Improving Single-Cell RNA-seq Clustering by Integrating Pathway

Chenxing Zhang [1], Lin Gao [1,*], Bingbo Wang [1,*] and Yong Gao[2]

[1] School of Computer Science and Technology, Xidian University, Xi'an 710071, China.

[2] Department of Computer Science, The University of British Columbia Okanagan, Kelowna, BC, V1V 1V5, Canada

Corresponding author:

Lin Gao, School of Computer Science and Technology, Xidian University, 710071, Xi'an, China. Tel.: +86-29-88202354; Email: lgao@mail.xidian.edu.cn

Bingbo Wang, School of Computer Science and Technology, Xidian University, Xi'an 710071, China. Tel.: 86-29-88202354; E-mail: bingbowang@xidian.edu.cn

ABSTRACT

Single-cell clustering is an important part of analyzing single-cell sequencing data. However, the accuracy and robustness of the existing methods are not satisfactory. We speculate that these methods ignore the relationship between genes, making them more susceptible to noise, resulting in low performance. Pathway is a collection of relationships between genes, integrating pathway-level features may improve the clustering method. In this work, we studied the impact on accuracy and robustness of single-cell clustering method by integrating pathway. We collected 10 state-of-art single-cell clustering methods, 26 scRNA-seq data and 4 pathway databases, combined the AUCell method and the similarity network fusion (SNF) to integrate pathway data and scRNA-seq data, and introduced 3 accuracy indicators, 2 noise generation strategies and robustness indicators. Experiments on this framework showed that, integrating pathway can significantly improve the accuracy and robustness of most single-cell clustering methods.

KEY WORDS

single-cell clustering; scRNA-seq; pathway; accuracy; robustness

INTRODUCTION

Cell type identification is aimed at analyzing single-cell sequencing data (e.g. single-cell RNA sequencing data, scRNA-seq) to understand cell heterogeneity [1–3]. As an important part of it, many single-cell clustering methods have been proposed. Stuart et al. introduced a single-cell clustering method named Seurat, that uses the modular optimization method on cell-cell networks constructed from single-cell RNA-seq data by high variance gene selection and principle component analysis[4]. Kiselev et al. designed a single-cell consensus clustering SC3, that combines multiple k-means clustering solutions into a hierarchical clustering[5]. Xu and Sun proposed SNN-Cliq, which clusters cells by clique-based methods on a shared nearest neighbor network constructed from the single-cell RNA-seq data[6]. Kiselev et al. systematically analyzed advantages and limitations of 14 single-cell clustering method, such as Seurat, SC3, CIDR, pcaReduce, SNN-Cliq, etc.[7].

However, as compared to the RNA-seq data obtained from bulk cell population, single-cell RNA-seq data are much noisier and sparser due to the particular sequencing techniques and experiment protocols[2]. A good example is the existence of a large number of drop-out events where a gene

This article is the authors original version (AOV) and has been accepted for publication in *Briefings in Bioinformatics*, published by Oxford University Press. The version of record *Chenxing Zhang, Lin Gao, Bingbo Wang, Yong Gao, Improving Single-Cell RNA-seq Clustering by Integrating Pathways, Briefings in Bioinformatics, 2021* is available online at: https://doi.org/10.1093/bib/bbab147.

expression is supposed to exist but not detected. The high level of noise and sparsity in the single-cell RNA-seq data creates significant difficulties for clustering methods that current single-cell clustering approaches are based on[7]. In addition, most single-cell clustering method only use genes as feature of cells, ignoring the relationship between genes. We speculated that it could make clustering methods more susceptible to noise, resulting in low accuracy and robustness.

A pathway is a collection of relationship between genes which regular same biological process[8]. The noise on a single gene may have a relatively small impact on the entire pathway, such as the neuronal differentiation pathway[9] and the oncogenic signaling pathways[10]. It has also been suggested that the relationship between genes may be beneficial to cell type identification[11,12]. These motivated us to consider the possibility of using pathway-level features to improve the single-cell clustering method. We noted that in recent work, Wang et al. showed that pathway signals extracted from single cell RNA-seq data can be used to effectively classify and cluster heterogenous cell populations[13].

In this work, we studied the effectiveness of integrating pathway and single cell RNA-seq data for single cell clustering, focusing on the clustering accuracy and robustness. To quantify the performance improvement the integrative approach can provide, we designed a framework, consisting of 10 state-of-art single-cell clustering methods, 26 scRNA-seq data, 4 pathway databases, 2 accuracy quantification indicators, 2 noise generation strategies and the corresponding robustness indicators. Using the framework, we studied the impact of integrating pathway on the cell-cell similarity matrices that a single-cell clustering uses as its input and the performance of clustering methods, we found that integrating pathway could significantly improve accuracy and robustness of single-cell clustering method. Our observations, together with our further analysis on ranking the methods before and after integrating pathway based on each indicator, provide a strong support for the finding reported in the literature that integrating pathway can potentially help provide more effective and stable cell type signals [9,14,15].

