METHOD

Uncovering hypergraphs of cell-cell interaction from single cell RNA-sequencing data

Koki Tsuyuzaki[1], Manabu Ishii[1] and Itoshi Nikaido[1,2*]

*Correspondence:
itoshi.nikaido@riken.jp
[2]Bioinformatics Course, Master's/Doctoral Program in Life Science Innovation (T-LSI), School of Integrative and Global Majors (SIGMA), University of Tsukuba, Wako, 351-0198, Saitama, Japan
Full list of author information is available at the end of the article

Abstract

Complex biological systems can be described as a multitude of cell-cell interactions (CCIs). Recent single-cell RNA-sequencing technologies have enabled the detection of CCIs and related ligand-receptor (L-R) gene expression simultaneously. However, previous data analysis methods have focused on only one-to-one CCIs between two cell types. To also detect many-to-many CCIs, we propose `scTensor`, a novel method for extracting representative triadic relationships (hypergraphs), which include (i) ligand-expression, (ii) receptor-expression, and (iii) L-R pairs. When applied to simulated and empirical datasets, `scTensor` was able to detect some hypergraphs including paracrine/autocrine CCI patterns, which cannot be detected by previous methods.

Keywords: Single-cell RNA-sequencing; Cell-cell interaction; Hypergraph; Dimension Reduction; Tensor Decomposition; Non-negative Tucker Decomposition; R/Bioconductor

Background

Complex biological systems such as tissue homeostasis [1, 2], neurotransmission [3, 4], immune response [5], ontogenesis [6], and stem cells niche [7, 8] are composed by cell-cell interaction (CCI). Many molecular biology studies have been decomposed the system into constituent parts (e.g., genes, proteins, and metabolites) to clarify

their functions. Nevertheless, more sophisticated methodologies are still required, because CCI is the difference between the whole system and sum of their parts.

Previous studies have investigated CCIs using technologies such as fluorescence microscopy [9–13], microdevise-based methods such as microwells, micropatterns, single-cell traps, droplet microfluidics, and micropillars [14–22], and transcriptome-based method [23–53]. In particular, the recent development of single-cell RNA-sequencing (scRNA-seq) technologies has enabled the detection of exhaustive cell-type-level CCIs based on ligand and receptor (L-R) coexpression.

To assess the coexpression of known L-R genes, circle plots [24, 33, 36, 44, 45, 50], bigraph/Sankey diagrams [29, 30, 34, 49, 52, 53], network diagrams [26–28, 31, 37, 40, 41, 44, 46, 54] and heatmaps [30, 32, 34, 39, 46, 47, 49] are often drawn. Some studies have also introduced more systematical approaches for quantifying the degree of CCIs based on the L-R coexpression, such as the number of coexpressed L-R pairs [23, 24, 26, 27], Spearman correlation coefficients between L-R expression profiles [26, 31], original interaction scores between L-R coexpression [39], or hypothetical test based on random cell-type label permutation [29, 32, 37, 55].

All the approaches described above implicitly suppose that CCIs are the one-to-one relationships between two cell types and that the corresponding L-R coexpression is observed as the cell-type-specific manner. In real empirical data, however, the situation can be more complex; CCIs often exhibit many-to-many relationships involving many cell types, and an particular L-R pair can also function across multiple cell-type pairs. Therefore, in this work, we propose `scTensor`, which is a novel method based on a tensor decomposition algorithm. Our method regards CCIs as hypergraphs and extracts some representative triadic relationship from the data tensor, which includes (i) ligand-expressing cell types, (ii) receptor-expressing cell types, and (iii) L-R pairs.

CCI as hypergraph (*CaH*) and CCI-tensor

The simplest CCI representation is perhaps a directed graph, where each node represents a cell type and each edge represents the coexpression of all L-R pairs (Figure 1a, left). The direction of the edge is set as ligand → receptor. Such a data structure can also be described as an asymmetric adjacency matrix, in which each row and column represents a ligand and receptor, respectively. If some combinations of cell types are regarded as interacting, corresponding elements of the matrix are filled with 1 and otherwise 0. If the degree of CCI is not a binary relationship, weighted graphs and corresponding weighted adjacent matrices may also be used. The previous analytical methods are categorized within this approach [23, 24, 26–34, 36, 37, 39–41, 44–47, 49, 50, 52, 53, 55].

In contrast, this work describes CCIs as directed hypergraphs (CCI as hypergraph; *CaH*), where each node is a cell type but the edges are distinguished from each other by the different related L-R pair sets (Figure 1a, right). Such a context-aware edge is called a hyperedge and is described as multiple different adjacency matrices and the set of matrices is called a higher-order matrix or tensor. In contrast with the simple adjacency matrix, tensor contains considerable high-resolution information owing to its higher-order.

Prediction of many-to-many CCIs using tensor decomposition

Tensor data are constructed through the following steps (Figure 1b). Here, a scRNA-seq matrix and the cellular label specifying cell types are supposed to be provided by users. Firstly, Freeman-Tukey transformation [56] (FTT, $\sqrt{\boldsymbol{x}} + \sqrt{\boldsymbol{x}+1}$), which is a variance-stabilizing transformation, is performed to the data matrix. Next, the matrix is converted to a cell-type-level average matrix according to the label. Combined with a L-R database, two corresponding row-vectors of an L-R pair are extracted from the matrix. We originally developed the databases for 12 organisms (for more details, see Additional Files 1, 2 , Table 2 and 3). The outer product (direct

product) of the two vectors is then calculated, and a matrix is generated. The matrix can be considered as the similarity matrix of all possible cell-type combinations using the L-R pair. Finally, for each L-R pair, the matrix is calculated, and a tensor is generated as the combined matrices. In this work, this is called the CCI-tensor.

After the construction of a CCI-tensor, we perform non-negative Tucker decomposition (NTD [57, 58]), which is known as a tensor decomposition algorithm. We originally implemented this algorithm and confirmed its convergence within a realistic number of iterations (for more details, see Additional File 3). NTD assumes that the CCI-tensor can be approximated by the summation of some representative *CaH*s. NTD has the three rank parameters ($R1, R2, and\ R3$) and a *CaH* is calculated as the outer product of the column vectors of three factor matrices $\boldsymbol{A}^{(1)} \in \mathbb{R}^{J \times R1}$, $\boldsymbol{A}^{(2)} \in \mathbb{R}^{J \times R2}$, and $\boldsymbol{A}^{(3)} \in \mathbb{R}^{K \times R3}$ calculated by NTD (Figure 1c). Each *CaH*-strength is calculated by the core tensor $\mathcal{G}(r1, r2, r3) \in \mathbb{R}^{R1 \times R2 \times R3}$ of NTD. In this work, each *CaH* is termed $CaH(r1, r2, r3) = \boldsymbol{A}^{(1)}_{:r1} \circ \boldsymbol{A}^{(2)}_{:r2} \circ \boldsymbol{A}^{(3)}_{:r3} \in \mathbb{R}^{J \times J \times K}$, where r1, r2, and r3 are the indexes of the columns of three factor matrices. All *CaH*s are ordered by the size of elements of the core tensor, and the patterns explaining the top 40 % of cumulative core tensor value are reserved as representative *CaH*s. For more details on *CaH*, see the Materials and Methods. The *CaH*s are extracted in a data-driven way without the assumption of one-to-one CCIs. Therefore, it can also detect many-to-many CCIs according to the data complexity.

Results and discussion

Evaluation of multiple CCI prediction

Accuracy of the detection of CCIs and the related L-R pairs

Here, we demonstrate the efficacy of `scTensor` by using the two simulation datasets (Figure2a). Three different cell types are indicated as "A", "B", and "C". In the case I dataset, all CCIs represent the one-to-one relationships between two cell types. CCIs corresponding to A $\rightarrow$ B, B $\rightarrow$ C, and C $\rightarrow$ A are colored by red, blue, and

green, respectively. In contrast, the CCIs in the case II dataset represent many-to-many relationships involving many cell types such as A $\rightarrow$ B/C (red), C $\rightarrow$ A/B/C (blue), and A/B/C $\rightarrow$ B/C (green). To evaluate whether such ground truth combination of cell types and their related L-R pairs are enriched by `scTensor`, receiver operating characteristic (ROC) curves and their corresponding area under the curve (AUC) values were calculated (Figure2b). In the analysis of each L-R set, only the *CaH*s with the maximum AUC value for each ground truth L-R set were regarded as the corresponding *CaH* being accurately detected by `scTensor` (Figure2c, for more details on the simulation datasets, see the Materials and Methods).

Again, note that the *CaH*s detected by `scTensor` are not just CCIs, but sets of CCIs and their related L-R pairs. To the best of our knowledge, the label permutation method implemented in CellPhoneDB [37] is the only previous method that can detect CCIs and their related L-R pairs simultaneously. To demonstrate the efficacy of `scTensor` in terms of the detection of many-to-many CCIs, we also originally implemented this method and compared it with `scTensor` (for more details on the algorithm, see the Materials and Methods). Note that each combination of a CCI and its related L-R pair is separately extracted by NTD of `scTensor`, but under the label permutation method, the combinations are not separated and are just sorted in ascending order of their P-values. Combinations with low P-values indicate significant triadic relationships.

In the case I dataset, ground truth L-R sets are highly enriched according to the measures of both methods, and the AUC values show that there is no difference in their performance (Figure2b and c, left). On the other hand, in the case II dataset, the label permutation method cannot correctly detect blue or green L-R sets, and the AUC value becomes lower (Figure2b, right). In the `scTensor` analysis, red, blue, and green L-R sets are separately extracted as three *CaH*s (Figure2c, right), and the AUC values are still high. This is because, the label permutation

is implicitly hypothesized the CCI as a one-to-one relationship. Therefore, in the case II dataset, many-to-many CCIs such as the CCIs corresponding to green L-R sets, are hard to detect by the method. This is because for each L-R pair, mean values for any combination of cell types are basically high in such situations, and a P-value corresponding to a one-to-one CCI tends to be large (i.e., not significant); accordingly, the observed L-R coexpression and the null distribution calculated are hard to distinguish. In the analysis of real datasets presented later, however, the L-R gene expression pairs are not always the cell-type specific, and it is more natural that the CCI corresponding to the L-R has a many-to-many relationship. This simulation shows that `scTensor` is a more general method for detecting CCIs and their related L-R pairs at once, irrespective of whether a particular CCI is one-to-one or many-to-many.

Biological interpretation of real datasets

To demonstrate the efficacy of `scTensor` in the analysis of empirical datasets, we applied `scTensor` to four real scRNA-seq datasets (Table 1). First, we used the scRNA-seq data derived from fetal germ cells (FGCs (♀)) and their gonadal niche cells (Soma) from female human embryos (Germline_Female [25]). As a known *CaH* pattern, CCIs of Soma with FGCs (♀) involving the BMP signaling pathway are reported, and `scTensor` accurately detects the CCIs and CCI-related L-R pairs as *CaH*(4,2,5) (Figure 3). Moreover, `scTensor` was able to extract some putative *CaH*s such as *CaH*(3,4,14) and *CaH*(1,4,22). *CaH*(3,4,14) is the autocrine-type CCI within FGCs (♀) and L-R pairs such as Wnt (WNT5A/WNT6) and some growth factor genes (NGF/IGF/FGFR/VEGF), suggesting that this *CaH* is related to the proliferation and differentiation of Soma. *CaH*(1,4,22) is the CCI corresponding to FGCs (♀) and Soma, and most of the receptors are related to G protein-coupled receptor (GPCR). The conjugated ligands are some neuropeptides, and this suggests

that the peptides are related to the activation of GPCR in FGCs (♀) by some mechanism.

We also used the scRNA-seq data derived from FGCs (♂) and their gonadal niche cells from male human embryos (Germline_Male [25]). In contrast to the Germline_Female data, the original study reported that FGCs (♂) interact with Soma through AMH-BMPR interactions, and `scTensor` also detected the triadic relationship as $CaH(1,1,10)$ (Figure 4). The conjugate receptor AMHR2 is also detected in this CaH. In this dataset, like in Germline_Female, in this data, autocrine-type CCIs such as $CaH(3,3,9)$ and CCIs corresponding to FGCs (♂) and Soma ($CaH(1,3,16)$) were detected.

