Plant Biotechnology Journal
Open Access

Plant Biotechnology Journal (2022) **20**, pp. 2123–2134 doi: 10.1111/pbi.13893

PlantPhoneDB: A manually curated pan-plant database of ligand-receptor pairs infers cell–cell communication

Chaoqun Xu[1,†], Dongna Ma[1,†], Qiansu Ding[1], Ying Zhou[2,*] and Hai-Lei Zheng[1,*]

[1]*Key Laboratory of the Ministry of Education for Coastal and Wetland Ecosystems, College of the Environment and Ecology, Xiamen University, Xiamen, China*
[2]*National Institute for Data Science in Health and Medicine, School of Medicine, Xiamen University, Xiamen, China*

Received 18 November 2021;
revised 10 July 2022;
accepted 13 July 2022.
**Correspondence* (Tel +86 0592-2185175; fax +86 0592-2185175; email yingzhou@xmu.edu.cn (Y.Z.); Tel +86 0592-2181005; fax +86 0592-2185889; email zhenghl@xmu.edu.cn (H.-L.))
[†]These authors contributed equally to this article.

Keywords: ligand-receptor interactions, cell–cell communication, single-cell transcriptomics, plants, signalling pathway.

Summary

Ligand-receptor pairs play important roles in cell–cell communication for multicellular organisms in response to environmental cues. Recently, the emergence of single-cell RNA-sequencing (scRNA-seq) provides unprecedented opportunities to investigate cellular communication based on ligand-receptor expression. However, so far, no reliable ligand-receptor interaction database is available for plant species. In this study, we developed PlantPhoneDB (https://jasonxu.shinyapps.io/PlantPhoneDB/), a pan-plant database comprising a large number of high-confidence ligand-receptor pairs manually curated from seven resources. Also, we developed a PlantPhoneDB R package, which not only provided optional four scoring approaches that calculate interaction scores of ligand-receptor pairs between cell types but also provided visualization functions to present analysis results. At the PlantPhoneDB web interface, the processed datasets and results can be searched, browsed, and downloaded. To uncover novel cell–cell communication events in plants, we applied the PlantPhoneDB R package on GSE121619 dataset to infer significant cell–cell interactions of heat-shocked root cells in *Arabidopsis thaliana*. As a result, the PlantPhoneDB predicted the actively communicating AT1G28290-AT2G14890 ligand-receptor pair in atrichoblast–cortex cell pair in *Arabidopsis thaliana*. Importantly, the downstream target genes of this ligand-receptor pair were significantly enriched in the ribosome pathway, which facilitated plants adapting to environmental changes. In conclusion, PlantPhoneDB provided researchers with integrated resources to infer cell–cell communication from scRNA-seq datasets.

Introduction

In order to adapt to environmental changes, plants achieve controlled short and long ranges of cell–cell communication to perceive environmental cues in many ways, including mobile transcriptome, transcription factors, phytohormones, and small signalling peptides (Busch and Benfey, 2010; Murphy *et al*., 2012). In recent years, the importance of secreted signalling peptides in cell–cell communication has received massive attention in plants, coordinating cellular functions to sustain plant growth and development (Jeon *et al*., 2021; Oh *et al*., 2018; Takahashi *et al*., 2018; Zhong *et al*., 2022). Similar to mammals, plants have evolved a large number of secreted peptides, which are considered to be intercellular signalling molecules (Lease and Walker, 2006). Secreted peptide ligands have been considered as the first messenger to bind to cell surface receptors that are transmembrane proteins with extracellular and intracellular kinase domains for signalling transduction. For instance, Phytosulfokine (PSK) peptide may interact with PSK receptor gene 1 (PSKR1) and PSK receptor gene 2 (PSKR2) to regulate root growth in *Arabidopsis* (Kutschmar *et al*., 2009), and the pathway of AtPep3 peptide and membrane-receptor kinase gene PEPR1 is associated with salt tolerance in *Arabidopsis* (Nakaminami *et al*., 2018). Many cell surface receptors are composed of receptor-like proteins and receptor-like kinases, which contain more than 610 receptor-like kinase members in *Arabidopsis thaliana* (Shiu and Bleecker, 2001) and over 1000 receptor-like kinase members in *Oryza sativa* (Shiu *et al*., 2004). And the peptide-receptor interactions can activate a series of downstream physiological and biochemical processes. In brief, secreted peptides and corresponding cell surface receptors play important roles in cell–cell communication in plants (Chakraborty *et al*., 2019).

Plants are composed of different cell types that form a dynamic and complex cell–cell communication network to ensure functional connections. To better study cellular functions, it is necessary to understand how cells communicate with each other in response to their environment. The emergence of high-throughput single-cell RNA-sequencing (scRNA-seq) technologies provides unprecedented opportunities to characterize cellular compositions and activities at single-cell resolution. Compared with traditional bulk RNA-seq, scRNA-seq has significant advantages on gene dynamic expression in individual cell types. The scRNA-seq has been increasingly used to study transcriptional regulations and developmental mechanisms of plant tissues, responses of various cell types to different environmental stimuli, and finally cell–cell interactions (Jean-Baptiste *et al*., 2019; Liu *et al*., 2021; Thibivilliers and Libault, 2021; Xu *et al*., 2021).

Some software tools have been developed to infer cell–cell communication. For example, SingleCellSignalR uses a new regularized product score (LRscore) to account for variable levels of depth in scRNA-seq datasets and provides a cutoff value

(LRscore >0.5) to control the false discovery rate for ligand-receptor interactions based on two benchmarks (Cabello-Aguilar *et al*., 2020). Another software CellPhoneDB calculates ligand-receptor interaction scores using a permutation test by randomly shuffling the cluster labels (such as 100 times), and computes a *P*-value based on a null distribution of interactions scores. CellPhoneDB considers that ligand interacts with receptor if *P*-value <0.05 (Efremova *et al*., 2020). Also, scTensor adopts non-negative Tucker decomposition to detect some hypergraphs based on automatically generating 12 organisms' ligand-receptor pairs, including *Arabidopsis thaliana*. The scTensor algorithm includes five steps: construction of CCI-tensor, CANDECOMP/PARAFAC and tucker decomposition, non-negative tucker decomposition, extraction of CCIs as hypergraphs, and label permutation method (Kim and Choi, 2007; Tsuyuzaki *et al*., 2019; Zhou *et al*., 2014). However, most of them are specific to humans or mice, and no real ligand-receptor pairs databases are available for plants. Although scTensor supports the analysis of data from plants, the confidence level of the predicted ligand-receptor pairs by scTensor is not controlled (Cabello-Aguilar *et al*., 2020).

To address this problem, in this study, we created a PlantPhoneDB, a pan-plant database containing ligand-receptor pairs with controlled quality from seven resources. Based on ligand-receptor pairs, we developed an R package 'PlantPhoneDB', which provided optional four scoring approaches to calculate the score of ligand-receptor interactions to infer cell–cell communication between different cell types from scRNA-seq datasets. As a result, the PlantPhoneDB R package can predict downstream target genes regulated by ligand-receptor pairs that were involved in the signalling pathway in plants. Finally, we successfully developed a web interface, where users can search, browse and download the processed datasets.