RESULT

To evaluate the accuracy and robustness of single-cell clustering method by integrating pathway, we designed a framework, consisting of three parts:

1) Sufficient materials, including 26 scRNA-seq data, 4 pathway databases and 10 state-of-art single-cell clustering methods. (Details are in Table 1-3 and Supplement Text 1,2)

2) An integration strategy, integrating pathway and scRNA-seq data into the cell-cell similarity matrices that a single-cell clustering uses as its input. (Step 1 in Figure 1, Details are in Supplement Text 3)

3) A series of evaluation indicators, quantifying the performance of clustering. It contains 3 accuracy quantification indicators, 2 noise simulated strategies and the corresponding robustness indicators.

By comparing these indicators before and after integrating pathway, we observed significant improvement of accuracy and robustness of clustering method. (Step3 in Figure 1; Details are in Figure 3 and Supplement Text 4)

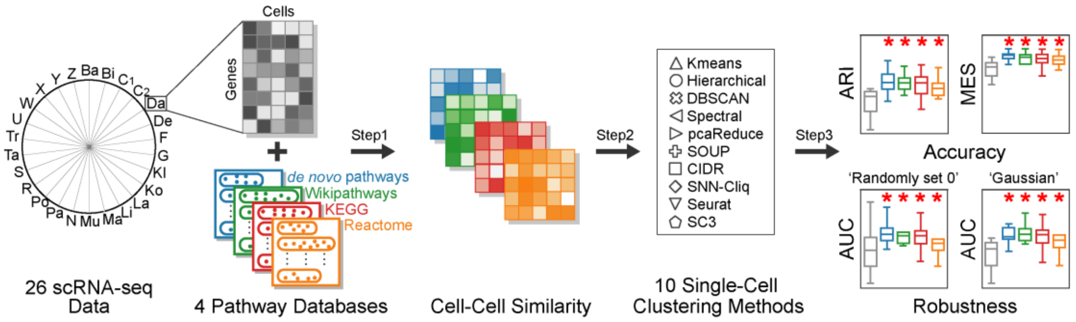

Figure 1. The framework of evaluating the single-cell clustering method by integrating pathway.

Step1. Integrating scRNA-seq and pathway into cell-cell similarity metrics (Details are in Supplement Text 3).

Step2. Inputting the cell-cell similarity metrics into single-cell clustering methods and get the clustering results.

Step3. Evaluating accuracy and robustness of single-cell clustering methods by integrating pathway (Details are in Figure 4).

Collection of Sufficient Materials

In our framework, we collected 4 pathway databases (Table 1; Supplement Text 1), 10 state-of-art single-cell clustering method (Table 2; Supplement Text 2) and 26 scRNA-seq data (Table 3). These materials will be used in our subsequent analysis of improvement on cell-cell similarity metrics and clustering methods.

pathway databases

Among the four pathway databases, three of them are public pathway databases and one is *de novo* pathway database. The details of these pathway databases are described in Supplement Text 1

Table 1. The detail of pathway datasets.

Pathway Database	Ref	# Items	
		Mouse	Human
KEGG[8]	(Kanehisa et al., 2017)	394	396
Reactome[16]	(Fabregat et al., 2018)	1623	2213
Wikipathways[17]	(Slenter et al., 2018)	220	601
de novo pathway[14]	(Ji and Ji, 2016)	150	

single-cell clustering method

Among the ten state-of-the-art clustering methods, we introduced four traditional clustering methods and six single-cell clustering methods. The details of these methods are described in Supplement Text 2.

Table 2. The detail of single-cell clustering methods.

Methods	Ref	Methods	Ref
Kmeans[18]	(Lloyd, 1982)	SOUP[19]	(Zhu et al., 2018)
Hierarchical[20]	(Ward, 1963)	CIDR[21]	(Lin et al., 2017)
Spectral[22]	(Shi et al., 2000)	pcaReduce[23]	(Žurauskien, 2016)
DBSCAN[24]	(Daszykowski et al, 2009)	SNN-Cliq[6]	(Xu and Su, 2015)
Seurat[4]	(Stuart et al., 2019)	SC3[5]	(Kiselev et al., 2017)

single-cell datasets

The 26 single-cell RNA-seq datasets were downloaded from the website maintained by Hemberg's lab (https://hemberg-lab.github.io/scRNA.seq.datasets/), including 12 datasets from human and 14 from mouse. The details are shown in the table below:

Table 3. The detail of scRNA-seq datasets.