Next, we used the scRNA-seq data derived from immune cells isolated from the metastatic melanoma patients (Melanoma [59]). The meta-analysis of scRNA-seq and TCGA datasets in the original published study claimed that T cell abundance and the expression of complement system-related genes in cancer-associated fibroblasts (CAFs) are correlated, and CCIs between T cells and CAFs were therefore inferred. `scTensor` detected these CCIs as $CaH(1,1,3)/CaH(1,1,9)$ and complement system-related genes such as C3, CXCL12, CFB, and C4A were also detected (Figure 5). `scTensor` was also able to capture a well-known CCI between T cells and B cells as $CaH(2,2,6)$, although this was not a focus of the original study; this CCI is the antigen-presenting from B cells to T cells by class II major histocompatibility complex (MHC) through the coexpression of CD8 (T cell receptor) and human leukocyte antigen (HLA) genes. `scTensor` also detected $CaH(4,1,13)$ and $CaH(1,4,8)$, which are the CCIs of macrophages with T cells and an autocrine-type CCI of known chemokine ligands with their receptors.

Finally, we used data derived from non-myocyte cells isolated from mouse heart (NonMyocyte [27]). The original study focused on the CCIs of pericytes/fibroblasts with macrophages through the L-R pairs Il34/Csf1 and Csf1r, and the correspond-

ing CCIs were detected as CaH(2,2,19) by `scTensor` (Figure 6). `scTensor` also detected CaH(1,2,17) and CaH(3,2,21), which are the autocrine-type CCI among macrophages and the CCI of NK cells with macrophages by known chemokine ligands and their receptors.

Application to minor organism

To demonstrate the applicability of using `scTensor` in the species that is not mouse or human, we used a scRNA-seq data derived from zebrafish habenular neurons (Habenular_Larva [60]). Although the original study did not focus on the CCIs among the neuronal cell types, `scTensor` detected some triadic relationships as CaH(3,3,4) (La_Hb01/03/07 with La_Hb02/08), CaH(2,1,3) (La_Hb09 with La_Hb02/08), and CaH(1,3,1) (La_Hb02/04-06/08/11-15/Olf with La_Hb02/08) (Figure 7). The spatial distribution of the cell types measured by RNA-fluorescence *in situ* hybridization (FISH) shows that the cell-type pairs detected as CaH(3,3,4) and CaH(2,1,3) are dorsally located and proximal to each other in the habenula. However, CaH(1,3,1) was a more global interaction related to the entire habenula regions. Although the spatial distribution of rare cell types La_Hb03/05/14 could not be determined by RNA-FISH in the original study, `scTensor` was able to assign the conjugated cell types of La_Hb03 as La_Hb02/08 in the dorsal region. This result suggests that `scTensor` may also be useful in spatial transcriptomics [61, 62].

scTensor and L-R database implementations as R/Bioconductor packages

All the algorithms and L-R lists are available as R/Bioconductor packages and a web application described below.

nnTensor and scTensor packages

NTD is implemented as the function of `nnTensor` R/CRAN package and internally imported in `scTensor`. `scTensor` constructs the CCI-tensor, decomposes the tensor by NTD, and generates an HTML report. `scTensor` is assumed to be used with

`LRBase.XXX.eg.db`, which are the L-R databases for multiple organisms. To enhance the biological interpretation of *CaH*s, a wide variety of gene information is assigned to the L-R lists through the other R/Bioconductor packages (Figure 8). For example, gene annotation is assigned by `biomaRt` [63] (Gene Name, Description, Gene Ontology (GO), STRING, and UniProtKB), `reactome.db` [64] (Reactome) and `MeSH.XXX.eg.db` [65] (Medical Subject Headings; MeSH), while the enrichment analysis (also known as over-representative analysis; ORA) is performed by `GOstats` [66] (GO-ORA), `meshr` [65] (MeSH-ORA), `ReactomePA` [67] (Reactome-ORA), and `DOSE` [68] (Disease Ontology; DO, Network of Cancer Gene; NCG, and DisGeNET-ORA). To validate that the detected gene expression of L-R gene pair is also consistently detected in the other data with tissue- or cell-type-level transcriptome data, the hyperlinks to RefEx [69], Expression Atlas [70], Single-Cell Expression Atlas [71], scRNASeqDB [72], PanglaoDB [73] are embedded in the HTML report, facilitating comparisions of the L-R expression with the data from large-scale genomics projects such as GTEx [74], FANTOM5 [75], NIH Epigenomics Roadmap [76], ENCODE [77], and Human Protein Atlas [78]. Additionally, in consideration of users who might want to experimentally investigate detected CCIs, we embedded the hyperlinks to Connectivity Map (CMap [79]), which provides the relationships between perturbation by the addition of particular chemical compounds/genetic reagents and succeeding gene expression change.

LRBase.XXX.eg.db-type packages

For data sustainability and the extension to the wide range of organisms, in this work, we originally constructed L-R databases as R/Bioconductor packages named `LRBase.XXX.eg.db` (where XXX represents the abbreviation for an organism, such as "Hsa" for Homo sapiens). `LRBase.XXX.eg.db` currently provides the L-R databases for 12 organisms (Table 2 and 3). The data process pipeline is almost the same as that of the FANTOM5 project for constructing the putative L-R lists.

Precise differences between `LRBase.XXX.eg.db` and FANTOM5 are summarized in Table S4 in Additional File 1.

LRBaseDbi package

All the `LRBase.XXX.eg.db` packages are generated by `LRBaseDbi`, which is the another R/Bioconductor package. `LRBaseDbi` generates the `LRBase.XXX.eg.db` packages from the CSV files, in which NCBI Gene IDs are saved as two columns describing the L-R relationship (we call this function as "meta"-packaging).

In addition to the `LRBase.XXX.eg.db` packages we summarized, the users may want to specify the user's original L-R list. For example, there are some other L-R databases such as IUPHAR [80], DLRP [81], FANTOM5 [75], CellPhoneDB [28], and Cell-Cell Interaction Database summarized by Gary Bader et. al. (http://baderlab.org/CellCellInteractions) (Additional File 1). Besides, if the users want to apply the `scTensor` to minor species, the L-R list may be constructed by the orthologous relationship with major species (e.g., human) [27]. The corresponding `LRBase.XXX.eg.db` is easily generated from the original L-R list by `LRBaseDbi` and can be used with `scTensor`.

CellCelldb

All analytical results `scTensor` are outputted as HTML reports. Hence, combined with a cloud web service such as Amazon Simple Storage Service (Amazon S3), reports can be used as simple web applications, enabling the user to share their results with collaborators or to develop an exhaustive CCI database. We have already performed `scTensor` analyses using a wide-variety of scRNA-seq datasets, including the five empirical datasets examined in this study (https://q-brain2.riken.jp/CellCelldb/).

Conclusions

In this work, CCIs were regarded as *CaH*s, which are hypergraphs that represent triadic relationships, and a novel algorithm `scTensor` for detecting such *CaH* was developed. In evaluations with empirical datasets from previous CCI studies, previously reported CCIs were also detected by `scTensor`. Moreover, some CCIs were detected only by `scTensor`, suggesting that the previous studies may have overlooked some CCIs. To extend the use of `scTensor` to a wide range of organisms, we also developed multiple L-R databases as `LRBase.XXX.eg.db`-type packages. When combined with `LRBase.XXX.eg.db`, `scTensor` can currently be applied to 12 organisms.

There are still some plans for improving both `LRBase.XXX.eg.db` and `scTensor` to build on the advantages of this current framework. For example, the range of corresponding organisms and the L-R lists can be extended with the spread of genome-wide researches. Additionally, the algorithm can be improved, for example, by utilizing acceleration techniques such as randomized algorithm/sketching methods [82], incremental algorithm/stochastic optimization [83, 84], or distributed computing with MapReduce/Hadoop on large-scale memory machines [85] for NTD, which is now available. Tensors are a very flexible way to represent heterogeneous biological data [86], and easily integrate the side information about genes or cell types with semi-supervised manner. Such information will improve the accuracy and extend the scope of the data.

We aim to tackle such problems and develop the framework further through updates of the R/Bioconductor packages. In the package registration process for R/Bioconductor, package source code is peer-reviewed via the Bioconductor single package builder system and assigned to a curator (`https://github.com/Bioconductor/packagebuilder`), and even after the package is accepted, the daily package builder tests the source code every day (`https://www.bioconductor.org/`

checkResults/). Furthermore, biannual updates of Bioconductor require that the internal data of all data packages are updated (https://www.bioconductor.org/developers/release-schedule/). Such strict check systems for source code and internal data improve the sustainability and usability of packages. Our team has maintained over one hundred R/Bioconductor packages since 2015 [65], and we are still organizing a system for the maintenance of the combined `LRBase.XXX.eg.db` and `scTensor` framework.

Materials and methods

Construction of L-R list

Public databases

To compare our database with other databases (Additional File 1 and 2), the data from FANTOM5 (http://fantom.gsc.riken.jp/5/suppl/Ramilowski$_$et$_$al$_$2015/data/PairsLigRec.txt), DLRP (http://dip.doe-mbi.ucla.edu/dip/dlrp/dlrp.txt), IUPHAR (http://www.guidetopharmacology.org/DATA/interactions.csv), and HPRD (ftp://ftp.ebi.ac.uk/pub/databases/genenames/new/tsv/locus$_$groups/protein-coding$_$gene.txt) were downloaded. The subcellular localization data from SWISSPROT and TrEMBL were downloaded from UniProtKB (https://www.uniprot.org/downloads). Protein-protein interaction (PPI) data of STRING (v-10.5) were downloaded from https://stringdb-static.org/cgi/download.pl. To unify the gene identifier as NCBI Gene ID, we retrieved the corresponding table from Biomart (Ensembl release 92). All the data were downloaded by RESTful access using the wget command and query.xml http://www.biomart.org/martservice.html.

Simulation datasets

The simulated single-cell gene expression data were sampled from the negative binomial distribution $NB\left(FC_{gc}m_g, \phi_g\right)$, where FC_{gc} is the fold-change (FC) for gene g and cell type c, and m_g and ϕ_g are the average gene expression and the dispersion

parameter of gene g, respectively. For the setting of differentially expressed genes (DEGs) and non-DEGs, FC_{gc} values were calculated based on the non-linear relationship of FC and the gene expression level $\log_{10} FC_{gc} = a \exp(-b \log_{10}(m_g + 1))$ as follows:

$$FC_{gc} = \begin{cases} 10^{4.42 \exp(-0.81 \log_{10}(m_g+1))} & \text{(DEG)} \\ 1 & \text{(non-DEG).} \end{cases}$$

The m_g and gene-wise variance v_g were calculated from the scRNA-seq dataset of human embryonic stem cells (hESCs) measured by Quartz-Seq [87], and the gene-wise dispersion parameter ϕ_g was estimated as $\phi_g = (v_g - m_g)/m_g^2$. The NB distribution reduces to Poisson when $\phi_g = 0$. To simulate the "dropout" phenomena of scRNA-seq experiments, we introduced the dropout probability $p_{dropout_{gc}} = \exp(-cFC_{gc}m_g^2)$, which is used in `ZIFA` [88] (default: c=1), and the expression values were randomly converted to 0 based on this dropout probability.

For the setting of the case I datasets, 150 × 150 × 500 CCI-tensor was constructed. For each cell type, 50 cells were established, and in total, three cell types were set. For L-R set 1 (red), 50 L-R pairs were established, and the cell-type-wise ligand and receptor patterns were (1,0,0) and (0,1,0), respectively (0, non-DEG; 1, DEG). For L-R set 2 (blue), 50 L-R pairs were established, and the cell-type-wise ligand and receptor patterns were (0,1,0) and (0,0,1), respectively. For L-R set 3 (green), 50 L-R pairs were established, and the cell-type-wise ligand and receptor patterns were (0,0,1) and (1,0,0), respectively. The other 350 L-R pairs were sampled randomly as non-DEGs.

For the setting of the case II datasets, 150 × 150 × 500 CCI-tensor were constructed. For each cell type, 50 cells were established, and in total, three cell types were established. For L-R set 1 (red), 50 L-R pairs were established, and the cell-

type-wise ligand and receptor patterns were (1,0,0) and (0,1,1), respectively. For L-R set 2 (blue), 50 L-R pairs were established, and the cell-type-wise ligand and receptor patterns were (0,0,1) and (1,1,1), respectively. For L-R set 3 (green), 50 L-R pairs were established, and the cell-type-wise ligand and receptor patterns were (1,1,1) and (1,0,1), respectively. The other 350 L-R pairs were sampled randomly as non-DEGs.

