Genome Biology

METHOD

Open Access

Dimension reduction, cell clustering, and cell–cell communication inference for single-cell transcriptomics with DcjComm

Qian Ding[1†], Wenyi Yang[1†], Guangfu Xue[1†], Hongxin Liu[1], Yideng Cai[1], Jinhao Que[1], Xiyun Jin[2], Meng Luo[1], Fenglan Pang[1], Yuexin Yang[1], Yi Lin[2], Yusong Liu[2], Haoxiu Sun[2], Renjie Tan[2], Pingping Wang[2*], Zhaochun Xu[2*] and Qinghua Jiang[1,2,3*]

†Qian Ding, Wenyi Yang and Guangfu Xue contributed equally to this work.

*Correspondence: wangpingping@hrbmu.edu.cn; zhaochunxu@hrbmu.edu.cn; qhjiang@hit.edu.cn

[1] Center for Bioinformatics, School of Life Science and Technology, Harbin Institute of Technology, Harbin 150000, China
[2] School of Interdisciplinary Medicine and Engineering, Harbin Medical University, Harbin 150076, China
[3] State Key Laboratory of Frigid Zone Cardiovascular Diseases (SKLFZCD), Harbin Medical University, Harbin 150076, China

Abstract

Advances in single-cell transcriptomics provide an unprecedented opportunity to explore complex biological processes. However, computational methods for analyzing single-cell transcriptomics still have room for improvement especially in dimension reduction, cell clustering, and cell–cell communication inference. Herein, we propose a versatile method, named DcjComm, for comprehensive analysis of single-cell transcriptomics. DcjComm detects functional modules to explore expression patterns and performs dimension reduction and clustering to discover cellular identities by the non-negative matrix factorization-based joint learning model. DcjComm then infers cell–cell communication by integrating ligand-receptor pairs, transcription factors, and target genes. DcjComm demonstrates superior performance compared to state-of-the-art methods.

Keywords: Single-cell, Cell clustering, Cell–cell communication, Joint learning, Non-negative matrix factorization

Background

The advancement of high-throughput single-cell RNA sequencing (scRNA-seq) technologies provides valuable insights into the diversity of cellular states and unravels their dynamic connections [1]. Systematically executing multiple single-cell data analysis tasks, such as functional gene modules detection, representative features selection, cell clustering, and cell–cell communication (CCC) inference, enhances the comprehensive characterization of gene expression patterns and cellular dynamic processes.

Genes and other biological molecules frequently demonstrate intricate interactions and modular organization, crucial for understanding cellular mechanisms, unraveling cellular functions, and exploring the complexities of heterogeneity. Various initiatives have been undertaken to identify the functionality and structure of gene modules, with

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

network topology-based methods widely adopted for module detection, effectively identifying these modules as highly connected subgraphs within larger biological networks. Among them, WGCNA [2] utilizes the overlapping topological structures of networks and hierarchical clustering methods to detect gene functional modules. Qcut [3] is a partition-based module recognition method, which is most likely to detect modules automatically from the network with thousands of nodes in a relatively short time. What is more, as a network density-based module detection method, MCODE [4] is used to perform network node weighting, module prediction, and module optimization processing. In addition, expression-based methods are another commonly used approach for module detection, as they capture genes that exhibit similar expression patterns across multiple samples. For example, Hwang et al. [5] proposed the commonly unsupervised clustering method, namely MCL, to detect gene functional modules by simulating the process of randomly walking in the gene co-expression network. Chen et al. introduced the non-negative matrix factorization (NMF) to identify local patterns of gene expression data [6]. By adopting a perspective centered on molecular modules rather than solely focusing on individual molecules, we can achieve a deeper understanding of the behaviors exhibited by complex biological systems.

Cellular processes are fundamental mechanisms that sustain normal life functions. Multicellular organisms consist of various cell types with distinct functions, facilitating our understanding of tissue structure and function. Various computational methods have been developed to identify these cell types from scRNA-seq data. For instance, Steinley et al. [7] introduced the Kmeans clustering method to assign each cell to different clusters by iteratively identifying cluster centers and selecting the nearest cluster. Becht et al. [8] proposed the uniform manifold approximation and projection (Umap) method to achieve a meaningful organization of cell clusters, preserving both local and global data structures effectively. The hierarchical clustering method is also used to identify cell types based on single-cell data. For instance, Jiang et al. [9] proposed the cell–pair differentiability correlation (Corr) method to measure the similarity and adopt hierarchical clustering of cells by their neighborhood information. Later, CIDR [10] and SC3 [11] further improved the hierarchical clustering method of scRNA-seq data to produce robust results, which overcome the limitations of individual cells. In addition, learning similarity metrics from single-cell data and clustering based on these similarities is also a common method in single-cell clustering. SIMLR [12] learns a cell–cell similarity metric that optimally captures the data structure by integrating multiple kernels. Furthermore, the integration of dimension reduction methods with techniques for measuring cell-to-cell similarity has been widely adopted to effectively capture intercellular relationships. As one of the widely employed computational approaches for dimension reduction, the principal component analysis (PCA) [13] method identifies the genes with the highest variance by projecting the high-dimensional data into a low-dimensional space. For example, Stuart et al. [14] first utilize PCA to map single-cell gene expression data into a lower-dimensional space and then combine a smart local moving algorithm or spectral clustering with the cell-to-cell similarity matrix learning measure to identify cell subpopulations. However, scRNA-seq data typically exhibit characteristics such as high dimensionality, sparsity, and significant noise. This poses challenges in computing similarity between cells, as some genes may be expressed at low levels or even absent

in many cells, yet these genes may be crucial for distinguishing cell types. Moreover, in high-dimensional spaces, traditional distance metrics for calculating cell-to-cell similarity may become less accurate, thereby impacting the accuracy and reliability of clustering results [15]. In addition, deep learning methods such as scCAEs [16] and scGNN [17] have been successfully applied to explore the clusters on single-cell data. These algorithms directly utilize gene expression profiles to identify cell types, which are sensitive to the noise of data and ignore latent features in scRNA-seq data. Non-negative matrix factorization (NMF) aims to cluster cells by representative cells in a well-separated latent space, successfully applied to cell clustering in scRNA-seq data to enhance performance and accelerate convergence. DRjCC [15] and SSNMDI [18] further joint dimension reduction and NMF for cell clustering of scRNA-seq data. The advantage of DRjCC and SSNMDI is that the dimensionality reduction process generates features under the guidance of individual cell clustering, while individual cell clustering selects appropriate features. Compared to other methods, NMF-based methods have the advantage of learning interpretable individual parts and detecting context-dependent patterns of gene expression [19], which is one of the motivations of this study.

In multicellular organisms, cellular communication facilitates the coordination of multiple cells, enabling the formation of tissues, organs, and the fulfillment of diverse biological functions. Unveiling the CCC network through quantifying ligand-receptor (L-R) pairs in single-cell transcriptomics represents an extraordinary opportunity. For instance, Efremova et al. [20] developed the CellPhoneDB method to quantify the contextual communications of different cell types and further reveal their physiological processes according to the collected novel repository including ligands, receptors, and their interactions. Jin et al. [21] developed the CellChat tool to infer and analyze cellular signaling networks of ligand-receptor (L-R) pairs from scRNA-seq data. Anthony et al. [22] presented the CellTalker method to predict putative cell extrinsic interactions by quantifying ligand and receptor expression. While previous studies have inferred CCCs by considering the intercellular signaling specificity of L-R pairs, the transmission and amplification of intracellular signals through receptor-transcription factor (R-TF) and transcription factor-target genes (TF-TG) also significantly contribute to intracellular communication. Addressing this issue, several methods have explored intracellular signaling pathways. For example, Browaeys et al. [23] proposed the NicheNet method to infer active ligands and their gene regulatory effects between interacting cells. Zhang et al. [24] developed CellCall to infer intercellular and intracellular communication pathways by using the information on L-R pairs, transcription factor activity, and their target genes. Baruzzo et al. [25] presented scSeqComm, a computational method to infer the ongoing intercellular and intracellular signaling from scRNA-seq data. Cheng et al. [26] proposed scMLnet, a scRNA-seq data-based multilayer network method to identify functional intercellular communications and intracellular gene regulatory networks. In addition, spatially resolved transcriptomics has provided profound insights in biology and biomedicine, greatly enhancing the accuracy and reliability of inferring spatial proximal CCCs [27, 28, 29, 30]. There are currently several tools available that are specifically designed for inferring CCCs from spatial transcriptomics data, including COMMOT [31], Scriabin [32], Giotto [33], NICHES [34], and SpaTalk [35].

Here, we develop a computational method called DcjComm to coherently perform multiple scRNA-seq data analysis tasks such as functional gene module detection, representative features selection, cell clustering, and CCC inference. First, DcjComm detects the functional gene modules and selects representative features according to the decomposed projected matrix generated by the NMF-based joint learning model. Meanwhile, DcjComm utilizes non-negative matrix factorization of the joint learning model to discover cellular subpopulations. Then, DcjComm uses the inference statistical model to infer CCCs by integrating intercellular and related intracellular signals. To comprehensively capture these signals, DcjComm constructs a comprehensive and dependable database that gathers L-R, R-TF, and TF-TG interactions based on Reactome and KEGG pathways. DcjComm also provides a rich suite of visualization outputs (Circos plot, heatmap plot, Sankey plot, bubble plot, ridge plot, etc.) to intuitively show the analysis results of CCCs. Furthermore, we evaluate the performance of DcjComm on several publicly available scRNA-seq datasets and compare it with other state-of-the-art methods. The outstanding performance of DcjComm indicates that it is a powerful tool for performing multiple scRNA-seq data analysis tasks coherently, including functional gene module selection, representative feature selection, cell clustering, and inference of CCC networks.

Results

Overview of DcjComm

DcjComm enables the coherent execution of multiple single-cell analysis tasks to systematically reveal gene expression behaviors and cellular transcriptional states. DcjComm comprises two main components: (i) the NMF-based joint learning model and (ii) the CCCs inference statistical model. In the first component, DcjComm takes the gene expression matrix as input (Fig. 1A) and then performs functional gene modules detection, representative features selection, and cell type identification (Fig. 1B). DcjComm utilizes the matrix **U** obtained by the process of projection matrix decomposition to detect functional gene modules and select representative features and further quantifies the importance of selected modules based on the factor matrix **S** (Fig. 1D). Simultaneously, DcjComm applies non-negative matrix factorization to the coefficient matrix **V** to obtain the basis matrix **B**, which is subsequently used for cell clustering analysis. The Umap method is used for visualization to enhance the clarity of clustering results (Fig. 1C).

For the second component, the biological model of CCCs can be described as follows: intercellular signals are transmitted from ligands to receptors and then the signals are transduced to the downstream TFs through a specific signaling pathway, consequently triggering the transcriptional response of the target genes (Fig. 1E). Based on this biological model and the results of cell clustering, we construct a statistical model for deciphering the CCCs by quantifying intercellular and intracellular communication through the integration of paired L-R and TF activity (Fig. 1F). The expression matrix of ligands and receptors are respectively defined as the mean expression values of send cells and receive cells. The intercellular signaling is further defined as a two-dimensional vector represented by the L-R pair. Furthermore, intracellular signaling is defined as the activity of downstream TFs, which is computed through the Fisher test. Finally, the CCC score

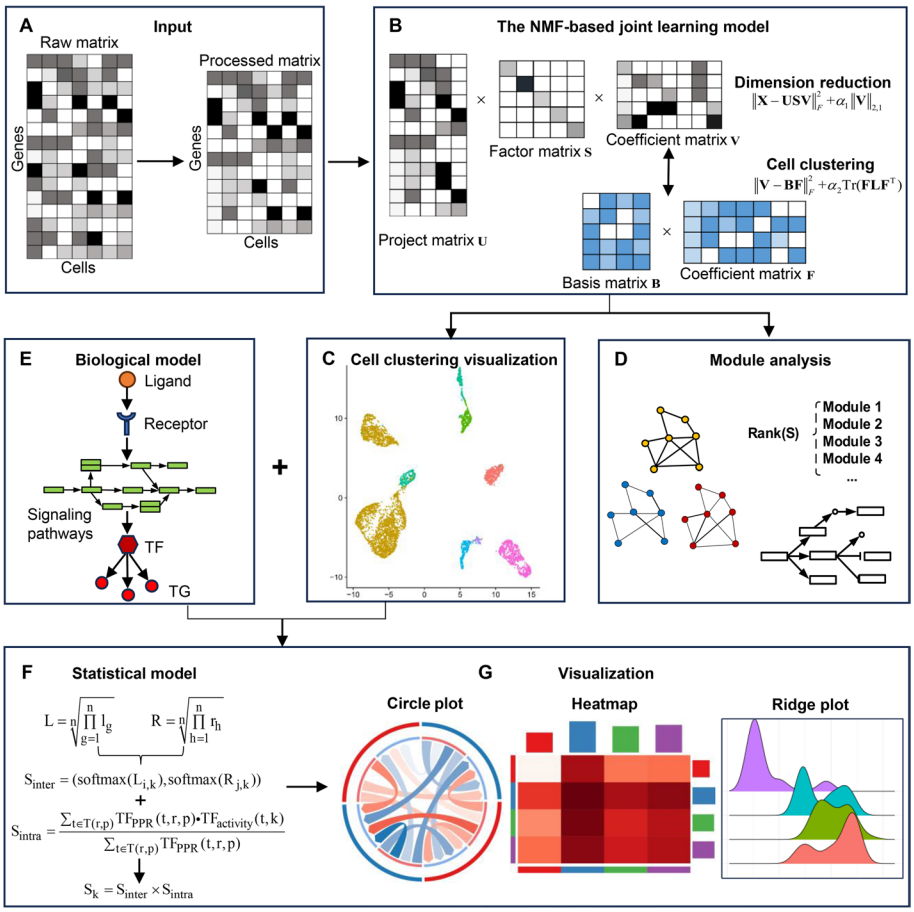

Fig. 1 Overview of DcjComm. **A** DcjComm takes a single-cell gene expression matrix as input and then processes it through a preprocessing step to obtain the preprocessed matrix. **B** The NMF-based joint learning model. DcjComm performs dimension reduction by projected matrix decomposition and cell clustering by non-negative matrix factorization. **C** Visualization of cell clustering results using the Umap method. **D** Selection and analysis of functional gene modules. It mainly contains the identification of functional gene modules and evaluation of their quality. **E** The biological model of CCCs. The prior knowledge includes links between ligands, receptors, signaling pathways, transcription factors, and target genes. **F** The inference statistical model of CCCs. DcjComm models the probability of communication between cells and identifies significant communications. **G** The visualization methods for the results of CCCs inference. It mainly includes circle plot, heatmap, and ridge plot

can be expressed as the product of the intercellular communication score and the intracellular communication score. DcjComm also provides a rich suite of visualization tools to intuitively demonstrate the prediction results of CCCs (Fig. 1G).

DcjComm performs functional gene module detection, dimension reduction, and cell clustering

Convergence and complexity analysis of DcjComm

The optimization of the NMF-based joint learning model comprises five sub-problems and introduces the extra variable **E** when solving sub-problem **V**. The complexities of the five sub-problems are induced as follows. In the process of dimension reduction, three variables are involved: **U**, **S**, and **V**. According to reference [36], the computational cost

of updating **U** is $O(tk_1mn)$. And the computational cost of updating **S** is also $O(tk_1mn)$. Then, the complexity of updating **V** is mainly focused on performing the process of singular value decomposition (SVD), which requires $O(m^3)$ [37]. Among them, t is the number of iterations, k_1 is the feature number after dimension reduction, m represents the number of genes, and n represents the number of cells in the input data. Similarly, the computational cost of updating **B** and **F** are $O(tk_1k_2n)$ and $O(tk_1k_2n)$ [36], where k_2 is the number of clusters. Then, the overall complexity of the DcjComm model is approximately $O(tk_1mn + tk_1k_2n + tm^3)$. Due to the $k_1, k_2 \ll n$, the complexity is basically $O(m^3)$ (The detailed parameter selection process is presented in the Additional file 1: Supplementary Note S1, and Additional file 1: Fig. S1, Fig. S2). For cell clustering, the running time and memory consumption of DcjComm and other comparison methods on fifteen single-cell datasets are respectively shown in Additional file 2: Table S1 and Table S2. Considering that deep learning methods typically require multiple rounds of pre-training and training, while hierarchical clustering methods often involve extensive computation of cell-to-cell similarities, these approaches can be time-consuming. Therefore, we only compared the runtime of DcjComm with eight other methods. Specifically, for the smaller cell dataset containing 366 cells (Wang), the runtime is approximately 1.8 s, with a memory requirement of 2514 MB; whereas for the larger dataset with 14,437 cells (Chen), the runtime is approximately 64.4 s, with a memory requirement of 3002 MB. For cell–cell communications inference, as shown in Additional file 2: Table S3, for the Wang dataset with 366 cells, cell–cell communications inference takes approximately 5 min, and for the Guerrero dataset with 12,951 cells, it takes approximately 10 min.

