

# Predicting cellular position in the *Drosophila* embryo from Single-Cell Transcriptomics data

Jovan Tanevski\*<sup>1,2</sup>, Thin Nguyen\*<sup>3</sup>, Buu Truong\*<sup>4</sup>, Nikos Karaiskos<sup>5</sup>, Mehmet Eren Ahsen<sup>6,7</sup>, Xinyu Zhang<sup>8</sup>, Chang Shu<sup>8</sup>, Ke Xu<sup>8</sup>, Xiaoyu Liang<sup>8</sup>, Ying Hu<sup>9</sup>, Hoang V.V. Pham<sup>4</sup>, Li Xiaomei<sup>4</sup>, Thuc D. Le<sup>4</sup>, Adi L. Tarca<sup>10</sup>, Gaurav Bhatti<sup>10</sup>, Roberto Romero<sup>11</sup>, Nestoras Karathanasis<sup>12</sup>, Phillipe Loher<sup>12</sup>, Yang Chen<sup>13</sup>, Zhengqing Ouyang<sup>14</sup>, Disheng Mao<sup>15</sup>, Yuping Zhang<sup>15</sup>, Maryam Zand<sup>16</sup>, Jianhua Ruan<sup>16</sup>, Christoph Hafemeister<sup>17</sup>, Peng Qiu<sup>18,19</sup>, Duc Tran<sup>20</sup>, Tin Nguyen<sup>20</sup>, Attila Gabor<sup>1</sup>, Thomas Yu<sup>21</sup>, Enrico Glaab<sup>22</sup>, Roland Krause<sup>23</sup>, Peter Banda<sup>23</sup>, DREAM SCTC Consortium<sup>†</sup>, Gustavo Stolovitzky<sup>24</sup>, Nikolaus Rajewsky<sup>‡5</sup>, Julio Saez-Rodriguez<sup>‡1,25</sup>, and Pablo Meyer<sup>‡24</sup>

<sup>1</sup>Institute for Computational Biomedicine, Faculty of Medicine, Heidelberg University Hospital and Heidelberg University, Heidelberg, Germany

<sup>2</sup>Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia

<sup>3</sup>Deakin University, Geelong, Australia

<sup>4</sup>University of South Australia, Mawson Lakes, Australia

<sup>5</sup>Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany

<sup>6</sup>Icahn School of Medicine at Mount Sinai, New York City, NY, USA

<sup>7</sup>University of Illinois, Urbana-Champaign, IL, USA

<sup>8</sup>Department of Psychiatry, Yale School of Medicine, New Haven, CT, USA

<sup>9</sup>Center for Biomedical Informatics & Information Technology, National Cancer Institute, MD, USA

<sup>10</sup>Wayne State University, Detroit, MI, USA

<sup>11</sup>Perinatology Research Branch, NICHD/NIH/DHHS, Bethesda, MD, and Detroit, MI, USA

<sup>12</sup>Computational Medicine Center, Thomas Jefferson University, Philadelphia, PA, USA

<sup>13</sup>The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA

<sup>14</sup>University of Massachusetts, Amherst, MA, USA

<sup>15</sup>University of Connecticut, CT, USA

<sup>16</sup>University of Texas at San Antonio, TX, USA

<sup>17</sup>New York Genome Center, New York City, NY, USA

<sup>18</sup>Georgia Institute of Technology, Atlanta, GA, USA

<sup>19</sup>Emory University, Atlanta, GA, USA

<sup>20</sup>University of Nevada, Reno, NV, USA

<sup>21</sup>Sage Bionetworks, Seattle, WA USA

<sup>22</sup>Biomedical Data Science Group, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur Alzette, Luxembourg

<sup>23</sup>Bioinformatics Core Group, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur Alzette, Luxembourg

<sup>24</sup>IBM T.J. Watson Research Center, Yorktown Heights, NY, USA

<sup>25</sup>Joint Research Centre for Computational Biomedicine, Faculty of Medicine, RWTH Aachen University, Aachen, Germany

---

\*These authors contributed equally

†DREAM SCTC Consortium authors and affiliations are listed in the supplementary material

‡To whom correspondence should be addressed; pmeyerr@us.ibm.com

## Abstract

Single-cell RNA-seq technologies are rapidly evolving but while very informative, in standard scRNAseq experiments the spatial organization of the cells in the tissue of origin is lost. Conversely, spatial RNA-seq technologies designed to keep the localization of the cells have limited throughput and gene coverage. Mapping scRNAseq to genes with spatial information increases coverage while providing spatial location. However, methods to perform such mapping have not yet been benchmarked. To bridge the gap, we organized the DREAM Single-Cell Transcriptomics challenge focused on the spatial reconstruction of cells from the *Drosophila* embryo from scRNAseq data, leveraging as gold standard genes with *in situ* hybridization data from the Berkeley *Drosophila* Transcription Network Project reference atlas. The 34 participating teams used diverse algorithms for gene selection and location prediction, while being able to correctly localize rare subpopulations of cells. Selection of predictor genes was essential for this task and such genes showed a relatively high expression entropy, high spatial clustering and the presence of prominent developmental genes such as gap and pair-ruled genes and tissue defining markers.

## 1 Introduction

1 The recent technological advances in single-cell sequencing technologies have revolutionized  
2 the biological sciences. In particular single-cell RNA sequencing (scRNAseq) methods allow  
3 transcriptome profiling in a highly parallel manner, resulting in the quantification of thousands of  
4 genes across thousands of cells of the same tissue. However, with a few exceptions [1, 2, 3, 4, 5]  
5 current high-throughput scRNAseq methods share the drawback of losing the information relative  
6 to the spatial arrangement of the cells in the tissue during the cell dissociation step.

7 One way of regaining spatial information computationally is to appropriately combine the single-  
8 cell RNA dataset at hand with a reference database, or atlas, containing spatial expression patterns  
9 for several genes across the tissue. This approach was pursued in a few studies [6, 7, 8, 9, 10].  
10 Achim *et al* identified the location of 139 cells using 72 reference genes with spatial information  
11 from whole mount *in situ* hybridization (WMISH) of a marine annelid and Satija *et al* developed  
12 the *Seurat* algorithm to predict position of 851 zebrafish cells based on their scRNAseq data and  
13 spatial information from *in situ*-hybridizations of 47 genes in ZFIN collection [11]. In both cases,  
14 cell positional predictions stabilized after the inclusion of 30 reference genes. Karaiskos *et al*  
15 reconstructed the early *Drosophila* embryo at single-cell resolution and while the authors were  
16 successful in their reconstruction, their approach did not lead to a predictive algorithm and mainly  
17 centered around maximizing the correlation between scRNAseq data and the expression patterns  
18 from *in situ*-hybridizations of 84 mapped genes in The Berkeley *Drosophila* Transcription Network  
19 Project (BDTNP). In this project, *in situ* hybridization data was collected resulting in a quantitative  
20 high-resolution gene expression reference atlas [12]. Indeed, Karaiskos *et al* showed that the  
21 combinatorial expression of these 84 BDTNP markers suffice to uniquely classify almost every cell  
22 to a position within the embryo.

23 In the absence of a reference database, it is also possible to regain spatial information compu-  
24 tationally solely from the transcriptomics data by leveraging general knowledge about statistical  
25 properties of spatially mapped genes against the statistical properties of the single-cell RNA dataset  
26 [13, 14]. Bageritz *et al.* were able to reconstruct the expression map of a *Drosophila* wing disc  
27 using scRNAseq data by correlation analysis. They exploited the coexpression of non-mapping  
28 genes to a few mapping genes with known expression patterns, to predict the spatial expression  
29 patterns of 824 genes [13]. Nitzan *et al.* exploited the knowledge of the distribution of distances  
30 between mapping genes in physical space to predict the possible locations of cells based on the  
31 distribution of distances between genes in the expression space. Following this approach, they  
32 were able to successfully reconstruct the locations of cells of the *Drosophila* and zebrafish embryos

33 from scRNAseq data [14]. Although these approaches have indicated important steps to reconstruct  
34 the position of a cell in a tissue from their RNAseq expression, a global assessment is needed  
35 to evaluate the methods used and the number and nature of the genes with spatial expression  
36 information required for correctly assigning a location to each cell.

