 2020	✓ report	✓ report
version 1 18 Oct 2019	? report	? report

1 **Roosevelt Alves Da Silva** , Universidade Federal de Goiás, Jataí, Brazil

2 **Pablo Meyer** , IBM research Computational Biology Center, New York, USA

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

Corresponding author: Luis Diambra (ldiambra@gmail.com)

Author roles: Alonso AM: Conceptualization, Formal Analysis, Writing – Original Draft Preparation; **Carrea A**: Conceptualization, Formal Analysis, Writing – Review & Editing; **Diambra L**: Conceptualization, Formal Analysis, Supervision, Writing – Original Draft Preparation

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

How to cite this article: Alonso AM, Carrea A and Diambra L. **Prediction of cell position using single-cell transcriptomic data: an iterative procedure [version 2; peer review: 2 approved]** F1000Research 2020, **8**:1775 (https://doi.org/10.12688/f1000research.20715.2)

First published: 18 Oct 2019, **8**:1775 (https://doi.org/10.12688/f1000research.20715.1)

REVISED **Amendments from Version 1**

This new version has benefited from all useful comments and criticism of the reviewers. We have tried to answer and contemplate all of the comments. The main modifications in this new version offer clarification in the Introduction and in the Method Sections. We have improved Figure 3 and added a suitable legend to Figure 4A, and two references requested by R1. We have also improved the English of the entire manuscript. Unfortunately, we have not been able to give a satisfactory a response as we would like, to the last point raised by R2. However, we believe that even so, our proposal can help to inspire other researchers to design better methods.

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

Introduction

Multicellular organisms show throughout their development a crescent cellular heterogeneity, distributed and organized in different organs and tissues. This spatial heterogeneity has been explored using different techniques, such as immunohistochemistry and single-molecule fluorescence *in situ* hybridization (FISH)[1]. These approaches allow quantification of gene expression in many cells but, unfortunately, these techniques can currently be assayed only over a small number of genes. The selection of these genes introduces a bias that limits the power of these studies. With the advent of emergent methods in genomics, it has become possible to assess the transcriptomic profile of complex tissues with unprecedented resolution, thereby allowing insights into complex processes such as: differentiation trajectories, cell fate decisions, and spatial relationships. In this sense, high-throughput single-cell RNA-seq (sc-RNA-seq) is becoming an established experimental technique[2]. The protocol of this technique includes the initial step of sample collection, during which solid tissue dissociation results in single cells. Removing cells from their native context results in the loss of spatial information. However, this information can be crucial when the goal is to study the molecular composition of individual cells in the context of spatial location, for example, in the context of primary cancer cells research[3]. Fortunately, some progress has been made to overcome limitations of spatial information loss associated to this technique. Computational methods, based on Principal Component Analysis, are able to partially recover the spatial structure of gene expression patterns[4]. More recently, several computational techniques coupled to *in situ* RNA patterns facilitate this reconstruction with better resolution[5–7].

In order to catalyze research on computational methods for the spatial reconstruction of single-cell gene expression data, a crowd-sourced competition was designed by the DREAM Consortium in collaboration with Nikos Karaiskos and Nikolaus Rajewsky from Max Delbruck Institute. Using sc-RNA-seq data from Rajewsky Lab, published in 7, and the expression patterns of driver genes as an expression reference atlas, three main subchallenges were designed. The particular aim was to predict the position of 1297 cells in the 3039 *Drosophila melanogaster* embryonic locations, or bins, for one half of an embryo in stage 6 (pre-gastrulation), based on scR-NAseq data. The prediction of the 1297 cell positions must be done using a limited number of genes selected from a pool of 84 expression patterns used as a reference atlas. In subchallenge 1 the prediction must be performed using 60 driver genes out of 84 genes, in subchallenge 2 using a subset of any expression patterns from 40 genes out of the 84, and in subchallenge 3 using a subset of any expression patterns from only 20 driver genes. The selection of the subset of genes used for the prediction poses an additional and interesting problem. In this paper we present a procedure for solving the cell-position problem posed in the DREAM SCTC. This challenge consists of predicting the positions of individual cells, based on an expression reference atlas and a small set of genes reported in single-cell studies.