Single-cell Data	#Type(#Cell)	Single-cell Data	#Type(#Cell)
Baron[25]	13(1886)	Muraro[26]	10(2126)
Biase[27]	5(52)	Nestorowa[28]	9(1656)
Camp1[29]	7(777)	Patel[30]	5(430)
Camp2[31]	6(734)	Pollen[32]	11(301)
Darmanis[33]	9(466)	Romanov[34]	7(2881)
Deng[35]	12(280)	Segerstolpe[36]	15(3514)
Fan[37]	6(66)	Tasic[38]	18(1679)
Goolam[39]	5(124)	Treutlein[40]	5(80)
Klein[41]	4(2717)	Usoskin[42]	11(622)
Kolodziejczyk[43]	3(704)	Wang[44]	8(635)
Lake[45]	16(3042)	Xin[46]	8(1600)
Li[47]	9(561)	Yan[48]	6(90)
Manno[49]	32(2150)	Zeisel[50]	9(3005)

Improvement of Cell-Cell Similarity Metrics by integrating pathway

In order to make the step of integrating pathway applicable for different clustering methods, we first integrated the scRNA-seq and pathway into cell-cell similarity metrics, and then bring it into different methods. Therefore, the similarity between the same type of cells and the dissimilarity between different types affect the performance of the clustering method. Combining this similarity and dissimilarity, in our framework, we called the quality of cell-cell similarity metrics. To quantify the quality of cell-cell similarity metrics, we introduced the existing indicators, such as Silhouette coefficient, Davies-Bouldin score and Calinski-Harabasz score (see Supplement Text 4 for the details). Although these indicators are used to evaluate accuracy based on clustering results and cell-cell similarity metrics, we replace the clustering results with known/true cell type labels, the higher value of these indicators represent that cells with known same types are more similar and cells with known different types are more dissimilar, that is, higher quality of cell-cell similarity metrics.

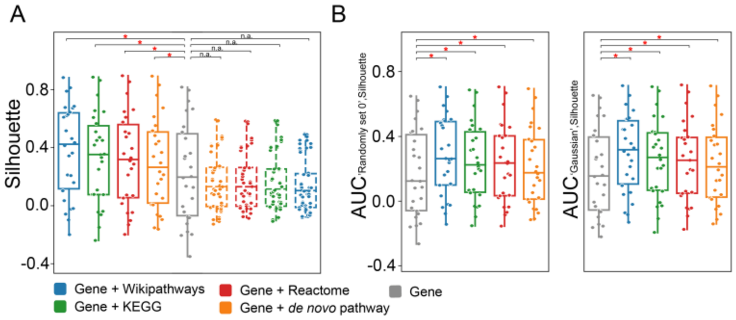

Figure 2. Improvement of cell-cell similarity metric for clustering. **(A)** Silhouette of cell-cell similarity metric (colored solid box) and that of which integrate permutated pathway (colored dashed box) on each scRNA-seq data. **(B)** The area under 'randomly set 0' (left side) and 'Gaussian' (right side) noise proportion – Silhouette curve on noisy cell-cell similarity metric which integrating pathway and noisy scRNA-seq data. The red star indicates significant improvement (P<0.05, Wilcoxon signed-rank test, one-sided). The 'n.a.' indicates nonsignificant improvement (P≥0.05). Each dot indicates a scRNA-seq data.

We compared the Silhouette coefficients of cell-cell similarity before and after integrating pathway on each scRNA-seq data, and found that, after integrating pathway, Silhouette coefficient has been significantly improved. the average improvement rates of *de novo* pathway, Wikipathways, KEGG and Reactome are 67.5% (P=9e-5), 47.3% (P=1e-3), 41.2% (P=1e-3) and 28.5% (P=1e-2),

This article is the authors original version (AOV) and has been accepted for publication in *Briefings in Bioinformatics*, published by Oxford University Press. The version of record *Chenxing Zhang, Lin Gao, Bingbo Wang, Yong Gao, Improving Single-Cell RNA-seq Clustering by Integrating Pathways, Briefings in Bioinformatics, 2021* is available online at: https://doi.org/10.1093/bib/bbab147.

respectively (Figure 4A, left side). In addition, we integrated 'random' pathway (permutating gene signature in the pathway) and each scRNA-seq data into cell-cell similarity, Silhouette coefficient is not significantly improved (Figure 4A, right side). CHscore and DBscore also show the similar improvement and significance (Supplement Figure 19). These results indicate that integrating pathway improve the quality of cell-cell similarity metrics.

