**File S1 c-CSN implementation details and additional notes**

**Note 1: Statistical model of cell-specific network**

Assume that there are *n* cells with *m* genes in scRNA-seq data. The statistical model of cell-specific network (CSN) can be described as following [1]. In CSN,

is used to measure the independency of genes *x* and *y*. If gene *x* and gene *y* are independent,. If gene *x* and gene *y* are dependent, . And the estimation of in cell k is

(S-1)

where, , is the number of neighborhoods of , and ).

And if *x* and *y* are independent of each other, the statistic approximately follows the normal distribution. CSN performs the following hypothesis test (one-side test) for gene association (e.g. gene *x* and gene *y* of cell *k*):

(null hypothesis): genes *x* and *y* are independent in cell *k*.

(alternative hypothesis): genes *x* and *y* are associated with each other in cell *k*.

If the statistic is above the significance level, the null hypothesis will be rejected and , otherwise .

**Note 2: Statistical model of c-CSN**

1. **Statistical model of c-CSN**

Assume that there are *n* cells with *m* genes in scRNA-seq data. Based on the statistical conditional independency in probability theory, we design a statistic for each gene pair *x* and *y* and the conditional gene *z* in each cell *k* as

(S-2)

Where is the statistic for genes *x*, *y*, *z* of cell *k* (the red dot in Figure S1 A), is the number of cells in the neighborhood of, and, is the number of cells in the neighborhood of (, (. is the number of cells in the neighborhood of ( and *n* is the total number of dots.

In Equation (S-2),are determined in advance ( < *n* ) for a given conditional gene *z*. Thus the statistic only changes with . As shown as Figure S1 A, we first considered the conditional gene *z*. Around the cell *k* (the red dot), we first draw two parallel planes ( and ), where *z*-axis is orthogonal with two planes and the projection of cell *k* in *z*-axis is contained in the space between the two planes. The dots contained in the space between the two planes are the neighbors of . The number is. Next, we can also get four parallel light shadow planes (, and , ) where two planes are orthogonal with *x*-axis, the other two are orthogonal with *y*-axis. The number of dots which are contained in the space surrounded by (,and ,) is . The number of dots which are contained in the space surrounded by (,and ,) is . The number of dots (Figure S1 C) which are contained in the space surrounded by (, and ,, ,) is .

We can identify the association of two genes by a statistical test based on the statistic.

**(2) Expectation and variance of the statistic**

Since,, are determined in advance, and follows the binomial distribution, we can get the probability:

where, and is the probability that a dot is located in the(,and ,), (,and ,), (, and ,, ,) respectively.

The expectation of the statistic is

The variance of the statistic is

**(3) The null hypothesis of the test**

If genes *x* and gene *y* are conditionally independent of each other given the gene *z*, that is, then,

The statistic can be normalized.

(S-3)

In probability theory, the conditional mutual information (CMI) [2-4] is, the expected value of the mutual information of two random variables given a third random variable.

(S-4)

The conditional mutual information of gene and gene given gene in cell k is. When, the gene and gene are independent given gene . When, it can be described that uncertainty of gene and gene compared with the condition of are increased given gene . So, we think gene and gene are not associated given gene , when. The statistic is not greater than 0, when the null hypothesis is true.

**(4) The distribution of the statistic**

Based on the numerical simulation, the distribution of the normalized statistic of Equation (S-3) tends to follow the standard Gaussian distribution as n increases.

**Note 3: The computational complexity of c-CSN**

c-CSN can be constructed for each conditional gene in each cell. To estimate direct association between a pair of genes in a cell, all the remaining *m*2 genes theoretically could be used as conditional genes, where *m* denotes the number of genes in our analysis. And *m*(*m*1)/2 gene pairs/associations should be estimated for each conditional gene in each cell. So c-CSN runs in the order of O, if all remaining *m*2 are set as conditional genes. Numerically, a small number of conditional genes are used to identify direct association between a pair of genes in a cell, which can significantly reduce computational cost. Then c-CSN runs in the order of O, where is the number of conditional genes.

**Note 4: The input, output, and application fields of our c-CSN method**

Input: Gene expression matrix (FPKM/TPM/RPKM/count)

Significance level (e.g. 0.001, 0.01, 0.05 …).

Box size: The box size is the size of neighborhood in our algorithm. (). Default = 1. Users can change this parameter as well.

*z*: the number of conditional genes. Default = 5. Users can change this parameter as well.

Output: conditional cell-specific network for each cell for a given conditional gene. (row= genes, column= cells)

Conditional network degree matrix (row = genes, column = cells)

Network flow entropy for all cells (1*n* vectors)

Application fields:

The number of cells should be 100 at least.

The clustering results of different parameters were shown in Figure S7.

**Note 5: Algorithms and their parameters used in clustering, visualization, and the number of the conditional genes used in datasets**

Preprocessing: GEM is preprocessed form initial gene expression matrix by normalization. After gene selection, CNDM is transformed from GEM and also logarithm transformed.

To benchmark c-CSN, we applied the method to several datasets, and compared the results with some existing methods: K-means, t-SNE + K-means, hierarchical, t-SNE + hierarchical, k-medoids. For t-SNE, we use Rtsne, an R package with 10000 iteration times, and the input is 15 PCs to 30 PCs. We repeated 100 times for these methods on each dataset.

For a fair comparison, ten different perplexities (from 10 to 70) were applied in t-SNE in each dataset. We then repeated 100 times for K-means and hierarchical after t-SNE for each perplexity to get stable results. And the adjusted random index (ARI) was used to evaluate the results of cells clustering.

In the SIMLR clustering, the R function SIMLR was used to perform clustering analysis. All parameters were as default. The number of clusters was set as the number of categories in the original dataset. In the SC3, the sc3 function was applied to perform clustering. The parameter gene\_filter was FALSE and the other parameters were as default.

In Seurat, we constructed a shared nearest neighbor (SNN) graph using the first 10 PCs. Finally, the clusters of cells can be identified by the Louvain algorithm based on the SNN. For the Tabula Muris\_droplet1 dataset, we first performed PCA and the top 20 PCs were used to perform t-SNE and clustering analysis.