We first adopt the relative error to demonstrate the convergence of DcjComm. Specifically, relative errors of DcjComm on the above fifteen datasets as the number of iterations increases from 1 to 100 are shown in Additional file1: Fig. S3A. It can be observed that as the number of iterations increases, the error value gradually decreases. This indicates that DcjComm converges on all datasets and can quickly converge within 100 iterations. Then, the non-smoothness of the objective function, and the update process of DcjComm involves seven variables (**U**, **S**, **V**, **E**, **T**, **B**, **F**). Generally, the seven variables are iteratively updated until the convergence condition of DcjComm is satisfied, which guarantees the convergence of the DcjComm method. Firstly, for variables **V**, **E**, and **T**, we adopt the ADMM algorithm to search for their optimal solutions. Previous studies have demonstrated that the iterative functions of **V**, **E**, and **T** are convergent [38]. Secondly, the convergence of variables **U** and **S** has been demonstrated in [36]. Moreover, the convergence of variables **B** and **F** has also been demonstrated by [15]. Additionally, we have included a comprehensive proof of DcjComm's convergence in Additional file 1: Supplementary Note S2. Hence, we can conclude that the DcjComm method is convergent.

DcjComm identifies functional gene modules

To comparatively study the performance of DcjComm in detecting functional gene modules, we implement DRjCC and NMF, which also perform module detection after decomposing the top 2000 variable gene expression matrix. Taking Deng and Tabula datasets as examples, we equally select the suitable threshold of DcjComm, DRjCC, and NMF methods on these two datasets, so as to compare the performance of DcjComm

against the other two module detection methods. The detailed information on the threshold selection process for DcjComm, DRjCC, and NMF methods during module detection is in Additional file 1: Supplementary Note S3 and Fig. S4.

We first validated DcjComm by comparing it with the DRjCC and NMF methods through the analysis of cosine similarity between the core modules they identified. The DRjCC and NMF methods respectively select the largest module as the core module, which has the largest number of cells. For the DcjComm method, the core module we select corresponds to the position where the diagonal element of the matrix **S** is the highest value. Figure 2A reports the values of cosine similarity of the comparing methods, which represent the importance of detected modules. It is evident that DcjComm has

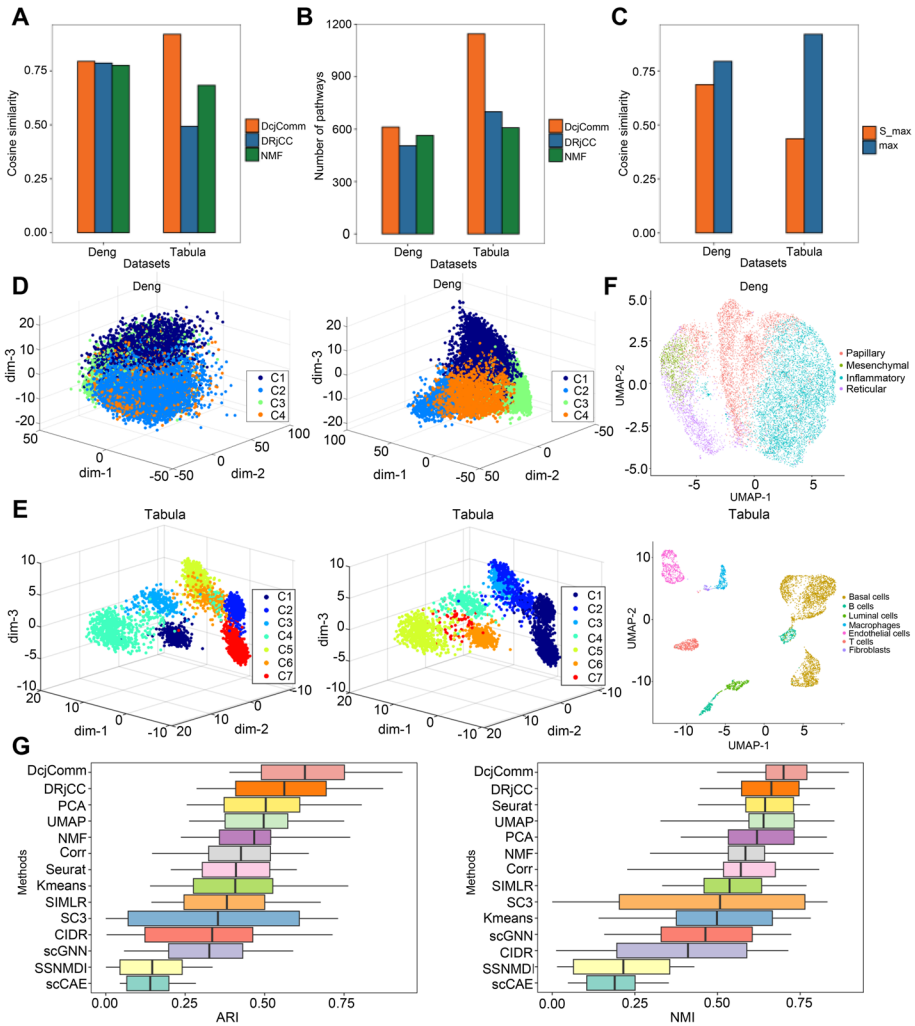

Fig. 2 Analysis of gene modules selection, dimension reduction, and cell clustering. **A** The values of cosine similarity of DcjComm and other comparison methods on Deng and Tabula datasets. **B** The number of significant pathways of DcjComm and other comparison methods on Deng and Tabula datasets. **C** Compare the values of cosine similarity of the core module and max module on Deng and Tabula datasets. **D**, **E** Visualize the cell distribution in the Deng (**D**) and Tabula (**E**) datasets based on raw data (left) and dimension reduction results (right). **F** Visualization results of cell clustering after cell type allocation of Deng (up) and Tabula (down) datasets. **G** Boxplots exhibiting the metrics of ARI and NMI for DcjComm and other comparison methods tested on 15 single-cell datasets

the highest cosine similarity among these methods, whereas the values of cosine similarity for Deng of DcjComm, DRjCC, and NMF are 0.7944, 0.7865, and 0.7757, respectively. In addition, the values of cosine similarity for the Tabula of DcjComm, DRjCC, and NMF are 0.9197, 0.4932, and 0.6834.

Furthermore, we also perform the pathway enrichment analysis of DcjComm, DRjCC, and NMF methods based on Deng and Tabula datasets to further demonstrate the validity of modules detected by DcjComm. We compare the performance of these three methods on GO, KEGG, Reactome, and WikiPathways databases. The number of significant pathways reflects the importance of the pathway enriched in modules. We select the significant pathways whose *P*-value (corrected by false discovery rate (FDR)) is smaller than 0.05 to analyze the performance of DcjComm, DRjCC, and NMF methods. The greater the number of important pathways, the higher the importance of the module. Figure 2B shows the number of significant pathways detected by these methods on Deng and Tabula datasets. Compared with DRjCC and NMF, DcjComm enriches more significant pathways on these two datasets. This shows that the modules detected by DcjComm are meaningful, and it has more important biological significance than the other methods. For the pathways enriched in the core modules detected by the Deng dataset, we specifically focus on the top 10 pathways with the lowest *P*-values (Additional file 2: Table S4). Among them, the cytosolic ribosome pathway is involved in regulating the upregulation of genes in human fibroblasts during fetal development, thereby influencing their growth and functional characteristics [39]. The selenocysteine synthesis pathway may be intricately related to collagen production and crosslinking, which are intimately associated with fibroblast redox homeostasis [40]. Among the top ten pathways enriched in the core modules detected by the Tabula dataset (Additional file 2: Table S5), the ribosome pathway and translation pathway are ranked highly in pathway enrichment analysis of differentially expressed genes detected in mammary gland samples [41]. The structural constituent of ribosome pathway as a downregulated molecule in mouse mammary glands may influence the expression of mammary genes [42].

Finally, we utilize the diagonal elements in **S** to evaluate the quality of the identified modules. The core modules of DcjComm are selected according to the values of the diagonal element of the matrix **S**. To demonstrate the importance of core modules detected by DcjComm, we also compare the cosine similarity values of the core modules with the largest modules; the results are shown in Fig. 2C. From Fig. 2C, the values of cosine similarity of the core modules are higher than the largest modules on these two data (i.e., Deng and Tabula datasets).

DcjComm improves the performance of dimension reduction

DcjComm jointly learns dimension reduction and cell clustering, and the process of cell clustering guides the feature selection process by dimension reduction. Here we evaluate whether the NMF-based joint learning model can improve the performance of dimensionality reduction. We first compare the dimension reduction performance of DcjComm, DRjCC, PCA, tSNE, and Umap according to calculating the mean square error (MSE), which is defined as the average distance between the predicted labels and the truth ground of cell types. The results of the above these 5 methods to calculate MSE on 15 datasets are shown in Additional file 1: Fig. S3B, which indicates that the MSE of

DcjComm is significantly lower than that of DRjCC and PCA. Among them, the MSEs of DRjCC, PCA, tSNE, and Umap are 1.83, 1.80, 1.74, and 1.82 times that of DcjComm, respectively. Thus, in the low-dimensional space generated by DcjComm, clusters are more compact and better separated.

Furthermore, taking the Deng and Tabula datasets as examples, Fig. 2D and Fig. 2E are respectively the schematic diagrams of dimensionality reduction of the raw data matrix and their dimensionality reduction results using the joint learning model of DcjComm. As shown in Fig. 2D and Fig. 2E, the cells in the raw data are mixed and not well separated. In contrast, the cells in the low-dimensional space are well separated, indicating that DcjComm chose features that are more distinguishable than dimensionality reduction methods. The NMF-based joint learning model of DcjComm improves the performance of dimensionality reduction. Except for the Chen dataset, the cell visualization results in the low-dimensional space before and after dimensionality reduction for other datasets are respectively shown in Additional file1: Fig. S5 and Fig. S6. Because the Chen dataset contains too many cell types, it is not shown here. After using the DcjComm method for dimensionality reduction, cells are better separated in the three-dimensional space obtained by the PCA method.

To further underscore the importance of dimension reduction, we ranked the low-dimensional matrix **U** from the DcjComm method by Laplacian scores [43] and identified the top 10 feature genes. These genes play crucial roles in the formation and development of human skin fibroblasts and mouse mammary gland epithelial cells through their specific expression. For the Deng dataset, the FBN1 gene encodes a member of the fibrillin family, which provides structural support for tensile strength in elastic and non-elastic connective tissues throughout the body [44]. The TRAM1 gene influences glycosylation and facilitates the translocation of secretory proteins across the endoplasmic reticulum membrane [45]. According to the records on the GeneCards website (https://www.genecards.org/), SNRPG, HIC1, TRAM1, ELANE, CCDC47, GPX4, HM13, and NOVA1 as coding genes, are respectively associated with spinal muscular atrophy, Miller–Dieker syndrome, Meckel syndrome, cyclic neutropenia and neutropenia, trichohepatoneuro developmental syndrome, spondylometaphyseal dysplasia, hepatitis C, and myoclonus. For the Tabula dataset, the Lmo4 gene plays a crucial role in promoting the development of mammary gland [46]. The Ptpn22 gene is a regulator of mammary gland differentiation, and inhibiting the expression of Ptpn22 can promote stem cell activity [47]. According to records on the GeneCards website, the genes Rgs10, Cisd2, Bhlhe40, Rpl3, Cab39, and Cnbp encode proteins associated with schizophrenia, Wolfram Syndrome, septal myocardial infarction, Diamond–Blackfan anemia, pancreatic cancer, and myotonic dystrophy, respectively. Similarly, based on records on the GeneCards website, the Slc35e4 gene is involved in transport mediated from the endoplasmic reticulum to the Golgi vesicles, while the cnih4 gene facilitates transmembrane transport.

DcjComm improves the accuracy of cell type discovery

In the process of cell clustering, determining the number of cell clusters (parameter k_2) is a fundamental and challenging issue. We introduce the gap decomposition method to predict the optimal number of clusters [48], which is a commonly used method for

analyzing single-cell data. That is, we respectively choose the number of eigenvalues close to 0 and the corresponding index when the maximum eigenvalue gap occurs as the lower and upper bounds of the number of clusters (Additional file 1: Fig. S2). The number of clusters estimated by DcjComm is in agreement with the true number in 11 out of 15 cases with a difference of no more than 1 (Additional file 2: Table S6). That is, the accuracy of the DcjComm method in identifying the number of clusters is 73.3%. In contrast, as shown in Table S6, the SC3 method has an accuracy of 18.2%, and the Corr method has an accuracy of 13.3%. Thus, DcjComm performs well in determining the number of cell types. Considering the effectiveness and fairness of the clustering process, the number of clusters provided in the original study is selected for the subsequent analysis.

To assess the performance of DcjComm on the identification of cell types, we select 15 biological datasets where the number of cells ranges from 366 to 14,437. To evaluate the clustering performance of DcjComm, several state-of-the-art clustering methods are selected for comparison, including three base clustering methods (Kmeans, PCA, and Umap), three NMF-based methods (NMF, DRjCC and SSNMDI), two graph-based methods (SIMLR and Seurat), two deep learning methods (scGNN, ScCAEs), and three hierarchical clustering methods (Corr, CIDR, SC3). Two common measurements are adopted as test statistics to characterize the performance of cell clustering, such as adjusted rand index (ARI) and normalized mutual information (NMI) (detailed definition in Additional file 1: Supplementary Note S4). The higher the value of the two metrics, the better the clustering performance. We compare the ARI and NMI results obtained using DcjComm with those from other cell clustering methods across fifteen scRNA-seq datasets (Additional file 1: Fig. S7 and Fig. S8). For further observation intuitively, the boxplot in Fig. 2G demonstrates the results of ARI and NMI, where DcjComm obtains the highest and most stable clustering performance on most scRNA-seq data compared with other methods in general. Among them, the SC3 method makes it difficult to handle the cell clustering problem of large-scale data (Baron, Deng, Guerrero and Chen) and the CIDR method cannot achieve convergence even after satisfying the maximum number of iterations for the Kolod dataset. From Additional file 1: Fig. S7 and Fig. S8, DcjComm significantly outperforms them on 12 datasets and has a similar performance with the best state-of-the-art methods on the other three datasets. For most methods, the experiment results of datasets with subtypes (i.e., Joost, Guerrero, Chen) are not as satisfactory as those with primitive cell types. While DcjComm achieves relatively acceptable clustering results on datasets with subtypes compared with other methods. Overall, across 15 datasets, DcjComm on average improves ARI by 24.55% and NMI by 19.66% over other comparison methods and up to 6.49% and 3.69% over the best comparison method for all datasets. These results further demonstrate that DcjComm is promising for cell-type assignment.

Based on the above clustering results, the following conclusions are easily obtained. Firstly, from Additional file 1: Fig. S7 and Fig. S8, it can be seen that the majority of methods exhibit relatively poor clustering performance for the single-cell data from the $10\times$ platform or with high resolution. By comparison, DcjComm obtains acceptable clustering performance among them. Secondly, the NMF-based methods, base clustering methods, and graph-based methods perform relatively well on the majority

of scRNA-seq datasets. Specifically, the NMF-based methods combine the dimensionality reduction with clustering to achieve better clustering results while graph-based methods consider the structural information between cells to improve the clustering performance. Thirdly, the base clustering methods and deep learning methods perform significant differences in clustering results on different datasets. Since the base clustering methods lack the ability to process contaminated data and deeply consider cell heterogeneity, for deep learning methods, the clustering results are not robust due to noisy data and parameter settings. Therefore, the DcjComm method demonstrates superior clustering accuracy compared to other state-of-the-art scRNA-seq clustering algorithms, as evidenced by comparative results.

Furthermore, we also compared DcjComm with other methods, such as Harmony [49], CIDER [50], and Seurat V3 [14]. The results demonstrate that DcjComm effectively eliminates batch effects in single-cell datasets while preserving excellent clustering performance. Furthermore, our analysis of various batch effect removal methods, including ComBat [51] and Limma [52], revealed that these methods have minimal impact on the clustering performance of DcjComm (Additional file 1: Supplementary Note S5 and Additional file2: Table S7 provide a detailed description of the above results of batch effects.).

In addition, as a widely used data visualization tool, the Umap method is employed to project high-dimensional data into a lower-dimensional space. To validate the subspace reconstruction ability of DcjComm, we respectively demonstrate the results of clustering visualization of these fifteen datasets by using Umap (Additional file 1: Fig. S9). From Fig. S9, the distance between the cells of different clusters increases and the boundaries of cells of the same species are more clearly defined. Therefore, DcjComm achieves better clustering results and corresponding visualization results, which demonstrates that DcjComm is robust in the improvement of feature extraction. Taking Deng and Tabula datasets as examples (Fig. 2F), we further assign cell types to different clusters according to the specific expression of marker genes in different clusters and the number of cells from different types contained in different clusters (Additional file 1: Supplementary Note S6, Fig. S10 and Fig. S11).