Results

Statistics of PlantPhoneDB

The current PlantPhoneDB website contains 3514 unique ligand-receptor pairs for *Arabidopsis thaliana*, which are curated from seven resources, including plant.MAP, Interactome v2.0, IntAct, BioGRID, Text-mining from literature, STRING, and Orthologs resources (Figure 1a). Ligand-receptor pairs in PlantPhoneDB include 574 ligands and 585 receptors in *Arabidopsis thaliana*, respectively. scTensor, an R package automatically generates 12 organisms' ligand-receptor pairs from the STRING database using 36 approaches. scTensor generates 3014 ligand-receptor pairs involving 671 ligands and 645 receptors for *Arabidopsis thaliana* (Figure S1a). Compared with scTensor, only 26.11% (787/3014) ligand-receptor pairs from scTensor were covered in the PlantPhoneDB, and 2727 ligand-receptor pairs in *Arabidopsis thaliana* were uniquely recorded in PlantPhoneDB but not in scTensor (Figure S1b). For further comparison, we filtered the ligand-receptor pairs (3014 pairs) provided by scTensor using a PPI combined score > 600 as the cutoff of filtering criteria. Among 818 ligand-receptor pairs obtained, 762 pairs were overlapped with STRING resource (1112 pairs) from PlantPhoneDB (Figure S1b). Also, by assigning orthologs of ligand-receptor pairs between *Arabidopsis thaliana* and other four plant species proteomes using the InParanoid algorithm (Sonnhammer and Östlund, 2015), the number of ligand-receptor pairs identified ranged from 1751 (*Solanum lycopersicum*) to 3762 (*Oryza sativa*) (Figure 1b).

In addition, in PlantPhoneDB, we manually reviewed and confirmed 23 peer-reviewed publications and four preprints, and collected the information of 29 scRNA-seq datasets, including ~560 000 cells of 15 tissues from five plant species, including *Arabidopsis thaliana*, *Oryza sativa*, *Populus alba x Populus glandulosa*, *Solanum lycopersicum* and *Zea mays* (Data S1). Among them, 14 scRNA-seq datasets were directly obtained from PlantscRNAdb (Chen *et al*., 2021) (http://ibi.zju.edu.cn/plantscrnadb/index.php). After processing, the qualified scRNA-seq datasets were used to perform cell–cell communication (filtering criteria see method). Of note, we will update our database once 10 pending scRNA-seq datasets are processed or new plant scRNA-seq datasets are available (Figure 1c).

Function of PlantPhoneDB

We designed several modules to display the analysis results, including ligand-receptor pairs, the processed scRNA-seq datasets, cell-type annotation, and cell–cell communication from single-cell transcriptomics. On the homepage, users can obtain the statistics of PlantPhoneDB that contains the number of ligand-receptor pairs and ~ 560 000 cells of 15 tissues from five plant species (Figure 2a). In the search tab, users can query the detailed information of a specific ligand or receptor using an accepted ID, such as Uniprot Accession, TAIR Locus identifier, or Rice locus. The References module provides 27 articles about plant scRNA-seq datasets. Users can directly review the title and abstract when they click on one article of interest. The Download module supports users with the demand for ligand-receptor pairs and single-cell-level expression matrices.

In the Explorer tab (Figure S2a), PlantPhoneDB allows detailed exploration of the cell–cell communication for a processed scRNA-seq dataset. We also provided R scripts with a full document under the Document module to help researchers analyse their own datasets locally. In the About tab, PlantPhoneDB welcomes any feedback by email.

Single scRNA-seq dataset exploration

In the aspect of visualization, users can upload each processed scRNA-seq dataset to the FASTGenomics platform (https://www.fastgenomics.org/) for visualization or explore it using Cellxgene (https://chanzuckerberg.github.io/cellxgene/) on local. Here, we adopted a MAESTRO tutorial (gene marker-based annotation method) to perform cell identity annotation of a scRNA-seq dataset with accession GSE114615 (Turco *et al*., 2019) and demonstrated cell-type compositions (Figure 2b) and gene expression distributions (Figure 2c). If users are interested in one specific cell type, such as the lateral root cell of *Arabidopsis thaliana*, they can choose the interesting cell type and other cell types for DEGs analysis. Users can optionally use the Wilcoxon rank-sum test to evaluate the statistical difference in gene expression between different cell types.

Compared with other cell types, we observed that AT2G43610 had the highest expression level in Lateral root cells (Figure 2c). Moreover, AT2G43610 (logFC = 1.78, FDR = 1.29×10^{-222}) was the top one DEG in the lateral root cells (Data S2) and a marker gene reported by PlantscRNAdb (Chen *et al*., 2021; Wendrich *et al*., 2020). Also, we can see that the UMAP plot and violin plot revealed specific expression patterns of AT2G43610 across different cell types in Lateral root tissue (Figure S2b; Figure 2c). In summary, the violin plot and UMAP

