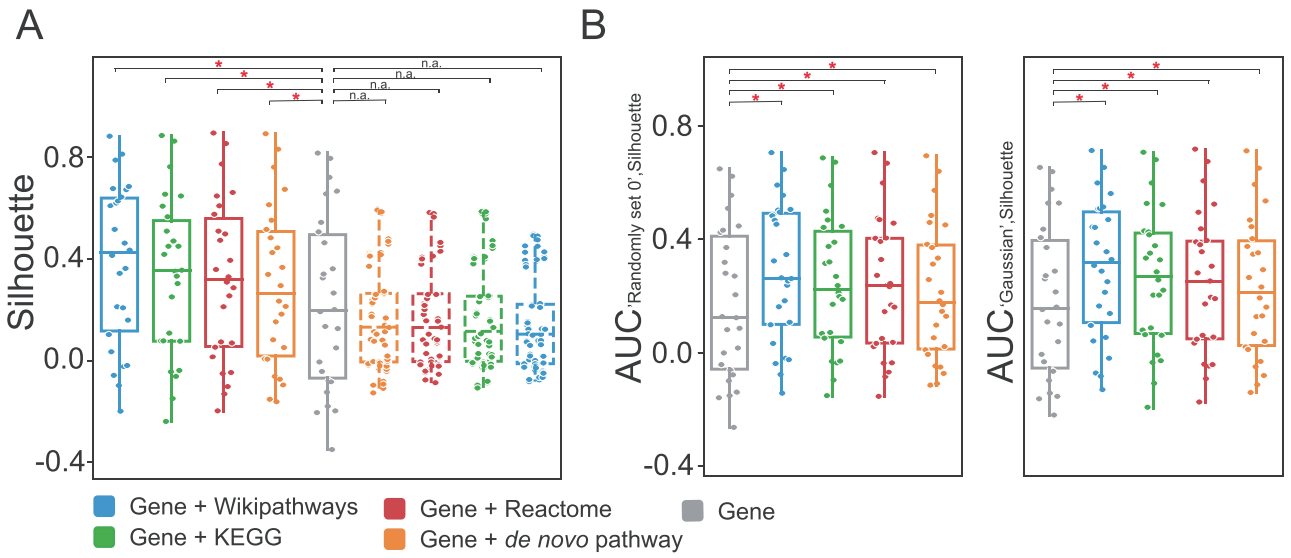

Figure 2. Improvement of cell–cell similarity metric for clustering. **(A)** Silhouette of cell–cell similarity metric (colored and solid box) and that of which integrate permutated pathway (colored and dashed box) on each scRNA-seq data. **(B)** The area under 'randomly set 0' (left side) and 'Gaussian' (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*