Real datasets

The gene expression matrix and cellular labels for Germline_Female and Germline_Male scRNA-seq data were retrieved from the GEO database (GSE86146), and only highly variable genes (HVGs : http://pklab.med.harvard.edu/scw2014/subpop_tutorial.html) with low P-values ($\leq$ 1E-7) were extracted. The gene expression matrix and cellular label of Melanoma scRNA-seq data were retrieved from the GEO database (GSE72056), and only HVGs with low P-values ($\leq$ 1E-10) were extracted. The gene symbols were converted to NCBI GeneIDs using the R/Bioconductor *Homo.sapiens* package. The gene expression matrix and cellular labels for NonMyocyte scRNA-seq data were retrieved from the ArrayExpress database (E-MTAB-6173), and only HVGs with low P-values ($\leq$ 1E-10) were extracted. The gene symbols are converted to NCBI GeneIDs using the R/Bioconductor *Mus.musculus* package. For each dataset, t-Distributed Stochastic Neighbor Embedding (t-SNE) with 40 perplexity was performed using the Rtsne R package.

scTensor algorithm details

Construction of CCI-tensor

Here we assume that a data matrix $\boldsymbol{Y} \in \mathbb{R}^{I \times H}$ is the gene expression matrix of scRNA-seq, where I is the number of genes and H is the number of cells. Next, the matrix $\boldsymbol{Y}$ is converted to cell-type mean matrix $\boldsymbol{X} \in \mathbb{R}^{I \times J}$, where J is the number of mean vectors for each cell type. The cell-type label is supposed to be specified by user's prior analysis such as clustering or confirmation of marker gene expression.

The relationship between the $\boldsymbol{X}$ and $\boldsymbol{Y}$ is described as below:

$$\boldsymbol{X} = \boldsymbol{Y}\boldsymbol{A}, \tag{1}$$

where the matrix $\boldsymbol{A} \in \mathbb{R}^{H \times J}$ converts cellular-level matrix $\boldsymbol{Y}$ to cell-type-level matrix X and each element of $\boldsymbol{A}$ is

$$\boldsymbol{A}_{hj} = \begin{cases} 1/n_j & \text{(h-th cell belongs to j-th cell type)} \\ 0 & \text{(otherwise)}, \end{cases}$$

where n_j is the number of cells belonging to j's cell type.

Next, NCBI gene IDs of each L-R pair stored in `LRBase.XXX.eg.db` are searched in the row names of matrix X and if both IDs are found, corresponding J-length row-vectors of the ligand and receptor genes ($\boldsymbol{x}_L$ and $\boldsymbol{x}_R$) are extracted.

Finally, a $J \times J$ matrix is calculated as the outer product of $\boldsymbol{x}_L$ and $\boldsymbol{x}_R$ and incrementally stored as a sub-tensor (frontal slice) of the CCI-tensor $\chi \in \mathbb{R}^{J \times J \times K}$ as below:

$$\chi_{::k} = \boldsymbol{x}_{L(k)} \circ \boldsymbol{x}_{R(k)} \tag{2}$$

where K is the number of L-R pairs found in the row names of matrix $\boldsymbol{X}$.

CANDECOMP/PARAFAC and Tucker decomposition

Here, we suppose that the CCI-tensor has some representative triadic relationship. To extract the triadic relationships from a CCI-tensor, here we consider performing some tensor decomposition algorithms. There are two typical decomposition methods; CANDECOMP/PARAFAC (CP) and Tucker decomposition [57, 58].

In CP decomposition, CCI-tensor χ is decomposed as follows:

$$\begin{aligned}\chi &= \Lambda \times_1 \boldsymbol{A}^{(1)} \times_2 \boldsymbol{A}^{(2)} \times_3 \boldsymbol{A}^{(3)} \\ &= \sum_{r=1}^{R} \lambda_r \boldsymbol{A}^{(1)}_{:r} \circ \boldsymbol{A}^{(2)}_{:r} \circ \boldsymbol{A}^{(3)}_{:r} \\ &\text{subject to } \|\boldsymbol{A}^{(1)}_{:r}\| = \|\boldsymbol{A}^{(2)}_{:r}\| = \|\boldsymbol{A}^{(3)}_{:r}\| = 1,\end{aligned} \tag{3}$$

where $\times_n$ is mode-n product, R is the rank of χ, and Λ is diagonal cubical tensor, in which the element λ_r on the superdiagonal can be non-zero. $\boldsymbol{A}^{(1)} \in \mathbb{R}^{J \times R}$, $\boldsymbol{A}^{(2)} \in \mathbb{R}^{J \times R}$, and $\boldsymbol{A}^{(3)} \in \mathbb{R}^{K \times R}$ are factor matrices. $\boldsymbol{a}_r \circ \boldsymbol{b}_r \circ \boldsymbol{c}_r$ is rank-1 tensor, and the scalar λ_r is the size of rank-1 tensor. The rank-1 tensor indicates the triadic relationship described above, and CP model suppose that CCI-tensor is approximated by the summation of R rank-1 tensor. There are some algorithms for optimizing CP decomposition problem such as alternative least squares (ALS) or power method [58]. Despite its wide use, CP decomposition has a drawback when using the problem in this work; the number of columns of three factor matrices must be a common number R, and the correspondence of $\boldsymbol{A}^{(1)}_{:r}$, $\boldsymbol{A}^{(2)}_{:r}$, and $\boldsymbol{A}^{(3)}_{:r}$ in each r is one-to-one. This constraint is sometimes too strict and unnatural for biological applications. For example, in the CCI-tensor case, the number of ligand expression, receptor expression, and L-R-pair patterns are commonly R, and all of them must correspond to each other in each r. Thus, if an L-R pair assigned in a CCI, this L-R pair cannot be part of other CCIs.

To deal with this problem, the application of Tucker decomposition can be considered next. In Tucker decomposition, a CCI-tensor is decomposed as follows:

$$\begin{aligned}\chi &= \mathcal{G} \times_1 \boldsymbol{A}^{(1)} \times_2 \boldsymbol{A}^{(2)} \times_3 \boldsymbol{A}^{(3)} \\ &= \sum_{r1=1}^{R1} \sum_{r2=1}^{R2} \sum_{r3=1}^{R3} \mathcal{G}(r1, r2, r3) \boldsymbol{A}^{(1)}_{:r1} \circ \boldsymbol{A}^{(2)}_{:r2} \circ \boldsymbol{A}^{(3)}_{:r3} \\ &\text{subject to } \|\boldsymbol{A}^{(1)}_{:r1}\| = \|\boldsymbol{A}^{(2)}_{:r2}\| = \|\boldsymbol{A}^{(3)}_{:r3}\| = 1,\end{aligned} \tag{4}$$

where $R1$, $R2$, and $R3$ are the rank of mode-1,2, and 3, respectively. Unlike the CP model, the constraint conditions of the Tucker model are relaxed, that is, three factor matrices $\boldsymbol{A}^{(1)} \in \mathbb{R}^{J \times R1}$, $\boldsymbol{A}^{(2)} \in \mathbb{R}^{J \times R2}$, and $\boldsymbol{A}^{(3)} \in \mathbb{R}^{K \times R3}$ can differ in their numbers of columns and any combination of $\boldsymbol{A}^{(1)}_{:r1}$, $\boldsymbol{A}^{(2)}_{:r2}$, and $\boldsymbol{A}^{(3)}_{:r3}$ can be considered. This is because $\mathcal{G} \in \mathbb{R}^{R1 \times R2 \times R3}$ is a dense core tensor and any element, including non-diagonal elements, can have a non-zero value. There are some algorithms for optimizing Tucker decomposition, such as higher order singular value decomposition (HOSVD) or higher orthogonal iteration of tensors (HOOI) [58].

Non-negative Tucker decomposition

Despite its effectiveness, Tucker decomposition cannot be directly applied to the extraction of *CaH*s. This is because the factor matrices of the Tucker model can have negative value elements, and these make interpretation difficult. For example, if $\boldsymbol{A}^{(1)}_{:r1}$ contains very large positive elements and very small negative elements, we cannot determine which cell type is highly related to a ligand expression pattern and which cell type is not.

For the above reason, here we utilize NTD. Unlike Tucker decomposition based on singular value decomposition (SVD), NTD is based on non-negative matrix factorization (NMF), which is another matrix decomposition method. NMF is formalized as follows:

$$\begin{aligned} &\boldsymbol{X} = \boldsymbol{W}\boldsymbol{H} \\ &\text{subject to } \boldsymbol{W} \geq 0,\ \boldsymbol{H} \geq 0. \end{aligned} \tag{5}$$

The typical algorithm for optimizing the NMF problem is multiplicative updating (MU) [58]. Two widely used forms are considered. The first form is minimization problem of Euclidean distance ($min\|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_{Euclid}$), where $\boldsymbol{H}$ and $\boldsymbol{W}$ are iter-

atively updated by considering Gaussian noise:

$$\begin{aligned} \boldsymbol{H} &\leftarrow \boldsymbol{H} * \frac{\boldsymbol{W}^T\boldsymbol{X}}{\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H}} \\ \boldsymbol{W} &\leftarrow \boldsymbol{W} * \frac{\boldsymbol{X}\boldsymbol{H}^T}{\boldsymbol{W}\boldsymbol{H}\boldsymbol{H}^T}, \end{aligned} \tag{6}$$

where $*$ is the element-wise (Hadamard) product. The second form is a minimization problem of Kullback-Leibler (KL) divergence ($min\|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_{KL}$), where $\boldsymbol{H}$ and $\boldsymbol{W}$ are iteratively updated by considering Poisson noise:

$$\begin{aligned} \boldsymbol{H} &\leftarrow \boldsymbol{H} * \frac{\boldsymbol{W}^T\frac{\boldsymbol{X}}{\boldsymbol{W}\boldsymbol{H}}}{\boldsymbol{W}^T1} \\ \boldsymbol{W} &\leftarrow \boldsymbol{W} * \frac{\frac{\boldsymbol{X}}{\boldsymbol{W}\boldsymbol{H}}\boldsymbol{H}^T}{1\boldsymbol{H}^T}. \end{aligned} \tag{7}$$

These update rules are derived from the element-wise gradient descent with the spatial form of the learning rate [58]. Starting with a random non-negative initial value, update of $\boldsymbol{W}$ and $\boldsymbol{H}$ are updated iteratively until convergence. In this work, MU with the KL-form, which shows a stable convergence with simulation data, is used for following initialization step of NTD (for more details, see Additional File 3).

To extend the KL form of MU to NTD, we consider iterative updating $\boldsymbol{A}^{(n)}\mathcal{G}_n\boldsymbol{A}^{(-n)T}$, which is the matricized expression of Tucker decomposition. Here $\boldsymbol{A}^{(-n)}$ represents the factor matrices without $\boldsymbol{A}^{(n)}$. For example, if n=1, this part becomes $\boldsymbol{A}^{(2)T}\boldsymbol{A}^{(3)T}$. By considering a part of $\boldsymbol{A}^{(n)}\mathcal{G}_n\boldsymbol{A}^{(-n)T}$ as a variable and fixing other parts as constants, the KL form of MU can be performed to the matricized tensor, such as $\boldsymbol{A}^{(1)} \rightarrow \boldsymbol{A}^{(2)} \rightarrow \boldsymbol{A}^{(3)} \rightarrow \mathcal{G} \rightarrow \cdots$. Each updating rule for $\boldsymbol{A}^{(n)}$ is as follows:

$$\boldsymbol{A}^{(n)} \leftarrow \boldsymbol{A}^{(n)} * \frac{\frac{\boldsymbol{X}_{(n)}}{\boldsymbol{A}^{(n)}\mathcal{G}_A^{(n)}}\mathcal{G}_A^{(n)T}}{1\boldsymbol{z}^T}. \tag{8}$$

Additionally, the updating rule for core tensor $\mathcal{G}$ is:

$$\begin{aligned} \mathcal{G}^{(n)} &\leftarrow \max\{\chi \times_1 \boldsymbol{A}^{(1)T} \times_2 \boldsymbol{A}^{(2)T} \times_3 \boldsymbol{A}^{(3)T}, \epsilon\} \\ \mathcal{G}^{(n)} &\leftarrow \frac{\chi \times_1 \boldsymbol{A}^{(1)T} \times_2 \boldsymbol{A}^{(2)T} \times_3 \boldsymbol{A}^{(3)T}}{\mathcal{G} \times_1 \boldsymbol{A}^{(1)T}\boldsymbol{A}^{(1)} \times_2 \boldsymbol{A}^{(2)T}\boldsymbol{A}^{(2)} \times_3 \boldsymbol{A}^{(3)T}\boldsymbol{A}^{(3)}}, \end{aligned} \tag{9}$$

where ϵ is a small value included to avoid generating negative values. In the `nnTensor`, 1e-10 is used.