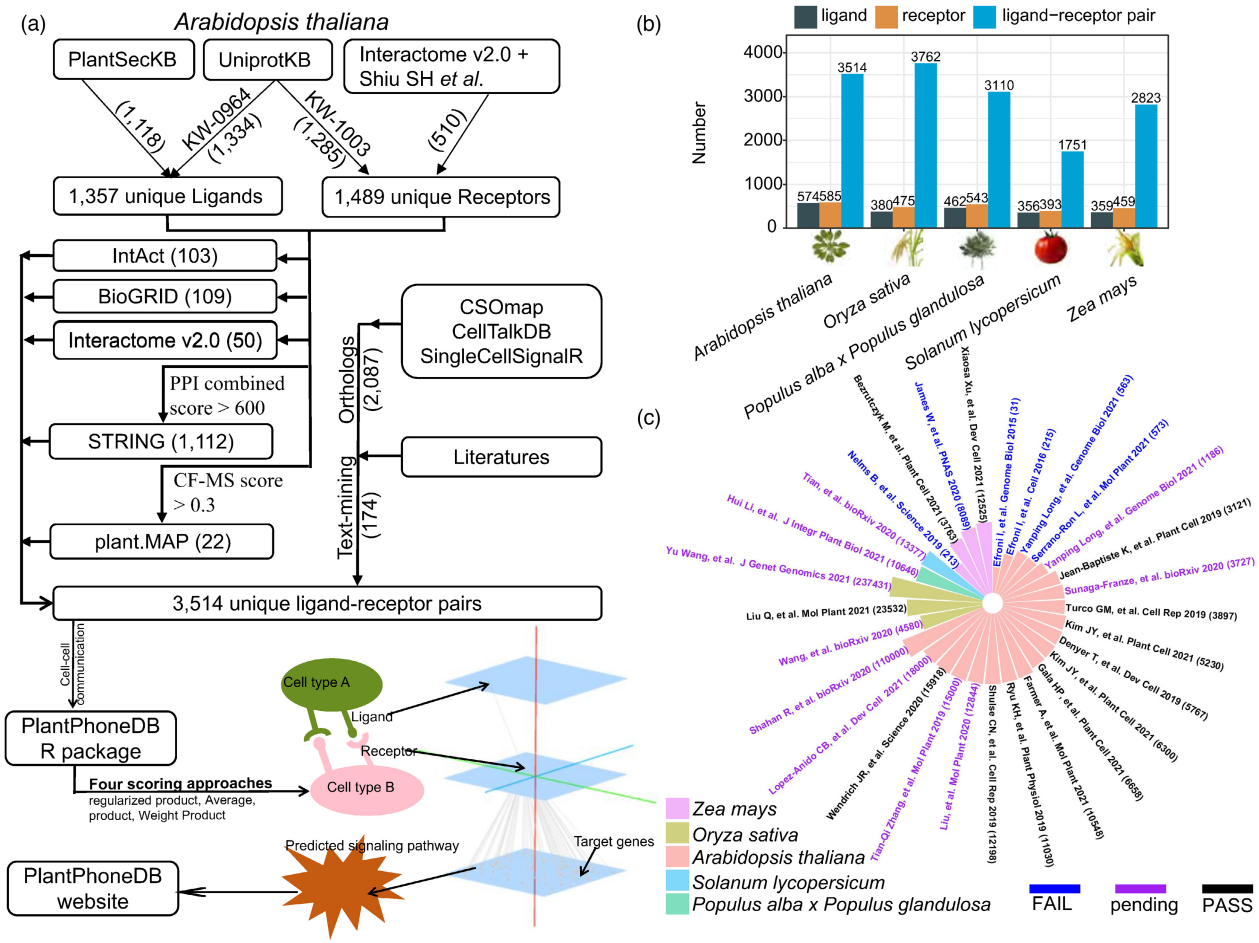

Figure 1 Statistics of PlantPhoneDB and summary of scRNA-seq datasets were analysed. (a) The number of ligand-receptor pairs curated from plant.MAP, Interactome v2.0, IntAct, BioGRID, Text-mining from literatures, STRING, and Orthologs resources in *Arabidopsis thaliana*. And 3514 unique ligand-receptor pairs are used to infer cell–cell communication. (b) The number of ligands, receptors and ligand-receptor pairs identified in 5 plant species, including *Arabidopsis thaliana*, *Oryza sativa*, *Populus alba x Populus glandulosa*, *Solanum lycopersicum*, and *Zea mays*. (c) PlantPhoneDB includes 29 scRNA-seq datasets information, covering ~560 000 cells of 15 tissues from 5 plant species. FAIL, PASS, and pending datasets are indicated as blue, black, and purple bar, respectively. PASS datasets indicate scRNA-seq datasets with ≥1000 high-quality cells; pending datasets indicate PASS datasets without available the expression matrix or datasets are too large to be analysed on our laptop. The rest of scRNA-seq datasets were considered to be FAIL datasets (<1000 high-quality cells). The number of cells (recorded by original paper) for each dataset is shown inside the parenthesis.

plot could show the expression patterns of AT2G43610 across different cell types and check whether AT2G43610 as one of the marker genes is helpful for cell identity annotation.

Selection of automatic cell identification methods

In order to select the better-fitting model to scRNA-seq datasets for cell-type annotation, here, we benchmarked the performance of 10 classifiers across 7 human peripheral blood mononuclear cells (PBMC) datasets (PbmcBench; Data S3) using 5-fold cross-validation. That is, each dataset was randomly split into 5 parts, 4 of 5 folds were used to train the classifiers, and the last fold was used to evaluate the performance of the classifiers. We repeated this procedure 5 times to obtain the F1-score and running time. In brief, we tested the performance of classifiers across datasets from different sequencing protocols (inter-datasets model; Figure 3a) and within a dataset (intra-dataset model; Figure 3b), respectively. As a result, we obtained 49 pairwise train-test combination results. Most notably, the best-performing classifier was MAESTRO classifier regardless of dataset type, which had a higher F1-score and lower running time (Figure 3c).

Besides, almost all classifiers performed well except for the index of cell identity (ICI) classifier, the mean F1-score was greater than 0.75 for all classifiers except for the ICI classifier (Figure 3c). Another exception is that the garnett classifier performed poorly on inDrop protocol, but well for other protocols. A classifier is actually required to predict cell identities for cross-datasets in the real scenario. Therefore, we evaluated the statistical difference in the performance of each classifier between inter-dataset and intra-dataset models using the F1-score, and concluded that there was no difference except for SingleR and scmap-cluster classifier (*P*-value <0.05; Figure S3a). We subsequently evaluated the performance of all classifiers on the inter-datasets model or intra-dataset model using the F1-score and obtained a significant difference in performance (*P*-value <0.05; Figure S3b).

Comparisons of scoring method

Subsequently, we used four scoring approaches to infer cell–cell communication from scRNA-seq datasets with cell types annotated by the MAESTRO classifier. Interestingly, the resulting heatmap (Figure 4a, b) showed a similar cell–cell communication

