Briefings in Bioinformatics, 22(6), 2021, 1–12

https://doi.org/10.1093/bib/bbab147
Problem Solving Protocol

Improving Single-Cell RNA-seq Clustering by Integrating Pathways

Chenxing Zhang, Lin Gao, Bingbo Wang and Yong Gao

Corresponding authors: Lin Gao, School of Computer Science and Technology, Xidian University, Xi'an 710071, China. Tel.: +86-29-88202354; E-mail: 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 RNA-sequencing data. However, the accuracy and robustness of existing methods are disturbed by noise. One promising approach for addressing this challenge is integrating pathway information, which can alleviate noise and improve performance. In this work, we studied the impact on accuracy and robustness of existing single-cell clustering methods by integrating pathways. We collected 10 state-of-the-art single-cell clustering methods, 26 scRNA-seq datasets and four pathway databases, combined the AUCell method and the similarity network fusion to integrate pathway data and scRNA-seq data, and introduced three accuracy indicators, three noise generation strategies and robustness indicators. Experiments on this framework showed that integrating pathways 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, which combines multiple k-means clustering results into a hierarchical clustering [5]. Xu and Sun proposed SNN-Cliq, which clusters cells by clique-based methods on a shared nearest neighbor (SNN) 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 with 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]. For example, there is a large number of drop-out events where a gene 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.

Chenxing Zhang is a PhD candidate of computer science and technology at Xidian University. His research interests include the development and improvement of computational methods on single-cell data analysis.
Lin Gao is a Professor of the School of Computer Science and Technology at Xidian University. She focuses on computational models and algorithms in analysis of single-cell data and omics-data.
Bingbo Wang is an Associate Professor of the School of Computer Science and Technology, Xidian University. His research interests are in bioinformatics and network medicine.
Yong Gao is a Professor of computer science at the University of British Columbia Okanagan (UBC Okanagan), Canada. His research interests include artificial intelligence, discrete algorithms, graph theory and their applications in network science and computational biology.
Submitted: 30 November 2020; **Received (in revised form):** 21 March 2021

A pathway is a collection of relationship between genes which regulate the 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]. Therefore, serval methods for evaluating pathway activation score in single-cell data have been proposed. For example, Zhang *et al*. presented a web-based platform scTPA to estimate the pathway activation score of each cell in scRNA-seq data and to identify and analyze cell-type-specific activation pathways [11]. Zhang *et al*. systematically evaluated the accuracy, stability and scalability of 11 widely used pathway activity transformation algorithms in the analysis of scRNA-seq data [12]. Frost developed a variance-adjusted mahalanobis (VAM) method to compute cell-specific pathway scores, which is to reduce the impact of high levels of technical noise and inflated zero count in scRNA-seq data [13]. Meanwhile, many pathway-based methods and tools have appeared to analyze scRNA-seq data and identify cell-type-specific pathways or functional modules. Ma *et al*. introduced a method, iDEA, to integrate differential expression analysis and pathway enrichment analysis in scRNA-seq data and found that its performance of differential expression analysis is nearly 64% better than other DE method [14]. Klimm *et al*. presented scPPIN method to detecting cell transcription-specific functional modules by integrating scRNA-seq data with a protein–protein interaction network [15]. DePasquale *et al*. created a cellHarmony to analysis, classify and compare cell types from different scRNA-seq data and to identify impacted pathways related to transcriptional difference [16]. DeTomaso *et al*. presented a tool, Vision, identify and analyze the differential gene signatures in cell similarity graph [17]. Dai *et al*. presented a method to construct a cell-species network on scRNA-seq data and apply to clustering and pseudo-trajectory analysis [18].

In addition, several pathway-based clustering methods have been presented, suggesting that the pathway may be beneficial to cell-type identification. Fan *et al*. developed pathway and gene set overdispersion analysis method, PAGODA, to cluster and identify cell-type-specific pathways [9]. Aibar *et al*. presented a computational method, SCENIC, for constructing gene regulatory network, identifying functional module and applying to cell-state identifications [19]. 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 [20]. Wegmann *et al*. developed CellSIUS to identify rare cell-type based on gene module from scRNA-seq data [21]. Since the addition of pathway information reduces the impact of noise, these methods have good performance in single-cell clustering. These motivated us to consider the possibility of using pathway-level features to improve the existing single-cell clustering method.

In this work, we studied the effectiveness of integrating pathways 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 performance evaluation framework, named sciPath (Single-cell Clustering by integrating Pathways), consisting of 10 state-of-the-art single-cell clustering methods, 26 scRNA-seq data, four pathway databases, two accuracy quantification indicators, three noise generation strategies and the corresponding robustness indicators. Using the framework, we studied the impact of integrating pathways 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 pathways could significantly improve the 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 pathways can potentially help provide more effective and stable cell-type signals [9, 19, 22].

Result

To evaluate the accuracy and robustness of single-cell clustering method by integrating pathways, we designed a framework, sciPath, consisting of the following three parts.

(i) Sufficient materials, including 26 scRNA-seq data, four pathway databases and 10 state-of-the-art single-cell clustering methods. (Details are in Tables 1–3 and Supplementary Text 1, 2 available online at https://academic.oup.com/bib)
(ii) An integration strategy, integrating pathways 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 Materials and Methods)
(iii) A series of evaluation indicators, quantifying the performance of clustering. It contains three accuracy quantification indicators, three noise simulated strategies and the corresponding robustness indicators.

By comparing these indicators before and after integrating pathways, we observed a significant improvement of accuracy and robustness of clustering method. (Step3 in Figure 1; details are in Figure 3 and Supplementary Text 3 available online at https://academic.oup.com/bib).

Collection of sufficient materials

In our framework, we collected four pathway databases (Table 1; Supplementary Text 2 available online at https://academic.oup.com/bib), 10 state-of-the-art single-cell clustering method (Table 2; Materials and Methods) 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 Table 1; Supplementary Text 2 and Supplementary Figure 1 available online at https://academic.oup.com/bib.

Single-cell clustering methods

Among the 10 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 Table 2 and Materials and Methods.

Single-cell RNA-seq 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 3.