We also compared quality of cell-cell similarity metrics before and after integrating pathway on noisy scRNA-seq data. In our framework, we generated two types of noise, randomly set 0 and Gaussian noise. These noises are added to the data in a specific proportion (5%, 10%, 15% and 20%) and to obtain the noisy scRNA-seq data respectively. We integrated noisy scRNA-seq data and pathway into noisy cell-cell similarity. As the proportion of noise increases, the quality of noisy cell-cell similarity metrics will decrease. For high quality of cell-cell similarity, this decreasing trend is relatively weak, but obvious for low quality. We use the area under the noise proportion – quality curve (AUC) to characterize this trend, which quantized the quality of cell-cell similarity under noise. Under the randomly set 0 noise, AUC is significantly improved after integrating pathway, the AUC average improvement rates of *de novo* pathway, Wikipathways, KEGG and Reactome are 63.4% (P=5e-5), 40.8% (P=3e-3), 35.6% (P=3e-3) and 23.5% (P=4e-2), respectively (Figure 4B, left side). And the same improvement phenomena are under the Gaussian noise, the AUC average improvement rates of *de novo* pathway, Wikipathways, KEGG and Reactome are 61.7% (P=2e-4), 41.9% (P=2e-3), 34.5% (P=5e-3) and 25.5% (P=2e-2), respectively (Figure 4B, right side). CHscore and DBscore also show the similar improvement and significance (Supplement Figure 20,21). These results indicate that integrating pathway improve the quality of cell-cell similarity under noise.

In summary, our analysis indicated that quality of cell-cell similarity metrics is improved by integrating pathway. Using the improved cell-cell similarity metrics as the input may improve performance of clustering methods. It would be the question to be answered in the next section, that is, whether the integrating pathway can improve the accuracy and robustness of clustering methods.

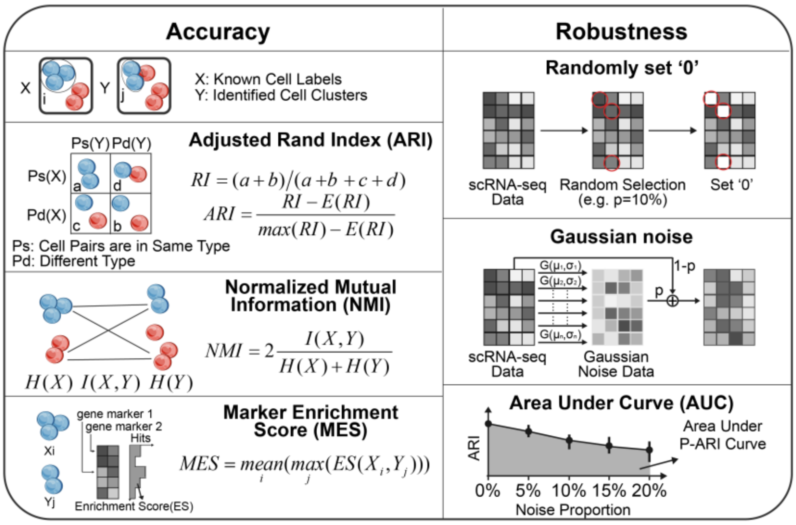

Figure 3. Evaluating accuracy and robustness of clustering method. **Accuracy** (left side), including Adjusted rand index (ARI), normalized mutual information (NMI) and marker enrichment score (MES). **Robustness** (right side), including the noise generation ('randomly set 0' noise and 'Guassian' noise) and robustness indicator (the area under curve, AUC).

Improvement in Accuracy and Robustness of Single-Cell Clustering Method

Improvement in clustering accuracy

Accuracy is the most basic and important performance criterion for a single-cell clustering method and is evaluated by the quality of the clustering results. In our study, we quantify the quality of the clustering results at both computational and biological levels. At the computational level, our accuracy indictors are based on the agreement between the cell grouping obtained by a clustering method and the true type labels of the cells. At the biological level, we used the gene marker indicator to define a marker enrichment score (MES) to measure the quality of a clustering result (Supplement Text 4).

For the clustering accuracy at the computational level, we used two performance indicators, the adjusted rand index (ARI) and the normalized mutual information (NMI). In our analysis, we found that, after integrating pathway, ARI indicator has been significantly improved in the overall clustering methods level (Figure 1A, left side) and the specific clustering method level (9/10 methods, Supplement Figure 2, top side). On the overall level, the ARI average improvement rates on *de novo* pathway, Wikipathways, KEGG and Reactome are 24.6% (P=2e-3, Wilcoxon signed-rank test, one-sided), 25.7% (P=2e-3), 25.0% (P=2e-3) and 17.0% (P=6e-3), respectively. We also observed that the ARI indicator of SC3, the highest accuracy in our evaluation framework, and Seurat, the most commonly used single-cell clustering method, has improved in most of the scRNA-seq data (17/26 for SC3 and 20/26 for Seurat; Figure 1A, right side). And the other methods have the same phenomenon (Supplement Figure 2, bottom side). The results on NMI indicator are consistent with ARI indicator, the average NMI improvement rates of *de novo* pathway, Wikipathways, KEGG and Reactome are 18.7% (P=2e-3), 19.4% (P=2e-3), 18.5% (P=2e-3) and 12.1% (P=6e-3), respectively (Supplement Figure 3A). And 9/10 clustering methods have been significantly improved, among them, the NMI of SC3 improved 17/26 of the scRNA-seq data, and Seurat is 21/26 (Supplement Figure 3B, 5).