Methods

DREAM challenge data

Expression patterns used as a reference atlas correspond to 84 driver genes obtained from *in situ* hybridization experiments; the data correspond to The Berkeley Drosophila Transcription Network Project (BDTNP)[8]. This gene expression data set is listed in the file bdtnp.csv at DVEX server. One half of the *Drosophila* embryo has 3039 cells locations, each location is specified by three coordinates (x, y and z) (geometry.txt at DVEX). Thus, the reference database consists of an expression matrix of 84 genes (columns) quantified across the 3039 embryonic locations (rows). These data were next binarized[7], sorted in the same order of cell location, and listed in an additional file (binarized_bdtnp.csv at DVEX server). The single-cell RNA sequencing data is provided as a matrix with 8924 genes as rows, and 1297 cells as columns. These data are divided by the total number of counts for that cell, in this step a pseudocount is added. The normalized values are obtained by taking the logarithm of the total counts. The normalized values are also binarized, i.e. a given gene is ON (OFF) if the normalized values are above (below) of a quantile value. Based on a distance minimization criterion, the quantile value was chosen as 0.23. The short sequences for each of the 1297 cells in the raw and normalized data are the barcodes of individual cells. Both normalized as well as binarized data were provided by the DREAM Challenge.

Selection of the gene sets

In order to select the gene sets to be used in each subchallenge, we take into account two criteria:

(i) Genes that have complementary expression patterns across the single-cell population. It is well known that many genes are co-expressed, that is, their expression profiles are highly correlated. This correlation introduces a degree of redundancy in the expression matrix, which frequently is reduced by clustering those genes with similar expression profiles. This step allows us to identify genes with complementary expression patterns.

(ii) Genes with expression levels broadly distributed across the single-cell population. This step is performed in order to select one gene per cluster. Those genes with many null expression values over a large part of the population are discarded, because they are associated with distributions with a large peak at zero.

To accomplish these criteria, we first perform an agglomerative clustering procedure over the expression matrix comprising the 84 genes (the same genes as the available in the *in situ* expression data) over the 1297 cells). We cluster genes with similar expression profiles across the cells, by means of using the Euclidean distance over the normalized gene expression levels, and the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) as a linkage method. Then, we cut the dendogram tree into 20, 40, or 60 groups depending on the subchallenge. Next, we need to select only one gene per cluster. This selection is performed based on the criterion of the broadest distribution. To this end, for each gene within a given cluster we compute the frequency distribution p_i, where p_i denotes the frequency of occurrence of expression levels within the bin *i*. Here, we set the bin size equal to 0.125. After that, we compute the associated entropy $H = -\sum_i p_i \ln p_i$. Then, we select the gene with the greatest entropy in each cluster, i.e. the gene within the cluster with the broadest expression distribution across the single-cell population. This selection procedure is performed with the R script named preprocessing.r, which uses the function selgen. R, both available at Zenodo (see *Data availability*). To assess this method for the gene selection, we compare the prediction performance obtained with the set of 20 genes selected in this way with the results obtained with different sets of genes sampled at random. For comparison, we consider the Mathews correlation coefficient (MCC)[9] between the 1297 cells and the 3039 bins. Then, the ten better scored bins are selected as putative position for each cell. As the true positions of the cells are not available, we take the bin with the highest MCC, obtained with the set that include all 84 genes, as the bin associated with the true position. Thus, we count cells with the ten best scores containing the true position as cells whose positions are well predicted. The percentage of the well-predicted positions will be our measure of the performance. Figure 1 depicts the histogram of percentage of cells with well-predicted positions, obtained with 200 sets of 20 randomly selected genes. In all cases, this percentage is quite lower than that obtained with 20 genes selected as indicated above, which is 33.46%.

We use this procedure to select an additional set of 100 genes from the 8924 genes measured by the single-cell technique, but excluding the genes from the 84 reference gene set. These 100 genes will be used in further steps during the iterative procedure, and will be denoted as the outgroup set hereafter. The 20, 40 and 60 selected genes used for each cell location prediction task are listed in Table S1 (see *Extended data*); we also include the outgroup set of genes.

Scoring functions

In order to predict the position of a given single cell, we use a score approach based on two similarity measures between the sc-RNA-seq data, and the reference atlas. One of these measures is the MCC computed between the binarized expression profiles,