Table 1. The detail of pathway datasets

Pathway database	Ref.	# Items	
		Mouse	Human
KEGG [8]	Kanehisa *et al.* (2017)	394	396
Reactome [23]	Fabregat *et al.* (2018)	1623	2213
Wikipathways [24]	Slenter *et al.* (2018)	220	601
de novo pathway [22]	Ji and Ji (2016)	150	

Table 2. The detail of single-cell clustering methods

Methods	Ref.	Methods	Ref.
Kmeans [25]	Lloyd (1982)	SOUP [26]	Zhu *et al.* (2018)
Hierarchical [27]	Ward (1963)	CIDR [28]	Lin *et al.* (2017)
Spectral [29]	Shi *et al.* (2000)	pcaReduce [30]	Žurauskien (2016)
DBSCAN [31]	Daszykowski *et al.* (2009)	SNN-Cliq [6]	Xu and Su (2015)
Seurat [4]	Stuart *et al.* (2019)	SC3 [5]	Kiselev *et al.* (2017)

Table 3. The detail of Single-cell RNA-seq datasets

Single-cell RNA-seq data	#Type(#Cell)	Single-cell data	#Type(#Cell)
Baron [32]	13(1886)	Muraro [33]	10(2126)
Biase [34]	4(56)	Nestorowa [35]	9(1656)
Camp1 [36]	7(777)	Patel [37]	5(430)
Camp2 [38]	6(734)	Pollen [39]	11(301)
Darmanis [40]	9(466)	Romanov [41]	7(2881)
Deng [42]	6(268)	Segerstolpe [43]	15(3514)
Fan [44]	6(66)	Tasic [45]	18(1679)
Goolam [46]	5(124)	Treutlein [47]	5(80)
Klein [48]	4(2717)	Usoskin [49]	4(622)
Kolodziejczyk [50]	3(704)	Wang [51]	8(635)
Lake [52]	16(3042)	Xin [53]	8(1600)
Li [54]	9(561)	Yan [55]	6(90)
Manno [56]	32(2150)	Zeisel [57]	9(3005)

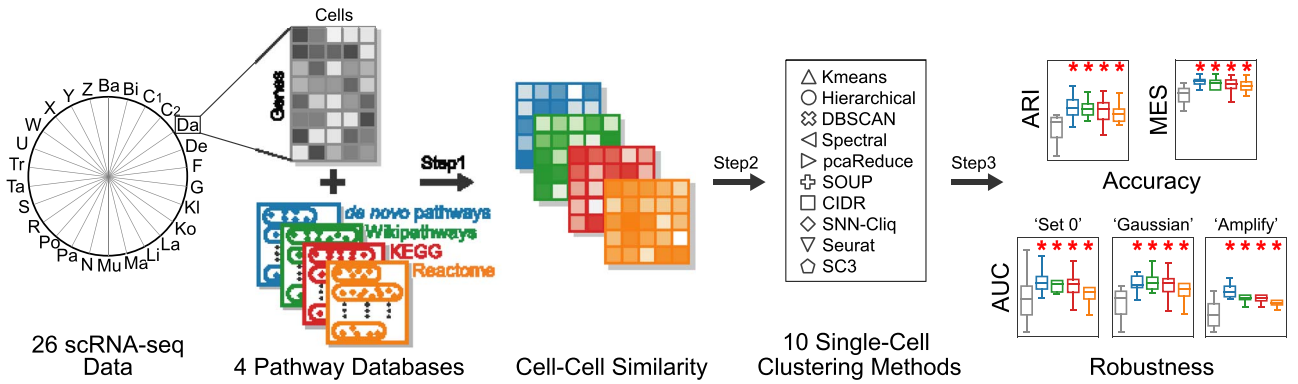

Figure 1. The framework of evaluating the single-cell clustering method by integrating pathways. Step1. Integrating scRNA-seq and pathway into cell–cell similarity metrics (details are in Supplementary Text 3 Materials and Methods available online at https://academic.oup.com/bib). Step2. Inputting the cell–cell similarity metrics into single-cell clustering methods and getting the clustering results. Step3. Evaluating accuracy and robustness of single-cell clustering methods by integrating pathways (details are in Figure 3).

Improvement of cell–cell similarity metrics by integrating pathways

In order to make the step of integrating pathways 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,

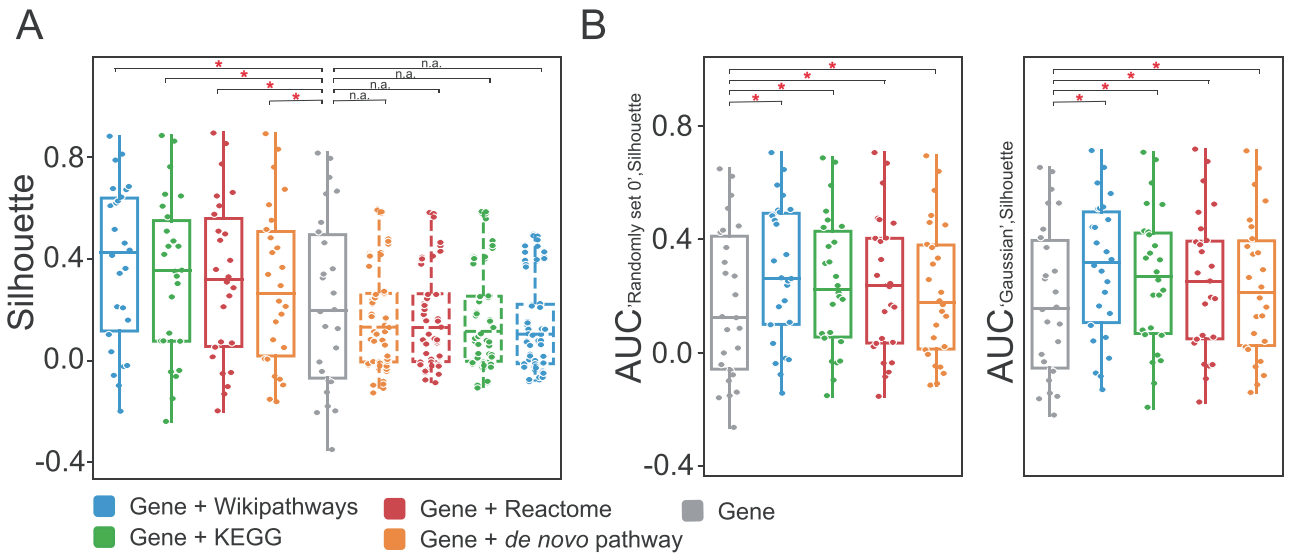

Figure 2. Improvement of cell–cell similarity metric for clustering. **(A)** Silhouette of cell–cell similarity metric (colored and solid box) and that of which integrate permutated pathway (colored and dashed box) on each scRNA-seq data. **(B)** The area under 'randomly set 0' (left side) and 'Gaussian' (right side) noise proportion—Silhouette curve on noisy cell–cell similarity metric which integrating pathways 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 ($P \geq 0.05$). Each dot indicates an scRNA-seq data.

Davies-Bouldin score and Calinski-Harabasz score (see Supplementary Text 3 available online at https://academic.oup.com/bib 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 represents 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.

We compared the Silhouette coefficients of cell–cell similarity before and after integrating pathways on each scRNA-seq data and found that Silhouette coefficient has been significantly improved after integrating pathways. 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% ($P = 1e{-}2$), respectively (Figure 2A, left side). In addition, we integrated 'random' pathway (permutating gene signature in the pathway, details are in Supplementary Text 3 available online at https://academic.oup.com/bib) and each scRNA-seq data into cell–cell similarity, Silhouette coefficient is not significantly improved (Figure 2A, right side). Calinski Harabasz score (CH score) and Davies Bouldin score (DB score) also show the similar improvement and significance (Supplementary Figure 3 available online at https://academic.oup.com/bib). These results indicate that integrating pathways improve the quality of cell–cell similarity metrics.