To further substantiate the biological significance of the clustering results, we performed differential gene expression (DEG) analysis and pathway enrichment analysis. Specifically, we utilized the FindAllMarkers function from the Seurat package to analyze differential gene expression based on the clustering results of the Deng and Tabula datasets, respectively. Subsequently, we employed the gprofiler2 package for pathway enrichment analysis of these detected DEGs (Additional file 1: Supplementary Note S7). This figure respectively presents the top ten DEGs detected in Deng (Additional file 1: Fig. S12A) and Tabula (Additional file 1: Fig. S12B) datasets, while Additional file 1: Fig. S13 illustrates the expression profiles of these identified marker genes. This analysis of DEGs is largely consistent with known marker genes, further validating the accuracy and effectiveness of the clustering results. Furthermore, pathway enrichment analysis was performed separately for the differentially expressed genes from the Deng (Additional file2: Table S8) and Tabula (Additional file2: Table S9) datasets, resulting in enrichment in 2875 and 5365 pathways, respectively. These pathways satisfy the *P*-value (FDR-adjusted) less than 0.05. The smaller the *P*-value, the higher its significance.

Among the top 10 pathways enriched with DEGs detected in the Deng dataset, extracellular space (GO:0005615) [53], extracellular space (GO:0005576) [54], extracellular exosome (GO:0070062) [55], and the SRP-dependent co-translational protein targeting to membrane pathway play important roles in human skin fibrosis processes [56]. Among the top ten pathways enriched with DEGs detected in the Tabula dataset, cellular component organization or biogenesis (GO:0071840) [57], cytoplasm (GO:000573 [58]), binding (GO:0005488) [42], organelle (GO:0043226) [59], and cellular component organization (GO:0016043) [60] are significantly enriched with differentially expressed genes in mouse mammary gland development, playing crucial regulatory roles in this process.

Comparison of the CCCs inference statistical model of DcjComm with other tools

DcjComm improves the performance of CCCs inference on single-cell datasets

To evaluate the performance of DcjComm in inferring CCCs, we compare it with that of nine other tools (scSeqComm, scMLnet, NicheNet, CellPhoneDB, CellChat, CellCall, CellTalker, NICHES, and iTALK) on three scRNA-seq datasets from the mouse skin (Guerrero), mouse skin (Joost), and human testicular (Wang). For evaluating the performance of CCCs inference, we perform the CCCs statistical inference model of DcjComm and other CCCs inference tools using the marker genes expression matrix and raw cell meta data. We first demonstrate that DcjComm can identify more comprehensive known L-R pairs. As shown in Fig. 3A–C (left), compared to other methods, the DcjComm method significantly identifies more L-R pairs on all three datasets by their own default cut-offs, respectively. Since the number of inferred L-R pairs is much larger than that of the other methods, for the downstream analysis, we select the top 500 L-R pairs from DcjComm of each cell–cell communication according to the CCC score.

Next, we infer that a more accurate method will have a larger proportion of overlapping predictions compared to other methods on average. Since the Jaccard coefficient reflects the overlap between different methods, a more accurate method should have a larger Jaccard coefficient compared to other methods (detailed definition of Jaccard coefficient in Additional file 1: Supplementary Note S8). As shown in Fig. 3D–F, the DcjComm method significantly identifies more commonly L-R pairs on all these three datasets. We found that DcjComm has the highest average rank based on the Jaccard coefficient of L-R pairs between any two methods of all three datasets, suggesting that DcjComm has the highest accuracy among these methods. We also evaluate the performance of DcjComm by simply comparing the overlap between the CCCs predicted by these CCC detection methods. Using the shared CCCs obtained by these CCC detection methods as the positive set, we further compare the F1 score for each method on these three datasets (Detailed definition of F1 score in Additional file 1: Supplementary Note S8). As shown in Fig. 3A–C (right), the DcjComm method shows high F1 score ranks on all these three datasets. Taken together, these results suggest that our DcjComm method highly ranked Jaccard coefficient and F1 scores in measuring the performance of different CCC detection methods and they are successful in finding more commonly identified communications.

Then, we compare the performance of DcjComm for inference of intracellular signal pathways of the receiver cells triggered by the CCCs with those of NicheNet, scMLnet,

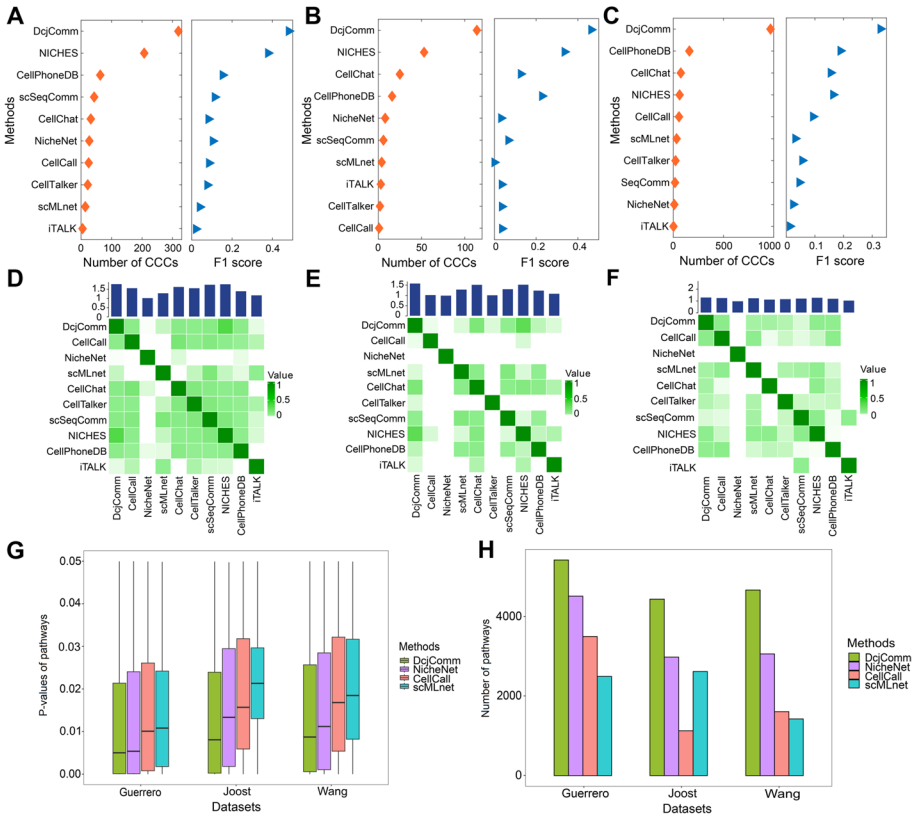

Fig. 3 The superior performance of DcjComm compared to other methods on Guerrero, Joost, and Wang datasets. **A**, **B**, **C** Comparison of the number of L-R pairs (left) and the value of the F1 score (right) between DcjComm and other methods on Guerrero (**A**), Joost (**B**), and Wang (**C**) datasets. **D**, **E**, **F** The heatmap of the Jaccard coefficient of L-R pairs between any two methods on Guerrero (**D**), Joost (**E**), and Wang (**F**) datasets. **G** and **H** Comparison of the *P*-values (**G**) and the number of pathways (**H**) between DcjComm and other methods on Guerrero, Joost, and Wang datasets

and CellCall that also infer the downstream target genes of CCC networks. We also infer that more accurate methods are more likely to utilize the downstream target genes inferred from receptor cells to enrich more receptor-related biological processes or pathways. To obtain more statistically significant pathways, we adopt the FDR-corrected method for pathway enrichment analysis with the GO, KEGG, Reactome, and WikiPathways databases. We compare the number and *P*-values of the intracellular pathways and biological processes triggered by the target genes obtained from all these methods. As shown in Fig. 3G and Fig. 3H, the DcjComm method enriches the most numerous and biologically significant pathways or biological processes. Our analyses demonstrate that the DcjComm method has an overall better performance in inferring the CCCs on all three datasets.

Considering that these methods utilize different default databases, which may affect the results of CCC inference, to compare fairly, we uniformly use the DcjComm-DB database as the reference database. Firstly, we compare the number of ligand–receptor pairs detected by DcjComm, CellCall, scMLnet, and NicheNet methods on Guerrero, Joost, and Wang datasets according to the DcjComm-DB database. These methods respectively obtain 319, 28, 46, and 81 ligand–receptor communication pathways on the

Guerrero dataset (Additional file 1: Fig. S14A (left)), 114, 39, 12, and 25 ligand–receptor communication pathways on the Joost dataset (Additional file 1: Fig. S14B (left)), and 969, 58, 47, and 21 ligand–receptor communication pathways on the Wang dataset (Additional file 1: Fig. S14C (left)). That is, compared with other methods, DcjComm detects more L-R pairs when using the DcjComm-DB database. Then, we compare the number of pathways detected by DcjComm, CellCall, scMLnet, and NicheNet methods. Based on the DcjComm-DB database, the number of pathways detected by the four methods on the Guerrero dataset are 5580, 3625, 2606, and 4259 (Additional file 1: Fig. S14A (right)) and 4567, 1326, 1410, and 3095 pathways on the Joost dataset (Additional file 1: Fig. S14B (right)). Similarly, the number of pathways detected by these four methods on the Wang dataset is 4803, 1656, 1715, and 3416 (Additional file 1: Fig. S14C (right)). What is more, we also record the median of *P*-values of pathways enriched by target genes related to TF of DcjComm, NicheNet, scMLnet, and CellCall methods based on the DcjComm-DB database in Additional file 1: Fig. S14D-F. We find that DcjComm detects the most pathways and also enriches the most effective pathways among these methods. Interestingly, when using the DcjComm-DB database instead of their default databases on Guerrero, Joost, and Wang datasets, all the above methods detect more L-R pairs and significant pathways (Additional file2: Table S10). Therefore, the effectiveness of the DcjComm method and the DcjComm-DB database we constructed are proven.

DcjComm improves the performance of CCCs inference on spatial transcriptomic datasets

To further provide additional validation of DcjComm, we apply it to two publicly available spatial transcriptomic datasets with single-cell resolution and use the information on cell spatial location, cell annotation, and gene expression matrices released by the original research. The first dataset is BZ5 tissue slices obtained from the medial prefrontal cortex of different mice using STARmap technology [61]. The second dataset is Bregma-0.04 obtained from the preoptic region of the mouse hypothalamus using MERFISH technology [62]. These two datasets respectively include 1049 and 5488 cells. Additional information about these two datasets is in Additional file 1: Supplementary Note S9.

Based on the assumptions that the adjacent cells should be more likely to have certain types of interactions than non-adjacent cells that are far away from each other, thus, given an L-R pair and two cell clusters, we should observe larger values of CCC scores S_k in close cells compared to distant cells. For intuitive observation, Additional file 1: Fig. S15A and Additional file 1: Fig. S15C demonstrate all L-R pairs between different clusters of the STARmap and MERFISH datasets. Additional file 1: Fig. S15B and Additional file 1: Fig. S15D further demonstrate the number of significant L-R pairs between different cell clusters (CCC score $S_k > 0.5$). For the STARmap dataset, the eL5 and L5 cells are mainly enriched in the L5 layer and also have a small distribution in the L2/3 layer, which leads to significant CCCs between eL5 and L5. In comparison, there are limited CCCs between eL6, Lhx6, Oligo, and other cell types, as these three cell types are primarily localized in the L6 layer and are spatially distant from other cell types. Nevertheless, there are still some interactions among these three cell types; for instance, there are effective interaction pairs including Lhx6-eL6 and Lhx6-Oligo. Then,

we further characterize the L-R pairs of different cell types on the MERFISH dataset. The majority of excitatory and inhibitory neurons are enriched in certain tissue regions while a small number of them are dispersed throughout the tissue. Thus, there are CCCs between these two cell types and other cell types. Specifically, there is an enrichment of Excitatory and Inhibitory neurons in both MPA and PVH regions, which leads to more significant L-R pairs between excitatory and inhibitory neurons. On the contrary, the ODMature and ependymal cells are respectively enriched in the fx and V3 regions. The distance between ODMature and the ependymal cells is relatively far, resulting in a few CCCs between them.

In addition, we compared the performance of DcjComm with nine other CCC detection tools on the STARmap and MERFISH datasets. For the STARmap dataset, DcjComm identifies three distinct L-R pairs. In contrast, apart from the iTALK method, which detects only one L-R interaction, the other eight tools fail to detect interactions between all cell types in the STARmap dataset. On the MERFISH dataset, DcjComm detects a total of 22 L-R interactions. In comparison, the CellChat, NicheNet, and NICHES methods detect 3, 7, and 2 L-R interactions, respectively, while the other tools do not detect interactions between all cell types. Thus, DcjComm outperforms the other nine CCC tools in detecting the most L-R pairs across both spatial transcriptomics datasets.

DcjComm infers CCCs among fibroblasts of human skin

To evaluate the CCC performance of DcjComm, we apply it to the Deng dataset on human skin fibroblasts. The marker genes expression matrix and the defined cell type assigned after cell clustering are imputed into DcjComm to determine the potential CCCs. Deng dataset covers the following four fibroblast subpopulations: reticular, papillary, mesenchymal, and inflammatory, which are known to signal to each other [63]. As is shown in Fig. 4A, we identify CCC signals between these cell types, including reticular–mesenchymal, reticular–papillary, reticular–reticular, reticular–inflammatory, papillary–mesenchymal, papillary–papillary, papillary–reticular, papillary–inflammatory, mesenchymal–mesenchymal, mesenchymal–papillary, mesenchymal–reticular, mesenchymal–inflammatory, inflammatory–mesenchymal, inflammatory–papillary, inflammatory–reticular, and inflammatory–inflammatory. Specifically, the darker the color of the edges in the circle plot (Fig. 4A), the higher the score of CCCs between cell–cell pairs. For observed intuitively, we further demonstrate the number of significant L-R pairs (CCC score $S_k > 0.5$) between different cell types (Fig. 4B), which involve 400, 373, 377, 281, 384, 345, 352, 277, 421, 412, 402, 385, 326, 259, 284, and 151 L-R interactions of the above cell–cell pairs, respectively. As shown in Fig. 4A and Fig. 4B, mesenchymal cells exhibit the highest number of detected L-R pairs with other cells, suggesting strong evidence of communications between mesenchymal cells and other cell types in CCCs, indicating a dominant role of mesenchymal cells. Similarly, the heatmap in Fig. 4C further emphasizes the crucial role of mesenchymal cells. The color intensity in the heatmap reflects the communication score between mesenchymal cells and other cell types, with darker colors indicating higher communication scores. To further illustrate the important role of mesenchymal cells in cell communication, Fig. 4D further visualized the communications between mesenchymal cells and other cell types.

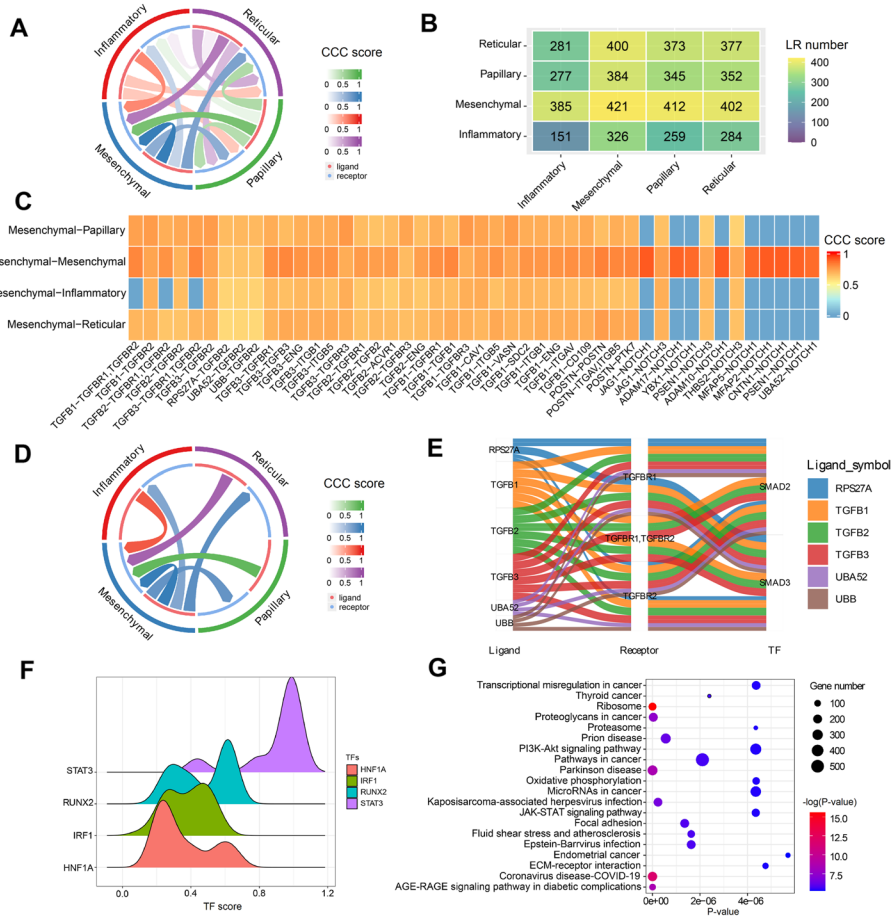

Fig. 4 CCCs analysis of Deng dataset. **A** Circos plot of CCCs among four cell types. **B** The number of CCCs is summarized by counting the number of significant L-R pairs in the form of a heatmap. **C** The heatmap illustrates the relative importance of mesenchymal cells based on significant CCCs from mesenchymal cells to other cell types. **D** The visualization highlights the important CCCs between mesenchymal cells and other cell types with a circos plot. **E** Sankey plot of CCC pathways of significant L-R-TF. **F** Ridge plot of significant TFs and their corresponding $TF_{activity}$ score. **G** Pathway enrichment analysis of differentially expressed genes regulated by TFs

The L-R pairs involved in Fig. 4C will affect signal transduction between these cell types. For instance, the increase of TGFB1 and TGFB2, the decrease of TGFB3, and the increase of TGFBR1 and TGFBR2 are known to be related to skin fibrosis [64]. CAV1 is a known inhibitor of TGFB1, and the loss of its expression can lead to skin fibrosis [65]. The binding of ITGB1 and ITGB5 with latent TGFB regulates the production and degradation of extracellular matrix and further leads to skin fibrosis [66]. CD109 is known to have the ability to activate TGFB, which is associated with fibrosis in many organs [67]. The combination of POSTN and PTK7 is known to activate the Wnt signaling pathway [68], which plays an important role in the development of fibrotic skin diseases [69]. The pathologic activation of NOTCH signaling is related to the pathogenesis of various fibrotic diseases [70], which is composed of four NOTCH receptors i.e., NOTCH1-NOTCH4, and several ligands such as JAG1 and JAG2.