Extraction of CCIs as hypergraphs

To extract the representative *CaH*s, `scTensor` estimates the NTD ranks by SVD performed for each matricized CCI-tensor ($X^{(n)}$, n=1,2, and 3). The eigenvalues and eigenvectors that explain the top 80% to 90% of variance are selected. With the estimated ranks of NTD ($\hat{R}1, \hat{R}2, \hat{R}3$), NTD is performed, and only the triads (r1,r2,r3) with large core tensor values are selected as representative *CaH*s. In its default mode, `scTensor` selects the *CaH*s that explain the top 40% of cumulative core tensor values. For each *CaH*, corresponding column vectors of factor matrices were selected as $CaH(\mathrm{r1}, \mathrm{r2}, \mathrm{r3}) = A^{(1)}_{:r1} \circ A^{(2)}_{:r2} \circ A^{(3)}_{:r3}$.

To enhance the interpretation, each column vector is binarized in advance by two-class hierarchical clustering using Ward's minimum variance method, and only large values are converted to 1, with other values becoming 0.

CCI-strength (cf. Figure 3, 4, 5, 6, 7) is calculated as the summation of mode-3 of reconstructed tensor from all *CaH*s. With the selected indexes in each mode ($r1' \in \{1..\hat{R}1\}, r2' \in \{1..\hat{R}2\}, r3' \in \{1..\hat{R}3\}$), CCI-strength is defined as follows:

$$\text{CCI-strength(i, j)} = \sum_{k=1}^{K} \left[\sum_{r1=1}^{\hat{R}1} \sum_{r2=1}^{\hat{R}2} \sum_{r3=1}^{\hat{R}3} \mathcal{G}(r1, r2, r3) \boldsymbol{A}^{(1)}_{:r1} \circ \boldsymbol{A}^{(2)}_{:r2} \circ \boldsymbol{A}^{(3)}_{:r3} \right]_{ijk} \tag{10}$$

Label permutation method

In this method, the cluster labels of all cells are randomly permuted 1000 times, and the average ligand expression level of a cluster and the average receptor expression level of a cluster are calculated [37]. For each L-R pair, the mean values of the averaged L-R expression level are calculated in all possible combinations of the cell types. This process generates 1,000 of synthetic L-R coexpression matrices and these are used to generate the null distribution, that is, in a combination of cell types, the proportion of the means which are "as or more extreme" than the observed mean is the calculated as P-value.

Availability and requirements

- `scTensor`: https://bioconductor.org/packages/devel/bioc/html/scTensor.html
- `nnTensor`: https://cran.r-project.org/web/packages/nnTensor/index.html
- `LRBase.Hsa.eg.db`: https://bioconductor.org/packages/release/bioc/html/LRBase.Hsa.eg.db.html
- `LRBase.Mmu.eg.db`: https://bioconductor.org/packages/release/bioc/html/LRBase.Mmu.eg.db.html
- `LRBase.Ath.eg.db`: https://bioconductor.org/packages/release/bioc/html/LRBase.Ath.eg.db.html
- `LRBase.Rno.eg.db`: https://bioconductor.org/packages/release/bioc/html/LRBase.Rno.eg.db.html
- `LRBase.Bta.eg.db`: https://bioconductor.org/packages/release/bioc/html/LRBase.Bta.eg.db.html
- `LRBase.Cel.eg.db`: https://bioconductor.org/packages/release/bioc/html/LRBase.Cel.eg.db.html
- `LRBase.Dme.eg.db`: https://bioconductor.org/packages/release/bioc/html/LRBase.Dme.eg.db.html

- LRBase.Dre.eg.db: https://bioconductor.org/packages/release/bioc/html/LRBase.Dre.eg.db.html
- LRBase.Gga.eg.db: https://bioconductor.org/packages/release/bioc/html/LRBase.Gga.eg.db.html
- LRBase.Pab.eg.db: https://bioconductor.org/packages/release/bioc/html/LRBase.Pab.eg.db.html
- LRBase.Xtr.eg.db: https://bioconductor.org/packages/release/bioc/html/LRBase.Xtr.eg.db.html
- LRBase.Ssc.eg.db: https://bioconductor.org/packages/release/bioc/html/LRBase.Ssc.eg.db.html
- LRBaseDbi: https://bioconductor.org/packages/release/bioc/html/LRBaseDbi.html
- Operating system: Linux, Mac OS X, Windows
- Programming language: R (v-3.5.0 or higher)
- License: Artistic-2.0
- Any restrictions to use by non-academics: For non-profit use only

Abbreviations

CCI: cell-cell interaction; scRNA-seq: single-cell RNA sequencing; L-R: ligand and receptor; *CaH*: CCI as hypergraph; FTT: Freeman-Tukey transformation; NTD: non-negative Tucker decomposition; ROC: receiver operating characteristic; AUC: area under the curve; FGCs: fetal germ cells; Soma: gonadal niche cells; GPCR: G protein-coupled receptor; CAFs: cancer-associated fibroblasts; MHC: major histocompatibility complex; HLA: human leukocyte antigen; RNA-FISH: RNA-fluorescence $in\ situ$ hybridization; PM: plasma membrane; PPI: protein-protein interaction; GO: Gene Ontology; MeSH: medical subject headings; ORA: over-representative analysis; DO: Disease Ontology; NCG: network of cancer gene; CMap: Connectivity Map; Amazon S3: Amazon simple storage service; FC: fold-change; DEGs: differentially expressed genes; human embryonic stem cells: hESCs; HVGs: highly variable genes; t-SNE: t-distributed stochastic neighbor embedding; CP: CANDECOMP/PARAFAC; ALS: alternative least squares; HOSVD: higher order singular value decomposition; HOOI: higher orthogonal iteration of tensors; NMF: non-negative matrix factorization; MU: multiplicative updating; KL: Kullback-Leibler

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by MEXT KAKENHI Grant Number 16K16152 and by JST CREST grant number JPMJCR16G3, Japan.

Authors' contributions

KT and IN designed the study. KT designed the algorithm and benchmark, retrieved and preprocessed the test data to evaluate the proposed method, implemented the source code, and performed all analyses. MI implemented the pipeline for bi-annual automatic update of the R/Bioconductor packages. All authors have read and approved the manuscript.

Acknowledgements

Some cell images used in Figure 1 were previously presented by © 2016 DBCLS TogoTV. We thank Dr. Yoshihiro Taguchi for discussions about the algorithms. We thank Mr. Akihiro Matsushima for their assistance with the IT infrastructure for the data analysis. We are also grateful to all members of the Laboratory for Bioinformatics Research, RIKEN Center for Biosystems Dynamics Research for their helpful advice.

Author details

[1]Laboratory for Bioinformatics Research RIKEN Center for Biosystems Dynamics Research, Wako, 351-0198, Saitama, Japan. [2]Bioinformatics Course, Master's/Doctoral Program in Life Science Innovation (T-LSI), School of Integrative and Global Majors (SIGMA), University of Tsukuba, Wako, 351-0198, Saitama, Japan.

References

1. Yu, Y., Elble, R.C.: Homeostatic signaling by cell–cell junctions and its dysregulation during cancer progression. Journal of Clinical Medicine **5(2)**, 26 (2016)
2. Livshits, G., Kobielak, A., Fuchs, E.: Governing epidermal homeostasis by coupling cell–cell adhesion to integrin and growth factor signaling, proliferation, and apoptosis. PNAS **109(3)**, 4886–4891 (2012)
3. Chao, D.L., Ma, L., Shen, K.: Transient cell-cell interactions in neural circuit formation. Nature Review Neuroscience, 262–271 (2009)
4. Kasukawa, T., Masumoto, K., Nikaido, I., Nagano, M., Uno, K.D., Tsujino, K., Hanashima, C., Shigeyoshi, Y., Ueda, H.R.: Quantitative expression profile of distinct functional regions in the adult mouse brain. PLOS ONE, 23228 (2011)
5. Miller, J.F.A.P., Mitchell, G.F.: Cell to cell interaction in the immune response v. target cells for tolerance induction. Journal of Experimental Medicine **131(4)**, 675–699 (1970)
6. Pieters, T., Roy, V.F.: Role of cell–cell adhesion complexes in embryonic stem cell biology. Journal of Cell Science **127**, 2603–2613 (2014)
7. Tweedell, K.S.: The adaptability of somatic stem cells: A review. Journal of Stem Cells and Regenerative Medicine **13(1)**, 3–13 (2017)
8. Plaks, V., Kong, N., Werb, Z.: The cancer stem cell niche: How essential is the niche in regulating stemness of tumor cells? Cell Stem Cell **16**, 225–238 (2015)
9. Hegerfeldt, Y., Tusch, M., Brocker, E.B., Friedl, P.: Collective cell movement in primary melanoma explants: Plasticity of cell-cell interaction, beta1-integrin function, and migration strategies. Cancer Research **62**, 2125–2130 (2002)
10. Hofschroer, V., Koch, K.A., Ludwig, F.T., Friedl, P., Oberleithner, H., Stock, C., Schwab, A.: Extracellular protonation modulates cell-cell interaction mechanics and tissue invasion in human melanoma cells. Scientific Reports **7(42369)** (2017)
11. Stein, J.V., Gonzalez, S.F.: Dynamic intravital imaging of cell-cell interactions in the lymph node. Mechanisms of allergic diseases **139(1)**, 12–20 (2016)
12. Reinhar-King, C.A., Dembo, M., Hammer, D.A.: Cell-cell mechanical communication through compliant substrates. Biophysical Journal **95**, 6044–6051 (2008)