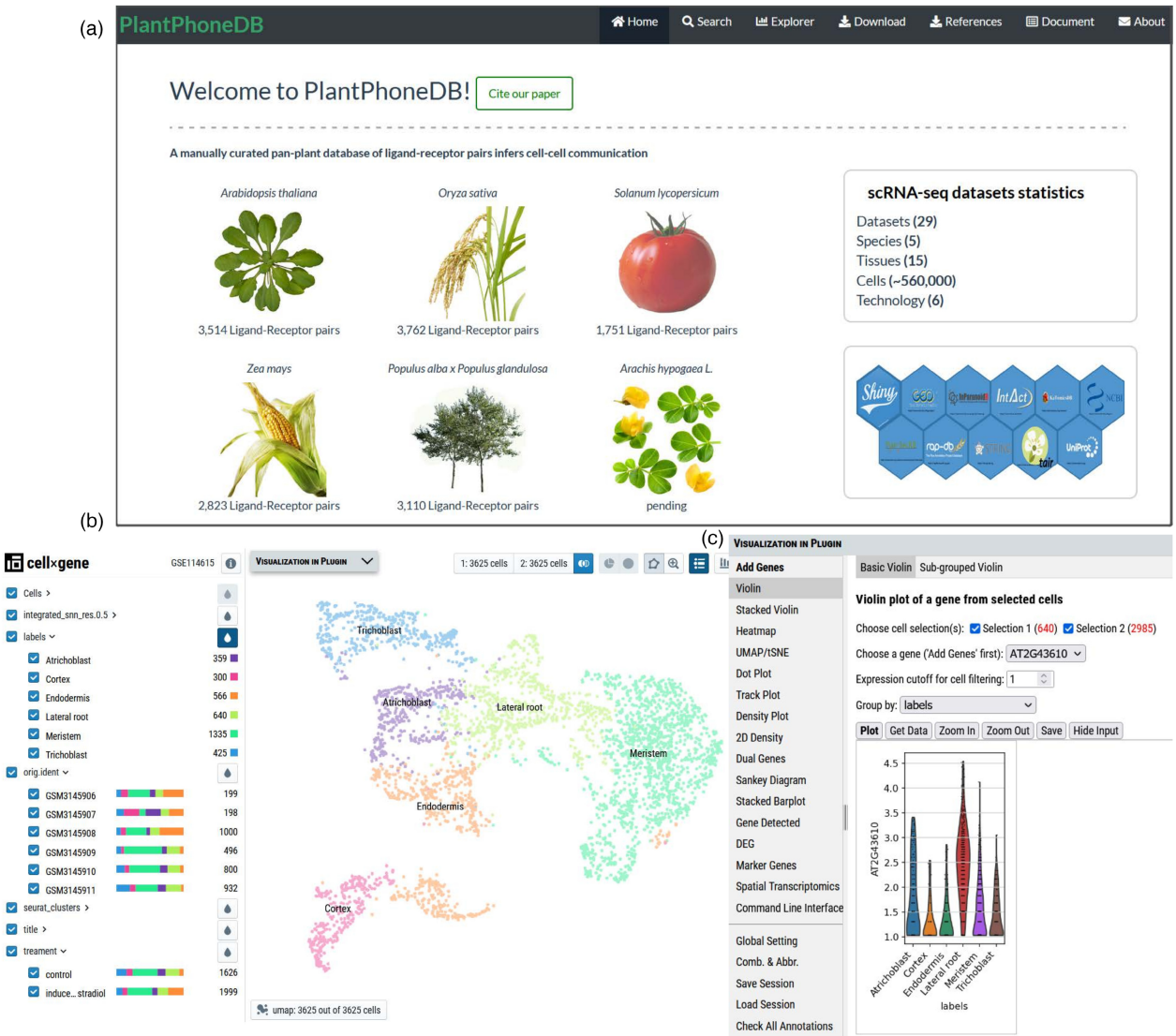

Figure 2 Functions of PlantPhoneDB web interface and a visualization example of AT2G43610 gene expression across cell types. (a) Overview of PlantPhoneDB. Seven modules are displayed on the navigation bar. Number of ligand-receptor pairs from five plant species are collected in PlantPhoneDB. The detailed information of scRNA-seq datasets and resources are showed in the box. (b) Visualization example of GSE114615 dataset using cellxgene software. The detailed meta-information for each dataset was displayed on the left, such as annotated cell identities and treatment conditions. On the right, the UMAP plot of GSE114615 dataset with cells coloured by trichoblast, atrichoblast, lateral root, meristem, endodermis, and cortex cell types. Each dot represents one cell. (c) A violin plot shows the distribution of AT2G43610 expression level across six different cell types.

network on 3 k or 8 k 10 × human peripheral blood mononuclear cells (PBMCs) dataset for four scoring approaches (LRscore, WeightProduct, Average, and Product). Furthermore, the similarity of the cell–cell communication network was evaluated using cosine similarity, and showed high similarity among the four scoring approaches (Figure S4a), suggesting that little performance difference among four scoring approaches. To evaluate whether the performance of the four scoring approaches was affected by the number of cells, the 8 k 10 × PBMCs dataset was subsampled to 10, 20, 30, 40, 50, 60, 70, 80, and 90% of its original size (8488 cells) in a stratified way. Using these cells in the dataset, four scoring approaches performed well regardless of the number of cells in this study (Figure S4b).

We next asked which cell–cell pair was communicating more frequently. An easy strategy was to count the number of ligand-receptor pairs for a given cell–cell pair, and then to normalize counts by dividing the total cell numbers of the corresponding cell–cell pair. Lastly, based on the ranking of normalized counts, we used the top 10 communicating cell type pairs identified to compare the performance among the four soring approaches. Our results indicated that the four scoring approaches could identify almost the same top communicating cell–cell pairs (Data S4). Nevertheless, users should pay attention to the difference in scoring approaches when highlighting their communication network of interest (Data S5). Therefore, we recommend using at least two soring approaches to infer cell–cell communication.

Application of PlantPhoneDB

To explore more applications of PlantPhoneDB, we next studied how cells communicate in plants under heat-shock stress. The processed scRNA-seq dataset (GSE121619) (Jean-Baptiste *et al.*, 2019) was