MFAP5 promotes angiogenesis and interacts with NOTCH1 by activating or inhibiting its activity, thereby promoting human skin fibrosis [71].

Since the CCC networks are regulated by master transcription factors (TFs) according to previous reports [72, 73], we further validate the biological significance of these intercellular communication pathways by analyzing their TFs downstream (Fig. 4E). For the convenience of observation, we only show the top 50 CCC pathways according to the correlation score between receptors and TFs in the Sankey plot, such as SMAD2, SMAD3, and SMAD4, which have been validated to be involved in skin fibrosis [74]. Notably, several master TFs of the fibroblast subpopulations, such as HNF1A, STAT3, IRF1 and NF1A, are known to be important in skin fibrosis [63]. To further describe the significant roles that TFs play in modulating the transcriptional response of their target genes, Fig. 4F shows the $TF_{activity}$ score between the master TFs and their corresponding target genes. Furthermore, we perform pathway enrichment analysis on differentially expressed genes regulated by TFs, Fig. 4G shows the top 20 KEGG pathways according to *P*-value (FDR corrected, *P*-value < 0.05). Among them, the ribosome pathway [75], AGE-RAGE signaling pathway [76], PI3K-Akt signaling pathway [77], JAK/STAT3 signaling pathway [78], ECM receptor interaction pathway [79], and focal adhesion [75], have been reported to be critical for skin fibrosis.

Next, to systematically evaluate the performance of DcjComm, we compare it to other nine CCC detection tools (i.e., scSeqComm, scMLnet, NicheNet, CellPhoneDB, CellChat, CellCall, CellTalker, NICHES, and iTALK) on the marker genes expression matrix and the defined cell type assigned after cell clustering of human skin cells. Firstly, we calculate the Jaccard coefficients of ligand-receptor pairs to reflect the overlap degree of any two methods, which is shown in the heatmap of Fig. 5A. The average values of the Jaccard coefficients between these methods are shown in the bar plot above the heatmap, and the shades of green reflect the size of the Jaccard coefficients. As is shown in Fig. 5C (left), the DcjComm method also achieves the highest F1 score. We find that DcjComm has the highest overlap proportion based on the Jaccard coefficients of L-R pairs between any two methods, suggesting that DcjComm has the highest accuracy. Then, we compare the number of detected L-R pairs between all cell types by their own default cutoffs (Fig. 5B and Fig. 5C (right)) and further count the number of CCCs with the CCC score $S_k > 0.5$ and $S_k > 0.8$ (Fig. 5D (left) and Fig. 5D (right)) between mesenchymal cells and other cell types. From Fig. 5C and Fig. 5D, DcjComm detects more L-R pairs than other methods. In addition, to demonstrate the effectiveness of intracellular signal transduction, we compare the performance of DcjComm with the other three methods (i.e., NicheNet, scMLnet, CellCall) that offer downstream analysis of TFs and target genes. Considering that the downstream target genes mediated by receptor cells may affect the significance of biological processes or pathway enrichment, we evaluate the inferred target genes for each L-R communication of DcjComm, NicheNet, scMLnet, and CellCall methods according to the significance of pathway enrichment. Here, the FDR-corrected method is adopted for pathway enrichment analysis with the GO, KEGG, Reactome, and WikiPathways databases on the activated genes in receiver cells. Generally, the number and *P*-value represent the significance of enriched pathways and biological processes. The more pathways there are and the smaller the *P*-value of the pathway, the pathway is more important. Specifically, we record the median *P*-values of pathways

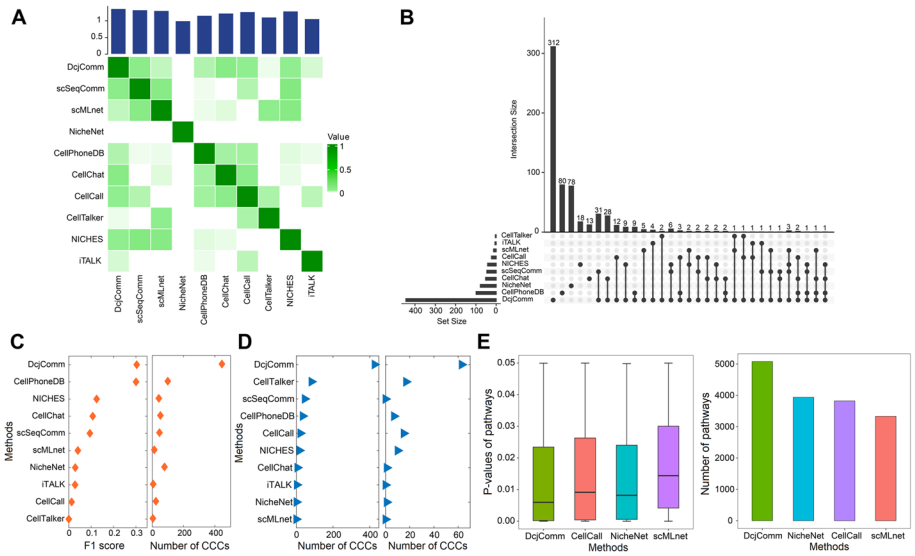

Fig. 5 Comparison analysis of DcjComm and other methods on the Deng dataset. **A** The heatmap of the Jaccard coefficients of L-R pairs between any two methods. The bar plot above the heatmap represents the average Jaccard coefficients between a given method and other methods. **B** UpSetR plot of predicted CCCs from the ten methods on the Deng dataset. **C** Comparison of the value of the F1 score (left) and the number of L-R pairs (right) between DcjComm and other methods. **D** Comparison of the number of L-R pairs from mesenchymal cells to other cell types obtained by DcjComm and other methods when threshold values are set to 0.5 (left) and 0.8 (right). **E** Comparison of the *P*-values (left) and the number of pathways (right) enriched by target genes obtained from DcjComm and other methods

enriched by target genes related to TFs in Fig. 5E (left). Besides, the number of pathways detected by DcjComm, NicheNet, scMLnet, and CellCall methods on the Deng dataset are 5103, 2603, 2259, and 3293, respectively (Fig. 5E (right)). From Fig. 5E, DcjComm obtains a smaller *P*-value and enriches more pathways compared to NicheNet, scMLnet, and CellCall methods. That is, DcjComm outperforms other compared methods for inferencing potential L-R pairs which mediate CCCs. To further assess the sensitivity of DcjComm to input data when inferring cell–cell communications, we employed the "geometric sketch" method [80] to subsample 80% and 90% of the total cells in the Deng dataset for validation. Subsequently, we compared the subsampled dataset with the input Deng dataset and computed the false positive rate (FPR). As shown in Additional file 1: Fig. S16, DcjComm exhibited the lowest FPR compared to CellCall, NicheNet, and scMLnet methods. Additionally, both DcjComm and CellCall demonstrated relative robustness during subsampling. According to reference [26], this may be attributed to the fact that both methods infer cell–cell communications based on cellular clustering. The relevant definition of FPR is detailed in Additional file 1: Supplementary Note S8.

Moreover, to compare fairly, we uniformly use the DcjComm-DB database as the reference database. Firstly, we compare the number of L-R pairs detected by DcjComm, NicheNet, scMLnet, and CellCall methods on the Deng dataset according to the DcjComm-DB database. These methods respectively obtain 445, 97, 20, and 15 L-R communication pathways (Additional file 1: Fig. S17A (left)). Then, we compare the number of pathways detected by DcjComm, NicheNet, scMLnet, and CellCall methods, i.e., these four methods respectively detect 5103, 4646, 2259, and 2847 (Additional file 1: Fig. S17A (right)). What is more, we also record the *P*-values of pathways enriched by target genes

related to TF of these four methods based on the DcjComm-DB database in Additional file 1: Fig. S17C. We find that the DcjComm method detects more L-R pairs and significant pathways among these methods. Interestingly, when using the DcjComm-DB database instead of their default databases on the Deng dataset, all other methods except for the CellCall method detect more pathways (Additional file 2: Table S11).

DcjComm infers CCCs among epithelial cells of mouse mammary gland

To further evaluate the effectiveness of DcjComm, we apply it to the Tabula dataset which contains seven cell types including T cells, B cells, macrophages, luminal cells, fibroblasts, endothelial cells, and basal cells. We input the marker genes expression matrix and the defined cell type assigned after cell clustering into DcjComm to infer the potential CCCs. For a better understanding of the CCCs between different cell types, we first record the CCC signals in Fig. 6A. Figure 6B further summarize the number of

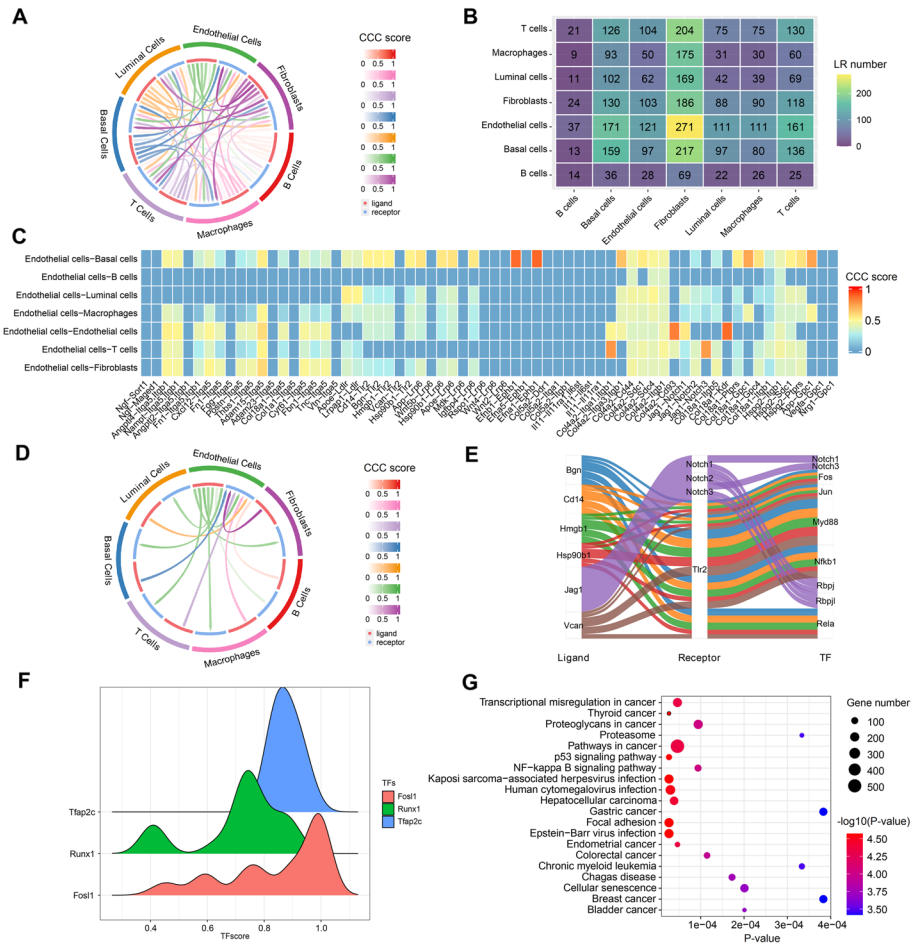

Fig. 6 CCCs analysis of Tabula dataset. **A** Circos plot of CCCs among seven cell types. **B** The number of CCCs is summarized by counting the number of significant L-R pairs in the form of a heatmap. **C** The heatmap illustrates the relative importance of mesenchymal cells based on significant CCCs from endothelial cells to other cell types. **D** The visualization highlights the important CCCs between endothelial cells and other cell types with a circos plot. **E** Sankey plot of CCC pathways of significant L-R-TF. **F** Ridge plot of significant TFs and their corresponding $TF_{activity}$ score. **G** Pathway enrichment analysis of differentially expressed genes regulated by TFs

significant L-R pairs (CCC score $S_k > 0.5$). As shown in Fig. 6A and Fig. 6B, compared to other cell types, endothelial cells exhibit significantly more frequent signal exchanges with other types of cells, indicating their dominant role in intercellular communication. To intuitively observe the role of these specific L-R pairs, we also record the interactions of these significant L-R pairs between endothelial cells and other cell types in the heatmap (Fig. 6C). Figure 6D further visualized the communications between endothelial cells and other cell types, elucidating their important roles in cell–cell communications.

What is more, recent studies have demonstrated that endothelial cells are a common cell type in many mammalian tissues, which interact with macrophages and play a crucial role in the tissue adaptation and tissue homeostasis of macrophages [81]. Thus, we focus on the endothelial cells in the next step. Considering the following eight ligands, i.e., Col5a2, Col4a2, Col18a1, Il11, Jag1, Ngf, and Hspg2 and six receptors (Ephb1, Gpc1, Itga5, Ldlr, Lrp6 and Tlr2) are known to be important in maintaining normal mouse mammary gland development and are associated with basal-like and triple-negative breast cancers [82, 83, 84, 85, 86].

Moreover, we also analyze the TFs activated downstream of these intercellular communication pathways to further validate their biological significance. The Sankey plot of Fig. 6E demonstrates TFs downstream of the top 50 CCC pathways according to the correlation score between receptors and TFs. From Fig. 6E, the specific L-R pairs mentioned above may regulate the following nine TFs: Notch1, Notch3, Jun, Fos, Myd88, Nfkb1, Rbpj, Rbpj1, and Rela. These TFs are known to be important in the development of mammary glands [87, 88, 89, 90]. Several master TFs of the endothelial cells, such as Fosl1, Runx1, and Tfap2c, are known to be important in defining cell identities. Figure 6F shows the $TF_{activity}$ score between the above three master TFs and their corresponding target genes to demonstrate the significant roles that master TFs play in modulating their target genes. Specifically, compromised Runx1 regulation is related to many cancers such as the blood, bone, and mammary glands [91]. The expression and protein levels of Fosl1 in human normal mammary gland cells and different breast cancer cells were significantly overexpressed [92]. Tfap2c has been identified as a prognostic factor for breast cancer and regulates the luminal epithelial phenotype in the development of the mammary gland [93]. To further validate the significance of these identified TFs, we perform pathway enrichment analysis on differentially expressed genes regulated by them. Figure 6G shows the top 20 KEGG pathways according to *P*-value (FDR corrected, $P\text{-value} < 0.05$). Among them, the p53 signaling pathway [94], NF-kappa B signaling pathway [94], proteoglycans in cancer [95], and chronic myeloid leukemia [96] have been reported to be critical for regulating the development of the mammary gland.