We also compared the quality of cell–cell similarity metrics before and after integrating pathways on noisy scRNA-seq data. 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—Silhouette 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 pathways, 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 2B, 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 2B, right side). CHscore and DBscore also show the similar improvement and significance (Supplementary Figures 4 and 5 available online at https://academic.oup.com/bib). These results indicate that integrating pathways improve the quality of cell–cell similarity under noise.

In addition, we observed the variance of Silhouette scores. Further correlation analysis found that the Silhouette coefficients is negatively correlated with the number of cells (PCC = −0.58, *P*-value = 0.001; Supplementary Figure 2A available online at https://academic.oup.com/bib). And there is no significant difference between human and mouse (*P*-value = 0.76, *t*-test; Supplementary Figure 2A available online at https://academic.oup.com/bib), means that the data quality of human and mouse is generally similar. For the area under noise proportion–silhouette curve (AUC), we found that the AUC average improvement rates fluctuates with integrating different pathway. Further correlation analysis found that the AUC average improvement rate is negatively correlated with the number of pathway items (PCC = −0.76 under randomly set 0 noise and PCC = −0.69 under Gaussian noise; '#Items' in Supplementary Figure 2B available online at https://academic.oup.com/bib) and its redundancy [PCC = −0.84 under randomly set 0 noise and PCC = −0.73 under Gaussian noise; 'avg(#items/gene)']. In addition, we can combine the pathway information in Supplementary Figure 1, available online at https://academic.oup.com/bib, and the results of Figure 2B to intuitively observe the same phenomenon. That is, the Reactome pathway, with the lowest improvement effect, has the highest redundancy and number of items. On the contrary, *de novo*

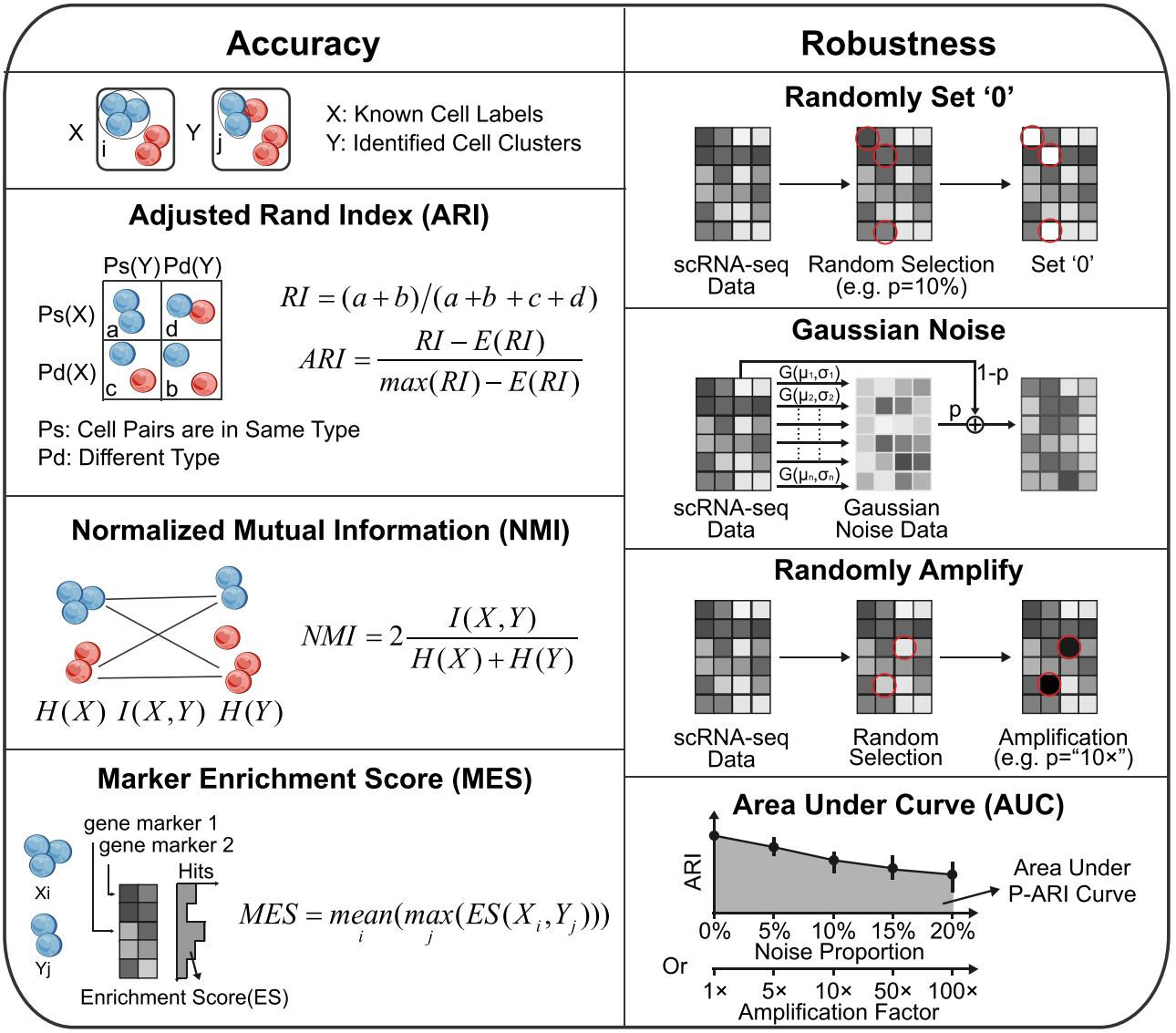

Figure 3. Evaluating accuracy and robustness of clustering methods. Accuracy (left side), including ARI, NMI and MES. Robustness (right side), including the noise generation ('randomly set 0' noise, 'Guassian' noise and 'amplify' noise) and robustness indicator (the area under curve, AUC).

pathway has the best improvement effect, with no redundancy and fewer items.

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

Improvement in accuracy and robustness of single-cell clustering methods

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 (Supplementary Text 3 available online at https://academic.oup.com/bib).