13. Dewji, N.N., Mukhopadhyay, D., Singer, S.J.: An early specific cell–cell interaction occurs in the production of beta-amyloid in cell cultures. PNAS **103(5)**, 1540–1545 (2006)
14. Konry, T., Sarkar, S., Sabhachandani, P., Cohen, N.: Innovative tools and technology for analysis of single cells and cell–cell interaction. The Annual Review of Biomedical Engineering **18**, 259–284 (2016)
15. Rothbauer, M., Zirath, H., Ertl, P.: Recent advances in microfluidic technologies for cell-to-cell interaction studies. Lab on Chip **18(2)**, 249–270 (2018)
16. Li, R., Lv, X., Zhang, X., Saeed, O., Deng, Y.: Microfluidics for cell-cell interactions: A review. 10(1) **10(1)**, 90–98 (2016)
17. Wiklund, M., Christakou, A.E., Ohlin, M., Iranmanesh, I., Frisk, T., Vanherberghen, V., Onfelt, B.: Ultrasound-induced cell–cell interaction studies in a multi-well microplate. Micromachines **5**, 27–49 (2014)
18. Tauriainen, J., Gustafsson, K., Gothlin, M., Gertow, J., Buggert, M., Frisk, T.W., Karlsson, A.C., Uhlin, M., Onfelt, B.: Single-cell characterization of in vitro migration and interaction dynamics of t cells expanded with il-2 and il-7. Frontiers in Immunology **6**, 196 (2015)
19. Merouane, A., Rey-Villamizar, N., Lu, Y., Liadi, I., Romain, G., Lu, J., Singh, H., Cooper, L.J.N., Varadarajan, N., Roysam, B.: Automated profiling of individual cell–cell interactions from high-throughput time-lapse imaging microscopy in nanowell grids (timing). Bioinformatics **31(19)**, 3189–3197 (2015)
20. Espulgar, W., Yamaguchi, Y., Aoki, W., Mita, D., Saito, M., Lee, J.K., Tamiya, E.: Single cell trapping and cell–cell interaction monitoring of cardiomyocytes in a designed microfluidic chip. Sensors and Actuators B: Chemical **207**, 43–50 (2015)
21. Sarkar, S., Sabhachandani, P., Stroopinsky, D., Palmer, K., Cohen, N., Rosenblatt, J., Avigan, D., Konry, T.: Dynamic analysis of immune and cancer cell interactions at single cell level in microfluidic droplets. Biomicrofluidics **10(5)**, 054115 (2016)
22. Dura, B., Dougan, S.K., Barisa, M., Hoehl, M.M., Lo, C.T., Ploegh, H.L., Voldman, J.: Profiling lymphocyte interactions at the single-cell level by microfluidic cell pairing. Nature Communication **6(5940)** (2015)
23. Ramilowski, J.A., Goldberg, T., Harshbarger, J., Kloppmann, E., Lizio, M., Satagopam, V.P., Itoh, M., Kawaji, H., Carninci, P., Rost, B., Forrest, A.R.R.: A draft network of ligand–receptor-mediated multicellular signalling in human. Nature Communication **22(6)**, 7866 (2015)
24. Camp, J.G., Sekin, K., Gerber, T., Loeffler-Wirth, H., Binder, H., Gac, M., Kanton, S., Kageyama, J., Damm, G., Seehofer, D., Belicova, L., Bickle, M., Barsacchi, R., Okuda, R., Yoshizawa, E., Kimura, M., Ayabe, H., Taniguchi, H., Takebe, T., Treutlein, B.: Multilineage communication regulates human liver bud development from pluripotency. Nature **546(7659)**, 533–538 (2017)
25. Li, L., Dong, J., Yan, L., Yong, J., Liu, X., Hu, Y., Fan, X., Wu, X., Guo, H., Wang, X., Zhu, X., Li, R., Yan, J., Wei, Y., Zhao, Y., Wang, W., Ren, Y., Yuan, P., Yan, Z., Hu, B., Guo, F., Wen, L., Tang, F., Qiao, J.: Single-cell rna-seq analysis maps development of human germline cells and gonadal niche interactions. Cell Stem Cell **20**, 858–873 (2017)
26. Zhou, J.X., Taramelli, R., Pedrini, E., Knijnenburg, T., Hunag, S.: Extracting intercellular signaling network of cancer tissues using ligand-receptor expression patterns from whole-tumor and single-cell transcriptomes. Scientific Reports **7(1)**, 8815 (2017)
27. A, S.D., Squiers, G.T., McLellan, M.A., Bolisetty, M.T., Robson, P., Rosenthal, N.A., Pinto, A.R.: Single-cell transcriptional profiling reveals cellular diversity and intercommunication in the mouse heart. Cell Reports **22(3)**, 600–610 (2018)
28. Pavlicev, M., Wagner, G.P., Chavan, A.R., Owens, K., Maziarz, J., Dunn-Fletcher, C., Lallapur, S.G., Muglia, L., Jones, H.: Single-cell transcriptomics of the human placenta: inferring the cell communication network of

the maternal-fetal interface. Genome Research **27**, 349–361 (2017)

29. Joost, S., Jacob, T., Sun, X., Annusver, K., La Manno, G., Sur, I., Kasper, M.: Single-cell transcriptomics of traced epidermal and hair follicle stem cells reveals rapid adaptations during wound healing. Cell Reports **25(3)**, 585–597 (2018)
30. Kramann, R., Machado, F., Wu, H., Kusaba, T., Hoeft, K., Schneider, R.K., Humphreys, B.D.: Parabiosis and single-cell rna sequencing reveal a limited contribution of monocytes to myofibroblasts in kidney fibrosis. JCI insight **3(9)**, 99561 (2018)
31. Cohen, M., Giladi, A., Gorki, A.D., Solodkin, D.G., Zada, M., Hladik, A., Miklosi, A., Salame, T.M., Halpern, K.B., David, E., Itzkovitz, S., Harkany, T., Knapp, S., Amit, I.: Lung single-cell signaling interaction map reveals basophil role in macrophage imprinting. Cell **175(4)**, 1031–1044 (2018)
32. Davidson, S., Efremova, M., Riedel, A., Mahata, B., Pramanik, J., Huuhtanen, J., Kar, G., Vento-Tormo, R., Hagai, T., Chen, X., Haniffa, M.A., Shields, J.D., Teichmann, S.A.: Single-cell rna sequencing reveals a dynamic stromal niche within the evolving tumour microenvironment. bioRxiv (2018). doi:10.1101/467225
33. Potter, S.S., Mucenski, M.L., Mahoney, R., Adam, M., Potter, A.S.: Single cell rna-seq study of wild type and hox9,10,11 mutant developing uterus. bioRxiv (2018). doi:10.1101/395574
34. Wu, H., Uchimura, K., Donnelly, E.L., Kirita, Y., Morris, S.A., Humphreys, B.D.: Comparative analysis and refinement of human psc-derived kidney organoid differentiation with single-cell transcriptomics. Cell Stem Cell **23(6)**, 869–881 (2018)
35. Chen, L., Lee, J.W., Chou, C.L., Nair, A.V., Battistone, M.A., Paunescu, T.G., Merkulova, M., Breton, S., Verlander, J.W., Wall, S.M., Brown, D., Burg, M.B., Knepper, M.A.: Transcriptomes of major renal collecting duct cell types in mouse identified by single-cell rna-seq. PNAS **114(46)**, 9989–9998 (2017)
36. Menon, R., Otto, E.A., Kokoruda, A., Zhou, J., Zhang, Z., Yoon, E., Chen, Y.C., Troyanskaya, O., Spence, J.R., Kretzler, M., Cebrian, C.: Single-cell analysis of progenitor cell dynamics and lineage specification in the human fetal kidney. Development **145(16)** (2018)
37. Vento-Tormo, R., Efremova, M., Botting, R.A., Turco, M.Y., Vento-Tormo, M., Meyer, K.B., Park, J.E., Stephenson, E., Polanski, K., Goncalves, A., Gardner, L., Holmqvist, S., Henriksson, J., Zou, A., Sharkey, A.M., Millar, B., Innes, B., Wood, L., Wilbrey-Clark, A., Payne, R.P., Ivarsson, M.A., Lisgo, S., Filby, A., Rowitch, D.H., Bulmer, J.N., Wright, G.J., Stubbington, M.J.T., Haniffa, M., Moffett, A., Teichmann, S.A.: Single-cell reconstruction of the early maternal–fetal interface in humans. Nature **563(7731)**, 347–353 (2018)
38. Biton, M., Haber, A.L., Rogel, N., Burgin, G., Beyaz, S., Schnell, A., Ashenberg, O., Su, C.W., Smillie, C., Shekhar, K., Chen, Z., Wu, C., Ordovas-Montanes, J., Alvarez, D., Herbst, R.H., Zhang, M., Tirosh, I., Dionne, D., Nguyen, L.T., Xifaras, M.E., Shalek, A.K., von Andrian, U.H., Graham, D.B., Rozenblatt-Rosen, O., Shi, H.N., Kuchroo, V., Yilmaz, O.H., Regev, A., Xavier, R.J.: T helper cell cytokines modulate intestinal stem cell renewal and differentiation. Cell **175(5)**, 1307–1320 (2018)
39. Kumar, M.P., Du, J., Lagoudas, G., Jiao, Y., Sawyer, A., Drummond, D.C., Lauffenburger, D.A., Raue, A.: Analysis of single-cell rna-seq identifies cell-cell communication associated with tumor characteristics. Cell Reports **25(6)**, 1458–1468 (2018)
40. Single cell analyses of the effects of amyloid-beta42 and interleukin-4 on neural stem/progenitor cell plasticity in adult zebrafish brain **XX(X)**, (20XX)
41. Verma, M., Asakura, Y., Murakonda, B.S.R., Pengo, T., Latroche, C., Chazaud, B., McLoon, L.K., Asakura, A.: Muscle satellite cell cross-talk with a vascular niche maintains quiescence via vegf and notch signaling. Cell Stem Cell **23(4)**, 530–543 (2018)
42. Jerby-Arnon, L., Shah, P., Cuoco, M.S., Rodman, C., Su, M.J., Melms, J.C., Leeson, R., Kanodia, A., Mei, S.,

Lin, J.R., Wang, S., Rabasha, B., Liu, D., Zhang, G., Margolais, C., Ashenberg, O., Ott, P.A., Buchbinder, E.I., Haq, R., Hodi, F.S., Boland, G.M., Sullivan, R.J., Frederick, D.T., Miao, B., Moll, T., Flaherty, K.T., Herlyn, M., Jenkins, R.W., Thummalapalli, R., Kowalczyk, M.S., Canadas, I., Schilling, B., Cartwright, A.N.R., Luoma, A.M., Malu, S., Hwu, P., Bernatchez, C., Forget, M.A., Barbie, D.A., Shalek, A.K., Tirosh, I., Sorger, P.K., Wucherpfennig, K., Van Allen, E.M., Schadendorf, D., Johnson, B.E., Rotem, A., Rozenblatt-Rosen, O., Garraway, L.A., Yoon, C.H., Izar, B., Regev, A.: A cancer cell program promotes t cell exclusion and resistance to checkpoint blockade. Cell **175(4)**, 984–997 (2018)

43. Kelleher, A.M., Milano-Foster, J., Behura, S.K., Spencer, T.E.: Uterine glands coordinate on-time embryo implantation and impact endometrial decidualization for pregnancy success. Nature Communications **9(1)**, 2435 (2018)
44. Yin, J., Li, Z., Yan, C., Fang, E., Wang, T., Zhou, H., Luo, W., Zhou, Q., Zhang, J., Hu, J., Jin, H., Wang, L., Zhao, X., Li, J., Qi, X., Zhou, W., Huang, C., He, C., Yang, H., Kristiansen, K., Hou, Y., Zhu, S., Zhou, D., Wang, L., Dean, M., Wu, K., Hu, H., Li, G.: Comprehensive analysis of immune evasion in breast cancer by single-cell rna-seq. bioRxiv (2018). doi:10.1101/368605
45. Biase, F.H., Kimble, K.M.: Functional signaling and gene regulatory networks between the oocyte and the surrounding cumulus cells. BMC Genomics **19(1)**, 351 (2018)
46. Thorsson, V., Gibbs, D.L., Brown, S.D., Wolf, D., Bortone, D.S., Ou Yang, T.H., Porta-Pardo, E., Gao, G.F., Plaisier, C.L., Eddy, J.A., Ziv, E., Culhane, A.C., Paull, E.O., Sivakumar, I.K.A., Gentles, A.J., Malhotra, R., Farshidfar, F., Colaprico, A., Parker, J.S., Mose, L.E., Vo, N.S., Liu, J., Liu, Y., Rader, J., Dhankani, V., Reynolds, S.M., Bowlby, R., Califano, A., Cherniack, A.D., Anastassiou, D., Bedognetti, D., Rao, A., Chen, K., Krasnitz, A., Hu, H., Malta, T.M., Noushmehr, H., Pedamallu, C.S., Bullman, S., Ojesina, A.I., Lamb, A., Zhou, W., Shen, H., Choueiri, T.K., Weinstein, J.N., Guinney, J., Saltz, J., Holt, R.A., Rabkin, C.E., Network, C.G.A.R., Lazar, A.J., Serody, J.S., Demicco, E.G., Disis, M.L., Vincent, B.G., Shmulevich, L.: The immune landscape of cancer. Immunity **48(4)**, 812–830 (2018)
47. Han, X., Chen, H., Huang, D., Chen, H., Fei, L., Cheng, C., Huang, H., Yuan, G.C., Guo, G.: Mapping human pluripotent stem cell differentiation pathways using high throughput single-cell rna-sequencing open access. BMC Genome Biology **19(1)**, 47 (2018)
48. Costa, A., Kieffer, Y., Scholer-Dahirel, A., Pelon, F., Bourachot, B., Cardon, M., Sirven, P., Magagna, I., Fuhrmann, L., Bernard, C., Bonneau, C., Kondratova, M., Kuperstein, I., Zinovyev, A., Givel, A.M., Parrini, M.C., Soumelis, V., Vincent-Salomon, A., Mechta-Grigoriou, F.: Fibroblast heterogeneity and immunosuppressive environment in human breast cancer. Cancer Cells **33(3)**, 463–479 (2018)
49. Hrvatin, S., Hochbaum, D.R., Nagy, M.A., Cicconet, M., Robertson, K., Cheadle, L., Zilionis, R., Ratner, A., Borges-Monroy, R., Klein, A.M., Sabatini, B.L., Greenberg, M.E.: Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. Nature Neuroscience **21(1)**, 120–129 (2018)
50. Suryawanshi, H., Morozov, P., Straus, A., Sahasrabudhe, N., Max, K.E.A., Garzia, A., Kustagi, M., Tuschl, T., Z, W.: A single-cell survey of the human first-trimester placenta and decidua. Science Advances **4(10)**, 4788 (2018)
51. Puram, S.V., Tirosh, I., Parikh, A.S., Patel, A.P., Yizhak, K., Gillespie, S., Rodman, C., Luo, C.L., Mroz, E.A., Emerick, K.S., Deschler, D.G., Varvares, M.A., Mylvaganam, R., Rozenblatt-Rosen, O., Rocco, J.W., Faquin, W.C., Lin, D.T., Regev, A., Bernstein, B.E.: Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. Cell **171(7)**, 1611–1624 (2017)
52. Ximerakis, M., Lipnick, S.L., Simmons, S.K., Adiconis, X., Innes, B.T., Dionne, D., Nguyen, L., Mayweather, B.A., Ozek, C., Niziolek, Z., Butty, V.L., Isserlin, R., Buchanan, S.M., Levine, S.R., Regev, A., Bader, G.D.,