To quantify the clustering accuracy at the biological level, we use cell type-specific gene markers to define a marker enrichment score (MES). A higher MES indicates that the gene marker is highly expressed in the corresponding cell cluster obtained by a clustering method, representing a potential cell type. At the overall clustering methods level and the specific clustering method level, we observed MES indicator is significantly improved, the ARI average improvement rates of de novo pathway, Wikipathways, KEGG and Reactome are 6.8% (P=2e-3), 6.0% (P=3e-3), 5.9% (P=1e-2) and 5.3% (P=8e-3), respectively. (Figure 1B, Supplement Figure 5,6). In addition, we noted that a parameter in Seurat R package, the number of the marker genes (set to 50 in our experiments). In order to eliminate parameter sensitivity, we repeated the above process with different parameter values (the number of marker genes = 25 and 100) and observed similar improvement rates (Supplement Figure 7-10).

In summary, our analysis indicated that integrating pathway significantly improves accuracy of single-cell clustering methods as measured by computational and biological indicator.

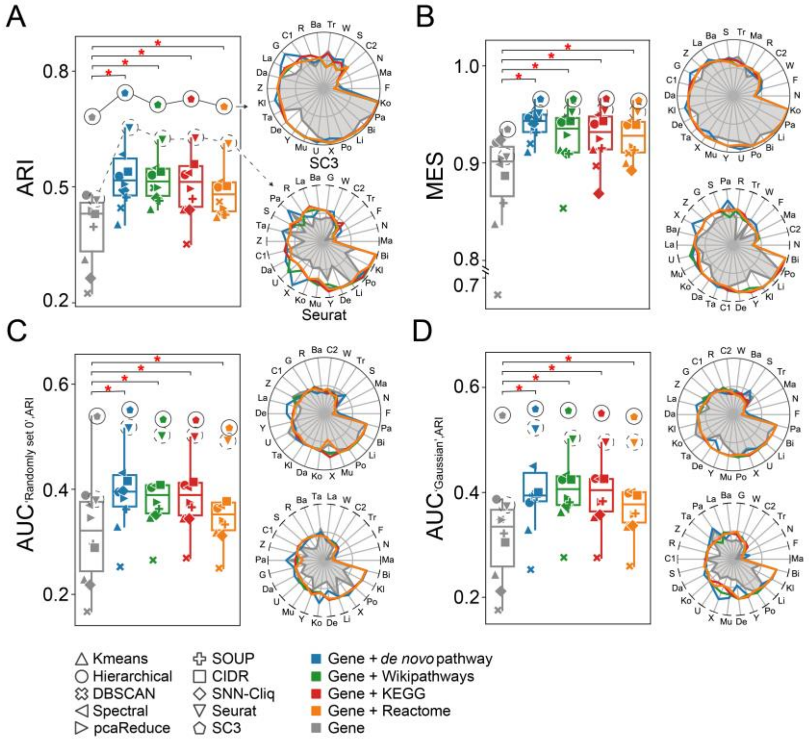

Figure 4. Improvement of accuracy (A,B) and robustness (C,D) of single-cell clustering. **(A)** Adjusted rand index (ARI), the performance of SC3 and Seurat on each scRNA-seq data (different angles in the radar chart) is shown on the right side. **(B)** Marker enrichment score (MES). **(C)** The area under 'randomly set 0' noise proportion – ARI curve (AUC). **(D)** The area under 'Gaussian' noise proportion – ARI curve (AUC). The red star indicates significant improvement (P<0.05, Wilcoxon signed-rank test, one-sided). The y coordinate of each point represents the average performance of corresponding method in all scRNA-seq.

Improvement in clustering robustness

Robustness characterizes the performance of method under noisy data. In our evaluation framework, we generated two types of noise, randomly set 0 and Gaussian noise. These noises are added to the data in a specific proportion (5%, 10%, 15% and 20%) and to obtain the noisy scRNA-seq data respectively. As the proportion of noise increases, the clustering accuracy will decrease. For robust methods, this decreasing trend is relatively weak, but obvious for methods with poor robustness. We use the area under the noise proportion – accuracy curve (AUC) to characterize this trend, which quantized the robustness of single-cell clustering method. Similar to the accuracy indicators, we also compared the robustness of clustering methods before and after integrating pathway under the noise from overall clustering methods level and the specific clustering method level (Supplement Text 4).