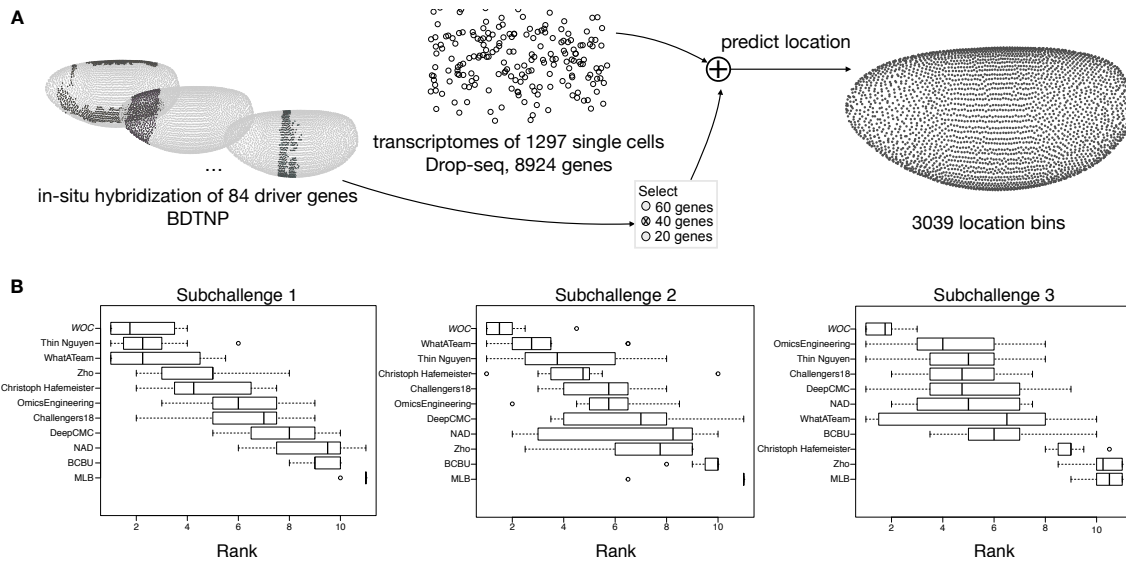
37 With this purpose in mind, and to catalyze the development of new methods to predict the  
38 location of cells from scRNAseq data we organized the DREAM Single cell transcriptomics  
39 challenge which ran from September through November 2018. DREAM challenges are a platform  
40 for crowdsourcing collaborative competitions[15] where a rigorous evaluation of each submitted  
41 solution allows for the comparison of their performance. The quality and reproducibility of each  
42 provided solution is also ensured. The combination of the individual solutions, i.e., the different  
43 approaches and insights to a common problem, leads to an overall wisdom-of-the-crowds (WOC)  
44 solution, with generally superior performance to any individual solution, from where collective  
45 insights can be garnered. We set up the challenge with 3 goals in mind. First, we used the data  
46 from Karaikos *et al* to foster the design of a variety of algorithms and objectively tested how  
47 well they could predict the localization of the cells. Second, we evaluated how the predictive  
48 performance of the algorithms was impacted by the number of reference genes from BDTNP  
49 with *in situ* hybridization information included in the predictions. Third, we investigated how the  
50 biological information carried in the selected genes was implemented in the algorithms to determine  
51 embryonic patterning.

52 The challenge, a first of its kind for single cell data, consisted of predicting the position of  
53 1297 cells among 3039 *Drosophila melanogaster* embryonic locations for one half of a stage 6  
54 pre-gastrulation embryo from their scRNAseq data (Figure 1A) [10]. At this stage cells in the  
55 embryo are positioned in a single two dimensional sheet following a bilateral symmetry, so that  
56 only positions in one half of the embryo were considered - accounting for the 3039 locations.  
57 Participants used the scRNAseq data for each of the 1297 cells obtained from the dissociation  
58 of 100-200 stage 6 embryos and the spatial expression patterns from *in situ*-hybridizations of 84  
59 genes in the BDTNP database [12]. Gene determinants of different tissues such as neurectoderm,  
60 dorsal ectoderm, mesoderm, yolk and pole cells were provided as a hint. To aid the development  
61 of prediction algorithms, we provided (when available) the regulatory relationship -positive or  
62 negative- between the 84 genes in the *in situ*-hybridizations and the rest of the genes. We asked  
63 participants to provide an ordered list of 10 most probable locations in the embryo predicted for  
64 each of the 1297 cells using the expression patterns from (i) 60 genes out of the 84 in subchallenge  
65 1, (ii) 40 genes out of the 84 in subchallenge 2, and (iii) 20 genes out of the 84 in subchallenge  
66 3. The predictions were compared to the ground truth location determined by calculating the  
67 maximum correlation using all 84 *in situs* [10]. We received submissions from 34 teams, and  
68 the overall analysis of the results showed that the selection of genes is essential for accurately  
69 locating the cells in the embryo. The most selected genes had a relatively high expression entropy,  
70 showed high spatial clustering and featured developmental genes such as gap and pair-ruled genes  
71 in addition to tissue defining markers.

## 72 **2 Results**

### 73 **2.1 Challenge setup**

74 A distinctive feature of the single cell transcriptomics challenge was the public availability of  
75 the entire dataset and the ground truth locations produced by DistMap, a method using the *in*  
76 *situ*-hybridizations available at BDTNP [12], published together with the data [10]. We took three  
77 actions to mitigate the issue of not having a blinded ground truth. First, for the purpose of predictor  
78 gene selection, we allowed the use of scRNA-Seq data and biological information from other



**Figure 1: Overview of the challenge and results.** **A.** In the DREAM Single-Cell Transcriptomics Challenge participants were asked to map the location of 1297 cells to 3039 location bins of an embryo of *Drosophila melanogaster*, by combining the scRNAseq measurements of 8924 genes for each cell and the spatial expression patterns from *in situ* hybridization of 60, 40 or 20 genes, for subchallenge 1, 2 and 3 respectively, for each embryonic location bin, selected from a total of 84 driver-genes. **B.** Ranking of the top 10 best performing teams and a wisdom of the crowds (WOC, in *italic*) solution, based on results from a post challenge cross-validated selection and prediction performance measured with three complementary scoring metrics. The boxplots show the distribution of ranks for each team on the 10 test folds. The rank for each fold is calculated as the average of the ranking on each scoring metric.

79 databases, but prohibited the use of *in situ* data. Second, to assess the quality of predictions, we  
 80 devised three scores (detailed in the Methods section) that were not disclosed to the participants  
 81 during the challenge. The scores measured not only the accuracy of the predicted location, but also  
 82 how well the expression in the cell at the predicted location correlates with the expression from the  
 83 reference atlas, the variance of the predicted locations for each cell, and how well the gene-wise  
 84 spatial patterns were reconstructed. Finally, we devised a post-challenge cross-validation scheme to  
 85 evaluate the soundness and robustness of the methods.

86 The challenge was organized in two rounds, a leaderboard round, and a final round. During the  
 87 leaderboard round the participants were able to obtain scores for five submitted solutions before  
 88 submitting a single solution in the final round. We received submissions from 40 teams in the  
 89 leaderboard round and 34 submissions in the final round. Out of the 34 teams that made submissions  
 90 in the final round, 29 followed up with public write-ups of their approaches and source code. For  
 91 subchallenges 1 and 3 we were able to determine a clear best performer, but for subchallenge 2,  
 92 there were two top ranked teams with statistically indistinguishable difference in performance (see  
 93 Supplementary Figures S1,S2 and S3).

94 As stated, given that the ground truth for this challenge was publicly available and to avoid  
 95 over-fitting, we decided to invite the top 10 performing teams to contribute to a post-challenge  
 96 collaborative analysis phase to assess the soundness and stability of their gene selection and  
 97 cell location prediction. Consequently, teams were tasked to provide predictions for a 10-fold  
 98 cross-validation (CV) scenario, under the same conditions as for the challenge phase. The folds  
 99 were extracted from the same RNA-seq dataset as in the challenge and every team used the same

Table 1: Best mean score for metrics  $s_1$ ,  $s_2$  and  $s_3$  achieved by the teams (Thin Nguyen, WhatATeam and OmicsEngineering) and the WOC solution. The standard deviation of scores across folds are in parenthesis. For more details on the scoring metrics see the Methods section.