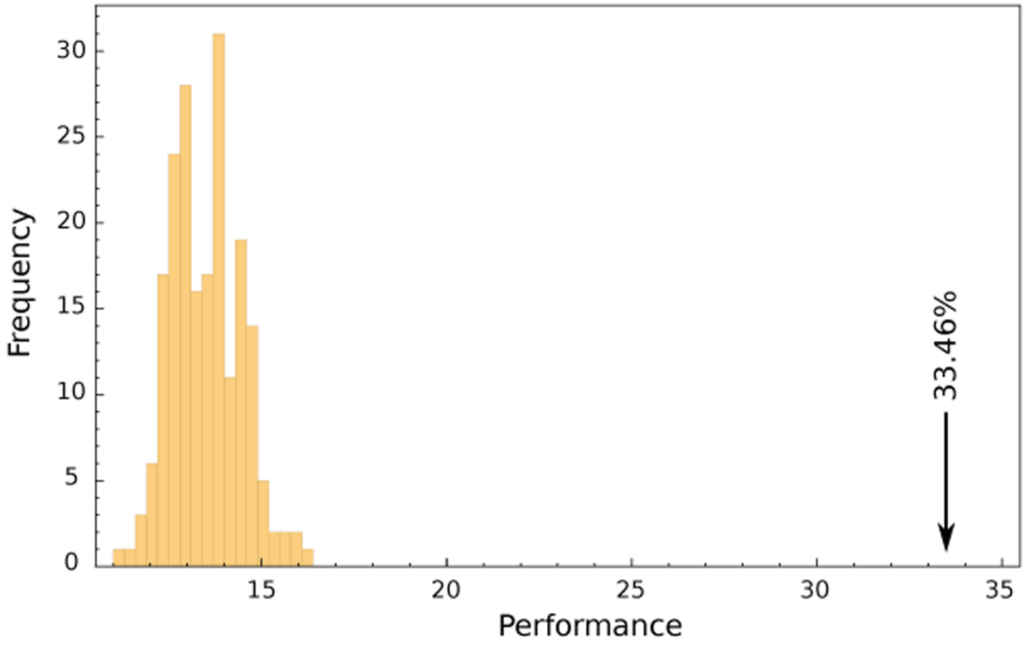

Figure 1. Performance of the gene selection procedure. Histogram of the performance obtained with 20 genes selected at random (yellow). The performance obtained with the set of 20 genes selected by the proposed method is indicated with a black arrow.

as proposed in 7. The MCC will be used in the initial step to assign putative bin positions for each single cell, and then to predict the spatial expression profile of the outgroup set of genes. The other measure is the overlap between the normalized expression vector of the single cells, and the projected vector corresponding to the predicted spatial expression profile. This vectorial space corresponds to the one spanned by the outgroup set of genes only. The overlap is defined by: $\cos(\theta) = \frac{\mathbf{u} \cdot \mathbf{v}_p}{\|\mathbf{u}\| \|\mathbf{v}_p\|}$, where **u** is the profile vector of the single cell, and $\mathbf{v}_p$ is the vector obtained by projecting the profile vector of the predicted profile on the subspace spanned by the non-null components of the profile vector **u**, as illustrated in Figure 2. The scoring functions are performed by the `R` script named functions.r, available at Zenodo (see *Data availability*).

Results

The proposed procedure is schematically illustrated in Figure 3. In the first step we select the set of N genes from the 84 driver genes to be used in the prediction, using the method described in *Selection of the gene sets* section. We also select an additional 100 genes (outgroup set of genes) from all genes measured in the sc-RNA seqexperiment, but excluding the driver genes. The name of the genes used are listed in *Extended data*: Table S1. Then, using the binarized expression data of the selected genes, we compute the MCC (measure 1) for each binarized single-cell vector against the 3039 binarized vectors associated with each positional bin of the reference atlas (BDTNP). By means of the MCC-based score, we predict the single-cell positions and build the putative expression patterns of the outgroup set of genes. In this sense, the expression level of gen g at the bin position i is given by the weighted average of the normalized gene expression across 10 putative positions corresponding to that bin, being the weight proportional to the associated MCC. Mathematically, $e_j^g = \sum_j^* c_{ij} e_j^g$, where c_{ij} are the MCC-based scores of the single cell j against position i, and e_j^g are the expression levels of gene g recorded in the individual cells j. The asterisk in the summation indicates that the 10 first better scored cells positions are included. The predicted expression patterns of the outgroup set of genes computed in this manner are used to compute the overlap (measure 2) with the corresponding expression level of each one of the 1297 single cells. Finally, using the measure 1 and the measure 2 we compute a composed score S, defined as $S = w_1 * c + w_2 * o$, where c is MCC-based score, o is the overlap-based score, and w_1 and w_2 are the respective weights. The score S is used to predict positions and improve the predicted expression patterns of the outgroup set of genes in each iteration. The last