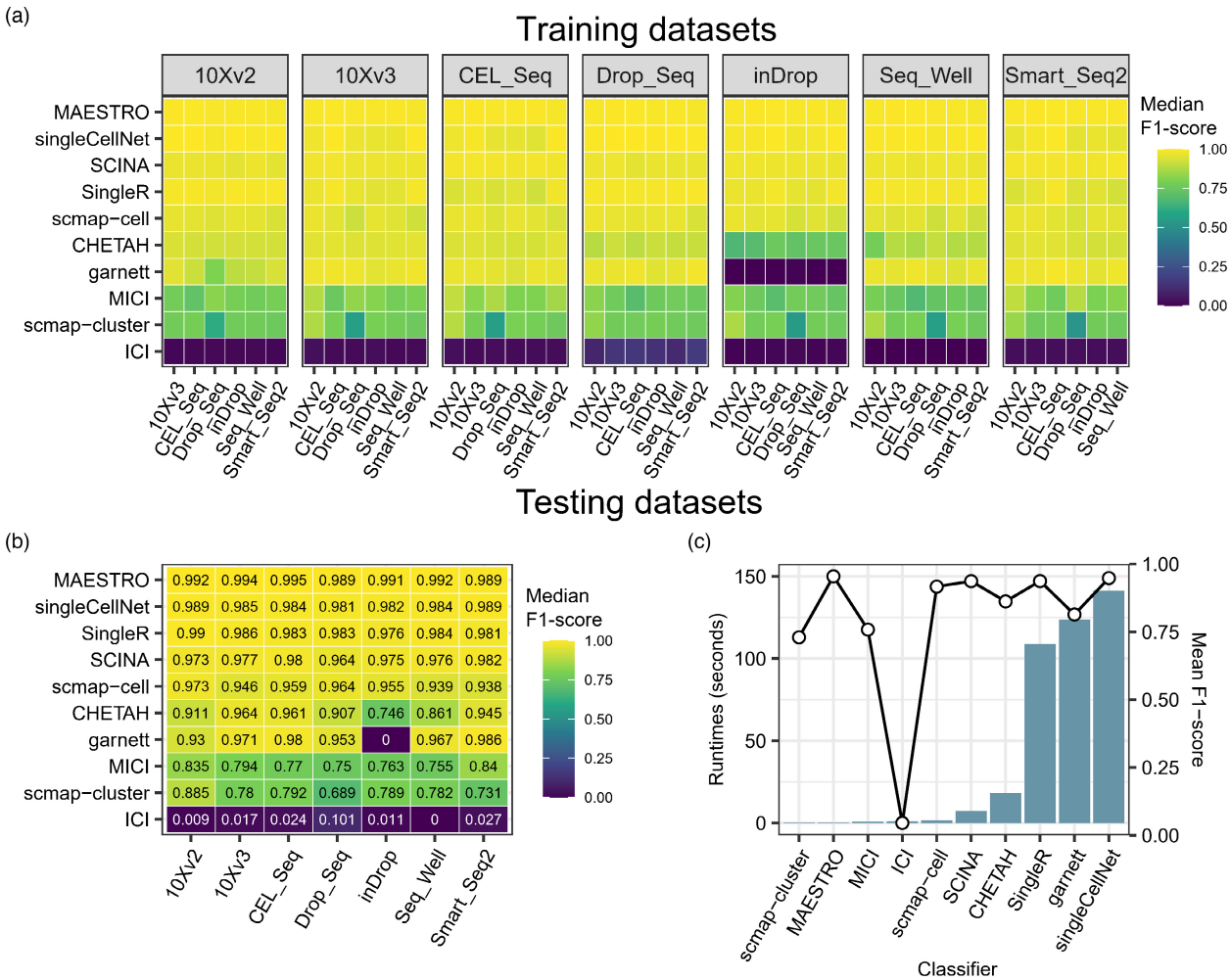

Figure 3 Benchmarks the performance of 10 classifiers across 7 PbmcBench datasets. (a) Heatmap shows the median F1-scores of 10 classifiers for 42 pairwise train-test combination across different protocols (inter-datasets model). Datasets on the top of the heatmap are used as training datasets, and testing datasets are indicated on the bottom of the heatmap. The inter-datasets model indicated trained scRNA-seq dataset from one sequencing protocol was used to predict cell type of scRNA-seq dataset from another sequencing protocol. (b) Median F1-scores of 10 classifiers within a dataset of different protocols (intra-dataset model), including 10 × v2, 10 × v3, CEL_Seq, Drop_Seq, inDrop, Seq_Well and Smart_Seq2 protocol. The intra-dataset model indicated trained scRNA-seq dataset from one sequencing protocol was used to predict cell type of scRNA-seq dataset from the same sequencing protocol. (c) Evaluates mean computation time and mean F1-scores of each classifier. Barplot indicates mean running time of each classifier (left); line plot indicates mean F1-scores (right).

used as input of the PlantPhoneDB R package, which contained 15 729 cells involving 9 cell types, namely Pericycle cells, Lateral root cells, Trichoblast cells, Cortex cells, Endodermis cells, Meristem cells, Phloem cells, Atrichoblast cells and Xylem cells (Figure 5a). The expression of DEGs (FDR <0.05, logFC > = 0.25) suggested that these cell types were correctly defined (Figure 5b). We then wanted to know whether all cells were able to exhibit heat-shock induction (Figure S5a). To do so, we calculated the proportion of different cell types from the control and heat-shock samples. Compared with control, atrichoblast cells, meristem cells, and cortex cells accounted for a higher proportion, which revealed that these cells were essential for heat-shock response (Figure S5b). A chi-square test was used to calculate the ratio of the observed and expected cell numbers ($R_{O/E}$) for each cell type. And these cell types displayed significant distinct preferences between the control and heat-shock samples (Figure 5c).

We also used PlantPhoneDB to identify a total of 1640 (including 439 experimental, 414 literatures-supported, and 787 predicted ligand-receptor pairs) significant ligand-receptor pairs between pairwise cell types using the Average scoring approach (Data S6), including 1457 paracrine ligand-receptor pairs and 183 autocrine ligand-receptor pairs (Figure 5d, Figure S5c, d). Herein, we focused on the top 10 ligand-receptor pairs ranked by score using the Average scoring method, which may play important roles in cell–cell communication. Notably, some ligand-receptor pairs were detected in most cell–cell pairs, such as AT3G53230–AT3G09840, AT3G53230–AT5G12110, and AT4G12420–AT2G45960; however, other ligand-receptor pairs were found in a few cell–cell pairs, such as AT4G15800–AT1G55330 and AT4G15800–AT3G13520 in atrichoblast–endodermis pair, implied different regulatory mechanisms of various ligand-receptor pairs (Figure 5e). In particular, 49 significant ligand-receptor pairs were detected in the biggest cell communication

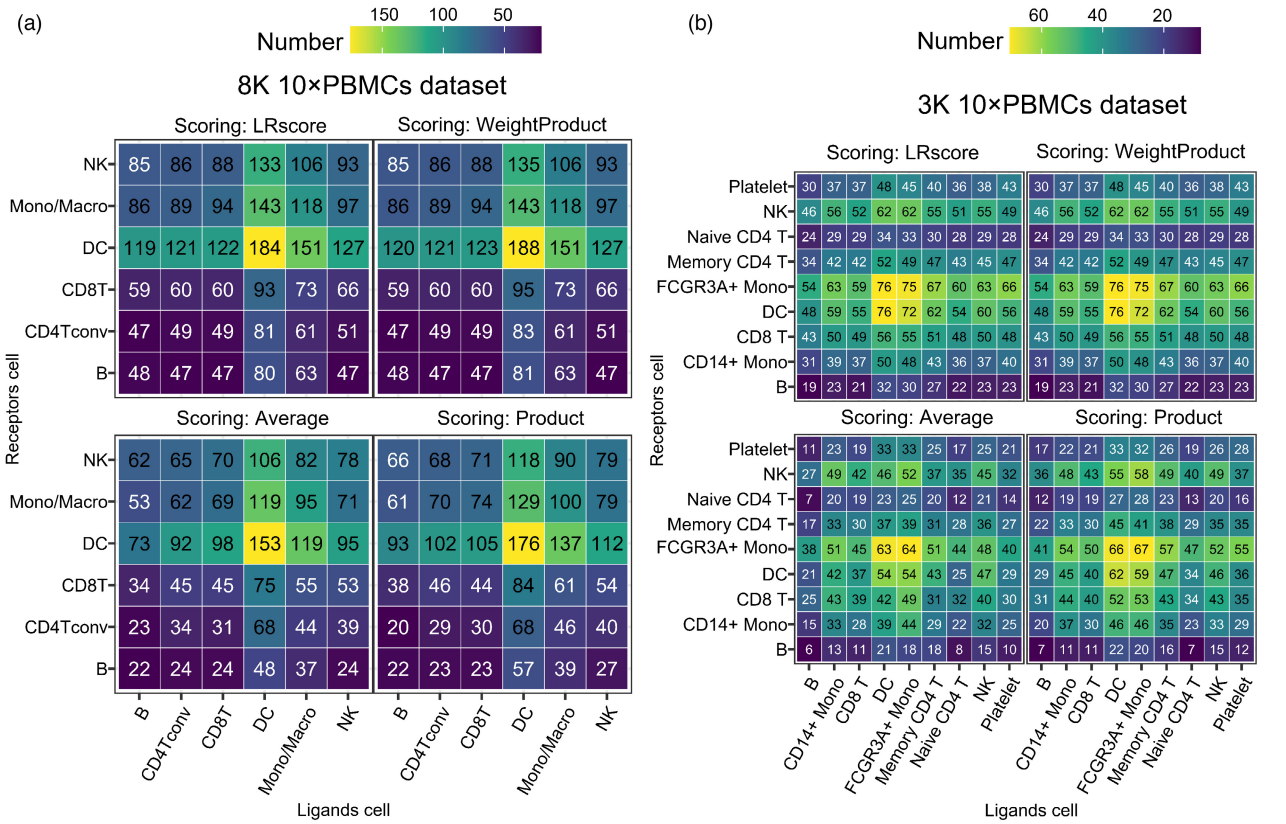

Figure 4 Comparison of four scoring approaches (LRscore, WeightProduct, Average, and Product). (a) The number of ligand-receptor pairs identified using four scoring approaches on 8 k 10 × PBMCs dataset. Rows represent cells expressing the receptors and columns represent cells expressing the ligands. Low and high number of ligand-receptor pair are showed by purple and yellow, respectively. (b) The number of ligand-receptor pairs identified using four scoring approaches on 3 k 10 × PBMCs dataset.