	$s_1$		$s_2$		$s_3$	
	Teams	WOC	Teams	WOC	Teams	WOC
Subchallenge 1	<b>0.76</b> (±0.04)	0.73(±0.04)	<b>2.52</b> (±0.28)	2.16(±0.20)	0.59(±0.01)	<b>0.62</b> (±0.01)
Subchallenge 2	0.69(±0.03)	<b>0.70</b> (±0.05)	1.16(±0.12)	<b>1.84</b> (±0.26)	<b>0.67</b> (±0.02)	0.65(±0.01)
Subchallenge 3	0.65(±0.05)	<b>0.68</b> (±0.03)	0.88(±0.13)	<b>1.42</b> (±0.16)	<b>0.79</b> (±0.02)	0.71(±0.01)

100 assignment of cells to folds. We evaluated the performance of the teams using the same scoring  
 101 approach as in the challenge. To ensure the validity of the findings we decided to perform all further  
 102 analysis and interpretation only from the results of the post-challenge phase.

## 103 2.2 Overview of results

104 Interestingly, for subchallenge 1 and 2, when participants had to use 60 or 40 genes for their  
 105 predictions, the ranking of the best performing teams in the CV scenario did not change significantly  
 106 compared to the challenge (Figure 1B cf. Figures S1 and S2). This was not the case in subchallenge  
 107 3 as no particular team from the top 10 outperformed in a statistically significant way the others  
 108 when using 20 genes for their predictions. The results from the cross-validation showed that the  
 109 approaches generalize well, i.e. the gene selection is performed consistently across the folds and  
 110 the variance of the achieved scores across the folds is small for all teams (Figure S4). For each  
 111 subchallenge we combined the gene selection and location predictions from the top 10 participants  
 112 into a WOC solution (see details below) that performed better compared to the individual solutions  
 113 (Figure 1B). The scores obtained by the best performing teams and the WOC solution are shown in  
 114 Table 1.

115 A summary of the methods used by participants for gene selection and location prediction can  
 116 be seen in Table S2. The most frequently used method by participants for location prediction was a  
 117 similarity based prediction, such as the maximum Matthews correlation coefficient between the  
 118 binarized transcriptomics and the *in situ* that was proposed by Karaikos et al. [10]. Another well  
 119 performing approach was combining the predictions of a machine learning model and the Matthews  
 120 correlation coefficients. The models were trained to predict either the coordinates of each cell or  
 121 the binarized values of the selected *in situ* given transcriptomics data as input. The predictions  
 122 were then made by selecting the location bins that corresponded to the nearest neighbors of the  
 123 predicted values.

124 The most frequently used method by participants for gene selection was unsupervised or  
 125 supervised feature importance estimation and ranking. For example, in a supervised feature  
 126 importance estimation approach a machine learning model is trained to predict the coordinates of  
 127 each cell, given the transcriptomics data at input, that is, the genes with available *in situ* hybridization  
 128 measurements or all genes. Different machine learning models were trained such as Random Forest  
 129 (BCBU, OmicsEngineering) or a neural network (DeepCMC, NAD). There were examples of  
 130 unsupervised feature importance estimation and ranking by expression based clustering (NAD,  
 131 Christoph Hafemeister, MLB), or a greedy feature selection based on predictability of expression  
 132 from other genes (WhatATeam). Background knowledge about location specific marker genes, or  
 133 the expected number of location clusters, was used by a small number of teams (WhatATeam and  
 134 NAD) to inform the gene selection. Given the diversity of approaches to gene selection, we focused  
 135 our analysis on better understanding the properties of frequently selected genes and providing  
 136 recommendations for future experimental designs.



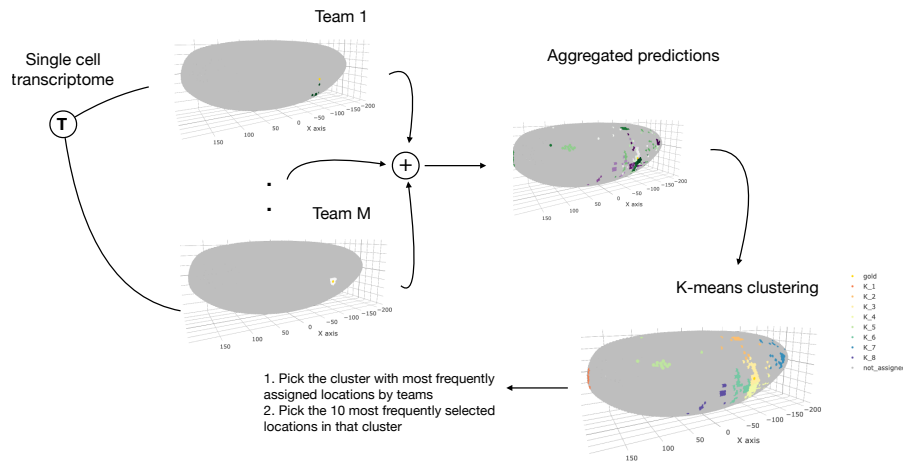


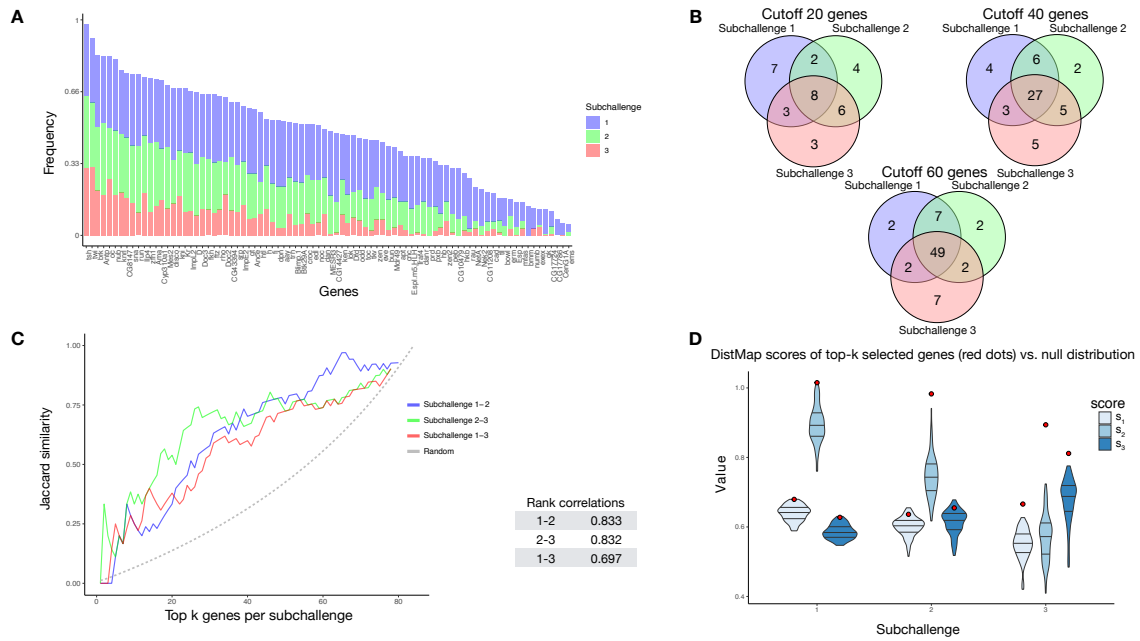
Figure 2: **Wisdom of crowds location prediction.** The location predictions for each cell by the top performing teams in the post-challenge cross-validation phase were aggregated in a wisdom of the crowds solution based on a k-means clustering approach.

### 137 2.3 Analysis of the location prediction

138 A recurrent observation across DREAM challenges is that an ensemble of individual predictions  
139 performs usually better and is more robust than any individual method [16, 17]. This phenomenon,  
140 common also in other contexts, is denoted as the wisdom-of-the-crowds (WOC) [15]. In a typical  
141 challenge, individual methods output a single probability reflecting the likelihood of occurrence of  
142 an event. The WOC prediction is then constructed in an unsupervised manner by averaging the  
143 predictions of individual methods.