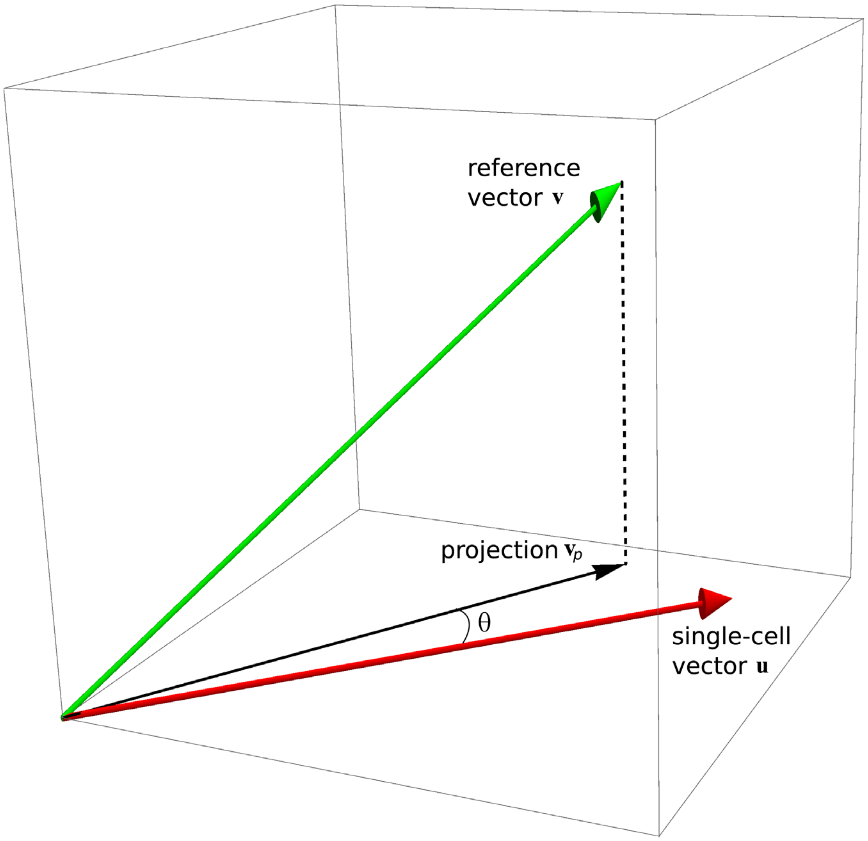
reference vector $\mathbf{v}$
projection $\mathbf{v}_p$
θ
single-cell vector $\mathbf{u}$

Figure 2. Overlap-based score. Low-dimensional representation of the angle between the expression vector **u**, and the projected expression vector $\mathbf{v}_p$.

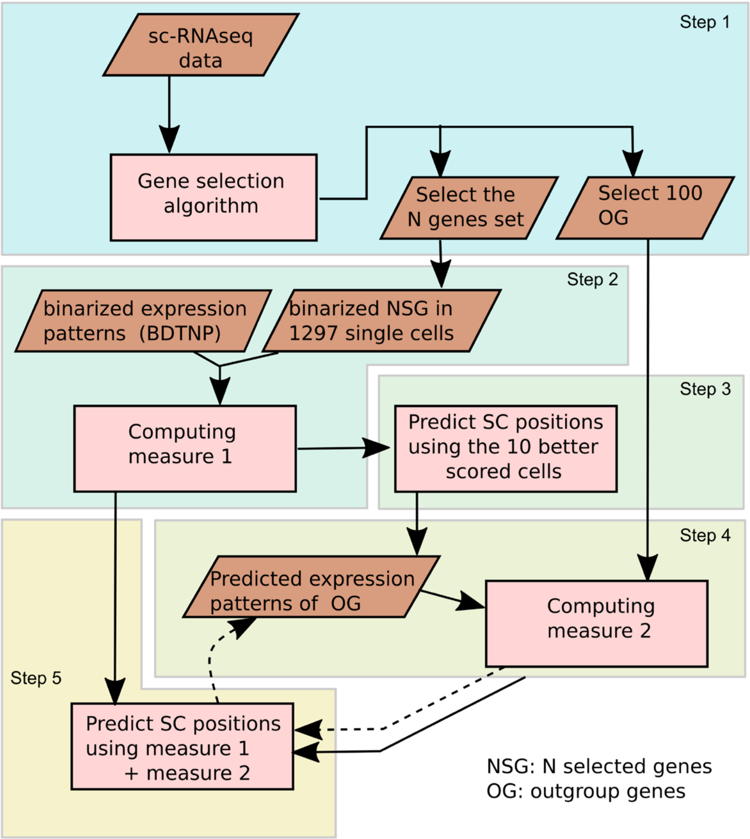
Step 1
sc-RNAseq data
Gene selection algorithm
Select the N genes set
Select 100 OG
Step 2
binarized expression patterns (BDTNP)
binarized NSG in 1297 single cells
Computing measure 1
Step 3
Predict SC positions using the 10 better scored cells
Step 4
Predicted expression patterns of OG
Computing measure 2
Step 5
Predict SC positions using measure 1 + measure 2
NSG: N selected genes
OG: outgroup genes