Levin, J.Z., Rubin, L.L.: Single-cell transcriptomics of the aged mouse brain reveals convergent, divergent and unique aging signatures. bioRxiv (2018). doi:10.1101/440032

53. Sivakamasundari, V., Bolisetty, M., Sivajothi, S., Bessonett, S., Ruan, D., Robson, P.: Comprehensive cell type specific transcriptomics of the kidney. bioRxiv (2017). doi:10.1101/238063
54. Yuanxin, W., Ruiping, W., Shaojun, Z., Shumei, S., Changying, J., Guangchun, H., Michael, W., Jaffer, A., Andy, F., Wang, L.: italk: an r package to characterize and illustrate intercellular communication. bioRxiv (2019). doi:10.1101/507871
55. Boisset, J.C., Vivie, J., Murano, M.J., Lyubimova, A., van-Oudenaarden, A.: Mapping the physical network of cellular interactions. Nature Methods (2018)
56. Freeman, M.F., Tukey, J.W.: Transformations related to the angular and the square root. The Annals of Mathematical Statistics **4(21)**, 607–611 (1950)
57. Kim, Y.-D., Choi, S.: Nonnegative tucker decomposition. In: In IEEE Conference on Computer Vision and Pattern Recognition (2007)
58. Cichocki, A., Zdunek, R., Amari, S.: Nonnegative Matrix and Tensor Factorizations. IEEE Signal Processing Magazine, ??? (2008)
59. Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H. II, Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., Fallahi-Sichani, M., Dutton-Regester, K., , L. Jia-Ren, Cohen, O., Shah, P., Lu, D., Genshaft, A.S., Hughes, T.K., Ziegler, C.G.K., Kazer, S.W., Gaillard, A., Kolb, K.E., Villani, A.C., Johannessen, C.M., Andreev, A.Y., Van Allen, E.M., Bertagnolli, M., Sorger, P.K., Sullivan, R.J., Flaherty, K.T., Frederick, D.T., Jane-Valbuena, J., Yoon, C.H., ROzenblatt-Rosen, O., Shalek, A.K., Regev, A., Garrayway, L.A.: Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. Science **352(6282)**, 189–196 (2016)
60. Pandey, S., Shekhar, K., Regev, A., Schier, A.F.: Comprehensive identification and spatial mapping of habenular neuronal types using single-cell rna-seq. Current Biology **28**, 1052–1065 (2018)
61. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., Regev, A.: Spatial reconstruction of single-cell gene expression data. Nature Computational Biology **33(5)**, 495–502 (2015)
62. Svensson, V., Teichmann, S.A., Stegle, O.: Spatialde: identification of spatially variable genes. Nature methods **15**, 343–346 (2018)
63. Durinck, S., Spellman, P., Birney, E., Huber, W.: Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. Nature Protocols **4**, 1184–1191 (2009)
64. Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Roca, C.D., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Viteri, G., Weiser, J., Wu, G., Stein, L., Hermjakob, H., D'Eustachio, P.: The reactome pathway knowledgebase. Nucleic Acids Research **46(D1)**, 649–655 (2018)
65. Tsuyuzaki, K., Morota, G., Ishii, M., Nakazato, T., Miyazaki, S., Nikaido, I.: Mesh ora framework: R/bioconductor packages to support mesh over-representation analysis. BMC Bioinformatics **16(45)** (2015)
66. Falcon, S., Gentleman, R.: Using gostats to test gene lists for go term association. Bioinformatics **23(2)**, 257–258 (2007)
67. Yu, G., He, Q.: Reactomepa: an r/bioconductor package for reactome pathway analysis and visualization. Molecular BioSystems **12(12)**, 477–479 (2016)
68. Yu, G., Wang, L., Yan, G., He, Q.: Dose: an r/bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics **31(4)**, 608–609 (2015)
69. Ono, H., Ogasawara, O., Okubo, K., Bono, H.: Refex, a reference gene expression dataset as a web tool for the

functional analysis of genes. Scientific Data **4**, 170105 (2017)

70. Papatheodorou, I., Fonseca, N.A., Keays, M., Tang, Y.A., Barrera, E., Bazant, W., Burke, M., Fullgrabe, A., Fuentes, A.M., George, N., Huerta, L., Koskinen, S., Mohammed, S., Geniza, M., Preece, J., Jaiswal, P., Jarnuczak, A.F., Huber, W., Stegle, O., Vizcaino, J.A., Brazma, A., Petryszak, R.: Expression atlas: gene and protein expression across multiple studies and organisms. Nucleic Acids Research **46(Database issue)**, 246–51 (2018)
71. Single Cell Expression Atlas Single cell gene expression across species. https://www.ebi.ac.uk/gxa/sc/home
72. Cao, Y., Zhu, J., Jia, P., Zhao, Z.: scrnaseqdb: A database for rna-seq based gene expression profiles in human single cells. Genes (Basel) **8(12)**, 368 (2017)
73. PanglaoDB. https://panglaodb.se/index.html
74. Carithers, L.J., Ardlie, K., Barcus, M., Branton, P.A., Britton, A., Buia, S.A., Compton, C.C., DeLuca, D.S., Peter-Demchok, J., Gelfand, E.T., Guan, P., Korzeniewski, G.E., Lockhart, N.C., Rabiner, C.A., Rao, A.K., Robinson, K.L., Roche, N.V., Sawyer, S.J., Segre, A.V., Shive, C.E., Smith, A.M., Sobin, L.H., Undale, A.H., Valentino, K.M., Vaught, J., Young, T.R., Moore, H.M., Consortium., G.: A novel approach to high-quality postmortem tissue procurement: The gtex project. Biopreservation and Biobanking **13(5)**, 311–319 (2015)
75. Collection — 29 August 2017 The FANTOM5 project. https://www.nature.com/collections/jcxddjndxy
76. Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., Farnham, P.J., Hirst, M., Lander, E.S., Mikkelsen, T.S., Thomson, J.A.: The nih roadmap epigenomics mapping consortium. Nature Biotechnology **28(10)**, 1045–1048 (2010)
77. Consotium, E.P.: An integrated encyclopedia of dna elements in the human genome. Nature **489(7414)**, 57–74 (2012)
78. Uhlen, M., Fagerberg, L., Hallstrom, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjostedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C.A., Odeberg, J., Djureinovic, D., Takanen, J.O., Hober, S., Alm, T., Edqvist, P.H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J.M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., F, P.: Proteomics. tissue-based map of the human proteome. Science **347(6220)**, 1260419 (2015)
79. Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.P., Subramanian, A., Ross, K.N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S.A., Haggarty, S.J., Clemons, P.A., Wei, R., Carr, S.A., Lander, E.S., Golub, T.R.: The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. Science **313(5795)**, 1929–1935 (2006)
80. Harding, S.D., Sharman, J.L., Faccenda, E., Southan, C., Pawson, A.J., Ireland, S., Gray, A.J.G., Bruce, L., Alexander, S.P.H., Anderton, S., Bryant, C., Davenport, A.P., Doerig, C., Fabbro, D., Levi-Schaffer, F., Spedding, M., Davies, N.-I. J A: The iuphar/bps guide to pharmacology in 2018: updates and expansion to encompass the new guide to immunopharmacology. Nucleic Acids Research **46(D1)**, 1091–1106 (2018)
81. Graeber, T.G., Eisenberg, D.: Bioinformatic identification of potential autocrine signaling loops in cancers from gene expression profiles. Nature Genetics **29(3)**, 295–300 (2001)
82. Wang, Y., Tung, H.-Y., Smola, A., Anandkumar, A.: Fast and guaranteed tensor decomposition via sketching. In: In NIPS, vol. 1, pp. 991–999 (2015)
83. Maehara, T., Hayashi, K., Kawarabayashi, K.: Expected tensor decomposition with stochastic gradient descent. In: AAAI'16, pp. 1919–1925 (2016)
84. Smith, S., Park, J., Karypis, G.: An exploration of optimization algorithms for high performance tensor completion. In: SC '16 Proceedings of the International Conference for High Performance Computing,

Networking, Storage and Analysis, vol. 31 (2016)

85. Shin, K., Sael, L., Kang, U.: Fully scalable methods for distributed tensor factorization. IEEE Transactions on Knowledge and Data Engineering **29(1)**, 100–113 (2017)
86. Tsuyuzaki, K., Nikaido, I.: Biological systems as heterogeneous information networks: A mini-review and perspectives. HeteroNAM'18 (2018)
87. Sasagawa, Y., Nikaido, I., Hayashi, T., Danno, H., Uno, K.D., Imai, T., Ueda, H.R.: Quartz-seq: a highly reproducible and sensitive single-cell rna sequencing method, reveals non-genetic gene-expression heterogeneity. BMC Genome Biology **14(4)**, 31 (2013)
88. Pierson, E., Yau, C.: Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. BMC Genome Biology **16(241)** (2015)