144 Given that in the single cell RNAseq prediction challenge participants had to submit 10 positions  
145 per cell, we developed a novel method that is based on k-means clustering to generate the WOC  
146 predictions. A diagram of the k-means approach is given in Figure 2 where for each single cell  
147 we first used k-means clustering to cluster the locations predicted by the individual teams [18]  
148 where the euclidean distance between the locations was used as the distance metric. In order to  
149 find the optimal  $k$ , we used the elbow method, i.e. we chose a  $k$  that saturates the sum of squares  
150 between clusters [19]. Note that each cluster consists of a group of locations and each location  
151 is predicted by one or more teams. Hence, for each cluster we calculated the average frequency  
152 that its constituent locations are predicted by individual teams. We then picked the cluster with  
153 the highest average frequency as our final cluster and ranked each location in this cluster based on  
154 how frequently it was predicted by individual methods. For each cell, the final prediction of the  
155 proposed WOC method consisted of the top 10 locations based on the above ranking. The k-means  
156 approach is based on the intuition that a single cell belongs to one location and its expression is  
157 mostly similar to that of cells in locations surrounding it.

158 The WOC location prediction approach does not take the genes used by the teams to make the  
159 predictions into account. However, after the WOC predictions are generated, in order to score them,  
160 we needed a list of genes for every subchallenge. To this end we used a WOC approach to gene  
161 selection (see the following section for more details) and used the most frequently selected genes  
162 per challenge. As reported above, the WOC solution performed better compared to the individual  
163 solutions (Figure 1B).



**Figure 3: Analysis of gene selection.** The results in all figures were generated from the genes that were selected by the top performing teams in the post-challenge cross-validation scenario. **A.** Frequency of selected genes in subchallenge 1 (blue), subchallenge 2 (green) and subchallenge 3 (red). The genes are ordered according to their cumulative frequency. **B.** Venn diagrams of the most frequently selected genes in the subchallenges with cutoff at 20, 40 and 60 most frequently selected genes, corresponding to the number of genes required for each subchallenge. **C. Left,** the similarity of most frequently selected genes for pairs of subchallenges. The Jaccard similarity measures the ratio of the size of the intersection and the union of two sets  $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$ . **Right,** table of correlations between gene rankings (by frequency) for pairs of subchallenges. **D.** Validation of the performances of the wisdom of the crowds (WOC) selection of genes, i.e the most frequently selected 60, 40 and 20 genes in the respective subchallenges. The violin plots represent null distribution of scores obtained by 100 randomly selected sets of 60, 40 and 20 genes using DistMap. The red dots represents the performance obtained by using DistMap with the WOC selection of genes.

## 164 2.4 Analysis of selected genes

165 The selection of a subset of *in situs* used for cell location prediction was the hallmark that differen-  
 166 tiated the subchallenges. It is unfeasible to evaluate all subsets of 20, 40 or 60 genes from the 84  
 167 due to the immense number of possible combinations of genes. Different approaches and heuristics  
 168 can be used to select a subset of genes and the most frequent among the top 10 ranked teams were  
 169 based on model based feature ranking algorithms, using normalized transcriptomics data (for more  
 170 details see Table S2). However, if a subset of genes is selected as a candidate for solving the general  
 171 task of location prediction, it should be consistently identified when similar sets of single cells are  
 172 used as inputs. Therefore, we analyzed the consistency of gene selection for each team across folds  
 173 by 10-fold cross-validation. More importantly, we were interested in subsets of genes that were  
 174 consistently selected by multiple teams as this could underlie biological relevance.

175 The approaches for selecting genes taken by the top 10 teams resulted in consistent selection  
 176 across folds, significantly better than random, for all subchallenges. Indeed, all of the pairwise  
 177 Jaccard similarities of sets of selected genes for all teams were significantly higher than the expected

178 Jaccard similarity of a random pair of subset of genes (see Supplementary Figure S4). Importantly,  
179 we measured an observable increase in variance and decrease of mean similarity as the number of  
180 selected genes decreased.

181 For each subchallenge we counted the number of times that the genes were selected by all teams  
182 in all folds. The genes, ordered by the frequency of selection in all subchallenges are depicted in  
183 Figure 3A. Forty percent of the top 20, 67% of the top 40 and 81% of the top 60 most frequently  
184 selected genes are the same for all three subchallenges (Figure 3B). The ranks assigned to all genes  
185 in the three subchallenges are highly correlated. Namely, the rank correlations range from 0.69  
186 between subchallenges 1 and 3, to 0.83 between subchallenges 1 and 2, and subchallenges 2 and 3.  
187 Figure 3C shows a plot of the Jaccard similarity of the sets of top-k most frequently selected genes  
188 for pairs of subchallenges. We observe that a high proportion of genes are consistently selected  
189 across subchallenges. The lists of most frequently selected 60, 40 and 20 genes in subchallenges 1,  
190 2 and 3 respectively are available in the supplementary material (Table S3).

191 We conclude that the gene selection is not only consistent by team across folds, but also across  
192 teams and subchallenges. This finding outlines a direction for further analysis, namely the validation  
193 of the predictive performance and analysis of the common properties of the most frequently selected  
194 genes.

#### 195 **2.4.1 Validation of frequently selected genes**

196 We defined a simple procedure to obtain a WOC gene selection for each of the subchallenges. It  
197 consisted on selecting the most frequently selected genes for each subchallenge (different colored  
198 bars in Figure 3A). For example, for subchallenge 1 we chose the 60 most frequently selected genes  
199 looking only at the heights of blue portion of the bar. Interestingly, the 20 most frequently selected  
200 genes in subchallenge 3 are included in the list of 40 most selected genes in subchallenge 2 (except  
201 for *Doc2*), conversely included in the list of 60 most selected genes in subchallenge 1.

202 To validate the predictive performance of the WOC gene selection, we predicted the cell  
203 locations using DistMap and scored the predictions using the same scoring metrics as for the  
204 challenge, estimating the significance of the scores through generated null distributions of scores  
205 for each subchallenge. The null distribution of the scores was generated by scoring the DistMap  
206 location prediction using 100 different sets of randomly selected genes. For each subchallenge and  
207 each score we estimated the empirical distribution function and then calculated the percentile of the  
208 values of the scores obtained with the WOC gene selection.

209 The null distributions and the values of the scores obtained with the WOC gene selection  
210 are shown in Figure 3D. All values of the scores for subchallenge 1 fall in the 99th percentile.  
211 For subchallenge 2  $s_1$  and  $s_3$  fall into the 92nd percentile and  $s_2$  in the 100th percentile. For  
212 subchallenge 3 all scores fall in the 100th percentile. Overall the performance of DistMap with the  
213 WOC selected genes performs significantly better than a random selection of genes. The actual  
214 values of the scores are on par with those achieved by the top 10 teams in the challenge.

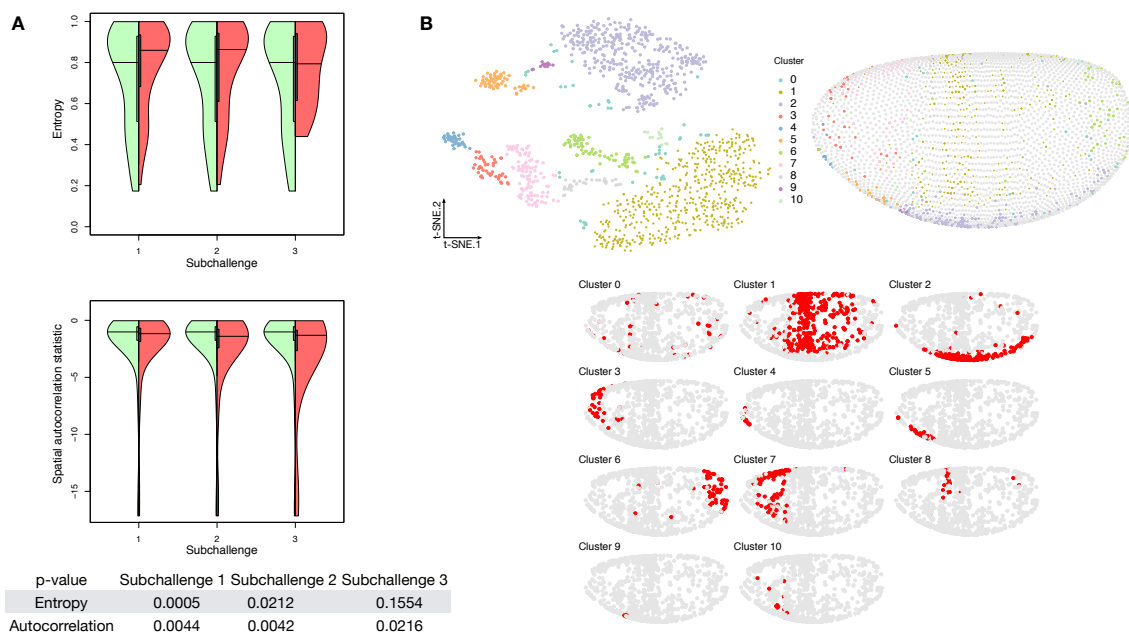
#### 215 **2.4.2 Properties of frequently selected genes**

216 We conjectured that the most frequently selected genes should carry enough information content  
217 collectively to uniquely encode a cell's location. Furthermore, genes should also contain location  
218 specific information, i.e. their expression should cluster well in space. To quantify these features,  
219 we calculated the entropy and the join count statistic for spatial autocorrelation of the *in situ* (see  
220 Figure 4A and Methods for description). We observed that most of the *in situ* genes have relatively  
221 high entropy as observed by the high density in the upper part of the plots and show high spatial  
222 clustering, i.e show values of the join count test statistic lower than zero.