Figure 3. Flow diagram of the proposed method. Step 1: The set of *N* genes and the additional 100 outgroup genes are selected from the sc-RNAseq data. Step 2: Using the binarized expression data of the *N* selected genes we compute measure 1 for the1297 single-cell vectors against the 3039 binarized vectors of the reference atlas. Step 3: We predict the single-cell positions using the positions of the 10 better scored cells. Step 4: We build the putative expression patterns of the outgroup set of genes and we compute measure 2 against the expression level of 1297 single-cells. Step 5: By means of using the composed score *S*, the predicted expression patterns of the outgroup set of genes is improved in each iteration. The last two steps are repeated (2 or 3 times).

two steps are repeated (2 or 3 times), as indicated in Figure 3 with dashed arrows.

The above scheme is applied to the subchallenges with 20, 40 and 60 genes using different weight values. First of all, we apply the procedure to the subchallenge 3. Using the 20 genes we compute the MCC for every cell-bin combination. The first iteration of this scoring procedure leads to a performance of 33.5% in assigning the putative positions to each single cell. By means of using the 20 highest coefficients, we predict the expression patterns of the outgroup set of genes. Then, we compute a scoring measure composed of two terms: the MCC computed in the first step (with a weight $w_1 = 0.7$), and the previously-defined overlap between the expression vector of each single cell and the projected expression vector of the reference atlas, being both vectors composed of the 100 outgroup genes (with a weight $w_2 = 0.3$). The score combining both measures is then used to predict the positions of each single cell, which leads to a performance of 36% in the second iteration, and 38% in the third iteration. Further iteration steps do not produce any additional improvement. Figure 4A depicts the performance evolution of the procedure using this gene set.

In order to select the set of 60 genes to be used in subchallenge 1, from the 84 genes available in the reference atlas, we perform the above-mentioned agglomerative clustering procedure. Then, the 60 genes with the greatest entropy within each cluster are selected. The names of the resulting genes are listed in the first column of *Extended data*: Table S1. As a first step, we compute the MCC for each binarized single-cell vector, and the corresponding 3039 binarized vectors associated with each positional bin of the reference atlas. By means of using the 20 highest MCC for each cell ($N = 20$), we compute the putative expression patterns of the outgroup set of genes. In this case, the used scoring measure was composed by MCC with a weight of 0.90; and the overlap of the single-cell expression profiles and the 3039 positions of the predicted expression patterns obtained in the previous step, with a weight of 0.10. After two iterations the performance obtained is 95.4%. Figure 4B shows the predicted expression pattern of the *ftz* gene obtained