The scRNA-seq data with 'randomly set 0' noise is generated by randomly set the expression value to zero with a specific proportion (Figure 2, 'randomly set 0'). Through the different proportion of noise (5%, 10%, 15% and 20%) and the accuracy of the clustering method on the noisy scRNA-seq data, we draw the noise proportion - accuracy curve and calculate the area under the curve (Figure 2, AUC). In our framework, we calculated AUC of each clustering method on each noisy scRNA-seq data. Comparing the AUC indicator before and after integrating pathway, we found that it has been significantly improved in the overall clustering methods level (Figure 1C, left side) and the specific clustering method level (7/10 methods, Supplement Figure 12, top side). On the overall level, the AUC average improvement rates on *de novo* pathway, Wikipathways, KEGG and Reactome are 24.8% (P=2e-3), 22.3% (P=3e-3), 22.4% (P=3e-3) and 13.6% (P=3e-2), respectively. We also observed that the AUC indicator of Seurat has improved in 21/26 of the scRNA-seq data (Figure 1C, right side). And the other methods have the same phenomenon (Supplement Figure 12, bottom side). In addition, we replace the accuracy indicator ARI with NMI to recalculate the AUC, and obtain similar results, the AUC average improvement rates on *de novo* pathway, Wikipathways, KEGG and Reactome are 21.9% (P=2e-3), 18.9% (P=2e-3), 18.3% (P=2e-3) and 11.8% (P=3e-2), respectively (Supplement Figure 13). And Seurat has improved in 21/26 of the scRNA-seq data (Supplement Figure 13,14).

The scRNA-seq data with Gaussian noise is generated by following three steps: First, fit a Gaussian distribution for each gene, subject to the mean and variance of its expression in all cells; Second, randomly generate Gaussian noise which satisfies these Gaussian distributions; Third, combine expression and noise with specific proportion to obtain noisy scRNA-seq data (Figure 2, ‘Gaussian noise’). These noisy data are also used to calculate the AUC indicator of each clustering methods. We observed the AUC at the overall clustering methods level and the specific clustering method level, found that AUC is significantly improved after integrating pathway, the AUC average improvement rates of *de novo* pathway, Wikipathways, KEGG and Reactome are 24.5% (P=3e-3), 26.0% (P=2e-3), 24.5% (P=2e-3) and 18.5% (P=3e-3), respectively (Figure 1D, Supplement Figure 15,16). We also observed that the AUC indicator of Seurat has improved in 20/26 of the scRNA-seq data (Figure 1D, right side). And the details of other methods or the results of AUC based on ARI have the same phenomenon (Supplement Figure 17,18).

In summary, these results indicate that integrating pathway could significantly improve robustness of single-cell clustering methods under ‘randomly set 0’ noise and Gaussian noise.

Ranking and Comparing of Clustering Methods

To further analyze the improvement of clustering methods, we ranked all the single-cell clustering methods before and after integrating pathway based on ARI, MES, AUC$_{\text{'random set 0'}}$ and AUC$_{\text{'Gaussian'}}$. The accuracy ranking is the average of ARI ranking and MES ranking. The robustness ranking is the average of AUC$_{\text{'random set 0'}}$ ranking and AUC$_{\text{'Gaussian'}}$ ranking. The overall ranking is the average of accuracy and robustness ranking. We normalized these rankings, that is, distributed in the range from 0 to 1 (1 indicates the top; 0 indicates the bottom) (Figure 5, Details of each methods are in Supplement Figure 22).

Combing these rankings, we found that clustering methods by integrating pathway are ranked higher both on accuracy and robustness. Although the accuracy and robustness of few methods are not significantly improved, their ranking is still improved after integrating pathway. For example, hierarchical clustering is not significantly improved in both accuracy and robustness, but the ranking is still raised 11 seats on accuracy (hierarchical+*de novo* pathway, from 27th to 16th) and 9 seats on robustness (hierarchical+KEGG, from 22th to 13th). Another example is SC3, the robustness improvement is not significant, but SC3+*de novo* pathway (Top 1 on robustness) and SC3+Wikipathways (Top 2 on robustness) is still ranked ahead of SC3 (Top 3 on robustness), and the ranking of SC3+KEGG (Top 4 on robustness) and SC3+Reactome (Top 5 on robustness) is very close to SC3's ranking. We speculate that, although the accuracy of SC3 is improved significantly (SC3+KEGG, P=0.041) and its ranking is raised 13 seats (from 14th to top 1), the robustness may have reached the ceiling, resulting in weak improvement.