223 To test our conjectures of high entropy and spatial correlation we tested the significance of the  
 224 shift of the values between the WOC selected genes and the non-selected genes from all *in situ* for  
 225 each subchallenge. Since the Shapiro-Wilk test of normality rejected the null-hypothesis for both  
 226 entropy and join count metrics ( $p < 2.3 \cdot 10^{-6}$  and  $p < 1.8 \cdot 10^{-15}$ ) that their values are distributed  
 227 normally for the *in situ* genes, we opted for a nonparametric, one sided Mann-Whitney U test. We  
 228 observed significant value shift for the autocorrelation statistic for all subchallenges 1 to 3 (see  
 229 bottom of Figure 4A right red part of violin plots and table). Although we see a decrease of the  
 230 statistical significance of the mean value shift for the distribution of values of the entropy of the  
 231 selected subsets of genes, the shift is significant for all subchallenges and at the same time, we  
 232 observe that tail of the distribution shortens.

233 To test whether the information relative to different cell types is retained with the selected  
 234 subset of 60, 40 or 20 WOC selected genes, we embedded the cells into 2D space using t-distributed  
 235 stochastic embedding (t-SNE) [20] aiming for high accuracy ( $\theta = 0.01$ ), Figure 4B and Figure S5.  
 236 We then clustered the t-SNE embedded data using density-based spatial clustering of applications  
 237 with noise (DBSCAN) [21]. DBSCAN determines the number of clusters in the data automatically  
 238 based on the density of points in space. The minimum number of cells in a local neighborhood was  
 239 set to 10 and the parameter  $\epsilon = 3.5$  was selected by determining the elbow point in a plot of sorted



**Figure 4: Properties of selected genes.** **A.** Double violin plots of the distribution of entropy and spatial autocorrelation statistic of *Left, green* all *in situ* calculated on all embryonic location bins and *Right, red* the most frequently selected 60, 40 and 20 genes in the respective subchallenges. [bottom table] p-values of a one sided Mann-Whitney U test of location shift comparing the selected (red part of the violin plot) genes vs the non-selected genes. **B.** *Top left*, visualization of the transcriptomics data containing only the most frequently selected 60 genes from subchallenge 1 by the top performing teams (embedding to 2D by t-SNE). Each point (cell) is filled with the color of the cluster that it belongs to (density-based clustering with DBSCAN). *Top right*, spatial mapping of the cells in the *Drosophila* embryo as assigned by DistMap using only the 60 most frequently selected genes from subchallenge 1. The color of each point corresponds to the color of the cluster from the t-SNE visualization. *Bottom*, highlighted (red) location mapping of cells in the *Drosophila* embryo for each cluster separately.

Table 2: Correlations of transcriptomics to *in situ* properties of the genes where both measurements are available.  $\sigma^2$  - variance of a gene across cells,  $c_v$  - coefficient of variation, 0 - number of cells with zero expression,  $H_b$  - entropy of binarized expression,  $H$  - entropy,  $Z$  - join count test statistic

	$\rho$	<i>in situ</i>	
		$H$	$Z$
scRNAseq $\sigma^2$		0.50	0.18
$c_v$		-0.69	0.26
0		-0.64	0.29
$H_b$		0.72	-0.30

240 distances of each cell to its 10th nearest neighbor. We found that the 9 prominent cell clusters  
 241 identified in the study by Karaiskos *et al.* [10] are preserved in our t-SNE embedding and clustering  
 242 experiments when considering the most frequently selected 60 or 40 genes from subchallenges 1  
 243 and 2. The number of clusters of cells with specific localization is reduced when considering the  
 244 most frequently selected 20 genes from subchallenge 3.

245 We next associated the properties of the *in situ* that were found to be indicative of good perfor-  
 246 mance in the task of location prediction with statistical properties of the genes in the transcriptomics  
 247 data. Our goal was to discover statistical properties of the transcriptomics data that might inform  
 248 future experimental designs when selecting target genes for *in situ* hybridizations. We calculated  
 249 statistical features across cells for the subset of genes from the transcriptomics data for which we  
 250 also have *in situ* measurements. These include the variance of gene expression  $\sigma^2$  across cells,  
 251 the coefficient of variation  $c_v = \frac{\sigma}{\mu}$ , the number of cells with expression zero 0 and the entropy of  
 252 binarized expression  $H_b$ . We then calculated the correlation across genes for each of these metrics  
 253 and the measured spatial properties of interest of the *in situ*, i.e entropy  $H$  and the value of the  
 254 joint count statistic  $Z$  (see Table 2). Although the selection of highly variable genes was one of  
 255 the approaches used by some of the top 10 teams, the variance for each gene in the scRNAseq  
 256 expression, although highly correlated to the entropy of the corresponding *in situ* measurements of  
 257 that gene, it is less correlated than other properties. Also, we observed that the positive correlation  
 258 of the entropy to the variance of each gene, becomes a negative correlation against their coefficient  
 259 of variation. This negative correlation can have two sources, the genes with high entropy may have  
 260 low standard deviation or high mean expression. Since we observe positive correlation of entropy  
 261 to the variance of expression, we can conclude that the negative correlation is a result of highly  
 262 expressed genes. Since a known drawback of scRNAseq is a high number of dropout events for  
 263 lowly expressed genes [22], this observation is further supported by the negative correlation of the  
 264 entropy and the number of cells with zero expression. We observed the highest correlation of *in*  
 265 *situ* entropy to the entropy of the binarized expression. Regarding the spatial autocorrelation, all  
 266 statistical features of the transcriptomics were only slightly positively correlated to the join count  
 267 statistic except for the entropy of binarized expression which had slightly negative correlation.

### 268 3 Discussion

269 In this paper we report the results of a crowdsourcing effort organized as a DREAM challenge,  
 270 around the issue of predicting the spatial arrangement of cells in a tissue from scRNAseq data.  
 271 Analysis of the top performing methods and their performance provided us a number of unbiased  
 272 insights. First, it unveiled a connection in the cell-to-cell variability in *Drosophila* embryo gene  
 273 expression and the selection of the best genes for predicting the localization of a cell in the embryo  
 274 from their scRNAseq expression. The most selected genes had a relatively high entropy, hence high  
 275 variance and expression while also showing high spatial clustering. The smaller the number of

276 selected genes, i.e going from subchallenge 1 to 3, the more these features became apparent. The  
 277 observed advantage of genes with high overall expression in cells might lead to less dropout counts  
 278 in the scRNAseq data, a known disadvantage of the technology, leading to more accuracy in the cell  
 279 placement. We also found that the 9 prominent spatially distinct cell clusters previously identified  
 280 [10] are preserved when considering the most frequently selected 60 or 40 genes, but the number  
 281 of clusters is reduced when considering only the most frequently selected 20 genes. This finding  
 282 is in line with the conclusions of Howe et al. [11] where in a related task of location prediction  
 283 the performance stabilized after the inclusion of 30 genes in a related experiment. The WOC gene  
 284 selection and the k-means clustered WOC model for cell localization performed comparably or  
 285 better than the participant's models, showing once more the advantage of the wisdom-of-the-crowds.  
 286 All these results can be explored in animated form at <https://dream-sctc.uni.lu/>.