network (from Atrichoblast to Cortex cells), which highlighted the importance of atrichoblast–cortex cell pair in response to heat-shock stress (Figure S5e**)**. Then, we constructed an internal signalling network regulated by each ligand-receptor pair from 49 ligand-receptor pairs of Atrichoblast-Cortex cell pair. The pathway analysis results showed that the downstream target genes of AT1G28290-AT2G14890 pair (FDR = 3.77×10^{-64}) were mainly involved in the ribosome pathway (ath03010; Data S7; Figure S5f). Intriguingly, a previous study supported that heat stress would give rise to the ribosome pausing phenomenon in *Arabidopsis thaliana* (Merret *et al*., 2015).

In addition, a rice scRNA-seq dataset (GSE146035) (Liu *et al*., 2021), including 10 968 cells and 12 564 cells from cultivar Nipponbare (Japonica) and 93–11 (Indica), respectively, was composed of six cell types, namely Columella cells, Cortex cells, Endodermis cells, Epidermis cells, Metaxylem cells and Stele cells (Figure 6a), was used to perform cell–cell communication. The expression of known marker genes revealed that these cell types were correctly annotated (Figure 6b). PlantPhoneDB R package not only can identify significant ligand-receptor pairs in single root cells of *Arabidopsis thaliana* under different environmental conditions, such as heat-shock stress, but also can compare the fractions of the number of interactions among different cell types by dividing by a total of interactions between two scRNA-seq datasets. We can see that using the fractions of the number of interactions as the quantification of the cell–cell communications highlights the importance of relative ranking in the cell–cell communications network for each cell group between the two rice cultivars. (Figure 6c). We also provided a network-graph view to visualize the different number of interactions among different cell types (Figure 6d, e) and the number of interactions for the cell–cell communication subnetwork when selecting one cell type of interest. We sometimes maybe more concerned on the cell–cell communication of certain cell types we were interested in (Figure S6). In summary, we can compare the difference in cell–cell interaction by comparing the two rice cultivars using the PlantPhoneDB R package.

Discussion

Ligand-receptor pairs are widely used to infer cell–cell communication from the single-cell transcriptome. The rapid increase in scRNA-seq datasets makes it possible to study how cell types of plant tissue communicate in response to environmental cues, such as heat-shock stress. Many software tools have been developed based on ligand-receptor pairs from human and model animals (Liu *et al*., 2022; Shao *et al*., 2021). However, no plant-specific ligand-receptor pair databases are available up to now. Therefore, it's necessary to develop a comprehensive and reliable ligand-receptor pairs database to study cell–cell communication for plants, especially *Arabidopsis thaliana*, being an important model plant. In this study, we developed PlantPhoneDB which contained a large number of high-confidence ligand-receptor pairs. Compared with scTensor, we identified 2727

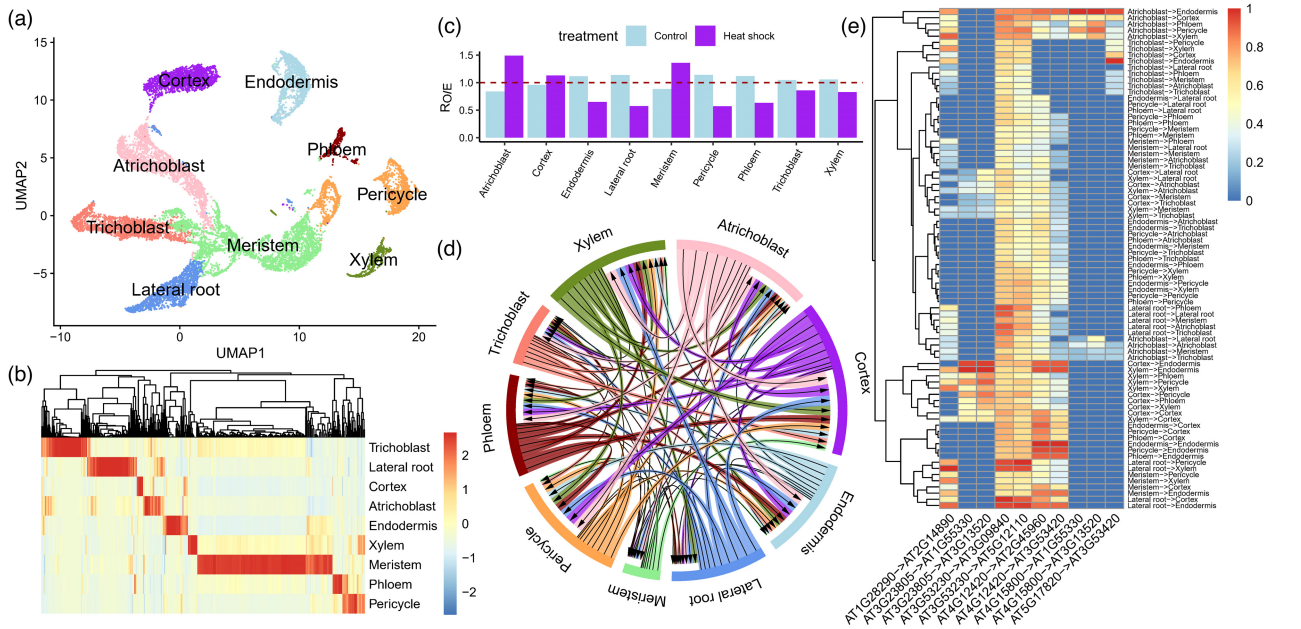

Figure 5 Significant cell–cell interactions of heat-shocked root cells in *Arabidopsis thaliana*. (a) UMAP plot of GSE121619 dataset with cells coloured by atrichoblast, cortex, endodermis, lateral root, meristem, pericycle, phloem, trichoblast, and xylem cell types. (b) The mean expression of signature genes for each cell type annotated by MAESTRO software. Low and high gene expression levels are showed by blue and read, respectively. (c) Preference of each cell type under heat-shock stress. $R_{O/E}$ above 1 indicates enrichment. (d) Chord diagram of cell–cell communication between pairwise cell types. The line width indicates the number of significant ligand-receptor pairs. (e) Top 10 ligand-receptor pairs with *P*-value <0.05 show different regulatory pattern. Columns are scaled by max ligand-receptor expression.