To further demonstrate the effectiveness of DcjComm, we also compare it with other CCC detection tools mentioned above on the marker genes expression matrix and the defined cell type assigned after cell clustering on the Tabula dataset. We first calculate the Jaccard coefficients (Fig. 7A) and F1 score (Fig. 7C (left)) to reflect the overlap degree between different methods, which indicates that DcjComm has the highest overlap proportion between any two methods. Then, compared with other methods, DcjComm obtains the most L-R pairs between all cell types (Fig. 7B and Fig. 7C (right)). Furthermore, we also count the number of CCCs with the CCC score $S_k > 0.5$ (Fig. 7D (left)) and $S_k > 0.8$ (Fig. 7D (right)) between the endothelial cells and other cell types

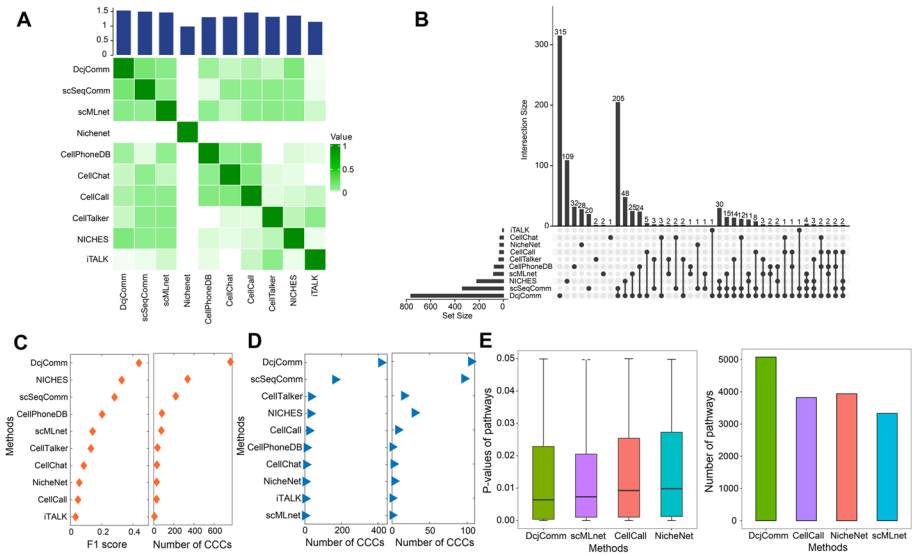

Fig. 7 Comparison analysis of DcjComm and other methods on the Tabula dataset. **A** The heatmap of the Jaccard coefficients of L-R pairs between any two methods. The bar plot above the heatmap represents the average Jaccard coefficients between a given method and other methods. **B** UpSetR plot of predicted CCCs from the ten methods on the Tabula dataset. **C** Comparison of the value of the F1 score (left) and the number of L-R pairs (right) between DcjComm and other methods. **D** Comparison of the number of L-R pairs from endothelial cells to other cell types obtained by DcjComm and other methods when threshold values are set to 0.5 (left) and 0.8 (right). **E** Comparison of the *P*-values (left) and the number of pathways (right) enriched by target genes obtained from DcjComm and other methods

on the Tabula dataset. Finally, we also compare the significance of intracellular signal transduction involved in DcjComm with the other three methods (i.e., NicheNet, scMLnet, CellCall) that offer downstream analysis of TFs and target genes. We evaluate the inferred target genes for each L-R communication of DcjComm, NicheNet, scMLnet, and CellCall methods according to the significance of pathway enrichment. Here, the FDR-corrected method is adopted for pathway enrichment analysis with the GO, KEGG, Reactome, and WikiPathways databases on the activated genes in receiver cells. Figure 7E also shows the *P*-values and the number of pathways enriched by target genes related to TFs. Compared to NicheNet, scMLnet, and CellCall methods, DcjComm obtains a smaller *P*-value (Fig. 7E (left)) and enriches more pathways (Fig. 7E (right)). Subsequently, we also subsampled 80% and 90% of the total cells in the Tabula dataset using the "geometric sketch" method. As shown in Additional file 1: Fig. S18, DcjComm also exhibited the lowest FPR on the Tabula dataset compared to CellCall, NicheNet, and scMLnet methods.

In addition, based on the DcjComm-DB database, we also compare the number of L-R pairs detected by DcjComm, NicheNet, scMLnet, and CellCall methods on the Tabula dataset. These methods respectively obtain 763, 465, 115, and 39 L-R communication pathways (Additional file 1: Fig. S17B (left)). Then, we compare the number of pathways detected by these four methods. Based on the DcjComm-DB database, the number of pathways detected by the four methods are 5080, 1166, 3608, and 4429, respectively (Additional file 1: Fig. S17B (right)). We also record the *P*-values of pathways enriched by target genes related to TF of these four methods based on the DcjComm-DB database in Supplementary Additional file 1: Fig.

S17D. We find that DcjComm detects the most pathways and also enriches the most effective pathways on the Tabula dataset among these methods (Additional file 2: Table S11). That is when using the DcjComm-DB database instead of their default databases, all the above methods detect more pathways.

DcjComm infers CCCs among spatial transcriptomics data of human breast *cancer*

To further validate the effectiveness of the DcjComm method in analyzing spatial transcriptomics data, we compared it with five other spatial methods, including COMMOT, Scriabin, SpaTalk, Giotto, and NICHES, for inferring CCCs on the human breast cancer dataset (BreastST [97]) downloaded from the 10 × Genomics website. The cell types in the BreastST dataset were previously annotated [97], as shown in Additional file 1: Fig. S19A. We utilized this annotation information as prior knowledge to differentiate between adjacent and distant cell types. The heatmap in Fig. S19B illustrates the Euclidean distances between different cell types in spatial locations. From this figure, we observed that basallike2 cells and mesenchymal cells are very close to each other, and mesenchymal cells are also in close proximity to stroma cells. In contrast, the distances between T cells and stroma cells, as well as between basallike2 cells and T cells, are relatively greater. We compared the CCC scores detected by DcjComm, COMMOT, Scriabin, SpaTalk, Giotto, and NICHES for these four cell types. As shown in Fig. S19, DcjComm identifies higher CCC scores for adjacent cell types (basallike2-mesenchymal and mesenchymal-stroma), whereas it detects lower CCC scores for distant cell types (T cells-stroma and basallike2-T cells). This is consistent with the ground truth of spatial transcriptomics data, which indicates that communications are more likely between adjacent cells than between distant cells. In addition to the Giotto and NICHES methods, DcjComm and other spatial approaches also reveal a consistent trend in spatial transcriptomics data, with neighboring cells showing higher CCC scores and distant cells exhibiting lower CC scores (Fig. S19C-Fig. S19H). Specifically, DcjComm detected relatively higher CCC scores for adjacent cells compared to the other methods. This demonstrates that DcjComm is a valuable tool for analyzing spatial transcriptomic data.

Next, we compared the number of L-R pairs and the F1 scores inferred by DcjComm and five other spatial methods using the BreastST dataset. As shown in Fig. S20A (left), DcjComm detected a greater number of L-R pairs compared to the other methods, highlighting its enhanced capability to capture spatial CCCs. Furthermore, DcjComm achieved the highest F1 score, outperforming the other methods and demonstrating superior accuracy in detecting CCCs (Fig. S20A (right)). Then, among these five comparison methods, only SpaTalk consider the downstream target genes involved in CCCs. To further validate the biological significance of DcjComm, we conducted a detailed analysis comparing the statistical significance (*P*-values) and the number of intracellular pathways and biological processes identified by DcjComm with those identified by SpaTalk. As illustrated in Fig. S20B and Fig. S20C, the pathways identified by DcjComm are more biologically significant and more numerous. Therefore, DcjComm still demonstrates superior performance compared to other methods for detecting CCCs in spatial transcriptomic data.

Discussion

Systematically characterizing gene expression patterns, exploring cell subpopulations, and exploring complex signaling patterns of scRNA-seq data can help reveal the dynamic phenomena within cells. Despite the development of various computational methods to address multiple single-cell analysis tasks, there is still a significant absence of comprehensive tools capable of effectively performing these tasks coherently.

Here, we develop the DcjComm, a toolkit to perform multiple scRNA-seq data analysis tasks coherently, such as detecting functional gene modules, selecting representative features, clustering cells, and inferring CCCs. We first propose the NMF-based joint learning method to detect functional gene modules, select representative features, and cluster cells simultaneously. By employing the projection matrix decomposition, DcjComm detects the functional gene modules to study the mechanisms of molecular actions and performs dimension reduction to select representative features. Meanwhile, DcjComm utilizes non-negative matrix factorization for cell clustering to decode the cell subpopulations. Then, we develop the CCCs inference statistical model by integrating paired L-R, R-TF, and TF-TG interactions to decipher intercellular and intracellular communications. In addition, to facilitate intuitive visualization and perform various downstream analysis tasks, DcjComm provides a rich suite of visualization options, i.e., circos plot, heatmap, sanky plot, ridge plot, and bubble plot. Compared with other state-of-the-art scRNA-seq data analysis methods, DcjComm achieves excellence in functional gene modules detection, representative features selection, cell clustering, and CCCs inference. DcjComm enables systematically performing multiple scRNA-seq data analysis tasks and offers valuable insights into a comprehensive understanding of the intricate communicative mechanisms across different conditions.

Despite DcjComm possesses specific advantages and characteristics in executing multiple scRNA-seq data analysis tasks, it still has certain limitations. First, the analysis of intercellular signaling mainly centers on transcript expression, neglecting the analysis of protein bioactivity and its post-translational mechanisms. This limitation could restrict the localized analysis of intracellular signaling networks and the comprehensive understanding of intercellular communications. While in intracellular signaling mechanisms, multiple signaling pathways may interfere with each other, impacting gene regulatory rules and potentially causing false positives or false negatives. Therefore, the results of the predicted CCCs still need further biological validation. Second, the communication between cells typically involves cellular distance information and their positional relationships. The lack of information about the spatial arrangement of cells within tissues or organs may hinder the understanding of cellular behavior under physiological and pathological conditions. In the future, the integration of other omics technologies, such as proteomics and glycomics, will be critical for further exploring the complexity and diversity of CCCs. The practical applicability of these data types can be explored in the future development of DcjComm. Furthermore, integrating spatial information with scRNA-seq data may provide new insights into CCCs. The current version of DcjComm offers a user-friendly tool for the systematic analysis of scRNA-seq data. In future work, we anticipate that the DcjComm method will enable the establishment of CCC networks on spatially resolved transcriptomic datasets by incorporating cellular spatial constraints.

Conclusions

In this study, we propose DcjComm, a novel computational method designed to perform multiple scRNA-seq data analytical tasks. By incorporating sparse penalty and graph regularization, DcjComm presents the NMF-based joint learning model to reduce the impact of noise and preserve algorithmic coherence. Compared to state-of-the-art methods, the NMF-based joint learning model demonstrates outstanding performance in functional gene module detection, dimension reduction, and cell clustering. Then, by integrating ligand-receptor pairs, transcription factors, and their target genes, DcjComm achieves accurate inference of CCCs. The superior performance of DcjComm applies to several publicly available scRNA-seq datasets, demonstrating that DcjComm extracts more biologically relevant modules and representative features, improves cell clustering, and enhances CCC network inference performance. The ability of DcjComm to perform multiple tasks coherently facilitates a comprehensive understanding of potential biological processes.

Methods

Data collection and preprocessing

We collect 15 scRNA-seq datasets of mice and humans to verify the accuracy of the NMF-based joint learning model of DcjComm. Specifically, this contains nine mouse scRNA-seq datasets including Tabula [98] for the mammary gland, Klein [99] for the embryonic stem cells, Park [100] for the kidney, Joost [101] and Guerrero [102] for the skin cells, Kolodziejczyk [103] for the embryo stem cell, Zeisel [104], Usoskin [105], and Chen [106] for the mouse brain, as well as six human datasets including Baron [107] and Segerstolpe [108] for the pancreas, Tirosh [109] for the melanoma cells, Wang [110] for the testicular cell, and Camp [111] and Deng [112] for the skin cells. And another dataset (Mammary [113]) with batch effect was used to analyze the ability of DcjComm to remove batch effects. Among them, Joost, Guerrero, and Chen datasets are with sub-types while others are with primitive cell types. These benchmark datasets are collected from recently published papers about scRNA-seq experiments and their detailed information is summarized in Additional file 1: Table S12).

Since highly variable genes play an important role in assigning cell types and providing more biological information [14], we select highly variable genes to prioritize for downstream analysis. First, we filter the low-quality cells that are not informative for cell clustering. That is, these genes are expressed in less than α% of cells or at least $1-\alpha$% of cells ($\alpha = 6$ by default) [48]. Then, principal component analysis (PCA) is performed on the filtered genes. Finally, we select the first 2000 principal components as highly variable genes and further perform log2 normalization for the downstream analysis. For datasets containing multiple batches (e.g., Mammary), it is advisable to use the ComBat method from the SVA package [51] to remove batch effects before selecting highly variable genes. A detailed explanation of the theoretical mechanism of batch effects is provided in the Additional file 1: Supplementary Note S5.

The NMF-based joint learning model of DcjComm

It is valuable to identify the most representative features by projecting $\mathbf{X}$ into a low-dimensional space due to the extensive gene repertoire in scRNA sequencing. The typical two-factor dimension reduction method decomposes the input matrix $\mathbf{X}$ into the basis matrix $\mathbf{U}$ and the loading matrix $\mathbf{V}$ (that is, $\mathbf{X} \approx \mathbf{UV}$). The three-factor dimension reduction method (that is, $\mathbf{X} \approx \mathbf{USV}$) also plays an important role of matrix factorization technique [6], which provides an extra factor $\mathbf{S}$ to absorb the different scales of $\mathbf{X}$, $\mathbf{U}$, and $\mathbf{V}$. Thus, we approximate the difference between the original and selected features as

$$\mathbb{Q}_{\mathbf{U},\mathbf{V},\mathbf{S}} = \|\mathbf{X} - \mathbf{USV}\|_F^2, \tag{1}$$

where $\|\bullet\|_F^2$ represents the Frobenius norm. The input single-cell gene expression matrix $\mathbf{X} \in \mathbf{R}^{m\times n}$ includes m genes and n cells. Besides, the project matrix $\mathbf{U} \in \mathbf{R}^{m\times k_1}$ and the coefficient matrix $\mathbf{V} \in \mathbf{R}^{k_1\times n}$ respectively represent the features of each gene and cell in the low-dimensional space, factor matrix $\mathbf{S} \in \mathbf{R}^{k_1\times k_1}$ provides additional degrees of freedom to ensure the accuracy of the projected matrix. k_1 is the feature number after dimensionality reduction. Due to the sparsity inherent in single-cell data, previous studies have incorporated the $l_{2,1}$-norm regularization in the objective function to facilitate cell clustering [114]. This approach imposes sparsity on the rows of the coefficient matrix, enabling more effective and meaningful cell clustering. Considering that the l_1 -norm focuses on the row sparsity and the l_2-norm is dedicated to eliminating the influence of noise and outliers, to combine their advantages, the $l_{2,1}$-norm first calculates the l_2-norm for the row vector, and then calculates the l_1-norm for the column vector:

$$\|\mathbf{V}\|_{2,1} = \sum_{i=1}^{m} \sqrt{\sum_{j=1}^{n} x_{ij}^2} = \sum_{i=1}^{m} \|x_i\|_2 \tag{2}$$

That is, the introduction of $l_{2,1}$-norm not only improves the interpretation and accuracy of algorithm through sparse representation but also reduces the impact of noise and outliers (the detailed proof of robustness is presented in the Additional file 1: Supplementary Note S10, Fig. S21, Fig. S22). Therefore, the objective of dimension reduction in Eq. (1) is reformulated as follows:

$$\mathbb{Q}_{\mathbf{U},\mathbf{V},\mathbf{S}} = \|\mathbf{X} - \mathbf{USV}\|_F^2 + \alpha_1 \|\mathbf{V}\|_{2,1}, \tag{3}$$

where parameter α_1 controls the relevant importance of $l_{2,1}$ constraint.

At present, great evidence has indicated that joint learning improves the accuracy and flexibility of algorithms [15]. To overcome the problem of the independence of dimension reduction and clustering of cells, we jointly learn the feature selection by projected matrix decomposition and cell type clustering by NMF. We assign cells into clusters according to the maximum coefficient of matrix $\mathbf{F}$, which is obtained by decomposing the matrix $\mathbf{V}$ into the two nonnegative matrices $\mathbf{B}$ and $\mathbf{F}$:

$$\mathbb{R}_{\mathbf{B},\mathbf{F}} = \|\mathbf{V} - \mathbf{BF}\|_F^2 \; s.t. \; \mathbf{B} \geq 0, \; \mathbf{F} \geq 0. \tag{4}$$

where $\mathbf{B} \in \mathbf{R}_{k_1 \times k_2}$ and $\mathbf{F} \in \mathbf{R}_{k_2 \times n}$ are respectively represented as the basis matrix and the feature matrix. k_2 is the number of clusters. Since $\mathbf{F}$ is expected to maintain the intrinsic geometrical structure of $\mathbf{V}$, while the NMF method cannot detect the inherent geometric structure of high-dimensional data such as manifolds, we introduce the graph regularization into the objective function of NMF:

$$\mathbb{R}_{\mathbf{B},\mathbf{F}} = \|\mathbf{V} - \mathbf{BF}\|_F^2 + \alpha_2 \mathrm{Tr}(\mathbf{FLF}^{\mathrm{T}}) s.t.\, \mathbf{B} \geq 0,\ \mathbf{F} \geq 0. \tag{5}$$

where the parameter α_2 balances the importance of graph regularization. The basic idea of graph regularization is to reconstruct the low-dimensional manifold structure embedded in high-dimensional ambient space. That is, if two cells v_i and v_j are close in the high-dimensional sample space, the corresponding representations in low-dimensional space, i.e., f_i and f_j, should be as close as possible, and vice versa. We construct a graph G to depict the closeness of cells in $\mathbf{V}$, each sampling point is used as a vertex and the similarity between a pair of cells denotes the edges. To quantify the edges, we define a symmetric weight matrix $\mathbf{W}$, whose element $\mathbf{W}_{i,j}$ represents the weights of the edges connecting cell i and cell j:

$$\mathbf{W}_{i,j} = \begin{cases} 1 \text{ if } \mathrm{v_i} \in \mathbf{N_K}(\mathrm{v_j}) \text{ or } \mathrm{v_j} \in \mathbf{N_K}(\mathrm{v_i}) \\ 0 \text{ otherwise} \end{cases}. \tag{6}$$

where $\mathrm{v_i}$ is the $i-$ th cell, $\mathbf{N_k}(\mathrm{v_i})$ is the set of K nearest neighbors of $\mathrm{v_i}$. Thus, the graph regularization can be formulated as follows:

$$\begin{aligned} &\min_{\mathbf{F}} \sum_{i,j} \left\|f_i - f_j\right\|^2 \mathbf{W}_{i,j} \\ &= \min_{\mathbf{F}} Tr(\mathbf{F}(\mathbf{D} - \mathbf{W})\mathbf{F}^T), \\ &= \min_{\mathbf{F}} Tr(\mathbf{FLF}^T) \end{aligned} \tag{7}$$

where $\mathbf{D}$ is a diagonal matrix in which elements are obtained by the sum of the rows or columns of $\mathbf{W}$. f_i and f_j are the low-dimensional representation of v_i and v_j. $\mathbf{L}$ is the Laplacian matrix of graph G, i.e., $\mathbf{L} = \mathbf{D} - \mathbf{W}$.