287 Given that it has been shown that positional information of the anterior-posterior (A-P) axis is  
 288 encoded as early in the embryonic development as when the expression of the gap genes occurs  
 289 [23, 24], we thought that it should be possible to implement in algorithms for this challenge the  
 290 information contained in the regulatory networks of *Drosophila* development [25]. Although only a  
 291 small number of participants, among them the best performers, directly used biological information  
 292 related to the regulation of the genes or their connectivity, the most frequently selected genes in  
 293 all 3 subchallenges have interesting biological properties. Indeed, gap genes such as *giant* (*gt*),  
 294 *kruppel* (*kr*), *knirps* (*kni*) were selected in all 3 subchallenges (see Figure 5 and Table S3 that also  
 295 includes *kni*-like *knrl*) although *tailless* (*tll*) and *hunchback* (*hb*) were not. Along the A-P axis,  
 296 maternally provided *bicoid* (*bcd*) and *caudal* (*cad*) first establish the expression patterns of gap and  
 297 terminal class factors, such as *hb*, *gt*, *kr* and *kni*. These A-P early regulators then collectively direct  
 298 transcription of A-P pair-rule factors, such as *even-skipped* (*eve*), *fushi-tarazu* (*ftz*), *hairy* (*h*), *odd*  
 299 *skipped*, (*odd*), *paired* (*prd*) and *runt* (*run*) which in turn cross-regulate each other. Not being part  
 300 of the *in situs*, neither *bcd*, nor *cad* were selected but *ama* sitting near *bcd* in the genome might  
 301 have been selected for its similar expression properties. Furthermore, we also find that pair-rule  
 302 genes were most prominently selected in subchallenges 1 (*eve*, *odd*, *prd*, the Paired-like *bcd* and  
 303 *bcd*) and 2 (*h*, *ftz* and *run*). A similar cascade of maternal and zygotic factors controls patterning

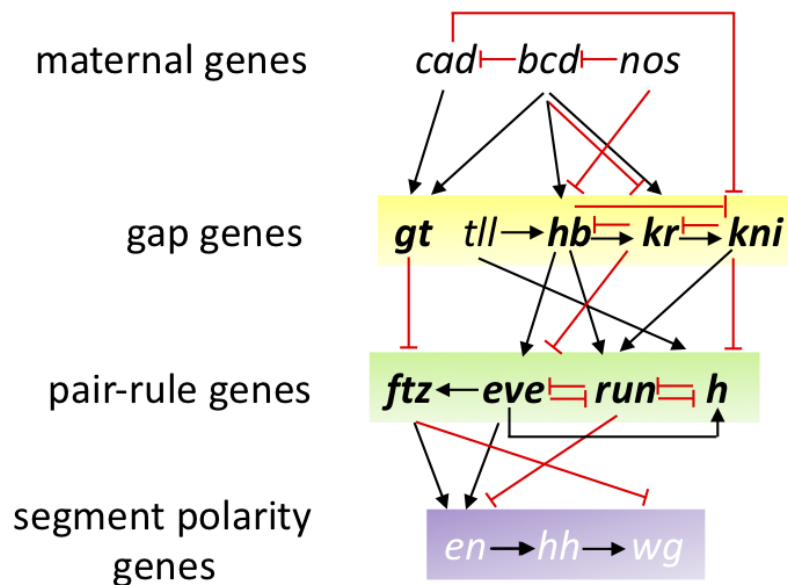


Figure 5: Gene regulatory network of early *Drosophila* development. Not all regulations are represented, nor pair ruled genes *odd* & *prd*. Frequently selected genes are represented in bold.

304 along the dorsal-ventral (D-V) axis were *dorsal* (*d*), *snail* (*sna*) and *twist* (*twi*) specify mesoderm  
305 and the pair rule factors *eve* and *ftz* specify location along the trunk of the A-P axis. Again, *sna* and  
306 *twi* were selected in all subchallenges and *d* in subchallenges 1 and 2. These selected transcription  
307 factors specify distinct developmental fates and can act via different cis-regulatory modules but  
308 their quantitative differences in relative levels of binding to shared targets correlates with their  
309 known biological and transcriptional regulatory specificities [26]. The rest of the selected genes  
310 were the homeobox genes (*nub*, *antp*) and differentiators of tissue such as mesoderm (*ama*, *mes2*,  
311 *zfh1*), ectoderm (*doc2* and *doc3*), neural tissue (*noc*, *oc*, *rho*) and EGFR pathway (*rho*, *edl*). The  
312 complete lists of most frequently selected genes are available in Table S3.

313 Since the ground truth of single cell locations was publicly available, the organization of this  
314 DREAM challenge brought risks that, given the importance of the scientific question asked, we  
315 thought worth taking. However, without the post-challenge phase it would have been impossible to  
316 distinguish the robust and sound methods from methods that were overfitting the results. Overall,  
317 the single cell transcriptomics challenges unveils not only the best gene-selection methods and  
318 prediction approaches to localize a cell in the *Drosophila* embryo, but also explains the biological  
319 and statistical properties of the genes selected for the predictions. Further identification of additional  
320 properties such as spatially autocorrelated genes might require the use of alternative scRNAseq  
321 focused approaches [13, 14]. However, we think that the approach defined here could be used or  
322 adapted when performing similar cell-placing tasks in other organisms, including human tissues.  
323 Given the importance of spatial arrangements for disease development and treatment, we foresee  
324 an application of these methods to medical questions as well.

## 325 References

- 326 [1] Anna K Casasent, Aislyn Schalek, Ruli Gao, Emi Sei, Annalyssa Long, William Pangburn,  
327 Tod Casasent, Funda Meric-Bernstam, Mary E Edgerton, and Nicholas E Navin. Multiclonal  
328 invasion in breast tumors identified by topographic single cell sequencing. *Cell*, 172(1-2):205–  
329 217, 2018.
- 330 [2] Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro,  
331 Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, et al.  
332 Visualization and analysis of gene expression in tissue sections by spatial transcriptomics.  
333 *Science*, 353(6294):78–82, 2016.
- 334 [3] Ditte Lovatt, Brittani K Ruble, Jaehee Lee, Hannah Dueck, Tae Kyung Kim, Stephen Fisher,  
335 Chantal Francis, Jennifer M Spaethling, John A Wolf, M Sean Grady, et al. Transcriptome in  
336 vivo analysis (tiva) of spatially defined single cells in live tissue. *Nature methods*, 11(2):190,  
337 2014.
- 338 [4] Samuel G Rodrigues, Robert R Stickels, Aleksandrina Goeva, Carly A Martin, Evan Murray,  
339 Charles R Vanderburg, Joshua Welch, Linlin M Chen, Fei Chen, and Evan Z Macosko. Slide-  
340 seq: A scalable technology for measuring genome-wide expression at high spatial resolution.  
341 *Science*, 363(6434):1463–1467, 2019.
- 342 [5] Yang Liu, Mingyu Yang, Yanxiang Deng, Graham Su, Cindy C Guo, Di Zhang, Dongjoo Kim,  
343 Zhiliang Bai, Yang Xiao, and Rong Fan. High-spatial-resolution multi-omics atlas sequencing  
344 of mouse embryos via deterministic barcoding in tissue. *bioRxiv*, page 788992, 2019.
- 345 [6] Kaia Achim, Jean-Baptiste Pettit, Luis R Saraiva, Daria Gavriouchkina, Tomas Larsson,  
346 Detlev Arendt, and John C Marioni. High-throughput spatial mapping of single-cell rna-seq  
347 data to tissue of origin. *Nature biotechnology*, 33(5):503, 2015.