For the clustering accuracy at the computational level, we used two performance indicators, adjusted rand index (ARI) and normalized mutual information (NMI). In our analysis, we found that, after integrating pathways, ARI indicator has been significantly improved in the overall clustering methods level (Figure 4A, left side) and the specific clustering method level (9/10 methods, Supplementary Figure 6 and 7, top side available online at https://academic.oup.com/bib). 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

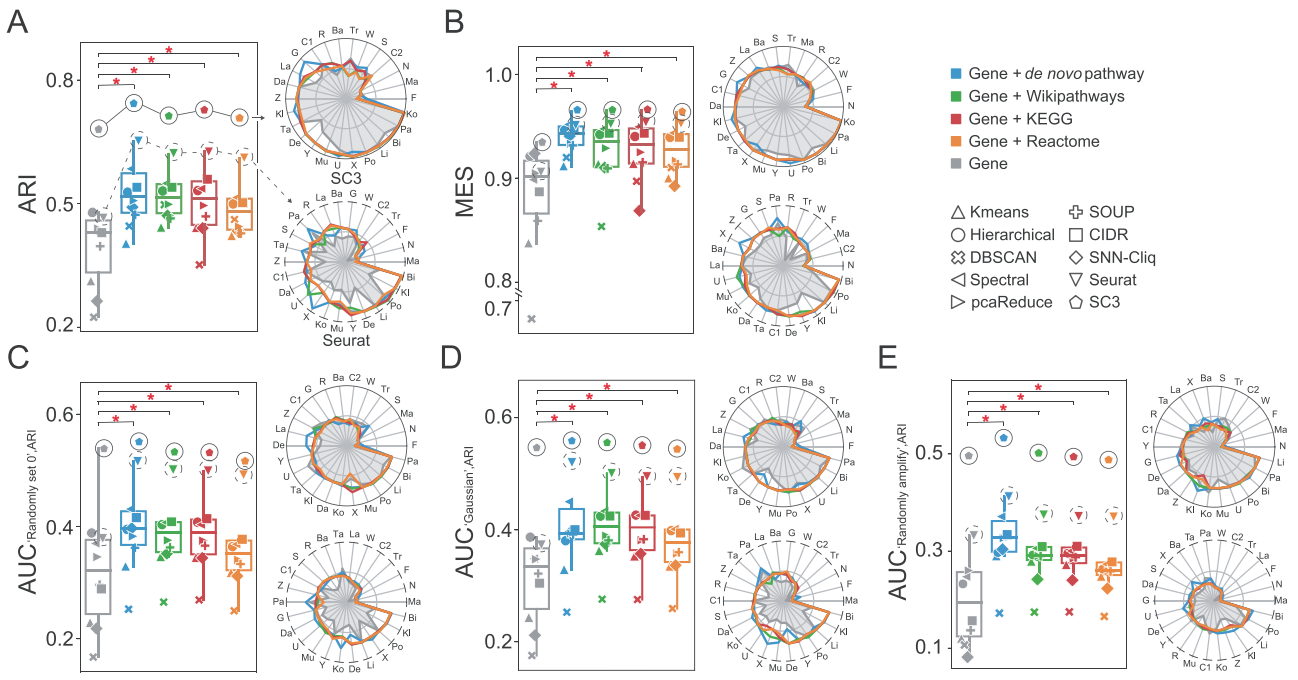

Figure 4. Improvement of accuracy (**A**, **B**) and robustness (**C**, **D**) of single-cell clustering. (**A**) ARI, the performance of SC3 and Seurat on each scRNA-seq data (different angles in the radar chart) is shown on the right side. (**B**) MES. (**C**) The area under 'randomly set 0' noise proportion—ARI curve (AUC). (**D**) The area under 'Gaussian' noise proportion—ARI curve (AUC). (**E**) The area under 'amplify' noise factor—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.

SC3, the highest accuracy in our evaluation framework, and Seurat, the most commonly used single-cell clustering method, have improved in most of the scRNA-seq data (17/26 for SC3 and 20/26 for Seurat; Figure 4A, right side). And the other methods have the same phenomenon (Supplementary Figure 7 available online at https://academic.oup.com/bib). 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 (Supplementary Figure 8A available online at https://academic.oup.com/bib). 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 (Supplementary Figures 8B and 9 available online at https://academic.oup.com/bib).

To quantify the clustering accuracy at the biological level, we used cell-type-specific gene markers to define an MES. A higher MES indicates that the gene markers are highly expressed in the corresponding cell cluster obtained by a clustering method, representing a potential cell type. At the overall and specific level of clustering methods, 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, Supplementary Figures 10 and 11 available online at https://academic.oup.com/bib). 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 (Supplementary Figures 12–15 available online at https://academic.oup.com/bib).

In addition, we used t-SNE to visualize the clustering results of each methods in each scRNA-seq data before and after integrating pathways. According to the visualization results, we observed that the cell-type boundary of results in most clustering methods is more obvious after integrating pathways. Such as the Seurat and SC3 methods in Klein, Baron, Kolodziejczyk and Usoskin scRNA-seq data, the boundary of the cell type is obscure before integrating pathways but is clear after integrating pathways (Supplementary Figures 16–20 available online at https://academic.oup.com/bib).

In order to avoid the known cell labels as the only accuracy evaluation materials, we adopt two evaluation strategies: (i) combine multiple clustering results as new cell labels and (ii) use evaluation indicators without known cell labels. For the first strategy, we applied the SAME method [58] to combine the known cell labels and the top five clustering results into a new cell labels and recalculated the accuracy indicators based on this cell labels. For the second strategy, we used WB-ratio [59] to evaluate accuracy based on clustering result and cell–cell distance in t-SNE visualization which is shown in Supplementary Figures 16–20 available online at https://academic.oup.com/bib. From the results of these two evaluation strategies, we observed the ARI, NMI and WB-ratio is also improved after integrating pathways (Supplementary Figure 21 available online at https://academic.oup.com/bib).

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

Improvement in clustering robustness

Robustness characterizes the performance of method under noisy data. In our evaluation framework, we generated three