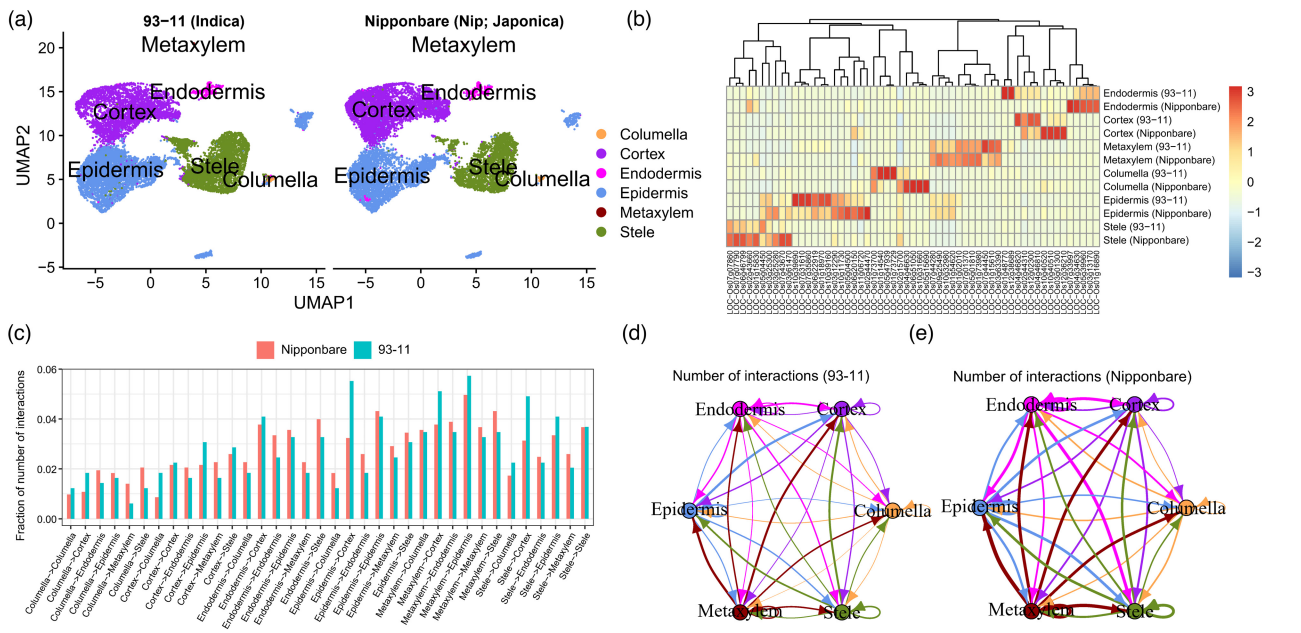

Figure 6 Comparison of the number of interactions of cell pairs between two rice cultivar datasets (93–11 and Nipponbare). (a) UMAP visualization of GSE146035 dataset, including 10 968 cells and 12 564 cells from two cultivar Nipponbare (Japonica) and 93–11 (Indica), respectively. Each dot represents one cell. (b) The mean expression of known marker genes for each cell type from two rice cultivars. (c) Difference of cell–cell interactions of each cell type in two rice cultivars, accounting for total cell number. (d) Identification of significant ligand-receptor pairs between pairwise cell types in rice cultivar 93–11. (e) Identification of significant ligand-receptor pairs between pairwise cell types in rice cultivar Nipponbare.

ligand-receptor pairs with certain confident criteria in *Arabidopsis thaliana* in PlantPhoneDB but not in scTensor. As mentioned above, only 787 ligand-receptor pairs were overlapped with the scTensor (Figure S1b). In addition, only 19.37% (74/382) literature-supported ligand-receptor pairs we collected was overlapped with the scTensor (Figure S4c). For further comparison,

93.15% (762/818) high-confidence ligand-receptor pairs and 1.14% ((787–762)/(3014–818)) low-confidence ligand-receptor pairs from scTensor were covered in the PlantPhoneDB. There are two possible reasons for the low overlap between scTensor and PlantPhoneDB: (1) The STRING database is constantly updated, and ligand-receptor pairs by scTensor are not the latest. (2) scTensor contains only ligand-receptor pairs from the STRING database in plants, but without other databases, such as BioGRID and IntAct, *etc*. As explained (Cabello-Aguilar *et al*., 2020), they could not use scTensor beyond its prepackaged example dataset. Despite considerable efforts scTensor made, it's necessary to develop a plant-specific tool for cell–cell communication. The PlantPhoneDB R package not only provides some visualization functions, including dot plot, heatmap plot, circular plot, and network of cell–cell communication but also supports four scoring approaches to estimate PPI strength (Table 1).

Recent scRNA-seq technologies have successfully solved cellular heterogeneity problems and promoted us to underline cell–cell communication in plant species at single-cell resolution. To gain the landscape of plant cellular communication, it's vital to identify cell-type identities. We evaluated the performance of 10 classifiers on 7 PbmcBench datasets based on the F1-score and computation time, and chose MAESTRO software to annotate the cell clusters. The ICI method performed worst on all PbmcBench datasets for cell identity annotation. We speculated that the ICI method is specific to the root tissue of *Arabidopsis thaliana* rather than other tissues or organisms. Wang *et al*. suggested the approach was likely not appropriate for the rice scRNA-seq dataset (Wang *et al*., 2021). In the present study, the PlantPhoneDB R package provided optional four scoring approaches to infer cellular communication. There was very little difference among the four scoring approaches except for running time.

In this study, we demonstrated two examples of how to use PlantPhoneDB in real-world scRNA-seq datasets from plants. On one hand, PlantPhoneDB could predict an important biological pathway regulated by AT1G28290-AT2G14890 pair, which was supported by a previous study (Merret *et al*., 2015). It could be an important regulatory mechanism, which facilitates plants adapting to environmental changes. To some extent, a previous study demonstrated that heat stress would give rise to the ribosome pausing phenomenon supported our result at the bulk RNA-Seq level. These findings provided important cues to further understand how cells communicate with each other in response to heat stress. However, further evidence is needed to support this finding. On the other hand, we compared cellular communication between two rice cultivars and revealed the importance of relative ranking in the cell–cell communications network for each cell group. PlantPhoneDB also provides multiple visualizations of the number of interactions among different cell types to compare differences in the communication network. However, there are some limitations to be noted. Firstly, we did not take into account the heteromeric interactions between ligand-binding receptors and respective co-receptors, which could serve as an important interaction platform (Smakowska-Luzan *et al*., 2018; Zhang *et al*., 2022). Secondly, due to the very low overlap of ligand-receptor pairs between non-literature-supported pairs (experimental and predicted pairs) and literature-supported pairs from PlantPhoneDB, we could not perform benchmarking analyses and accurately evaluate the likelihood of the cell–cell communications. Besides, plant hormones are also involved in many processes of plant growth and development, which trigger numerous transcriptional programs in response to environmental cues (Nemhauser *et al*., 2006). A possible mechanism is cross-talk between phytohormones and secreted signalling peptides to integrate cellular communication network and regulate physiological and biochemical processes. Lastly, scRNA-seq datasets are too large for memory and require high-power computing server. Therefore, for now, we only offer an R package for users to install and analyse their own datasets. In the future, we will build an application for visualization, comparison, and cell–cell communication for single-cell transcriptome datasets. It does not require associated expertise and expense, just uploading dataset, analysing and downloading analyse results. In the future, the use of spatial transcriptomics technologies on plant species will promote us to constantly update PlantPhoneDB.