Then, by combining Eq. (3) and Eq. (5), we propose the NMF-based joint learning model for simultaneously executing gene module selection, dimension reduction, and cell clustering. By formulating these three tasks as a constrained optimization problem, the DcjComm method effectively improves solution efficiency. DcjComm not only extracts essential feature information from the data but also significantly enhances the performance of cell clustering. The dimension reduction process generates features under the guidance of individual cell clustering, with the clustering of individual cells selecting appropriate features. The objective function of the NMF-based joint learning model is defined as the following optimization problem:

$$\begin{aligned} &\min \mathbb{N} = \mathbb{Q}_{\mathbf{U},\mathbf{S},\mathbf{V}} + \mathbb{R}_{\mathbf{B},\mathbf{F}} \\ &= \|\mathbf{X} - \mathbf{USV}\|_F^2 + \alpha_1 \|\mathbf{V}\|_{2,1} + \|\mathbf{V} - \mathbf{BF}\|_F^2 + \alpha_2 \mathrm{Tr}(\mathbf{FLF}^{\mathrm{T}}) \\ &\qquad s.t.\, \mathbf{B} \geq 0,\ \mathbf{F} \geq 0. \end{aligned} \tag{8}$$

The factor matrix **U** obtained from Eq. (8) guides us to construct gene functional modules, where features with relatively large values in each column are selected as members of the module. Specifically, we select the nodes with relatively large absolute values of the weighted factors **U** by calculating its z-score for each column vector $u_i (i = 1, ..., k)$ [6]:

$$z_{ij} = \frac{(\mathbf{U})_{ij} - \mu_{(\mathbf{U})_{\cdot j}}}{\sigma_{(\mathbf{U})_{\cdot j}}}. \tag{9}$$

where $\mu_{(\mathbf{U})_{\cdot j}} = \frac{1}{N} \sum (\mathbf{U})_{ij}$ and $\sigma^2_{(\mathbf{U})_{\cdot j}} = \frac{1}{N-1} \sum ((\mathbf{U})_{ij} - \mu_{(\mathbf{U})_{\cdot j}})^2$. Based on the above transformation, we assign $g(i)$ that satisfies the condition $z_{ij} > \theta$ as the $i-$ th module member. θ is a given threshold that enables the selected module members to have a significant signal. The element s_{ij} of the factor matrix **S** can be considered as the weight of $u_i v_j^T$ in the reconstruction of **X**. The larger the value of s_{ij}, the larger the elements of **X** for all combinations of selected features based on u_i and v_j. Since matrix **S** is a diagonal matrix, we use the value of **S** to represent the importance of the selected module.

Next, we assign cells to clusters based on the maximum coefficient of their matrix **F**, which is consistent with previous studies [15]. Furthermore, marker genes expressed in specific cell types play an important role in the identification of cell types. Therefore, we annotate cell clusters to known cell types based on the expression levels of the marker genes given in the original studies. In cases where clusters lack known marker expression, we assign the cell type containing the most cells to that cluster.

In addition, the NMF-based joint learning model contains five parameters k_1, k_2, α_1, α_2 and θ, where k_1 is the number of features for dimension reduction, k_2 is the number of clusters, α_1 and α_2 are regularization parameters, and θ is the threshold to select module members (Additional file 1: Supplementary Note S1 and S3).

Optimization algorithm

The iterative strategy is used to solve the non-convex problem of Eq. (8), it involves multiple variables, i.e., **U**, **S**, **V**, **B**, and **F**, we optimize one variable by fixing the others until the termination criterion is reached. The Lagrange function of Eq. (8) is formulated as follows:

$$\begin{aligned} L(\mathbf{U}, \mathbf{V}, \mathbf{S}, \mathbf{B}, \mathbf{F}) = \|\mathbf{X} - \mathbf{USV}\|_F^2 + \|\mathbf{V} - \mathbf{BF}\|_F^2 \\ + \alpha_1 \|\mathbf{V}\|_{2,1} + \alpha_2 \mathrm{Tr}(\mathbf{FLF}^{\mathrm{T}}) + \mathrm{Tr}(\Psi \mathbf{B}^T) + \mathrm{Tr}(\Phi \mathbf{F}^T). \end{aligned} \tag{10}$$

Let ψ_{ik} and ϕ_{ik} be Lagrange multipliers to constrain $b_{ik} \geq 0$ and $f_{jk} \geq 0$, where $\Psi = [\psi_{ik}]$ and $\Phi = [\phi_{ik}]$.

Firstly, we optimize the variable **U** by fixing **S** and **V**, the optimal value of **U** is given by

$$\mathbf{U} \leftarrow \arg\min_{\mathbf{U}} \|\mathbf{X} - \mathbf{USV}\|_F^2. \tag{11}$$

Taking the partial derivative of function L with respect to matrix **U**, when the derivative is 0, we obtain the optimal solution for **U**:

$$\mathbf{U} \leftarrow \mathbf{U}\frac{\mathbf{X}\mathbf{V}^T\mathbf{S}}{\mathbf{U}\mathbf{S}\mathbf{V}\mathbf{V}^T\mathbf{S}}. \tag{12}$$

Next, we optimize the variable $\mathbf{S}$ by fixing $\mathbf{U}$ and $\mathbf{V}$:

$$\mathbf{S} \leftarrow \arg\min_{\mathbf{S}} \|\mathbf{X} - \mathbf{U}\mathbf{S}\mathbf{V}\|_F^2. \tag{13}$$

We also taking the partial derivative of function L with respect to matrix $\mathbf{S}$, when the derivative is 0, we get the optimal solution for $\mathbf{S}$:

$$\mathbf{S} \leftarrow \mathbf{S}\frac{\mathbf{U}^T\mathbf{X}\mathbf{V}^T}{\mathbf{V}\mathbf{V}^T\mathbf{U}^T\mathbf{U}\mathbf{S}}. \tag{14}$$

Then, for optimizing variable $\mathbf{V}$, let

$$J(\mathbf{V}) = \|\mathbf{X} - \mathbf{U}\mathbf{S}\mathbf{V}\|_F^2 + \|\mathbf{V} - \mathbf{B}\mathbf{F}\|_F^2 + \alpha_1\|\mathbf{V}\|_{2,1}. \tag{15}$$

Since $J(\mathbf{V})$ includes the conductive portion $\|\mathbf{X} - \mathbf{U}\mathbf{S}\mathbf{V}\|_F^2 + \|\mathbf{V} - \mathbf{B}\mathbf{F}\|_F^2$ and the non-conductive portion $\alpha_1\|\mathbf{V}\|_{2,1}$, we redefine the above problem as

$$J(\mathbf{V}) = \|\mathbf{X} - \mathbf{U}\mathbf{S}\mathbf{V}\|_F^2 + \|\mathbf{V} - \mathbf{B}\mathbf{F}\|_F^2 + \alpha_1\|\mathbf{E}\|_{2,1}\ s.t.\ \mathbf{E} - \mathbf{V} = 0. \tag{16}$$

The augmented Lagrange function of Eq. (16) can be formulated as:

$$\begin{aligned} &L(\mathbf{V}, \mathbf{E}; \mathbf{T}) = \\ &\|\mathbf{X} - \mathbf{U}\mathbf{S}\mathbf{V}\|_F^2 + \|\mathbf{V} - \mathbf{B}\mathbf{F}\|_F^2 + \alpha_1\|\mathbf{E}\|_{2,1} + \sigma\|\mathbf{E} - \mathbf{V}\|_F^2 + \langle \mathbf{T}, \mathbf{E} - \mathbf{V}\rangle \\ &= \|\mathbf{X} - \mathbf{U}\mathbf{S}\mathbf{V}\|_F^2 + \|\mathbf{V} - \mathbf{B}\mathbf{F}\|_F^2 + \alpha_1\|\mathbf{E}\|_{2,1} + \sigma\left\|\mathbf{E} - \mathbf{V} + \frac{\mathbf{T}}{\sigma}\right\|_F^2. \end{aligned} \tag{17}$$

where $\sigma > 0$ is the penalty parameter and $\mathbf{T} \in \mathbf{R}^{k_1 \times m}$ is the Lagrange multiplier. Then, we optimize the following sub-problems of Eq. (17) by performing ADMM:

$$\begin{cases} \mathbf{V} \leftarrow \arg\min_{\mathbf{V}} \|\mathbf{X} - \mathbf{U}\mathbf{S}\mathbf{V}\|_F^2 + \|\mathbf{V} - \mathbf{B}\mathbf{F}\|_F^2 + \sigma\left\|\mathbf{E} - \mathbf{V} + \frac{\mathbf{T}}{\sigma}\right\|_F^2 \\ \mathbf{E} \leftarrow \arg\min_{\mathbf{E}} \sigma\left\|\mathbf{E} - \mathbf{V} + \frac{\mathbf{T}}{\sigma}\right\|_F^2 + \alpha_1\|\mathbf{E}\|_{2,1} \\ \mathbf{T} \leftarrow \mathbf{T} + \sigma\mathbf{E} - \mathbf{V} \end{cases}. \tag{18}$$

Specifically, optimizing the subproblems of $\mathbf{V}$ can be approximated through the following iterative process:

$$\mathbf{V} \leftarrow \mathbf{V}\frac{\mathbf{S}\mathbf{U}^T\mathbf{X} + \mathbf{B}\mathbf{F} + \sigma\mathbf{E} + \mathbf{T}}{\mathbf{U}^T\mathbf{U}\mathbf{S}\mathbf{S}\mathbf{V} + (\mathbf{I} + \sigma)\mathbf{V}}. \tag{19}$$

The sub-problem of optimizing $\mathbf{E}$ is given by

$$\mathbf{E} \leftarrow S_{\alpha_1/\sigma}(\mathbf{E} - \frac{\mathbf{T}}{\sigma}) + 2\alpha_1\mathbf{D}_1\mathbf{E}, \tag{20}$$

and S is a soft threshold function, which is defined as follows:

$$S_\varepsilon[x] = \begin{cases} x + \varepsilon \text{ if } x > \varepsilon \\ x - \varepsilon \text{ if } x < -\varepsilon \\ 0 \text{ otherwise} \end{cases}. \tag{21}$$

Furthermore, the process of optimizing the variable **B** is given below

$$\mathbf{B} \leftarrow \arg\min_{\mathbf{B}} \|\mathbf{V} - \mathbf{BF}\|_F^2 + \mathrm{Tr}(\Psi \mathbf{B}^T). \tag{22}$$

We calculate the partial derivative of function L with respect to matrix **B**. Incorporating the known KKT conditions [115], $\Psi \mathbf{B}^T = 0$, when the derivative of function L equals 0, we obtain the optimal solution for matrix for **B**:

$$\mathbf{B} \leftarrow \mathbf{B} \frac{\mathbf{VF}^T}{\mathbf{BFF}^T}. \tag{23}$$

In addition, Eq. (24) describes the optimization process for the variable **F**. By incorporating the known conditions $\mathbf{L} = \mathbf{D} - \mathbf{E}$ and $\Phi \mathbf{F}^T = 0$, and setting the partial derivative of function L with respect to the matrix **F** to 0, we derive the optimal solution for the matrix as given in Eq. (25).

$$\mathbf{F} \leftarrow \arg\min_{\mathbf{F}} \|\mathbf{V} - \mathbf{BF}\|_F^2 + \alpha_2 \mathrm{Tr}(\mathbf{FLF}^T) + \mathrm{Tr}(\Phi \mathbf{F}^T). \tag{24}$$

$$\mathbf{F} \leftarrow \mathbf{F} \frac{\mathbf{B}^T\mathbf{V} + \alpha_2 \mathbf{FW}}{\mathbf{B}^T\mathbf{BF} + \alpha_2 \mathbf{FD}}. \tag{25}$$

Finally, based on the matrix **F** obtained from the iterative optimization process, cells are automatically assigned to different cell clusters based on the largest coefficients in this matrix.

The CCCs inference statistical model of DcjComm

In this process, ligands function as transmitter cells, while receptors act as receiver cells. The binding of ligands alters the conformation of the receptors, subsequently influencing the expression levels of downstream transcription factors and target genes. Therefore, an extensive and reliable database is essential for inferring cellular interactions.

Here, we have developed the DcjComm-DB database by collecting and integrating multiple complementary data sources to serve as evidence for inferring CCCs. Specifically, we have compiled and integrated information from three layers: L-R, TF-TG, and R-TF pairs. The DcjComm-DB is available for both human and mouse. Firstly, we collect human L-R interactions derived from literature data: Jin [21], Shao [116], Cabello[117], Hou [86], Ramilowski [118], Zhang [119], Gao [120], and mouse L-R interactions are collected from the following literature data: Shao [116], Baccin [121], Jin [21], Zhang [119], Cain [122], Ding [123], Hu [81], Sheikh [124], Skelly [125], and Yuzwa [126]. Secondly, we use the R-TF database constructed by Baruzzo et al. [25], which includes 1533 receptors associated with 411 transcription factors for humans, and 731 receptors associated with 369 transcription factors for mice. In addition to receptors and their corresponding TF, the R-TF database also includes scores of the degree of association between a given receptor and a given TF in

a specific pathway, which are obtained using the PageRank algorithm based on KEGG or Reactome pathway databases. Finally, to provide an accurate and comprehensive repository of human and mouse TF-TG interactions, we include a merged version of human HTRIdb, TRRUST v2, RegNetwork "medium" confidence, and RegNetwork "high" confidence, resulting in 1537 transcription factors and 18,421 regulated gene, and a merged version of mouse TRRUST v2, RegNetwork "medium" confidence, and RegNetwork "high" confidence, resulting in 1649 transcription factors and 14,570 regulated genes [25].

To quantify the CCCs between different cell types, we have proposed the CCCs inference statistical model. In this model, we define the S_k as the product of the intercellular communication score (S_{inter}) and the ongoing intracellular signaling (S_{intra}):

$$S_k = \mathrm{S}_{\text{inter}} \times \mathrm{S}_{\text{intra}}. \tag{26}$$

The intercellular signaling score S_{inter} is evaluated by the l_2- norm of the L-R interaction LR_k, that is, $\mathrm{S}_{inter} = \left\| \overrightarrow{\mathrm{LR}_k} \right\|_2$. LR_k is a two-dimensional vector represented by the normalized expression value of the ligand and receptor for L-R interaction k:

$$\mathrm{LR}_k = (\text{soft}\max(L_{i,k}), \text{soft}\max(R_{j,k})), \tag{27}$$

where $\text{soft}\max(\bullet)$ is the softmax function to obtain the normalized expression values. $L_{i,k}$ and $R_{j,k}$ are respectively the mean expression values of ligand and receptor in cell type i and cell type j.

In addition, if the ligand L contains n subunits and l_g represents the expression value of subunit g, then we define L as the geometric mean of the expression value of all subunits:

$$L = \sqrt[n]{\prod_{g=1}^{n} \mathrm{l}_g}. \tag{28}$$

Similarly, if the receptor R contains n subunits and r_h represents the expression value of subunit h, R is defined as the geometric mean of the expression value of all subunits:

$$R = \sqrt[n]{\prod_{h=1}^{n} r_h}. \tag{29}$$

The intracellular signaling score S_{intra} is evaluated according to the R-TF scores and the interaction of transcription factors and target genes. To evaluate the interaction of receptor and TF, we introduce the $\mathrm{TF}_{\mathrm{PPR}}(t, r, p)$ score to measure the association of the given receptor r and the given TF t in the pathway p [25]. Besides, to provide the transcriptomic evidence of cell communication effects, we score the activity of TFs within the pathway measuring changes in expression levels of their regulated genes. Then, the activity of each TF t in the cluster k is computed through a Fisher test as follows:

$$P_{\text{Fisher}} = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}}. \tag{30}$$

where $a = |T_{\mathrm{DEG}} \cap L_{\mathrm{TF}}|$, $b = |T_{\mathrm{DEG}}| - a$, $c = |L_{\mathrm{TF}}| - a$, $d = |L_{\mathrm{all}}| - (a + b + c)$. T_{DEG} is the gene set of differentially expressed genes, L_{TF} is the target genes list corresponding to each TF, $|L_{\mathrm{all}}|$ is a list of all annotated genes. In addition, the Mann–Whitney U test is chosen for the identification of differentially expressed genes due to its excellent performance and low computational burden in identifying differentially expressed genes of single-cell data.