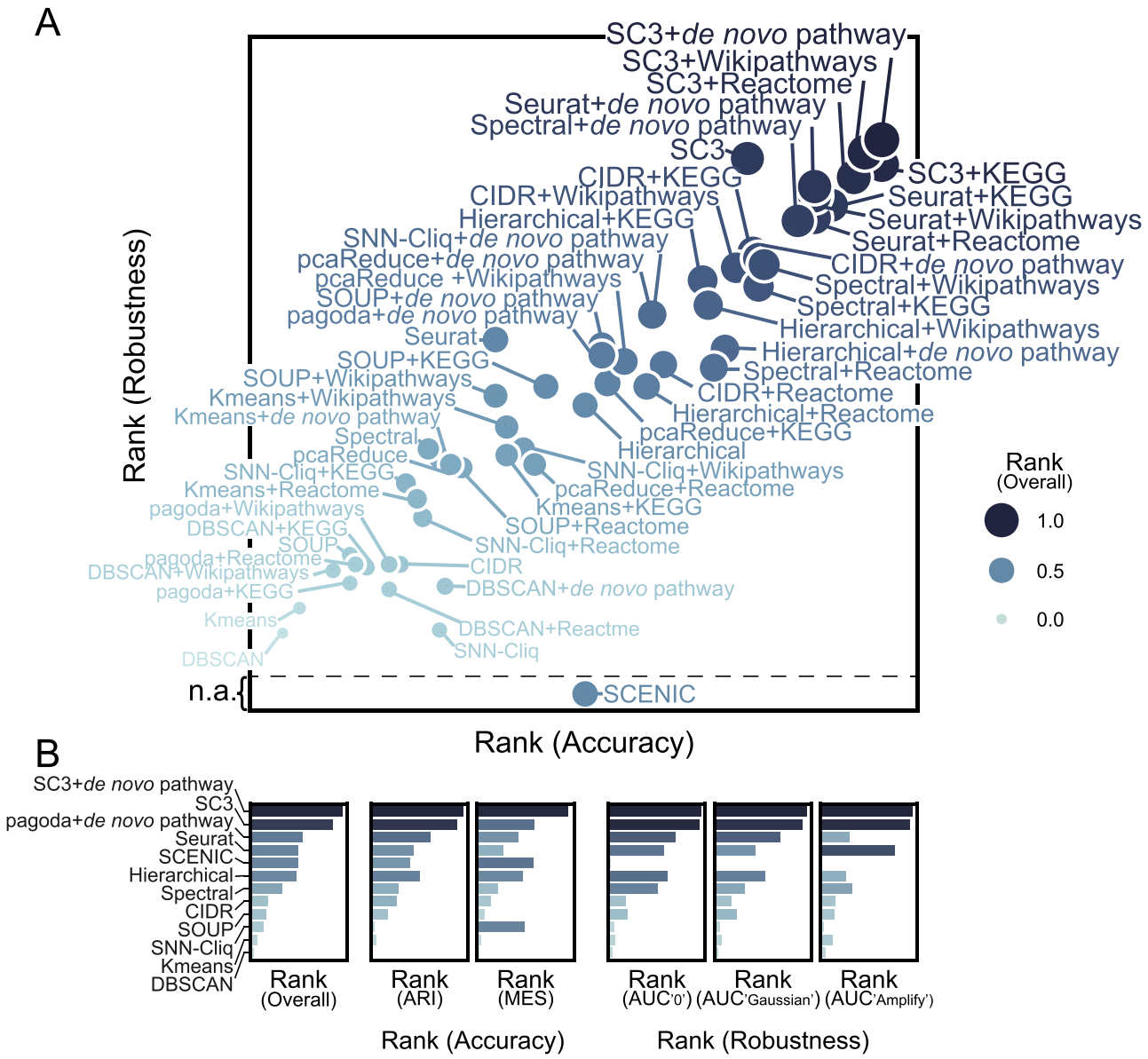

Figure 5. Ranking and comparing. (**A**) Rankings of clustering methods based on accuracy (*x*-axis) and robustness (*y*-axis). The size of dots and the depth of color both indicate the overall rank of the method. The 'n.a.' indicates that not available for evaluating robustness of SCENIC because of expensive time consumption. (**B**) The ranking details of top 1 method, the other 10 single-cell clustering methods and the two clustering methods that integrate pathway information.

types of noise, randomly set 0, Gaussian noise and amplify noise. These noises are added to the data in a specific proportion (5, 10, 15 and 20%) or amplify factor (5×, 10×, 50× and 100×) 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 used 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 pathways under the noise from the overall and specific level of clustering method (Supplementary Text 3 available online at https://academic.oup.com/bib).

The scRNA-seq data with 'randomly set 0' noise is generated by randomly set the expression value to zero with a specific proportion (Figure 4, 'randomly set 0'). Through the different proportion of noise (5, 10, 15 and 20%) and the accuracy of the clustering methods on the noisy scRNA-seq data, we draw the noise proportion—accuracy curve and calculate the area under the curve (AUC, Figure 4C). In our framework, we calculated AUC of each clustering method on each noisy scRNA-seq data. Comparing the AUC indicators before and after integrating pathways, we found that it has been significantly improved in the overall clustering methods level (Figure 4C) and the specific clustering method level (7/10 methods, Supplementary Figure 23 available online at https://academic.oup.com/bib). 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 (Supplementary Figure 22 and 23

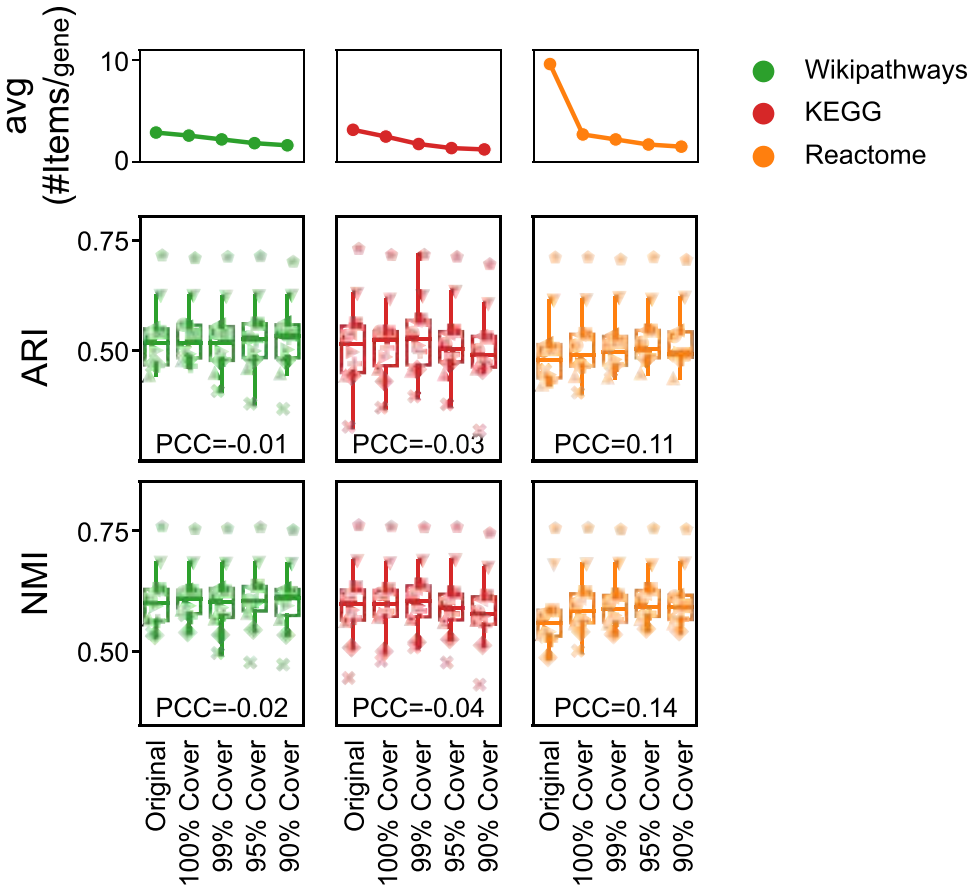

Figure 6. The improvement performance with reducing pathway redundancy. The 'avg(#items/gene)' on the top figure is a quantitative indicator of pathway redundancy, indicating the average number of pathway items that each gene participates in. The 'Original' indicates the original pathway without redundancy reduction processing; the '100% Cover', '99% Cover', '95% Cover' and '90% Cover' are the threshold for redundancy reduction method, indicating the proportion of genes covered by the ultimately retained pathway items to the original pathway. The smaller the proportion, the less redundancy of the pathway items retained.

available online at https://academic.oup.com/bib). And the other methods have the same phenomenon (Supplementary Figure 22 and 23 available online at https://academic.oup.com/bib). In addition, we replaced the accuracy indicator ARI with NMI to recalculate the AUC, and obtained 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 (Supplementary Figure 24 available online at https://academic.oup.com/bib). And Seurat has improved in 21/26 of the scRNA-seq data (Supplementary Figures 24 and 25 available online at https://academic.oup.com/bib).