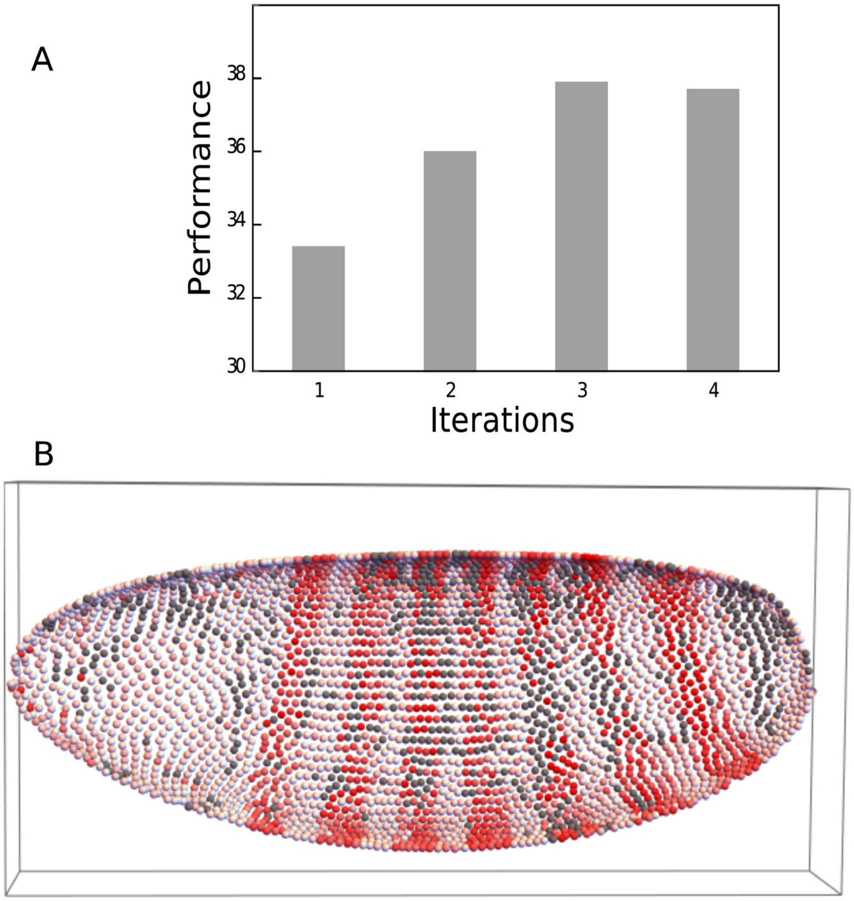

Figure 4. Prediction performance. Panel **A**: Performance obtained by means of using the iterative procedure with 20 genes. Panel **B**: Predicted expression pattern of the *ftz* gene obtained with 60 genes after two iteration steps. The expression level of each nuclei is given in white-red scale. Gray nuclei correspond to positional bins without prediction.

using the set of 60 genes. The same procedure is used to predict the positions of single cells by considering a set of 40 genes. Again, these genes are selected as described in Methods. The names of the resulting genes are listed in the second column of *Extended data*: Table S1. In this case, the performance obtained reaches 71.4%.

Discussion

We present three innovations that could represent improvements in regard to the original proposal[7]. One of these innovations is the method for selecting the set of genes to be used as reference in the cell-positions prediction task. This set of genes is a good starting point in the presented strategy for position prediction, although we have not explored this method in depth. For example, the Jaccard distance[10] could be used in the clustering procedure instead of the Euclidean distance. We noticed that MCC can overestimate false negatives due to the fact that sc-RNA-seq are not able to record expression of many genes. This results in profiles with many zeros, even in cases of moderate expression levels. For that reason, our second proposed innovation is an alternative way to make the comparison between profiles, as we used in subchallenge 3. Last but not least, the third innovation is the iterative procedure, which improves the performance of any of the alternative strategies presented here. In addition, we noticed that the iterative procedure does not necessarily converge to the correct solution, may be due to error propagation on the predicted patterns.

Data availability

Underlying data

Challenge documentation, including the detailed description of the Challenge design, overall results, scoring scripts, can be found at: http://www.synapse.org/#!Synapse:syn15665609/wiki. Data for this Challenge can be downloaded from http://shiny.mdc-berlin.de/DVEX/.

Zenodo: Prediction of cell position using single-cell transcriptomic data: an iterative procedure, https://doi.org/10.5281/zenodo.34700619[11].

This project contains code and documentation underlying the methods.

Data are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

Extended data

Zenodo: Prediction of cell position using single-cell transcriptomic data: an iterative procedure, https://doi.org/10.5281/zenodo.3470061[11].

This project contains the following extended data:

- Table S1: **Selected genes:** first, second and third columns list the name of genes used in the subchallenges 1, 2 and 3, respectively. The last column lists the names of the outgroup genes.

Data are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

Acknowledgements

We are grateful to the DREAM SCTC for curation of the challenge and the evaluation of the models.

References

1. Battich N, Stoeger T, Pelkmans L: **Image-based transcriptomics in thousands of single human cells at single-molecule resolution.** *Nat Methods.* 2013; **10**(11): 1127–33.
PubMed Abstract | Publisher Full Text
2. Shapiro E, Biezuner T, Linnarsson S: **Single-cell sequencing-based technologies will revolutionize whole-organism science.** *Nat Rev Genet.* 2013; **14**(9): 618–30.
PubMed Abstract | Publisher Full Text
3. Sierant MC, Choi J: **SingleSingle-Cell Ssequencing in Cancer: Recent Applications to Immunogenomics and Multi-omics Tools.** *Genomics Inform.* 2018; **16**: e17.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Durruthy-Durruthy R, Gottlieb A, Hartman BH, *et al.*: **Reconstruction of the mouse otocyst and early neuroblast lineage at single-cell resolution.** *Cell.* 2014; **157**(4): 964–78.
PubMed Abstract | Publisher Full Text | Free Full Text
5. 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
6. 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
7. 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
8. Fowlkes CC, Hendriks CL, Keränen SV, *et al.*: **A quantitative spatiotemporal atlas of gene expression in the *Drosophila* blastoderm.** *Cell.* 2008; **133**(2): 364–74.
PubMed Abstract | Publisher Full Text
9. Matthews BW: **Comparison of the predicted and observed secondary structure of T4 phage lysozyme.** *Biochim Biophys Acta.* 1975; **405**(2): 442–451.
PubMed Abstract | Publisher Full Text
10. Rogers DJ, Tanimoto TT: **A Computer Program for Classifying Plants.** *Science.* 1960; **132**(3434): 1115–1118.
PubMed Abstract | Publisher Full Text
11. Alonso A, Carrea A, Diambra L: **Prediction of cell position using single-cell transcriptomic data: an iterative procedure**. 2019.
http://www.doi.org/10.5281/zenodo.3470061