1 **Figures**

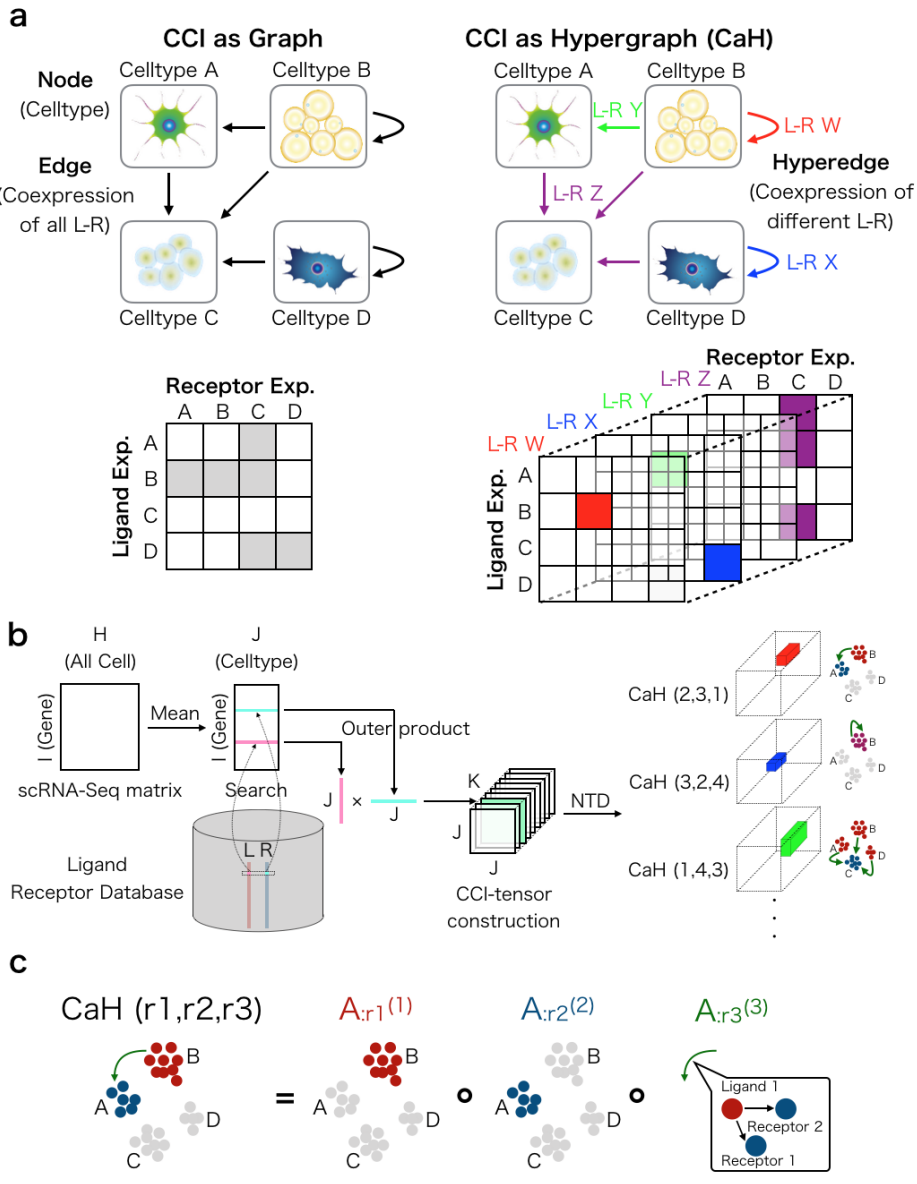

Figure 1 CCI as a hypergraph (a) Previous studies have regarded CCIs as graphs, and the corresponding data structure is expressed as an adjacency matrix (left). In this work, CCIs are regarded as hypergraphs, and the corresponding data structure is a tensor. (b) The CCI-tensor is generated by users' scRNA-Seq matrices, cell-type labels, and L-R databases. NTD is used to extract *CaH*s from the CCI-tensor. (c) Each *CaH*(r1,r2,r3) is equal to the outer product of three vectors. $A^{(1)}_{:r1}$ represents the ligand expression pattern, $A^{(2)}_{:r2}$ represents the receptor expression pattern, and $A^{(3)}_{:r3}$ represents the related L-R pairs pattern.

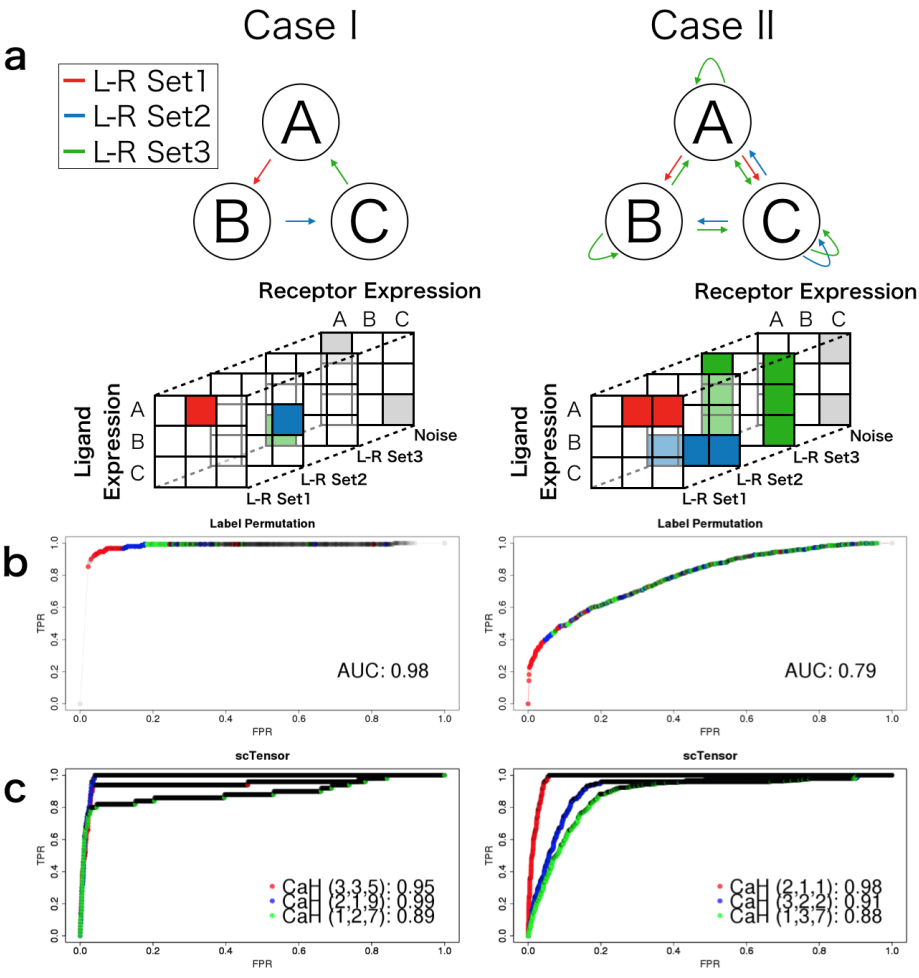

Figure 2 Comparison between the label permutation method and `scTensor` using simulated datasets (a) Two datasets simulating two cases were constructed. In the first case, all CCIs consisted of one-to-one relationships between two cell types. In the second case, CCIs consisted of many-to-many relationships involving many cell types. (b) The ROC curve for the label permutation method. (c) The ROC curves of the *CaH*s detected by `scTensor`.

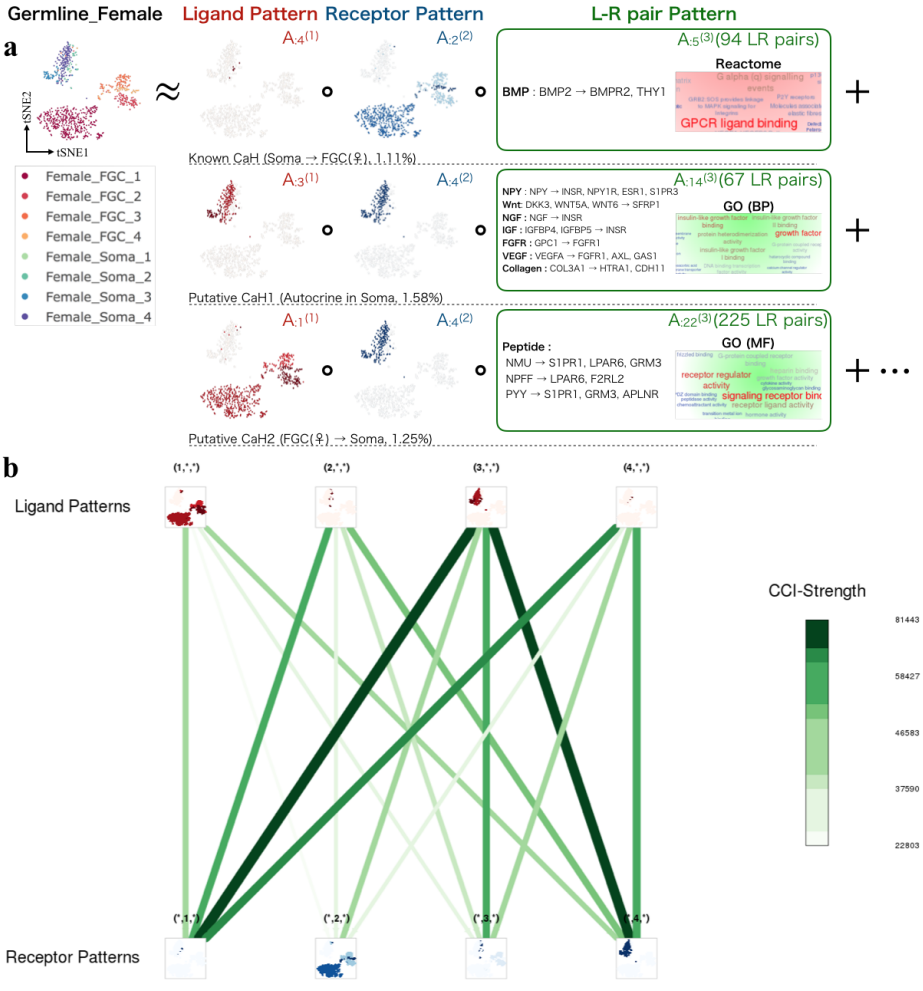

Figure 3 Results of `scTensor` for an empirical dataset (Germline_Female) (a) Detected *CaH*s. (b) Overview of the L-R patterns.

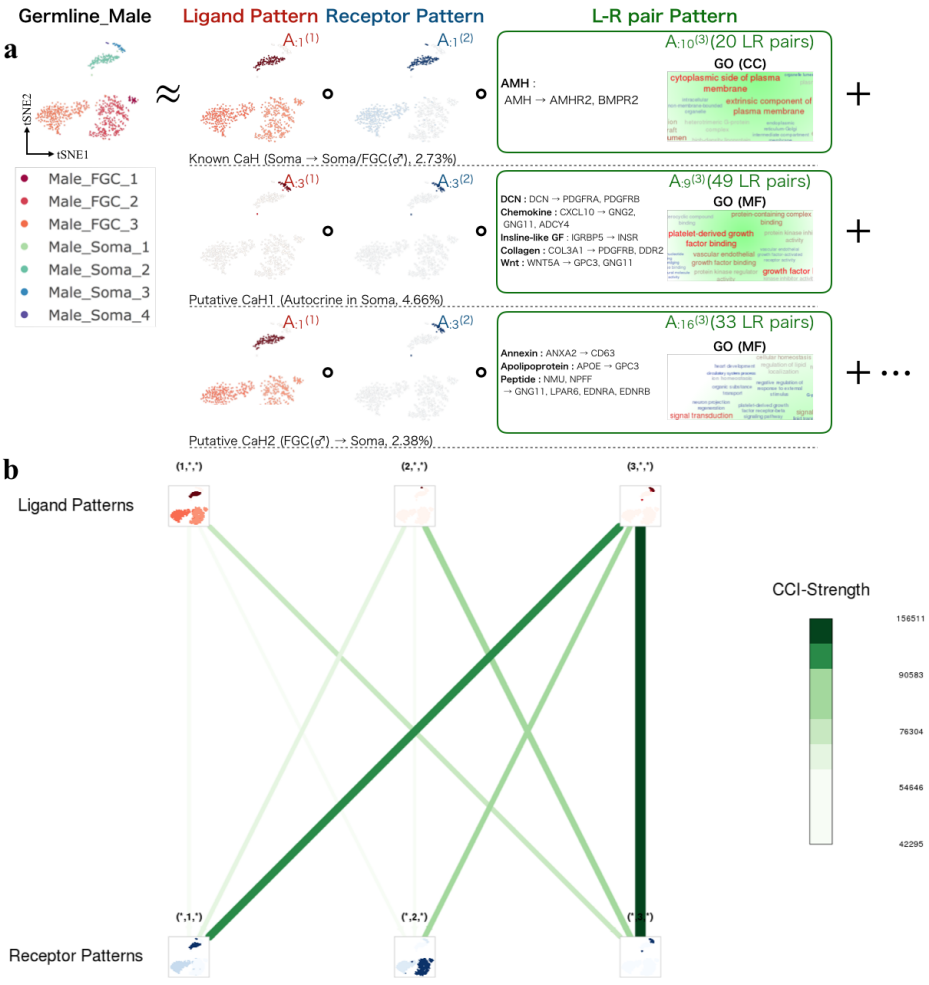

Figure 4 Results of `scTensor` for an empirical dataset (Germline_Male) (a) Detected *CaH*s. (b) Overview of the L-R patterns.