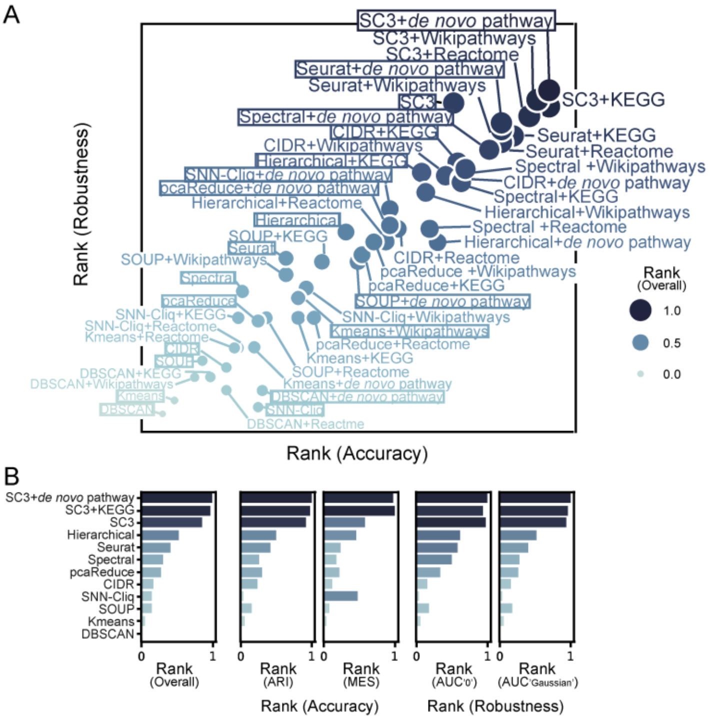

Figure 5. Ranking and comparing. **(A)** Rankings of clustering methods based on accuracy (x-axis) and robustness (y-axis). The size of dots and the depth of color both indicate the overall rank of the method. The box indicates original methods or the optimal improvement of methods by integrating pathway. **(B)** The ranking details of top 2 methods and other 10 single-cell clustering methods.

DISCUSSION

To analyze the improvement of single-cell clustering methods by integrating pathway, we designed a framework, including 10 state-of-art single-cell clustering methods, 26 scRNA-seq data, 4 pathway databases, 2 accuracy quantification indicators, 2 noise generation strategies and the corresponding robustness indicators. This framework can systematically quantified and compared the quality of cell-cell similarity metrics and performance of clustering methods before and after integrating pathway. Our analysis showed that integrating pathway can significantly improve the accuracy and robustness of most single-cell clustering methods. In addition, by ranking the methods under each indicator, we found that even if some methods are not significantly improved on accuracy or robustness, but their ranking are still raised by integrating pathway.

Our framework can potentially be used as a general tool for single cell clustering. By ranking single-cell clustering methods, an optimal method can be identified. For example, in our analysis among the four pathway databases, the *de novo* pathway database has the best effect on improving clustering method. We speculate this is because that its gene coverage is much higher than that of other manually-constructed pathway databases, as has been suggested in the literature[12,15,51,52]. In addition, new single-cell clustering methods can be evaluated using our framework and obtained some possible improvement strategies.

REFERENCE

1. Potter SS. Single-cell RNA sequencing for the study of development, physiology and disease. Nat. Rev. Nephrol. 2018; 14:479–492

2. Lähnemann D, Köster J, Szczurek E, et al. Eleven grand challenges in single-cell data science. Genome Biol. 2020; 21:31

3. Keller L, Pantel K. Unravelling tumour heterogeneity by single-cell profiling of circulating tumour cells. Nat. Rev. Cancer 2019;

4. Stuart T, Butler A, Hoffman P, et al. Comprehensive Integration of Single-Cell Data. Cell 2019; 1–15

5. Kiselev VY, Kirschner K, Schaub MT, et al. SC3: Consensus clustering of single-cell RNA-seq data. Nat. Methods 2017; 14:483–486

6. Xu C, Su Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. Bioinformatics 2015; 31:1974–1980

7. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat. Rev. Genet. 2018 2019; 1

8. Kanehisa M, Furumichi M, Tanabe M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017; 45:D353–D361

9. Fan JBJ, Salathia N, Liu R, et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nat. Methods 2016; 13:241–244

10. Sanchez-Vega F, Mina M, Armenia J, et al. Oncogenic Signaling Pathways in The Cancer Genome Atlas. Cell 2018; 173:321-337.e10

11. Dai H, Li L, Zeng T, et al. Cell-specific network constructed by single-cell RNA sequencing data. Nucleic Acids Res. 2019; 1–14

12. Wegmann R, Neri M, Schuierer S, et al. CellSIUS provides sensitive and specific detection of rare cell populations from complex single-cell RNA-seq data. Genome Biol. 2019; 20:142

13. Wang H, Sham P, Tong T, et al. Pathway-based Single-Cell RNA-Seq Classification, Clustering, and Construction of Gene-Gene Interactions Networks Using Random Forests. IEEE J. Biomed. Heal. Informatics 2019; PP:1–1

14. Ji Z, Ji H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. Nucleic Acids Res. 2016; 44:e117

15. Aibar S, González-Blas CB, Moerman T, et al. SCENIC: Single-cell regulatory network inference and clustering. Nat. Methods 2017; 14:1083–1086