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 2

Reviewer Report 14 April 2020

https://doi.org/10.5256/f1000research.25587.r62174

✓

Pablo Meyer
IBM research Computational Biology Center, New York, NY, USA

Authors have addressed my concerns.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computational biology and systems biology

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.

Reviewer Report 14 April 2020

https://doi.org/10.5256/f1000research.25587.r62173

✓

Roosevelt Alves Da Silva
Física, Universidade Federal de Goiás, Jataí, Brazil

All requested revisions have been made. The article is now in the format suitable for the reader.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Biological physics, computational science, protein structures.

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 08 November 2019

https://doi.org/10.5256/f1000research.22783.r55478

Pablo Meyer
IBM research Computational Biology Center, New York, NY, USA

Alonso *et al* propose a nice idea for solving the gene to position problem posed in the DREAM single cell transcriptomics challenge, but at the end fall short while finding ways to improve their pipeline. The explanations of why they used such approach are also scarce. Although maybe obvious, it would be important that they explain why they chose genes with maximum entropy. why 100 outgroup genes, how the values of w1 and w2 were selected and why doing a 2 step optimization. Overall the rationale of their approach should be summarized in a couple of sentences in order to make it clearer, unfortunately the diagram does not clarify the rationale either. Also, given that this criteria did not improve significantly while iterating, they should have tried other selection criteria such as high expression level, or maybe other metrics such as the ones proposed in the challenge. Maybe trying to optimize the position and the gene expression correlation as interlocked steps is the source of the problem. Some discussion regarding this should be added.

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?
Yes

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computational biology and systems biology

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, however I have significant reservations, as outlined above.

Author Response 31 Mar 2020

Luis Diambra, Universidad Nacional de La Plata, La Plata, Argentina

Alonso et al propose a nice idea for solving the gene to position problem posed in the DREAM single cell transcriptomics challenge, but at the end fall short while finding ways to improve their pipeline.

Response: We would like to thank the Referee 2 for reviewing our paper, and for his positive comments. Below we gives him a point by point response for all his comments.

The explanations of why they used such approach are also scarce. Although maybe obvious, it would be important that they explain why they chose genes with maximum entropy.
why 100 outgroup genes,
how the values of w1 and w2 were selected
and why doing a 2 step optimization.
Response: The approach for our gene set selection has two steps. We agree that these steps have not been justified in our previous version. Now we have added a rationale for our procedure. It is well known that many genes co-express, that is, their expression profiles are highly correlated. This correlation introduces a degree of redundancy in the expression matrix, which is oftenly reduced by clustering those genes with similar expression profiles. We assume that the genes within each cluster would not provide much more information than just one in this group. Since the challenge requires choosing a small set of genes, we must choose one gene that represents each cluster. The question is then, which one?
This selection is made with the second criterion, which discards those genes with many null values of expression over a large part of the population, because they are associated with distributions with a large peak at zero. Of course, one could also use a criterion that selects the gene with higher expression level, as you proposed (see last response). Regarding the number of genes used as the outgroup genes, we did not make any computation to determine this number, or to answer if there is an optimal number of genes. We just chose a number of genes greater than the 86 genes, which seems to be good enough for the correct prediction of the cell positions. Regarding the values of w1 and w2, unfortunately we cannot imagine other method to determine w1 and w2 different from the brute force. In this sense, we began our study with a value of 0.5 for each weight. That gave quite bad results. Then, we increased the weight w1 and decreased w2 by 0.10 at each step, obteining improvements in the performance, until we reached the weights shown in the paper. We did not try to find the better weitghs. Initially, we assumed that an iterative procedure could be a kind of self-consistent method that would lead to an improvement in performance at every step. Unfortunately, that was not verified. The iterative procedure improved the performance only in the first steps. For this resubmission, we have made several modifications and performed perdictions, testing other hypotheses. However, we did not manage to substantially improve our own results (see the last point).