Then, the activity of each TF t in the cluster k is defined as follows:

$$\mathrm{TF}_{\mathrm{activity}}(t,k) = 1 - P_{\mathrm{Fisher}}(t,k). \tag{31}$$

Finally, combining the R-TF score and the TF-TG score, we obtain the intracellular signaling score:

$$S_{\mathrm{intra}}(r,p,k) = \frac{\sum_{t \in T(r,p)} \mathrm{TF}_{\mathrm{PPR}}(t,r,p)\mathrm{TF}_{\mathrm{activity}}(t,k)}{\sum_{t \in T(r,p)} \mathrm{TF}_{\mathrm{PPR}}(t,r,p)}. \tag{32}$$

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03385-6.

Additional file 1: Supplementary Notes S1-10 and Figures S1-S22. Fig. S1. The selection of the corresponding rank according to the local minimum value of the stable distance. Fig. S2. Prediction of the number of clusters. Fig. S3. Comparison of convergence and MSE values about fifteen datasets. Fig. S4. Performance in terms of with different of Deng and Tabula datasets. Fig. S5. Performance of DcjComm on twelve datasets, the visualization of cells in the projected low-dimensional space before dimension reduction. Fig. S6. Performance of DcjComm on twelve datasets, the visualization of cells in the projected low-dimensional space after dimension reduction. Fig. S7. The performance comparison of ARI of different methods on fifteen scRNA-seq datasets. Fig. S8. The performance comparison of NMI of different methods on fifteen scRNA-seq datasets. Fig. S9. Visualization of the cells in fifteen datasets based on the clustering results of the DcjComm method. Fig. S10. Violin plots of marker gene expression distributions within each cluster of the Deng dataset. Fig. S11. Violin plots of marker gene expression distributions within each cluster of the Tabula dataset. Fig. S12. The top 10 significantly differentially expressed genes detected across all clusters in the Deng and Tabula datasets. Fig. S13. The differential expression of marker genes in the Deng and Tabula datasets. Fig. S14. DcjComm demonstrates superior performance in detecting CCCs before cell clustering, according to the DcjComm-DB. Fig. S15. CCCs analysis of spatial transcriptomics data dataset. Fig. S16. The robustness of comparing inferred communications from the subsampled dataset and the original Deng dataset. Fig. S17. DcjComm demonstrates superior performance in detecting CCCs after cell clustering, according to the DcjComm-DB. Fig. S18. The robustness of comparing inferred communications from the subsampled dataset and the original Tabula dataset. Fig. S19. DcjComm detects spatially CCCs for human breast cancer from BreastST data. Fig. S20. The superior performance of DcjComm to detect spatially CCCs for human breast cancer from BreastST data. Fig. S21. Illustration of different estimators and corresponding influence functions. Fig. S22. Recovery capability comparison of DcjComm, DRjCC and NMF on synthetic datasets.

Additional file 2: Supplementary Tables S1-11. Table S1. The comparison of running time (s) for cell clustering on fifteen single-cell datasets. Table S2. The comparison of memory consumption (MB) for cell clustering on fifteen single-cell datasets. Table S3. The comparison of running time (s) for inferring cell-cell communications on fifteen single-cell datasets. Table S4. The top 10 pathways enriched by the core modules of the Deng dataset. Table S5. The top 10 pathways enriched by the core modules of the Tabula dataset. Table S6. Prediction of the number of clusters by DcjComm and other compared methods. Table S7: The performance comparison of cell clustering of different methods on the multi-batch dataset. Table S8. The top 10 pathways enriched by the differential expression genes of the Deng dataset. Table S9. The top 10 pathways enriched by the differential expression genes of the Tabula dataset. Table S10. Comparison of CCCs inference performance between the raw database and DcjComm-DB before cell clustering. Table S11. Comparison of CCCs inference performance between the raw database and DcjComm-DB after cell clustering. Table S12. Details of scRNA-seq datasets used in experiments.

Additional file 3: Review history.

Acknowledgements
Not applicable

Peer review information
Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history
The review history is available as Additional file 3.

Authors' contributions
Q.J., Z.X., and P.W. conceived and designed the study. Q.D., W.Y., G.X., Y.C., and H.L. performed the research. Q.D., F.P., Y. Y., Y.L., and X.J collected the benchmark datasets and prepared the figures. Q.D. and W.Y. constructed the models. Q.D., J.Q., Y.L., P.W., M.L., and H.S. completed the downstream analysis work. Q.D. and Q.J. wrote the paper. All authors edited and approved the final manuscript.

Funding
This work was supported by the National Natural Science Foundation of China (nos. T2325009, 62032007, and 32270789), National Science and Technology Major Project of China (no. 2022ZD0117702), and the Science, Technology & Innovation Project of Xiongan New Area in China (No.2022XAGG0117).

Availability of data and materials
The R package of DcjComm is freely available at GitHub (https://github.com/Ginnay/DcjComm) [127] and Zenodo (https://zenodo.org/records/12666949) [128]. The source code is released under GPL-3.0 license. Additionally, we collected fifteen real and public scRNA-seq datasets across multiple platforms. These datasets come from different organs in humans and mouse, such as the brain, pancreas, skin, and others. These datasets include Tabula (GSE109774) [98], Klein (GSE65525) [99], Park (GSE107585) [100], Joost (GSE67602) [101], Guerrero [102], Kolodziejczyk [103], Zeisel (GSE60361) [104], Usoskin (GSE59739) [105] and Chen (GSE87544) [106], Baron (GSE84133) [107], Segerstolpe (E-MTAB-5016) [108], Tirosh (GSE72056) [109], Wang (GSE106487) [110], Camp (GSE81252) [111], and Deng (GSE163973) [112]. The Mammary dataset is downloaded from https://doi.org/10.6084/m9.figshare.20499630.v2 [113]. The details of these scRNA-seq data sets including the number of cells, number of cell types, and organs are tabulated in Supplementary Table S1 and can be downloaded from https://hemberg-lab.github.io/scRNA.seq.datasets/ and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi. The STARmap dataset is downloaded from http://sdmbench.drai.cn/ [61]. The MERFISH dataset is available at https://datadryad.org/stash/dataset/10.5061/dryad.8t8s248 [62]. The BreastST dataset is downloaded from https://www.10xgenomics.com/resources/datasets [97]. All other relevant data supporting the key findings of this study are available within the article and its Supplementary Information files or from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 9 January 2024 Accepted: 30 August 2024
Published online: 09 September 2024

References

1. Armingol E, Officer A, Harismendy O, Lewis NE. Deciphering cell-cell interactions and communication from gene expression. Nat Rev Genet. 2021;22:71–88.
2. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
3. Ruan J, Zhang W: Identifying network communities with a high resolution. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* 2008, 77**:**016104.
4. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics. 2003;4:2.
5. Hwang W, Cho Y-R, Zhang A, Ramanathan M. A novel functional module detection algorithm for protein-protein interaction networks. Algorithms for Molecular Biology. 2006;1:24.
6. Chen J, Zhang S. Discovery of two-level modular organization from matched genomic data via joint matrix tri-factorization. Nucleic Acids Res. 2018;46:5967–76.
7. Steinley D. K-means clustering: a half-century synthesis. Br J Math Stat Psychol. 2006;59:1–34.
8. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, Ginhoux F, Newell EW. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2018;37:38–44.
9. Jiang H, Sohn LL, Huang H, Chen L. Single cell clustering based on cell-pair differentiability correlation and variance analysis. Bioinformatics. 2018;34:3684–94.
10. Lin P, Troup M, Ho JW. CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data. Genome Biol. 2017;18:59.
11. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Reik W, Barahona M, Green AR, Hemberg M. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods. 2017;14:483–6.
12. Wang B, Zhu J, Pierson E, Ramazzotti D, Batzoglou S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nat Methods. 2017;14:414–6.
13. Bro R, Smilde AK. Principal component analysis. Analytical methods. 2014;6:2812–31.

14. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive integration of single-cell data. Cell. 2019;177:1888–902.
15. Wu W, Ma X. Joint learning dimension reduction and clustering of single-cell RNA-sequencing data. Bioinformatics. 2020;36:3825–32.
16. Hu H, Li Z, Li X, Yu M, Pan X: ScCAEs: deep clustering of single-cell RNA-seq via convolutional autoencoder embedding and soft K-means. *Briefings in Bioinformatics* 2022, 23:bbab321.
17. Wang J, Ma A, Chang Y, Gong J, Jiang Y, Qi R, Wang C, Fu H, Ma Q, Xu D. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. Nat Commun. 1882;2021:12.
18. Qiu Y, Yan C, Zhao P, Zou Q. SSNMDI: a novel joint learning model of semi-supervised non-negative matrix factorization and data imputation for clustering of single-cell RNA-seq data. Brief Bioinform. 2023;24:149.
19. Rajapakse M, Tan J, Rajapakse J: Color channel encoding with NMF for face recognition. In *2004 International Conference on Image Processing, 2004 ICIP'04*. IEEE; 2004: 2007–2010.
20. Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. Cell PhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. Nat Protoc. 2020;15:1484–506.
21. Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan C-H, Myung P, Plikus MV, Nie Q. Inference and analysis of cell-cell communication using Cell Chat. Nat Commun. 2021;12:1088.
22. Cillo AR, Kürten CH, Tabib T, Qi Z, Onkar S, Wang T, Liu A, Duvvuri U, Kim S, Soose RJ. Immune landscape of viral- and carcinogen-driven head and neck cancer. Immunity. 2020;52:183–99.
23. Browaeys R, Saelens W, Saeys Y. NicheNet: modeling intercellular communication by linking ligands to target genes. Nat Methods. 2020;17:159–62.
24. Zhang Y, Liu T, Hu X, Wang M, Wang J, Zou B, Tan P, Cui T, Dou Y, Ning L, et al. Cell Call: integrating paired ligand–receptor and transcription factor activities for cell–cell communication. Nucleic Acids Res. 2021;49:8520–34.
25. Baruzzo G, Cesaro G, Di Camillo B. Identify, quantify and characterize cellular communication from single-cell RNA sequencing data with scSeqComm. Bioinformatics. 2022;38:1920–9.
26. Cheng J, Zhang J, Wu Z, Sun X. Inferring microenvironmental regulation of gene expression from single-cell RNA sequencing data using scMLnet with an application to COVID-19. Brief Bioinform. 2021;22:988–1005.
27. Yuan Z, Li Y, Shi M, Yang F, Gao J, Yao J, Zhang MQ. SOTIP is a versatile method for microenvironment modeling with spatial omics data. Nat Commun. 2022;13:7330.
28. Yuan Z, Pan W, Zhao X, Zhao F, Xu Z, Li X, Zhao Y, Zhang MQ, Yao J. SODB facilitates comprehensive exploration of spatial omics data. Nat Methods. 2023;20:387–99.
29. Yuan Z. MENDER: fast and scalable tissue structure identification in spatial omics data. Nat Commun. 2024;15:207.
30. Yuan Z, Zhao F, Lin S, Zhao Y, Yao J, Cui Y, Zhang X-Y, Zhao Y. Benchmarking spatial clustering methods with spatially resolved transcriptomics data. Nat Methods. 2024;21:712–22.
31. Cang Z, Zhao Y, Almet AA, Stabell A, Ramos R, Plikus MV, Atwood SX, Nie Q. Screening cell–cell communication in spatial transcriptomics via collective optimal transport. Nat Methods. 2023;20:218–28.
32. Wilk AJ, Shalek AK, Holmes S, Blish CA. Comparative analysis of cell–cell communication at single-cell resolution. Nat Biotechnol. 2024;42:470–83.
33. Dries R, Zhu Q, Dong R, Eng C-HL, Li H, Liu K, Fu Y, Zhao T, Sarkar A, Bao F: Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biology* 2021, 22.
34. Raredon MSB, Yang J, Kothapalli N, Lewis W, Kaminski N, Niklason LE, Kluger Y: Comprehensive visualization of cell–cell interactions in single-cell and spatial transcriptomics with NICHES. *Bioinformatics* 2023, 39:btac775.
35. Shao X, Li C, Yang H, Lu X, Liao J, Qian J, Wang K, Cheng J, Yang P, Chen H, et al. Knowledge-graph-based cell-cell communication inference for spatially resolved transcriptomic data with SpaTalk. Nat Commun. 2022;13:4429.
36. Liu H, Wu Z, Li X, Cai D, Huang TS, Intelligence M. Constrained nonnegative matrix factorization for image representation. IEEE Transactions on Pattern Analysis. 2011;34:1299–311.
37. Suda R, Kuriyama S. Another preprocessing algorithm for generalized one-dimensional fast multipole method. J Comput Phys. 2004;195:790–803.
38. Liu Z, Wang J, Liu G, Zhang L. Discriminative low-rank preserving projection for dimensionality reduction. Appl Soft Comput. 2019;85: 105768.
39. Ma J, Lü X, Huang Y. Genomic analysis of cytotoxicity response to nanosilver in human dermal fibroblasts. J Biomed Nanotechnol. 2011;7:263–75.
40. Kreindl C, Soto-Alarcón SA, Hidalgo M, Riveros AL, Añazco C, Pulgar R, Porras O. Selenium compounds affect differently the cytoplasmic thiol/disulfide state in dermic fibroblasts and improve cell migration by interacting with the extracellular matrix. Antioxidants. 2024;13:159.
41. Zhong B-L, Bian L-J, Wang G-M, Zhou Y-F, Chen Y-Y, Peng F. Identification of key genes involved in HER2-positive breast cancer. European Review for Medical Pharmacological Sciences. 2016;20:664–72.
42. Cheng AA, Li W, Hernandez LL. Investigating the effect of positional variation on mid-lactation mammary gland transcriptomics in mice fed either a low-fat or high-fat diet. PLoS ONE. 2021;16: e0255770.
43. Zheng R, Li M, Liang Z, Wu F-X, Pan Y, Wang J. SinNLRR: a robust subspace clustering method for cell type detection by non-negative and low-rank representation. Bioinformatics. 2019;35:3642–50.
44. Yao Z, Jaeger JC, Ruzzo WL, Morale CZ, Emond M, Francke U, Milewicz DM, Schwartz SM, Mulvihill ER. A Marfan syndrome gene expression phenotype in cultured skin fibroblasts. BMC Genomics. 2007;8:1–13.
45. Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, Habuka M, Tahmasebpoor S, Danielsson A, Edlund KJM. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Molecular Cellular Proteomics. 2014;13:397–406.
46. Wang N, Kudryavtseva E, Ch'en IL, McCormick J, Sugihara TM, Ruiz R, Andersen B. Expression of an engrailed-LMO4 fusion protein in mammary epithelial cells inhibits mammary gland development in mice. Oncogene. 2004;23:1507–13.
47. Zhang Z, Christin JR, Wang C, Ge K, Oktay MH, Guo W. Mammary-stem-cell-based somatic mouse models reveal breast cancer drivers causing cell fate dysregulation. Cell Rep. 2016;16:3146–56.