The scRNA-seq data with Gaussian noise are 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; and third, combine expression and noise with specific proportion to obtain noisy scRNA-seq data (Figure 4D). These noisy data are also used to calculate the AUC indicator of each clustering methods. We observed the AUC at the overall and specific level of clustering methods and found that AUC is significantly improved after integrating pathways; 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 4D, Supplementary Figures 26 and 27 available online at https://academic.oup.com/bib). We also observed that the AUC indicator of Seurat has improved in 20/26 of the scRNA-seq data (Supplementary Figure 27 available online at https://academic.oup.com/bib). And the details of other methods or the results of AUC based on NMI have the same phenomenon (Supplementary Figures 28 and 29 available online at https://academic.oup.com/bib).

The scRNA-seq data with 'amplify' noise is generated by randomly selected 5% gene expressions and amplify (i.e. multiply) these expressions by 5, 10, 50 and 100 (i.e. amplify factor). Through the different factors (5×, 10×, 50× and 100×) and the accuracy of the clustering method on the noisy scRNA-seq data, we drew the amplify factor-accuracy curve and calculated the AUC (Figure 2, AUC 'amplify'). At the result, the AUC average improvement rates on *de novo* pathway, Wikipathways, KEGG and Reactome are 55.3% ($P = 9e{-}4$), 40.9% ($P = 9e{-}4$), 39.6% ($P = 1e{-}3$) and 30.2% ($P = 4e{-}3$), respectively (Figure 4E, Supplementary Figures 30 and 31 available online at https://academic.oup.com/bib). In addition, we also observed the improvement under the situation when the 'amplify' noise is generated by randomly selected 1% gene expressions; the AUC has also improved significantly (Supplementary Figures 32 and 33 available online at https://academic.oup.com/bib).

In summary, these results indicate that integrating pathways could significantly improve the robustness of single-cell clustering methods under 'randomly set 0' noise, Gaussian noise and 'amplify' 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 pathways (also includes two pathway-based clustering method, PAGODA and SCENIC) based on ARI, MES, $AUC_{\text{'random set 0'}}$, $AUC_{\text{'Gaussian'}}$ and $AUC_{\text{'amplify'}}$. 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, $AUC_{\text{'Gaussian'}}$ ranking and $AUC_{\text{'amplify'}}$ 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 Supplementary Figure 34 available online at https://academic.oup.com/bib).

Combing these rankings, we found that clustering methods by integrating pathways are ranked higher on accuracy and robustness. Although the accuracy and robustness of few methods are not significantly improved, their rankings are still improved after integrating pathways. For example, hierarchical clustering is not significantly improved on accuracy and robustness, but the ranking is still raised 13 seats on accuracy (hierarchical+*de novo* pathway, from 29th to 16th) and 16 seats on robustness (hierarchical+KEGG, from 31th to 15th). 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) are 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) are very close to SC3's ranking. We speculated that, although the accuracy of SC3 is improved significantly (SC3 + KEGG, $P = 0.041$) and its ranking is raised 12 seats (from 14th to top 2), the robustness may have reached the ceiling, resulting in weak improvement. For the two pathway-based single-cell clustering methods, PAGODA and SCENIC (due to the slow running efficiency, we neglected to evaluate its robustness, i.e. robustness = n.a.), we found that the performance of PAGODA is better than most (9/10) clustering methods without integrated pathway information, and the accuracy of SCENIC is also outperform most (9/10) existing methods (Figure 5B).

The impact of pathway redundancy on improvement performance

Furthermore, in order to observe the influence of the redundancy of the pathway on the improvement, we designed the following analysis. We used a pathway redundancy reduction method (proposed by Stoney *et al.* [60]) to gradually reduce the redundancy of the pathway, which the degree of redundancy is controlled by the threshold of gene coverage proportion. The smaller the proportion, the less redundancy of the pathway items retained. We selected the default set of thresholds (gene coverage proportion = 100, 99, 95 and 90%) and obtained the pathway databases under these thresholds (the details are in the Supplementary Text 4 available online at https://academic.oup.com/bib). Then, we integrated gene expression with these pathways and input in 10 single-cell clustering method. We observed the performance on 26 scRNA-seq datasets. Meanwhile, we used the same indicator introduced by Stoney *et al.* [60] to quantify the redundancy of pathway, the average number of pathway items that each gene participates in (abbreviated as 'avg(#Items/gene)' and shown at the top of the Figure 6). We observed some differences when reducing pathway redundancy. For the Reactome pathway databases, the result has a slight increase with the reducing pathway redundancy. On the contrary, the result is not obvious on Wikipathways and KEGG pathway databases (Figure 6). We got that the high-redundancy pathway databases (e.g. Reactome) have a slight negative effect on improvement, while the low-redundancy pathway databases could not.

DISCUSSION

To analyze the improvement of single-cell clustering methods by integrating pathways, we designed a framework, including 10 state-of-the-art single-cell clustering methods, 26 scRNA-seq data, four pathway databases, two accuracy quantification indicators, three noise generation strategies and the corresponding robustness indicators. This framework can systematically quantified and compared the quality of cell–cell similarity metrics and the performance of clustering methods before and after integrating pathways. Our analysis showed that integrating pathways 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 pathways.

We found that the most clustering method is effective for identifying cell subpopulation. However, there is a lot of noise in single-cell data, including dropout events, over-expression and other technical noise, which makes clustering methods confuse the boundaries between cell types. Integrating pathways information could effectively alleviate this problem, making cells with the same state more similar, and the boundaries between cells with different states clearer (Figure 2). The several pathway-based single-cell clustering methods, such as PAGODA and SCENIC, are better than most clustering methods on accuracy and robustness (Figure 5), which is also imply a positive effect of pathways on single-cell clustering. In addition, considering the information of pathways or functional modules could also capture the signals of rare cell types [19], which implies that the pathway-level information may be able to find cell population that cannot be identified by gene-level information. Other works based on pathways and functional modules to analyze single-cell data could still find its advantage, such as differential expression analysis [14] and pseudo-time reconstruction in scRNA-seq [22].

We also found that there are variants in the performance and improvement of different clustering methods after integrating pathways. For those clustering methods developed for single-cell data, they used preprocessing and other processes to reduce the noise of scRNA-seq data. Such as SC3, it considers the multiple similarities of cells in multiple dimensions to improve the accuracy and robustness. Therefore, they perform better than others without noise reduction process. Since the integrated pathway information has the same effect on reducing noise, the space for further improvement is compressed. On the contrary, the traditional clustering methods, such as DBSCAN without any preprocessing operations for single-cell data, are affected by a large amount of noise. Because integrating pathways could reduce the impact of noise on clustering, these traditional clustering methods improved more performance.