Overall the rationale of their approach should be summarized in a couple of sentences in order to make it clearer, unfortunately the diagram does not clarify the rationale either.
Response: We have remade the diagram of Fig. 3, and we have added a short description to each

step in the legend of this figure.

Also, given that this criteria did not improve significantly while iterating, they should have tried other selection criteria such as high expression level, or maybe other metrics such as the ones proposed in the challenge. Maybe trying to optimize the position and the gene expression correlation as interlocked steps is the source of the problem. Some discussion regarding this should be added.

Response: We have evaluated the performance of the approach by considering several modifications. In particular, those proposed by the referee:
1.- We selected the genes with higher expression levels within each cluster, instead of the genes with broadest level
distribution. This gene selection strategy gave similar results to the one presented in the original version.
2.- We used the metrics based only on MCC, instead of the combination of two metrics.
In this case, the predicted expression patterns computed in the first iteration are used to compute the MCC of the of the outgroup genes with the corresponding expression level of each one of the 1297 single cells. Finally, we computed a composed score S, defined as S = w1* c + w2* o, where c is MCC score computed in the first step (with the 20 genes), and o is the MCC score computed
with the outgroup set of genes, and w1 and w2 are the respective weights. We do not understand the reasons why subsequent iterations do not improve performance, significantly. However, we believe that the predicted expression pattern of the outgroup set of genes could be used to improve, in subsequent steps, any other method with better performance than the one we have proposed.

Competing Interests: No competing interests were disclosed.

Reviewer Report 24 October 2019

https://doi.org/10.5256/f1000research.22783.r55476

? **Roosevelt Alves Da Silva**
Física, Universidade Federal de Goiás, Jataí, Brazil

In their manuscript entitled "Prediction of cell position using single-cell transcriptomic data: an iterative procedure " Alonso *et al.* apply a method to predict the cell position combining single-cell transcriptomic data with a reference atlas. The study is submitted as part of the SCTC DREAM Challenge, which aims to evaluate different algorithms for this prediction task using 20, 40 or 60 genes as reference. The authors test their approach on SCTC challenge data-sets. They have three improvements on a method previously published (ref 7[1]): gene selection, a new scoring function, and an iterative procedure for improve the predictions. The paper is generally well-written, with a concise description of the problem and challenge. However it has room for improvement, especially in the manner information is presented. My criticisms are related to the clarification of details I thought to be missing from Methods and Results sections.

Major comments:

1. I suggest to authors to add a sentence that help to reader to understand the aim of the paper in the Introduction.

2. The fluxogram of Fig.3 seems has not a initial point. Why? I suggest to authors divide the procedure in layers. Also authors must give a description in the legend of this figure.

3. Panel A of Fig. 4 has not legend, please add one.

Minor comments:

1. The description of acronyms MCC is duplicated. The author must introduce it in the first time that it is mentioned.

2. Please, add adequate references for Mathews correlation coefficient (in Methods) and for Jaccard distance (in Discussion).

References

1. Karaiskos N, Wahle P, Alles J, Boltengagen A, et al.: TheDrosophila embryo at single-cell transcriptome resolution. *Science*. 2017; **358** (6360): 194-199 Publisher Full Text

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?

Yes

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

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Biological physics, computational science, protein structures.

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, however I have significant reservations, as outlined above.

Author Response 31 Mar 2020

Luis Diambra, Universidad Nacional de La Plata, La Plata, Argentina

Response: We would like to thank Referee 1 for reviewing our paper, and for his/her positive comments.
Below, we give him/her a point by point response for all his/her comments.

Major comments:

I suggest to authors to add a sentence that help to reader to understand the aim of the paper in the Introduction.
Response: We agree. We have added two sentences at the end of the Introduction in order to clarify the aim of our work.

The fluxogram of Fig.3 seems has not a initial point. Why? I suggest to authors divide the procedure in layers. Also authors must give a description in the legend of this figure.
Response: In this new version of the manuscript we have added a description to Fig. 3, and modified its initial point and the layers in order to make them clearer.

Panel A of Fig. 4 has not legend, please add one.
Response: In this new version of the manuscript we have added a legend to Fig. 4A.

Minor comments:

The description of acronyms MCC is duplicated. The author must introduce it in the first time that it is mentioned.
Response: The referee is correct, we have already fixed it.

Please, add adequate references for Mathews correlation coefficient (in Methods)
and for Jaccard distance (in Discussion).
Response: We have added two new references: ref10: 10.1016/0005-2795(75)90109-9 and ref11:
10.1038/234034a0, respectively.

Competing Interests: No competing interests were disclosed.