48. Wang S, Karikomi M, MacLean AL, Nie Q. Cell lineage and communication network inference via optimization for single-cell transcriptomics. Nucleic Acids Res. 2019;47: e66.
49. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M. Loh P-r, Raychaudhuri S: Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16:1289–96.
50. Hu Z, Ahmed AA, Yau C. CIDER: an interpretable meta-clustering framework for single-cell RNA-seq data integration and evaluation. Genome Biol. 2021;22:337.
51. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118–27.
52. Smyth GK, Speed TJM. Normalization of cDNA microarray data. Methods. 2003;31:265–73.
53. Kulebyakina M, Basalova N, Butuzova D, Arbatsky M, Chechekhin V, Kalinina N, Tyurin-Kuzmin P, Kulebyakin K, Klychnikov O, Efimenko A. Balance between pro-and antifibrotic proteins in mesenchymal stromal cell secretome fractions revealed by proteome and cell subpopulation analysis. Int J Mol Sci. 2023;25:290.
54. Pasanen I, Lehtonen S, Sormunen R, Skarp S, Lehtilahti E, Pietilä M, Sequeiros RB, Lehenkari P, Kuvaja P. Breast cancer carcinoma-associated fibroblasts differ from breast fibroblasts in immunological and extracellular matrix regulating pathways. Exp Cell Res. 2016;344:53–66.
55. Wang W-Z, Cao X, Bian L, Gao Y, Yu M, Li Y-T, Xu J-G, Wang Y-H, Yang H-F, You D-Y. Analysis of mRNA-miRNA interaction network reveals the role of CAFs-derived exosomes in the immune regulation of oral squamous cell carcinoma. BMC Cancer. 2023;23:591.
56. Farhangniya M, Farsani FM, Salehi N, Samadikuchaksaraei A. Integrated bioinformatic analysis of differentially expressed genes associated with wound healing. Cell J. 2023;25:874.
57. Chen W, Gu X, Lv X, Cao X, Yuan Z, Wang S, Sun W. Non-coding transcriptomic profiles in the sheep mammary gland during different lactation periods. Frontiers in veterinary science. 2022;9: 983562.
58. Verma AK, Ali SA, Singh P, Kumar S, Mohanty AK. Transcriptional repression of MFG-E8 causes disturbance in the homeostasis of cell cycle through DOCK/ZP4/STAT signaling in buffalo mammary epithelial cells. Frontiers in Cell Developmental Biology. 2021;9: 568660.
59. Bhat SA, Ahmad SM, Ibeagha-Awemu EM, Bhat BA, Dar MA, Mumtaz PT, Shah RA, Ganai NA. Comparative transcriptome analysis of mammary epithelial cells at different stages of lactation reveals wide differences in gene expression and pathways regulating milk synthesis between Jersey and Kashmiri cattle. PLoS ONE. 2019;14: e0211773.
60. Conte G, Giordani T, Vangelisti A, Serra A, Pauselli M, Cavallini A, Mele M. Transcriptome adaptation of the ovine mammary gland to dietary supplementation of extruded linseed. Animals. 2021;11:2707.
61. Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, Evans K, Liu C, Ramakrishnan C, Liu J: Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *STARmap Resources*, http://sdmbench.drai.cn/ (2018).
62. Moffitt JR, Bambah-Mukku D, Eichhorn SW, Vaughn E, Shekhar K, Perez JD, Rubinstein ND, Hao J, Regev A, Dulac C: Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *DRYAD*, https://datadryad.org/stash/dataset/doi:10.5061/dryad.8t8s248 (2018).
63. Deng CC, Hu YF, Zhu DH, Cheng Q, Gu JJ, Feng QL, Zhang LX, Xu YP, Wang D, Rong Z, Yang B. Single-cell RNA-seq reveals fibroblast heterogeneity and increased mesenchymal fibroblasts in human fibrotic skin diseases. Nat Commun. 2021;12:3709.
64. Bock O, Yu H, Zitron S, Bayat A, Ferguson MW, Mrowietz U. Studies of transforming growth factors beta 1–3 and their receptors I and II in fibroblast of keloids and hypertrophic scars. Acta Derm Venereol. 2005;85:216–20.
65. Pavlides S, Whitaker-Menezes D, Castello-Cros R, Flomenberg N, Witkiewicz AK, Frank PG, Casimiro MC, Wang C, Fortina P, Addya S, et al. The reverse Warburg effect: aerobic glycolysis in cancer associated fibroblasts and the tumor stroma. Cell Cycle. 2009;8:3984–4001.
66. van Caam A, Aarts J, van Ee T, Vitters E, Koenders M, van de Loo F, van Lent P, van den Hoogen F, Thurlings R, Vonk MC, van der Kraan PM. TGFβ-mediated expression of TGFβ-activating integrins in SSc monocytes: disturbed activation of latent TGFβ? Arthritis Res Ther. 2020;22:42.
67. Li Z, Belozertseva E, Parlakian A, Bascetin R, Louis H, Kawamura Y, Blanc J, Gao-Li J, Pinet F, Lacy-Hulbert A, et al: Smooth muscle αv integrins regulate vascular fibrosis via CD109 downregulation of TGF-β signalling. *European Heart Journal Open* 2023, 3:oead010.
68. Wasik A, Ratajczak-Wielgomas K, Badzinski A, Dziegiel P, Podhorska-Okolow M: The role of periostin in angiogenesis and lymphangiogenesis in tumors. *Cancers (Basel)* 2022, 14.
69. Griffin MF, Huber J, Evan FJ, Quarto N, Longaker MT. The role of Wnt signaling in skin fibrosis. Med Res Rev. 2022;42:615–28.
70. Kim J-E, Lee J-H, Jeong K-H, Kim GM, Kang H: Notch intracellular domain expression in various skin fibroproliferative diseases. *ad* 2014, 26:332–337.
71. Vaittinen M, Kolehmainen M, Schwab U, Uusitupa M, Pulkkinen L. Microfibrillar-associated protein 5 is linked with markers of obesity-related extracellular matrix remodeling and inflammation. Nutr Diabetes. 2011;1: e15.
72. D'Alessio AC, Fan ZP, Wert KJ, Baranov P, Cohen MA, Saini JS, Cohick E, Charniga C, Dadon D, Hannett NM. A systematic approach to identify candidate transcription factors that control cell identity. Stem Cell Reports. 2015;5:763–75.
73. Xu J, Du Y, Deng H. Direct lineage reprogramming: strategies, mechanisms, and applications. Cell Stem Cell. 2015;16:119–34.
74. Khalil H, Kanisicak O, Prasad V, Correll RN, Fu X, Schips T, Vagnozzi RJ, Liu R, Huynh T, Lee S-J. Fibroblast-specific TGF-β–Smad2/3 signaling underlies cardiac fibrosis. J Clin Investig. 2017;127:3770–83.
75. Marthandan S, Priebe S, Groth M, Guthke R, Platzer M, Hemmerich P, Diekmann S. Hormetic effect of rotenone in primary human fibroblasts. Immunity & Ageing. 2015;12:11.
76. Soman S, Raju R, Sandhya VK, Advani J, Khan AA, Harsha HC, Prasad TSK, Sudhakaran PR, Pandey A, Adishesha PK. A multicellular signal transduction network of AGE/RAGE signaling. Journal of Cell Communication and Signaling. 2013;7:19–23.

77. Lu Y, Azad N, Wang L, Iyer AK, Castranova V, Jiang B-H, Rojanasakul Y. Phosphatidylinositol-3-kinase/akt regulates bleomycin-induced fibroblast proliferation and collagen production. American journal of respiratory cell molecular biology. 2010;42:432–41.
78. Wang L, Luo J, He S. Induction of MMP-9 release from human dermal fibroblasts by thrombin: involvement of JAK/STAT3 signaling pathway in MMP-9 release. BMC Cell Biol. 2007;8:14.
79. Marthandan S, Menzel U, Priebe S, Groth M, Guthke R, Platzer M, Hemmerich P, Kaether C, Diekmann S. Conserved genes and pathways in primary human fibroblast strains undergoing replicative and radiation induced senescence. Biol Res. 2016;49:34.
80. Hie B, Cho H, DeMeo B, Bryson B, Berger B. Geometric sketching compactly summarizes the single-cell transcriptomic landscape. Cell Syst. 2019;8:483–93.
81. Hu Y, Peng T, Gao L, Tan K: CytoTalk: de novo construction of signal transduction networks using single-cell transcriptomic data. *Science Advances* 2021, 7**:**eabf1356.
82. Xiao Y, Li Y, Tao H, Humphries B, Li A, Jiang Y, Yang C, Luo R, Wang Z. Integrin α5 down-regulation by miR-205 suppresses triple negative breast cancer stemness and metastasis by inhibiting the Src/Vav2/Rac1 pathway. Cancer Lett. 2018;433:199–209.
83. Lindvall C, Zylstra CR, Evans N, West RA, Dykema K, Furge KA, Williams BO. The Wnt co-receptor Lrp6 is required for normal mouse mammary gland development. PLoS ONE. 2009;4: e5813.
84. Chakravarthy R, Mnich K, Gorman AMJB. Nerve growth factor (NGF)-mediated regulation of p75NTR expression contributes to chemotherapeutic resistance in triple negative breast cancer cells. Biochemical biophysical research communications. 2016;478:1541–7.
85. Kalscheuer S, Khanna V, Kim H, Li S, Sachdev D, DeCarlo A, Yang D, Panyam J. Discovery of HSPG2 (Perlecan) as a therapeutic target in triple negative breast cancer. Sci Rep. 2019;9:12492.
86. Hou R, Denisenko E, Ong HT, Ramilowski JA, Forrest AR. Predicting cell-to-cell communication networks using NATMI. Nat Commun. 2020;11:5011.
87. Hu C, Diévart A, Lupien M, Calvo E, Tremblay G, Jolicoeur P. Overexpression of activated murine Notch1 and Notch3 in transgenic mice blocks mammary gland development and induces mammary tumors. Am J Pathol. 2006;168:973–90.
88. Baravalle C, Silvestrini P, Cadoche MC, Beccaria C, Andreotti CS, Renna MS, Pereyra EAL, Ortega HH, Calvinho LF, Dallard BE. Intramammary infusion of Panax ginseng extract in bovine mammary gland at cessation of milking induces changes in the expression of toll-like receptors, MyD88 and NF-kB during early involution. Res Vet Sci. 2015;100:52–60.
89. Raafat A, Lawson S, Bargo S, Klauzinska M, Strizzi L, Goldhar AS, Buono K, Salomon D, Vonderhaar BK, Callahan R. Rbpj conditional knockout reveals distinct functions of Notch4/Int3 in mammary gland development and tumorigenesis. Oncogene. 2009;28:219–30.
90. Brantley DM, Yull FE, Muraoka RS, Hicks DJ, Cook CM, Kerr LD. Dynamic expression and activity of NF-κB during post-natal mammary gland morphogenesis. Mech Dev. 2000;97:149–55.
91. Hong D, Fritz AJ, Gordon JA, Tye CE, Boyd JR, Tracy KM, Frietze SE, Carr FE, Nickerson JA, Van Wijnen A. RUNX1-dependent mechanisms in biological control and dysregulation in cancer. J Cell Physiol. 2019;234:8597–609.
92. Li L, Wang N, Xiong Y, Guo G, Zhu M, Gu Y. Transcription factor FOSL1 enhances drug resistance of breast cancer through DUSP7-mediated dephosphorylation of PEA15. Mol Cancer Res. 2022;20:515–26.
93. Cyr AR, Kulak MV, Park JM, Bogachek MV, Spanheimer PM, Woodfield GW, White-Baer LS, O'Malley YQ, Sugg SL, Olivier AK. TFAP2C governs the luminal epithelial phenotype in mammary development and carcinogenesis. Oncogene. 2015;34:436–44.
94. Majdalawieh AF, Massri M, Ro H-S. AEBP1 is a novel oncogene: mechanisms of action and signaling pathways. Journal of oncology. 2020;2020:8097872.
95. Wu Y, Zhao L, Qin Y: Comprehensive RNA-seq profiling to evaluate the rabbit mammary gland transcriptome after mastitis. *Journal of Animal Science* 2023, 101:skad110.
96. Zhao H, Huang M, Chen Q, Wang Q, Pan Y. Comparative gene expression analysis in mouse models for identifying critical pathways in mammary gland development. Breast Cancer Res Treat. 2012;132:969–77.
97. Li H, Ma T, Hao M, Guo W, Gu J, Zhang X, Wei L: Decoding functional cell–cell communication events by multi-view graph learning on spatial transcriptomics. *Briefings in Bioinformatics* 2023, 24**:**bbad359.
98. Tabula Muris C, Overall c, Logistical c, Organ c, processing, Library p, sequencing, Computational data a, Cell type a, Writing g, et al: Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Datasets* Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE109774 (2018).
99. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW: Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Datasets*, Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65525 (2015).
100. Park J, Shrestha R, Qiu C, Kondo A, Huang S, Werth M, Li M, Barasch J, Susztak K: Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Datasets*, Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE107585 (2018).
101. Joost S, Zeisel A, Jacob T, Sun X, La Manno G, Lönnerberg P, Linnarsson S, Kasper M: Single-cell transcriptomics reveals that differentiation and spatial signatures shape epidermal and hair follicle heterogeneity. *Datasets*, Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67602 (2016).
102. Guerrero-Juarez CF, Dedhia PH, Jin S, Ruiz-Vega R, Ma D, Liu Y, Yamaga K, Shestova O, Gay DL, Yang Z, et al: Single-cell analysis reveals fibroblast heterogeneity and myeloid-derived adipocyte progenitors in murine skin wounds. *Datasets*, Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE113854 (2019).
103. Kolodziejczyk AA, Kim JK, Tsang JCH, Ilicic T, Henriksson J, Natarajan KN, Tuck AC, Gao X, Bühler M, Liu P, et al: Single Cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Datasets*, ArrayExpress. https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-2600 (2015).

104. Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, Marques S, Munguba H, He L, Betsholtz C, et al: Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Datasets*, Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60361 (2015).
105. Usoskin D, Furlan A, Islam S, Abdo H, Lonnerberg P, Lou D, Hjerling-Leffler J, Haeggstrom J, Kharchenko O, Kharchenko PV, et al: Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Datasets*, Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE59739 (2015).
106. Chen R, Wu X, Jiang L, Zhang Y: Single-cell RNA-Seq reveals hypothalamic cell diversity. *Datasets* 2017, Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE87544 (2017).
107. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, Ryu JH, Wagner BK, Shen-Orr SS, Klein AM, et al: A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Datasets*, Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84133 (2016).
108. Segerstolpe A, Palasantza A, Eliasson P, Andersson E-M, Andréasson A-C, Sun X, Picelli S, Sabirsh A, Clausen M, Bjursell MK, et al: Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Datasets*, ArrayExpress. https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-5061 (2016).
109. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, 2nd, Treacy D, Trombetta JJ, Rotem A, Rodman C, Lian C, Murphy G, et al: Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Datasets*, Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72056 (2016).
110. Wang M, Liu X, Chang G, Chen Y, An G, Yan L, Gao S, Xu Y, Cui Y, Dong J, et al: Single-cell RNA sequencing analysis reveals sequential cell fate transition during human spermatogenesis. *Datasets*, Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE106487 (2018).
111. Camp JG, Sekine K, Gerber T, Loeffler-Wirth H, Binder H, Gac M, Kanton S, Kageyama J, Damm G, Seehofer D, et al: Multilineage communication regulates human liver bud development from pluripotency. *Datasets*, Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81252 (2017).
112. Deng CC, Hu YF, Zhu DH, Cheng Q, Gu JJ, Feng QL, Zhang LX, Xu YP, Wang D, Rong Z, Yang B: Single-cell RNA-seq reveals fibroblast heterogeneity and increased mesenchymal fibroblasts in human fibrotic skin diseases. *Datasets*, Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE163973 (2021).
113. Yu X, Xu X, Zhang J, Li X. Batch alignment of single-cell transcriptomics data using deep metric learning. Nat Commun. 2023;14:960.
114. Nie F, Huang H, Cai X, Ding C: Efficient and robust feature selection via joint ℓ2,1-norms minimization. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2*. pp. 813–1821. Vancouver, British Columbia, Canada: Curran Associates Inc.; 2010:813–1821.
115. Wang Y, Yin W, Zeng J. Global convergence of ADMM in nonconvex nonsmooth optimization. J Sci Comput. 2019;78:29–63.
116. Shao X, Liao J, Li C, Lu X, Cheng J, Fan X: CellTalkDB: a manually curated database of ligand-receptor interactions in humans and mice. *Briefings in Bioinformatics* 2021, 22:bbaa269.
117. Cabello-Aguilar S, Alame M, Kon-Sun-Tack F, Fau C, Lacroix M, Colinge J. SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. Nucleic Acids Res. 2020;48: e55.
118. Ramilowski JA, Goldberg T, Harshbarger J, Kloppmann E, Lizio M, Satagopam VP, Itoh M, Kawaji H, Carninci P, Rost B, Forrest AR. A draft network of ligand-receptor-mediated multicellular signalling in human. Nat Commun. 2015;6:7866.
119. Zhang C, Dang D, Cong L, Sun H, Cong X. Pivotal factors associated with the immunosuppressive tumor microenvironment and melanoma metastasis. Cancer Med. 2021;10:4710–20.
120. Gao S, Feng X, Wu Z, Kajigaya S, Young NS. Cell CallEXT: analysis of ligand-receptor and transcription factor activities in cell-cell communication of tumor immune microenvironment. Cancers (Basel). 2022;14:4957.
121. Baccin C, Al-Sabah J, Velten L, Helbling PM, Grunschlager F, Hernandez-Malmierca P, Nombela-Arrieta C, Steinmetz LM, Trumpp A, Haas S. Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. Nat Cell Biol. 2020;22:38–48.
122. Cain MP, Hernandez BJ, Chen J: Quantitative single-cell interactomes in normal and virus-infected mouse lungs. *Disease models & mechanisms* 2020, 13:dmm044404.
123. Ding C, Li Y, Guo F, Jiang Y, Ying W, Li D, Yang D, Xia X, Liu W, Zhao Y, et al. A cell-type-resolved liver proteome. Mol Cell Proteomics. 2016;15:3190–202b.
124. Sheikh BN, Bondareva O, Guhathakurta S, Tsang TH, Sikora K, Aizarani N, Sagar, Holz H, Grun D, Hein L, Akhtar A: Systematic identification of cell-cell communication networks in the developing brain. *iScience* 2019, 21**:**273–287.
125. Skelly DA, Squiers GT, McLellan MA, Bolisetty MT, Robson P, Rosenthal NA, Pinto AR. Single-cell transcriptional profiling reveals cellular diversity and intercommunication in the mouse heart. Cell Rep. 2018;22:600–10.
126. Yuzwa SA, Yang G, Borrett MJ, Clarke G, Cancino GI, Zahr SK, Zandstra PW, Kaplan DR, Miller FD. Proneurogenic ligands defined by modeling developing cortex growth factor communication networks. Neuron. 2016;91:988–1004.
127. Ding Q, Yang W, Xue G, Liu H, Cai Y, Que J, Jin X, Luo M, Pang F, Yang Y, et al: Dimension reduction, cell clustering and cell-cell communication inference for single-cell transcriptomics with DcjComm. *Source Code on Github* 2024, https://github.com/Ginnay/DcjComm.
128. Ding Q, Yang W, Xue G, Liu H, Cai Y, Que J, Jin X, Luo M, Pang F, Yang Y, et al: Dimension reduction, cell clustering and cell-cell communication inference for single-cell transcriptomics with DcjComm. *Zenodo* 2024, https://zenodo.org/records/12666949.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.