In our framework, the step to integrate pathway information and gene information could be further improved. We used the

similarity network fusion (SNF) method for integration step and found that as the number of cells increases, the time consumption becomes huge (The time complexity is $O(n^3)$ and the actual running time is shown in Supplementary Figure 35 available online at https://academic.oup.com/bib), which is an important issue that limits our consideration of more single-cell datasets and pathway databases and our analysis of more types of noise that may be present in single-cell data. Therefore, an efficient and effective integration method is needed in our framework.

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 that 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 [19, 21, 61, 62]. In addition, new single-cell clustering methods can be evaluated using our framework and obtained some possible improvement strategies.

Materials and methods

Single-cell clustering method

Among the 10 state-of-the-art clustering methods are four traditional clustering methods and six single-cell clustering methods. These methods and the source of their implementation are described below.

K-means [25]. Given the number of clusters, updating the cluster center and the attribution of samples until both are stable. The 'kmeans' function in 'stats' (v3.5.2) R package was used.

Hierarchical clustering [27]. Starting from the situation where each sample is assigned to its own cluster, and iteratively joining the two closest samples until there is only one cluster (i.e. a tree). Given the number of clusters, the number of different clusters (i.e. subtree) corresponding to this is divided. The 'hclust' function with 'ward.D' agglomeration method in 'fastcluster' (v1.1.25) R package used in this paper is available at https://cran.r-project.org/web/packages/fastcluster/index.html.

Spectral clustering [29]. Clustering is performed using the eigenvectors of the Laplace matrix from the sample similarity matrix. The 'spectralClustering' function in 'SNFtool' (v2.3.0) R package was used at https://cran.r-project.org/web/packages/SNFtool/index.html.

DBSCAN [31]. Given the size of the epsilon neighborhood (eps) and number of minimum points in the eps region (minPts), the core points are those that satisfy the number of neighbors in the eps radius that exceed minpts. The iteratively joining each two core points if they are density-reachable (i.e. one core point is the epsilon neighborhood of the other) until there is no new pair of core points to be joined. The 'dbscan' function in 'dbscan' (v1.1-3) R package was used at https://cran.r-project.org/web/packages/dbscan/index.html. The parameter 'eps' in 'dbscan' function indicates the radius of neighbor selection, obtained by extracting the 5–95% (with 5% interval) quantile knn distance('kNNdist' function with $k = 5$) between cells.

Seurat [4]. The modularity optimization of cell–cell network which is constructed by SNN after feature selection (high variance gene selection) and feature extraction (principal component analysis). The 'Seurat' (v.3.0.0.9100) used in this paper is available at https://cran.r-project.org/web/packages/Seurat/index.html. The 'FindClusters' function was used to cluster cells and all processes and parameters remain default.

SOUP [26]. A semisoft clustering which could find pure cells (represent one cell type) and intermediate cells (represent Intermediate state between two cell types) by defining a purity score. The 'SOUP' (v0.0.0.9) R package used in this paper is available at https://github.com/lingxuez/SOUPR and all parameters remain default.

CIDR [28]. An ultrafast hierarchical clustering algorithm through imputation and dimensionality reduction. The 'cidr' (v0.1.5) R package used in this paper is available at https://github.com/VCCRI/CIDR and all processes and parameters remain default.

pcaReduce [30]. An agglomerative algorithm considering principal component analysis and hierarchical clustering at the same time. The 'pcaReduce' (v1.0) R package used in this paper is available at https://github.com/JustinaZ/pcaReduce. The parameters 'nbt', 'q' and 'method' indicate runs, number of cluster and merging strategy, respectively. ('nbt' = 5, 'q' = #cluster-1 and 'method' = 's' in our experiment).

SNN-Cliq [6]. A partition graph algorithm by merging Cliq (subgraph with fully connected) in a graph constructed by SNN. The 'SNN-cliq' source code used in this paper can be downloaded at http://bioinfo.uncc.edu/SNNCliq. The parameters 'r_cutoff' and 'merge_cutoff' indicate the radius of nearest neighbor and the merge threshold of merging each pair of Cliq, respectively ('r_cutoff' = 0.7 and 'merge_cutoff' = 0.5 in our experiment, and the other parameters remain default).

SC3 [5]. A single-cell consensus clustering considering different distance, dimension reduction methods and dimensions. The 'SC3' (v1.10.1) used in this paper is available at http://bioconductor.org/packages/sc3. The 'gene_filter' parameter is the bool value that controls whether gene filtering is performed in data preprocessing ('gene_filter' = False) and other processes and parameters remain default.

Integrating pathways

Pathway scoring

AUCell [19] is a method to score pathway on individual cell based on gene set enrichment analysis. The AUCell score is the AUC where the *x*-axis is the gene rank of its expression value in a cell and the *y*-axis is the number of genes hit in the pathway. For a pair of cell and pathway, a higher AUCell score indicates that most genes in this pathway are expressed higher than genes that are not in this pathway. From a pair of pathway database and scRNA-seq dataset, we can construct a cell-pathway AUCell score matrix. We used 'AUCell_calcAUC' function in 'AUCell' R package (v1.8.0) to calculate AUCell score and kept all parameters default.

Similarity network fusion

SNF [63] is an integration method based on network fusion. It is designed to fuse sample-sample similarity matrix in multi-omics data. It uses local correlation (i.e. the number of neighbors shared between nodes in the network) to convey information between different omics (i.e. network) and can potentially be used in different context. Therefore, we applied SNF for integrating pathway-level features to clustering method.

For pathway integration, SNF is used to integrate gene-level features (cell-gene expression matrix) and pathway-level features (cell-pathway AUCell score matrix). Let $W(i, j)$ represent the

correlation between cell i and cell j, the definition of normalized weight matrix and local similarity matrix is as follows:

$$P(i,j) = \begin{cases} \frac{W(i,j)}{2\sum_{k\neq i} W(i,k)}, j \neq i \\ 1/2, j = i \end{cases} \quad S(i,j) = \begin{cases} \frac{W(i,j)}{\sum_{k\in N_i} W(i,k)}, j \in N_i \\ 0, \text{otherwise} \end{cases},$$

where N_i indicates the neighbors of x_i.

The SNF performs an iterative operation between cell-gene expression matrix and cell-pathway AUCell score matrix and obtains the updated similarity matrix as follows:

$$P^{(1)}_{t+1} = S^{(1)} \times P^{(2)}_t \times \left(S^{(1)}\right)^T$$
$$P^{(2)}_{t+1} = S^{(2)} \times P^{(1)}_t \times \left(S^{(2)}\right)^T,$$

where $P^{(1)}_t$ and $S^{(1)}$ are normalized weight matrix and local similarity matrix of cell-gene expression matrix, and $P^{(2)}_t$ and $S^{(2)}$ are that of cell-pathway AUCell score matrix.