- 348 [7] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial  
349 reconstruction of single-cell gene expression data. *Nature biotechnology*, 33(5):495, 2015.
- 350 [8] Keren Bahar Halpern, Rom Shenhav, Orit Matcovitch-Natan, Beáta Tóth, Doron Lemze,  
351 Matan Golan, Efi E Massasa, Shaked Baydatch, Shanie Landen, Andreas E Moor, et al.  
352 Single-cell spatial reconstruction reveals global division of labour in the mammalian liver.  
353 *Nature*, 542(7641):352, 2017.
- 354 [9] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi,  
355 William M Mauck III, Yuhao Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija.  
356 Comprehensive integration of single-cell data. *Cell*, 2019.
- 357 [10] Nikos Karaiskos, Philipp Wahle, Jonathan Alles, Anastasiya Boltengagen, Salah Ayoub,  
358 Claudia Kipar, Christine Kocks, Nikolaus Rajewsky, and Robert P Zinzen. The drosophila  
359 embryo at single-cell transcriptome resolution. *Science*, 358(6360):194–199, 2017.
- 360 [11] Douglas G Howe, Yvonne M Bradford, Tom Conlin, Anne E Eagle, David Fashena, Ken  
361 Frazer, Jonathan Knight, Prita Mani, Ryan Martin, Sierra A Taylor Moxon, et al. Zfin, the  
362 zebrafish model organism database: increased support for mutants and transgenics. *Nucleic  
363 acids research*, 41(D1):D854–D860, 2012.
- 364 [12] Charless C Fowlkes, Cris L Luengo Hendriks, Soile VE Keränen, Gunther H Weber, Oliver  
365 Rübél, Min-Yu Huang, Sohail Chatoor, Angela H DePace, Lisa Simirenko, Clara Henriquez,  
366 et al. A quantitative spatiotemporal atlas of gene expression in the drosophila blastoderm.  
367 *Cell*, 133(2):364–374, 2008.
- 368 [13] Josephine Bageritz, Philipp Willnow, Erica Valentini, Svenja Leible, Michael Boutros, and  
369 Aurelio A Teleman. Gene expression atlas of a developing tissue by single cell expression  
370 correlation analysis. *Nature Methods*, 16(8):750, 2019.
- 371 [14] Mor Nitzan, Nikos Karaiskos, Nir Friedman, and Nikolaus Rajewsky. Charting a tissue from  
372 single-cell transcriptomes. *bioRxiv*, page 456350, 2018.
- 373 [15] Julio Saez-Rodriguez, James C Costello, Stephen H Friend, Michael R Kellen, Lara Man-  
374 gravite, Pablo Meyer, Thea Norman, and Gustavo Stolovitzky. Crowdsourcing biomedical  
375 research: leveraging communities as innovation engines. *Nature Reviews Genetics*, 17(8):470–  
376 486, 2016.
- 377 [16] Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M  
378 Camacho, Kyle R Allison, Andrej Aderhold, Richard Bonneau, Yukun Chen, et al. Wisdom  
379 of crowds for robust gene network inference. *Nature methods*, 9(8):796, 2012.
- 380 [17] Michael P Menden, Dennis Wang, Mike J Mason, Bence Szalai, Krishna C Bulusu, Yuanfang  
381 Guan, Thomas Yu, Jaewoo Kang, Minji Jeon, Russ Wolfinger, et al. Community assessment  
382 to advance computational prediction of cancer drug combinations in a pharmacogenomic  
383 screen. *Nature communications*, 10(1):2674, 2019.
- 384 [18] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm.  
385 *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.
- 386 [19] Trupti M Kodinariya and Prashant R Makwana. Review on determining number of cluster in  
387 k-means clustering. *International Journal*, 1(6):90–95, 2013.
- 388 [20] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of  
389 Machine Learning Research*, 9:2579–2605, 2008.



- 390 [21] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm  
391 for discovering clusters in large spatial databases with noise. In *Proceedings of the Second*  
392 *International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226–  
393 231, 1996.
- 394 [22] Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell  
395 differential expression analysis. *Nature methods*, 11(7):740, 2014.
- 396 [23] Julien O Dubuis, Gašper Tkačik, Eric F Wieschaus, Thomas Gregor, and William Bialek. Posi-  
397 tional information, in bits. *Proceedings of the National Academy of Sciences*, 110(41):16301–  
398 16308, 2013.
- 399 [24] Mariela D Petkova, Gašper Tkačik, William Bialek, Eric F Wieschaus, and Thomas Gregor.  
400 Optimal decoding of cellular identities in a genetic network. *Cell*, 176(4):844–855, 2019.
- 401 [25] David Umulis, Michael B O'Connor, and Hans G Othmer. Robustness of embryonic spatial  
402 patterning in drosophila melanogaster. *Current topics in developmental biology*, 81:65–111,  
403 2008.
- 404 [26] Stewart MacArthur, Xiao-Yong Li, Jingyi Li, James B Brown, Hou Cheng Chu, Lucy Zeng,  
405 Brandi P Grondona, Aaron Hechmer, Lisa Simirenko, Soile VE Keränen, et al. Developmental  
406 roles of 21 drosophila transcription factors are determined by quantitative differences in  
407 binding to an overlapping set of thousands of genomic regions. *Genome biology*, 10(7):R80,  
408 2009.
- 409 [27] Andrew D. Cliff and John K. Ord. *Spatial autocorrelation*. Pion London, 1973.
- 410 [28] Robert R. Solak and Neal L. Oden. Spatial autocorrelation in biology: 1. methodology.  
411 *Biological Journal of the Linnean Society*, 10(2):199–228, 1978.

## 412 **4 Methods**

### 413 **4.1 Scoring**

414 We scored the submissions for the three subchallenges using three metrics  $s_1$ ,  $s_2$  and  $s_3$ .  $s_1$  measured  
415 how well the expression of the cell at the predicted location correlates to the expression from the  
416 reference atlas and included the variance of the predicted locations for each cell. While  $s_2$  measured  
417 the accuracy of the predicted location and  $s_3$  measured how well the gene-wise spatial patterns  
418 were reconstructed.

419 Let  $c$  represent the index of a cell, given in the transcriptomics data in the challenge where  
420  $1 \leq c \leq 1297$ . Each cell  $c$  is located in a bin  $\varepsilon_c \in \{1..3039\}$  at a position with coordinates  
421  $r(\varepsilon_c) = (x_c, y_c, z_c)$ . Each cell is associated with a binarized expression profile  $t_c = (t_{c1}, t_{c2}, \dots, t_{cE})$ ,  
422 where  $1 \leq E \leq 8924$ , and a corresponding binarized *in situ* profile  $f_c = (f_{c1}, f_{c2}, \dots, f_{cK})$ , where  
423 the maximum possible value of  $K$  for which we have *in situ* information is  $K = 84$ . For different  
424 subchallenges we consider  $K \in \{20, 40, 60\}$ . Using  $K$  selected genes the participants were asked to  
425 provide an ordered list of 10 most probable locations for each cell. We represent with the mapping  
426 function  $A(c, i, K)$  the value of the predicted  $i$ -th most probable location for cell  $c$  using  $K$  *in situs*.

427 For the first scoring metric  $s_1$  we calculated the weighted average of the Mathews correlation  
428 coefficient (MCC) between the *in situ* profile of the ground truth cell location  $f_{\varepsilon_c}$  and the *in situ*  
429 profile of the most probable predicted location  $f_{A(c,1,K)}$  for that cell

$$s_1 = \sum_{c=1}^N \frac{p_K(c, A)}{\sum_{i=1}^N p_K(i, A)} MCC(f_{A(c,1,K)}, f_{\varepsilon_c}),$$

430 where  $N$  is the total number of cells with predicted locations.

431 The Matthews correlation coefficient, or  $\phi$  coefficient, is calculated from the contingency table  
 432 obtained by correlating two binary vectors. The MCC is weighted by the inverse of the distance of  
 433 the predicted most probable locations to the ground truth location  $p_K(c)$ . The weights are calculated  
 434 as  $p_K(c, A) = \frac{\widetilde{d_{84}(c, A)}}{d_K(c, A)}$ , where  $d_K(c, A) = \frac{1}{10} \sum_{i=1}^{10} \|r(A(c, i, K)) - r(\epsilon_c)\|_2$ ,  $\widetilde{d_{84}(c, A)}$  is the value of  
 435  $d_K(c, A)$  using the ground truth most probable locations assigned with  $K = 84$  using DistMap, and  
 436  $\|\cdot\|_2$  is the Euclidean norm.

The second metric  $s_2$  is simply the average inverse distance of the predicted most probable locations to the ground truth location

$$s_2 = \frac{1}{N} \sum_{c=1}^N p_K(c, A).$$

437 Finally, the third metric  $s_3$  measures the accuracy of reconstructed gene-wise spatial patterns

$$s_3 = \sum_{s=1}^K \frac{MCC(t_{cs}, f_{\epsilon_{cs}})_{\forall c}}{\sum_{i=1}^K MCC(t_{ci}, f_{\epsilon_{ir}})_{\forall c}} MCC(t_{cs}, f_{A(c, 1, K)_s})_{\forall c},$$

438 where  $\forall c$  denotes that the  $MCC$  is calculated cell wise for each gene.