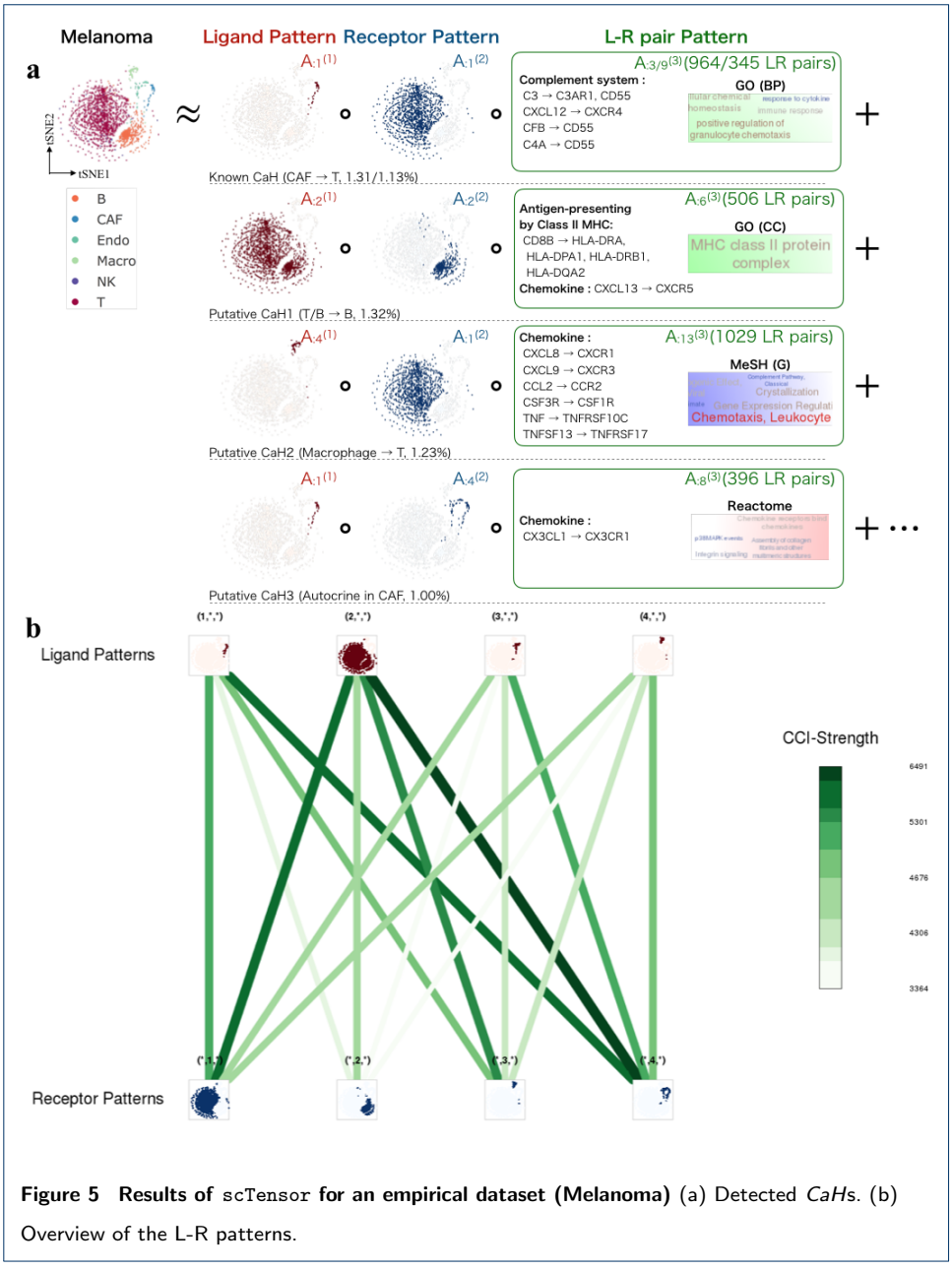

Figure 5 Results of `scTensor` for an empirical dataset (Melanoma) (a) Detected *CaH*s. (b) Overview of the L-R patterns.

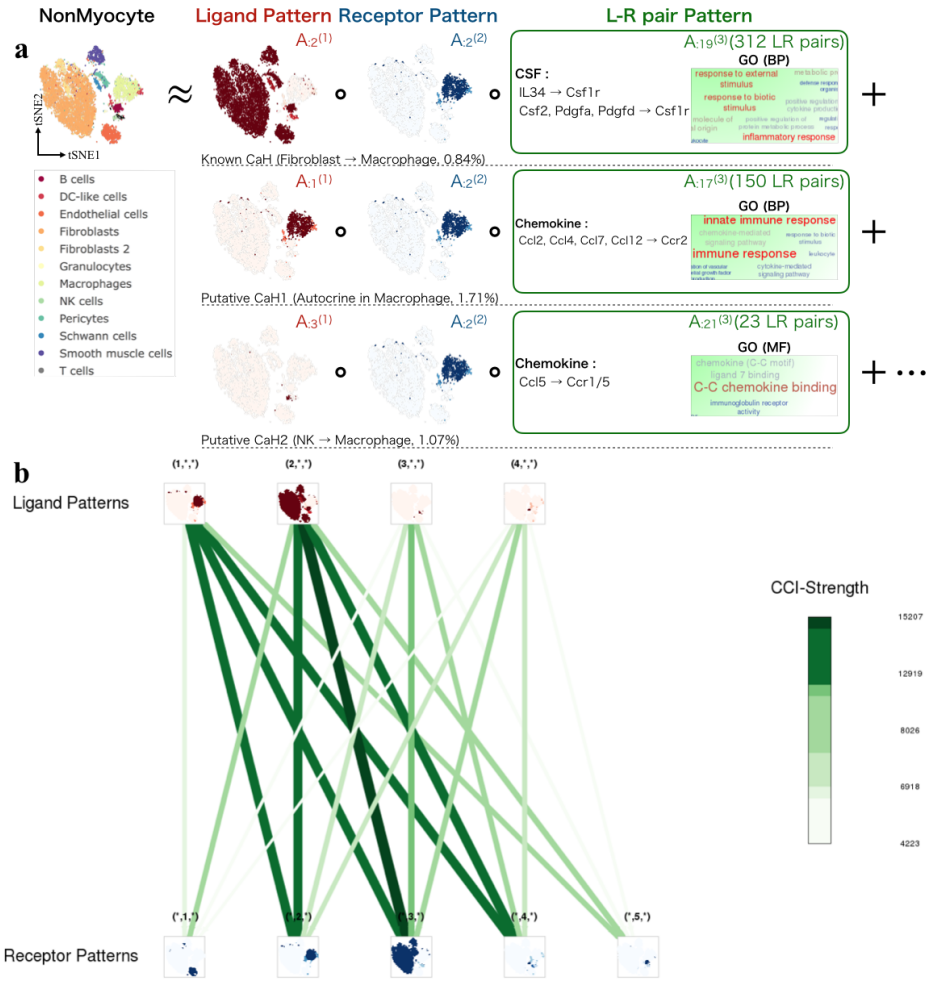

Figure 6 Results of `scTensor` for an empirical dataset (NonMyocyte) (a) Detected *CaH*s. (b) Overview of the L-R patterns.

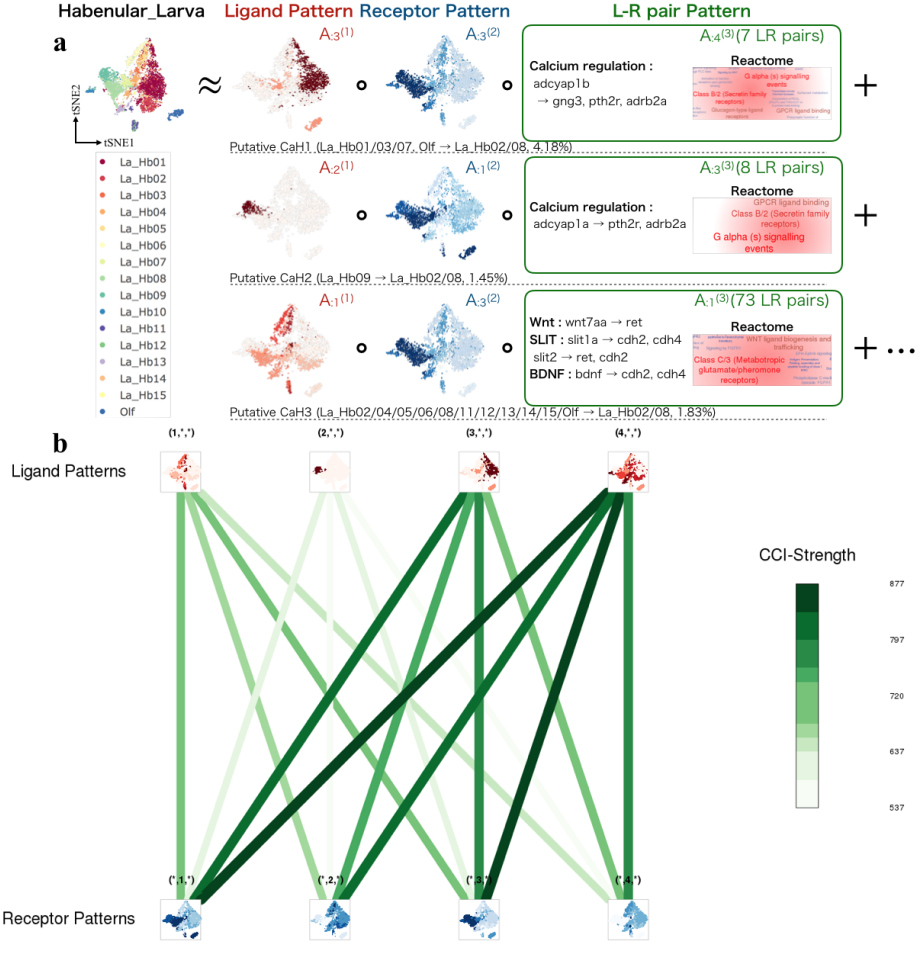

Figure 7 Results `scTensor` for an empirical dataset (Habenular_Larva) (a) Detected *CaH*s. (b) Overview of the L-R patterns.

Figure 8 Correspondence table containing the available L-R gene information for each organism For each detected *CaH*, gene annotation, enrichment analysis, tissue- or cell-type-specific gene expression, and chemical-gene expression relationships are assigned.

1 **Tables**

Table 1 Empirical datasets

Name	Organisms	ID	# Gene	# Cell	# Cell type	Unit
Germline_Female [25]	*Homo sapiens*	GSE86146	2717	992	8	TPM
Germline_Male [25]	*Homo sapiens*	GSE86146	2790	852	7	TPM
Melanoma [59]	*Homo sapiens*	GSE72056	6603	2250	6	TPM
NonMyocyte [27]	*Mus musculus*	E-MTAB-6173	2587	10519	12	UMI count
Habenular_Larva [60]	*Danio rerio*	GSE109158	15206	4233	16	UMI count

Table 2 Summary of `LRBase.XXX.eg.db` for 12 organisms (1/2)

XXX	Organisms	SWISSPROT (Secreted / Membrane)	TrEMBL (Secreted / Membrane)	STRING (PPI)
Hsa	*Homo sapiens*	1592 / 2269	176 / 334	18838
Mmu	*Mus musculus*	1309 / 1806	325 / 1555	19715
Ath	*Arabidopsis thaliana*	1260 / 1001	244 / 80	24174
Rno	*Rattus norvegicus*	643 / 983	232 / 1229	19963
Bta	*Bos taurus*	517 / 448	192 / 390	18349
Cel	*Caenorhabditis elegans*	198 / 247	28 / 60	13545
Dme	*Drosophila melanogaster*	249 / 333	89 / 148	11903
Dre	*Danio rerio*	119 / 169	318 / 376	21746
Gga	*Gallus gallus*	175 / 173	185 / 154	13084
Pab	*Pongo abelii*	80 / 134	212 / 211	16691
Xtr	*Xenopus Silurana tropicalis*	57 / 83	141 / 114	15338
Ssc	*Sus scrofa*	223 / 153	202 / 445	18683

Table 3 Summary of `LRBase.XXX.eg.db` for 12 organisms (2/2)

XXX	# L-R Pairs (SWISSPROT × STRING)	# L-R Pairs (TrEMBL × STRING)
Hsa	21882	472
Mmu	16386	476
Ath	8697	94
Rno	5270	65
Bta	2220	237
Cel	106	1
Dme	384	9
Dre	99	432
Gga	140	105
Pab	34	184
Xtr	19	107
Ssc	277	130

Additional Files

Additional file 1 — Development of L-R databases for multiple organisms (PDF 2.2 MB)

Additional file 2 — Distributions of and correlations among 8 STRING-scores (ZIP 18.5 MB)

Additional file 3 — Convergence of NTD with toy model and empirical data (PDF 9.5 MB)