The final fusion matrix is computed as follows:

$$P^{(c)} = \frac{P^{(1)}_t + P^{(2)}_t}{2}.$$

Key Points

- We designed a framework to study the accuracy and robustness of existing single-cell clustering by integrating pathways.
- Pathway information can reduce the impact of single-cell data noise and improve the accuracy and robustness of most single-cell clustering methods.
- Ten state-of-the-art single-cell clustering methods, 26 scRNA-seq data and four pathway databases, pathway integration method and a complete set of evaluation indicators are collected in our framework.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Data Availability

The computer codes are available at https://github.com/GaoLabXDU/sciPath

Acknowledgements

We thank all members in the Prof. Gao's lab at Xidian University for their valuable suggestions. We also thank the HPC system of Xidian University for providing computing resources.

Funding

National Key R&D Program of China (2018YFC0910400 to L.G.); National Natural Science Foundation of China (61532014 to L.G., 61772395 to B.W.); NSERC Discovery Grant (RGPIN-2019-04904 to Y.G.).

References

1. Potter SS. Single-cell RNA sequencing for the study of development, physiology and disease. *Nat Rev Nephrol* 2018;**14**:479–92.
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;**19**:553–67.
4. Stuart T, Butler A, Hoffman P, *et al.* Comprehensive integration of single-cell data. *Cell* 2019;**177**:1888–1902.e21.
5. Kiselev VY, Kirschner K, Schaub MT, *et al.* SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* 2017;**14**:483–6.
6. Xu C, Su Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* 2015;**31**:1974–80.
7. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet* 2019;**20**:273–82.
8. Kanehisa M, Furumichi M, Tanabe M, *et al.* KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;**45**:D353–61.
9. Fan JBJ, Salathia N, Liu R, *et al.* Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat Methods* 2016;**13**:241–4.
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. Zhang Y, Zhang Y, Hu J, *et al.* scTPA: a web tool for single-cell transcriptome analysis of pathway activation signatures. *Bioinformatics* 2020;**36**:4217–9.
12. Zhang Y, Ma Y, Huang Y, *et al.* Benchmarking algorithms for pathway activity transformation of single-cell RNA-seq data. *Comput Struct Biotechnol J* 2020;**18**:2953–61.
13. Frost HR. Variance-adjusted Mahalanobis (VAM): a fast and accurate method for cell-specific gene set scoring. *Nucleic Acids Res* 2020;**48**:e94.
14. Ma Y, Sun S, Shang X, *et al.* Integrative differential expression and gene set enrichment analysis using summary statistics for scRNA-seq studies. *Nat Commun* 2020;**11**:1585.
15. Klimm F, Toledo EM, Monfeuga T, *et al.* Functional module detection through integration of single-cell RNA sequencing data with protein–protein interaction networks. *BMC Genomics* 2020;**21**:756.
16. DePasquale EAK, Schnell D, Dexheimer P, *et al.* cellHarmony: cell-level matching and holistic comparison of single-cell transcriptomes. *Nucleic Acids Res* 2019;**47**:e138.
17. DeTomaso D, Jones MG, Subramaniam M, *et al.* Functional interpretation of single cell similarity maps. *Nat Commun* 2019;**10**:4376.
18. Dai H, Li L, Zeng T, *et al.* Cell-specific network constructed by single-cell RNA sequencing data. *Nucleic Acids Res* 2019;**47**:e62.
19. Aibar S, González-Blas CB, Moerman T, *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* 2017;**14**:1083–6.
20. 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;**24**:1814–22.

21. 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.
22. Ji Z, Ji H. TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res* 2016;**44**:e117.
23. Fabregat A, Jupe S, Matthews L, *et al*. The reactome pathway knowledgebase. *Nucleic Acids Res* 2018;**46**:D649–55.
24. Slenter DN, Kutmon M, Hanspers K, *et al*. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res* 2018;**46**: D661–7.
25. Lloyd S. Least squares quantization in PCM. *IEEE Trans Inf Theory* 1982;**28**:129–37.
26. Zhu L, Devlin B, Lei J, *et al*. Semisoft clustering of single-cell data. *Proc Natl Acad Sci* 2018;**116**:466–71.
27. Ward JH. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc* 1963;**58**:236–44.
28. 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.
29. Shi J, Malik J. Normalized cuts and image segmentation. *IEEE Trans Pattern Anal Mach Intell* 2000;**22**:888–905.
30. Žurauskien J, Yau C. pcaReduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinformatics* 2016;**17**:140.
31. Daszykowski M, Walczak B. Density-based clustering methods. *Compr Chemom* 2009;**2**:635–54.
32. 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.
33. Muraro MJ, Dharmadhikari G, Grün D, *et al*. A single-cell transcriptome atlas of the human pancreas. *Cell Syst* 2016;**3**:385–394.e3.
34. 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**:1787–96.
35. 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**:e20–31.
36. Camp JG, Sekine K, Gerber T, *et al*. Multilineage communication regulates human liver bud development from pluripotency. *Nature* 2017;**546**:533–8.
37. Patel AP, Tirosh I, Trombetta JJ, *et al*. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science (80-)* 2014;**344**:1–9.
38. 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**:15672–7.
39. 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**:1053–8.
40. 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**:7285–90.
41. Romanov RA, Zeisel A, Bakker J, *et al*. Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. *Nat Neurosci* 2017;**20**:176–88.
42. 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**:193–6.
43. 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.
44. 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.
45. Tasic B, Menon V, Nguyen TN, *et al*. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat Neurosci* 2016;**19**:335–46.
46. 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.
47. Treutlein B, Brownfield DG, Wu AR, *et al*. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* 2014;**509**:371–5.
48. Klein AM, Mazutis L, Akartuna I, *et al*. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 2015;**161**:1187–201.
49. 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**:145–53.
50. Kolodziejczyk AA, Kim JK, Tsang JCH, *et al*. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* 2015;**17**:471–85.
51. Wang YJ, Schug J, Won K-J, *et al*. Single-cell transcriptomics of the human endocrine pancreas. *Diabetes* 2016;**65**:3028–38.
52. 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**:1586–90.
53. Xin Y, Kim J, Okamoto H, *et al*. RNA sequencing of single human islet cells reveals type 2 diabetes genes. *Cell Metab* 2016;**24**:608–15.
54. 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**:708–18.
55. 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**:1131–9.
56. 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.
57. 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**:1138–42.
58. Huh R, Yang Y, Jiang Y, *et al*. SAME-clustering: single-cell aggregated clustering via mixture model ensemble. *Nucleic Acids Res* 2020;**48**:86–95.
59. Li X, Chen W, Chen Y, *et al*. Network embedding-based representation learning for single cell RNA-seq data. *Nucleic Acids Res* 2017;**45**:e166–6.
60. Stoney RA, Schwartz J-M, Robertson DL, *et al*. Using set theory to reduce redundancy in pathway sets. *BMC Bioinformatics* 2018;**19**:386.
61. Kamburov A, Stelzl U, Lehrach H, *et al*. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res* 2013;**41**:D793–800.
62. Rodchenkov I, Babur O, Luna A, *et al*. Pathway commons 2019 update: integration, analysis and exploration of pathway data. *Nucleic Acids Res* 2019;**48**:D489–97.
63. Wang B, Mezlini AM, Demir F, *et al*. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* 2014;**11**:333–7.