439 For 287 out of the 1297 cells, the ground truth location predictions were ambiguous, i.e., the  
 440  $MCC$  scores were identical for multiple locations. These cells were removed both from the ground  
 441 truth and the submissions before calculating the scores.

442 The teams were ranked according to each score independently. The final assigned rank  $r_t$   
 443 for team  $t$  was calculated as the average rank across scores. Teams were ranked based on the  
 444 performance as measured by the three scores on 1000 bootstrap replicates of the submitted solutions.  
 445 The three scores were calculated for each bootstrap. The teams were then ranked according to  
 446 each score. These ranks were then averaged to obtain a final rank for each team on that bootstrap.  
 447 The winner for each subchallenge was the team that achieved the lowest ranks. We calculated the  
 448 Bayes factor of the bootstrap ranks for the top performing teams. Bayesian factor of 3 or more was  
 449 considered as a significantly better performance. The Bayes factor of the 1000 bootstrapped ranks  
 450 of teams  $T_1$  and  $T_2$  was calculated as

$$BF(T_1, T_2) = \frac{\sum_{i=1}^{1000} \mathbf{1}(r(T_1)_i < r(T_2)_i)}{\sum_{i=1}^{1000} \mathbf{1}(r(T_1)_i > r(T_2)_i)},$$

451 where  $r(T_1)_i$  is the rank of team  $T_1$  on the  $i$ -th bootstrap,  $r(T_2)_i$  is the rank of team  $T_2$  on the  $i$ -th  
 452 bootstrap, and  $\mathbf{1}$  is the indicator function.

## 453 4.2 Entropy and spatial autocorrelation

The entropy of a binarized *in situ* measurements of gene  $G$  was calculated as

$$H(G) = -p \log_2 p - (1 - p) \log_2 (1 - p),$$

454 where  $p$  is the probability of gene  $G$  to have value 1. In other words,  $p$  is the fraction of cells where  
 455  $G$  is expressed.

456 The join count statistic is a measure of a spatial autocorrelation of a binary variable. We will  
 457 refer to the binary expression 1 and 0 as black ( $B$ ) and white ( $W$ ). Let  $n_B$  be the number of bins  
 458 where  $G$  is expressed ( $G = B$ ), and  $n_W = n - n_B$  the number of bins where  $G$  is not expressed  
 459 ( $G = W$ ). Two neighboring spatial bins can form join of type  $J \in \{WW, BB, BW\}$ .

460 We are interested in the distribution of BW joins. If a gene has a lower number of BW joins  
 461 that the expected number of BW, then the gene is positively spatially autocorrelated, i.e., the gene is  
 462 highly clustered. Contrarily, higher number of BW joins points towards negative spatial correlation,  
 463 i.e. dispersion.

Following Cliff and Ord [27] and Sokal and Oden [28], the expected count of BW joins is

$$\mathbb{E}[BW] = \frac{1}{2} \sum_i \sum_j \frac{w_{ij} n_B^2}{n^2},$$

where the spatial connectivity matrix  $w$  is defined as

$$w_{ij} = \begin{cases} 1 & \text{if } i \neq j \text{ and } j \text{ is in the list of 10 nearest neighbors of } i \\ 0 & \text{otherwise} \end{cases}$$

The variance of BW joins is

$$\sigma_{BW}^2 = \mathbb{E}[BW^2] - \mathbb{E}[BW]^2.$$

where the term  $\mathbb{E}[BW^2]$  is calculated as

$$\mathbb{E}[BW^2] = \frac{1}{4} \left( \frac{2x_2 n_B n_W}{n^2} + \frac{(x_3 - 2x_2) n_B n_W (n_B + n_W - 2)}{n^3} + \frac{4(x_1^2 + x_2 - x_3) n_B^2 n_W^2}{n^4} \right),$$

464 where  $x_1 = \sum_i \sum_j w_{ij}$ ,  $x_2 = \frac{1}{2} \sum_i \sum_j (w_{ij} - w_{ji})^2$ ,  $x_3 = \sum_i (\sum_j w_{ij} + \sum_j w_{ji})^2$ .

465 Note that the connectivity matrix  $w$  can also be asymmetric, since it is defined by the nearest  
 466 neighbor function.

Finally, the observed BW counts are

$$BW = \frac{1}{2} \sum_i \sum_j w_{ij} (G_i - G_j)^2.$$

The join counts test statistic is then defined as

$$Z(BW) = \frac{BW - \mathbb{E}[BW]}{\sqrt{\sigma_{BW}^2}},$$

467 which is assumed to be asymptotically normally distributed under the null hypothesis of no spatial  
 468 autocorrelation. Negative values of the  $Z$  statistic represent positive spatial autocorrelation, or  
 469 clustering, of gene  $G$ . Positive values of the  $Z$  statistic represent negative spatial autocorrelation, or  
 470 dispersion, of gene  $G$ .

### 471 4.3 Implementation details

472 The challenge scoring was implemented and run in R version 3.5, the post analysis was performed  
 473 with R version 3.6 and the core tidyverse packages. We used the publicly available implemen-  
 474 tation of DistMap (<https://github.com/rajewsky-lab/distmap>). MCC calculated  
 475 with R package mccr (0.4.4). t-SNE embedding and visualization produced with R package  
 476 Rtsne (0.15). DBSCAN clustering with R package dbscan (1.1-4).

### 477 4.4 Code availability

478 <https://github.com/dream-sctc/Scoring>

## 479 **4.5 Data description**

480 **Reference Database** The reference database comes from the Berkeley *Drosophila* Transcription  
481 Network Project. The *in situ* expression of 84 genes (columns) is quantified across the 3039  
482 *Drosophila* embryonic locations (rows) for raw data and for binarized data. The 84 genes were  
483 binarized by manually choosing thresholds for each gene.

484 **Spatial coordinates** One half of *Drosophila* embryo has 3039 cells places as x, y and z (columns)  
485 for a total of 3039 embryo locations (rows) and a total of 3039·3 coordinates.

486 **Single cell RNA sequencing** The single-cell RNA sequencing data is provided as a matrix with  
487 8924 genes as rows and 1297 cells as columns. In the raw version of the matrix, the entries are the  
488 raw unique gene counts (quantified by using unique molecular identifiers – UMI). The normalized  
489 version is obtained by dividing each entry by the total number of UMIs for that cell, adding a  
490 pseudocount and taking the logarithm of that. All entries are finally multiplied by a constant. For a  
491 given gene and only considering the Drop-seq cells expressing it we computed a quantile value  
492 above (below) which the gene would be designated ON (OFF). We sampled a series of quantile  
493 values and each time the gene correlation matrix based on this binarized version of normalized data  
494 versus the binarized BDTNP atlas was computed and compared by calculating the mean square  
495 root error between the elements of the lower triangular matrices. Eventually, the quantile value  
496 0.23 was selected, as it was found to minimize the distance between the two correlation matrices.

497 The short sequences for each of the 1297 cells in the raw and normalized data are the cell  
498 barcodes.

## 499 **5 Acknowledgments**

500 This research was funded in part by PROACTIVE 2017 “From Single-Cell to Multi-Cells Informa-  
501 tion Systems Analysis” (C92F17003530005 Department of Information Engineering, University of  
502 Padova) for B.D.C.; National Institutes of Health grant number U54CA21729 for J.R.; ICMR JRF  
503 (Indian Council of Medical Research - Junior Research Fellowship) for S.A. and X.W. was funded  
504 by the National Natural Science Foundation of China (No.61702421 and No.61772426).

## 505 **6 Author contributions**

506 Conceptualization, N.K., N.R., J.S.R., G.S., and P.M.; Methodology, J.S., M.E.A., G.S., and P.M.;  
507 Software, J.T., and M.E.A.; Formal Analysis, J.T., M.E.A., G.S. and P.M.; Writing - Original Draft,  
508 J.T. and P.M.; Writing - Supervision, J.S.R., G.S., and P.M. - R.K, E.G. and P.B produced animated  
509 figures of results at <https://dream-sctc.uni.lu/>

## 510 **7 Competing interests**

511 The authors declare no competing interests.

## 512 **8 Materials and Correspondence**

513 Requests for data, resources, and or reagents should be directed to Pablo Meyer ([pmeyerr@us.ibm.com](mailto:pmeyerr@us.ibm.com)).