16. Fabregat A, Jupe S, Matthews L, et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2018; 46:D649–D655

17. Slenter DN, Kutmon M, Hanspers K, et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res. 2018; 46:D661–D667

18. Lloyd S. Least squares quantization in PCM. IEEE Trans. Inf. Theory 1982; 28:129–137

19. Zhu L, Devlin B, Lei J, et al. Semisoft clustering of single-cell data. Proc. Natl. Acad. Sci. 2018; 116:466–471

20. Ward JH. Hierarchical Grouping to Optimize an Objective Function. J. Am. Stat. Assoc. 1963; 58:236–244

21. Lin P, Troup M, Ho JWK. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. Genome Biol. 2017; 18:1–11

22. Jianbo Shi, Malik J. Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 2000; 22:888–905

23. Žurauskien J. pcaReduce : hierarchical clustering of single cell transcriptional profiles. BMC Bioinformatics 2016; 1–11

24. Daszykowski M, Walczak B. Density-Based Clustering Methods. Compr. Chemom. 2009; 2:635–654

25. Baron M, Veres A, Wolock SL, et al. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. Cell Syst. 2016; 3:346-360.e4

26. Muraro MJ, Dharmadhikari G, Grün D, et al. A Single-Cell Transcriptome Atlas of the Human Pancreas. Cell Syst. 2016; 3:385-394.e3

27. Biase FH, Cao X, Zhong S. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. Genome Res. 2014; 24:1787–1796

28. Nestorowa S, Hamey FK, Pijuan Sala B, et al. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. Blood 2016; 128:e20–e31

29. Camp JG, Sekine K, Gerber T, et al. Multilineage communication regulates human liver bud development from pluripotency. Nature 2017; 546:533–538

30. Patel AP, Tirosh I, Trombetta JJ, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science (80-.). 2014; 344:1–9

31. Camp JG, Badsha F, Florio M, et al. Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. Proc. Natl. Acad. Sci. 2015; 112:15672–15677

32. Pollen AA, Nowakowski TJ, Shuga J, et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat. Biotechnol. 2014; 32:1053–1058

33. Darmanis S, Sloan SA, Zhang Y, et al. A survey of human brain transcriptome diversity at the single cell level. Proc. Natl. Acad. Sci. 2015; 112:7285–7290

34. Romanov RA, Zeisel A, Bakker J, et al. Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. Nat. Neurosci. 2017; 20:176–188

35. Deng Q, Ramskold D, Reinius B, et al. Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells. Science (80-.). 2014; 343:193–196

36. Segerstolpe Å, Palasantza A, Eliasson P, et al. Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. Cell Metab. 2016; 24:593–607

37. Fan X, Zhang X, Wu X, et al. Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. Genome Biol. 2015; 16:148

38. Tasic B, Menon V, Nguyen TN, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat. Neurosci. 2016; 19:335–346

39. Goolam M, Scialdone A, Graham SJL, et al. Heterogeneity in Oct4 and Sox2 Targets Biases Cell Fate in 4-Cell Mouse Embryos. Cell 2016; 165:61–74

40. Treutlein B, Brownfield DG, Wu AR, et al. Reconstructing lineage hierarchies of the distal lung

epithelium using single-cell RNA-seq. Nature 2014; 509:371–375

41. Klein AM, Mazutis L, Akartuna I, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 2015; 161:1187–1201

42. Usoskin D, Furlan A, Islam S, et al. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. Nat. Neurosci. 2015; 18:145–153

43. Kolodziejczyk AA, Kim JK, Tsang JCH, et al. Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. Cell Stem Cell 2015; 17:471–485

44. Wang YJ, Schug J, Won K-J, et al. Single-Cell Transcriptomics of the Human Endocrine Pancreas. Diabetes 2016; 65:3028–3038

45. Lake BB, Ai R, Kaeser GE, et al. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Science (80-.). 2016; 352:1586–1590

46. Xin Y, Kim J, Okamoto H, et al. RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes. Cell Metab. 2016; 24:608–615

47. Li H, Courtois ET, Sengupta D, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat. Genet. 2017; 49:708–718

48. Yan L, Yang M, Guo H, et al. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat. Struct. Mol. Biol. 2013; 20:1131–1139

49. La Manno G, Gyllborg D, Codeluppi S, et al. Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. Cell 2016; 167:566-580.e19

50. Zeisel A, Munoz-Manchado AB, Codeluppi S, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science (80-.). 2015; 347:1138–1142

51. Kamburov A, Stelzl U, Lehrach H, et al. The ConsensusPathDB interaction database: 2013 update. Nucleic Acids Res. 2013; 41:D793–D800

52. Rodchenkov I, Babur O, Luna A, et al. Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. Nucleic Acids Res. 2019; 48:D489–D497