In summary, PlantPhoneDB provides numerous high-confidence ligand-receptor pairs in five plant species. And we constructed a user-friendly website for systematically searching, browsing, and downloading the processed datasets, facilitating the exploration of cell–cell communication at single-cell resolution in plants. In addition, the PlantPhoneDB R package provides some functions using R (version: 4.0.2), such as LRscore,

Table 1 Comparison of PlantPhoneDB R package with other software tools

Features	PlantPhoneDB	SingleCellSignalR	CellPhoneDB	scTensor
Number of species	5	2	1	12
plant-specific	Y	N	N	N
preprocessed data	Y	Y	N	Y
Complete pipeline	Y	Y	N	N
Scoring approach	Regularized product/Average/ product/Weight Product	Regularized product	Average	Linear decomposition
Coding language	R	R	Python	R
Intracellular signalling	Y	Y	N	N
Dot plot	Y	N	Y	N
Heatmap plot	Y	N	Y	N
Circular plots	Y	Y	N	N
Tables	Y	Y	Y	Y
Web interface	Y	N	Y	Not available
Network of cell–cell communication	Y	Y	N	Y

heatmap_count, CCI_circle, CCI_network, and LR_pathway, to infer and construct cell–cell communication network and intracellular signalling pathway.

Materials and methods

PlantPhoneDB-a curated database based on ligand-receptor interactions

To construct a ligand-receptor interaction database, we searched secreted proteins and plasma membrane proteins from the UniProtKB/Swiss-Prot (Boutet *et al*., 2016) database using the keywords KW-0964 (secreted) and KW-1003 (cell membrane), respectively, according to the protocol of CellPhoneDB v.2.0. We also manually reviewed open access databases, for instance, TAIR (Berardini *et al*., 2015) database includes receptor kinase-like gene family and PlantSecKB (http://proteomics.ysu.edu/secretomes/plant/index.php) database, a knowledgebase for plant secretomes, and peer-reviewed publications or preprints with the term 'ligand and receptor' in the title or abstract and developed PlantPhoneDB. For *Arabidopsis thaliana*, we downloaded protein–protein interactions (PPIs) curated from literature from BioGRID (Oughtred *et al*., 2021), Interactome v2.0 (https://www.arabidopsis.org/download/index.jsp), IntAct (Hermjakob *et al*., 2004), plant.MAP (McWhite *et al*., 2020), and STRING (Szklarczyk *et al*., 2019) databases. Of which, STRING database also included a large number of experimental and predicted associations between proteins. Currently, all the annotation of PPIs from BioGRID, Interactome v2.0, IntAct, and plant.MAP databases were extracted from the literature. In plant.MAP, while CF-MS scores represent PPI strength, CF-MS scores >0.5 correspond to ~90% true positive rate, and scores >0.2 correspond to ~50% true positive rate. We kept pairs of which CF-MS score >0.3. In STRING, we selected pairs if PPI combined score >600. Finally, we matched ligands with receptors taken from UniProtKB using reliable PPIs.

Besides, PlantPhoneDB integrates some existing resources (CellTalkDB (Shao *et al*., 2021), SingleCellSignalR, and CSOmap (Ren *et al*., 2020)) that contain human ligand-receptor interactions. To expand our ligand-receptor interaction database, we did orthologs assignment between *Homo sapiens* and *Arabidopsis thaliana* proteomes using the InParanoid algorithm (Hou *et al*., 2020; Sonnhammer and Östlund, 2015) to transfer annotations of known PPI. Moreover, to the run PlantPhoneDB R package on other plant species, we extracted the homologues of ligand-receptor interactions based on *Arabidopsis thaliana* ortholog mappings using the InParanoid algorithm. The improved genome sequences and annotation files of *Populus alba*×*Populus glandulosa* (Huang *et al*., 2021) were obtained from https://doi.org/10.6084/m9.figshare.12369209. Protein sequences of other species were downloaded from UniProtKB/Swiss-Prot database.

Especially, for non-model plants, there are not available ligand-receptor pairs for cell–cell communication analysis. Therefore, we also provided a workflow to perform computational identifications of secreted proteins, receptor-like kinases (RLKs), or receptor-like proteins (RLPs) and their interaction based on protein sequences (Figure S2c). The workflow consists of three main steps: (1) identification of secreted proteins, (2) identification of RLKs/RLPs, and (3) prediction of protein–protein interactions. In detail, firstly, the SignalP 5.0 software was used for secretory signal peptide prediction (Almagro Armenteros *et al*., 2019). The accuracy of secretome prediction could be further improved by combing signalP 5.0 with other software, including Phobius (Käll *et al*., 2007), TMHMM (Krogh *et al*., 2001), and TargetP (Boos *et al*., 2018). Therefore, a predicted protein that has a secretory signal peptide by at least three software, including signalP 5.0 and TMHMM. And these predicted proteins without transmembrane helix and endoplasmic reticulum (ER) retention signals were considered to be secreted proteins. Then, the workflow provided by Restrepo-Montoya *et al*. was used for the computational identification of RLK/RLP and their structural domains in legumes, which can be applied to the proteomes of other plant species (Restrepo-Montoya *et al*., 2020). The Pfam_scan program was used to identify target domains. And those proteins with transmembrane helix and extracellular domain, and without Nucleotide-Binding domain shared by plant resistance gene products (NB-ARC) domain were considered to be RLPs. While those proteins with transmembrane helix, extracellular domain, intracellular domain and Pkinase domain, and without NB-ARC domain were considered to be RLKs. Lastly, computational prediction of protein–protein interaction was performed using CAMP (or other optional software), a sequence-based deep learning framework for the multifaceted prediction of peptide-protein interactions (Lei *et al*., 2021). Figure S2c showed filtering criteria, and for the detailed usage of software in the workflow, users can read documents on their published manuscripts.

Collection and processing of scRNA-seq datasets

Plant scRNA-seq datasets reported previously (Chen *et al*., 2021) were downloaded from PlantscRNAdb (http://ibi.zju.edu.cn/plantscrnadb/index.php), which includes 26 326 marker genes, 128 different cell types, 15 tissues from four plant species (*Oryza sativa*, *Arabidopsis thaliana*, *Zea mays*, and *Solanum lycopersicum*). In addition, we searched plant scRNA-seq datasets from the Gene Expression Omnibus (GEO) (NCBI Resource Coordinators, 2018) and ArrayExpress (Athar *et al*., 2019) using the keyword 'plant single cell or scRNA-seq'. Then, we manually confirmed and curated each dataset with available the expression matrix of the raw count, FPKM, TPM, or normalized count, and metadata information. Overall, a total of 29 plant single-cell transcriptome datasets across five plant species (Data S1) were obtained initially for further analysis.

Admittedly, we adopted a standard preprocessing analysis workflow based on Seurat v.4.0.3 (Hao *et al*., 2021) to perform quality control, data normalization, data scaling, integration, dimensional reduction, cell clustering, and differential expression analysis for the datasets we collected. Due to the technical noise in scRNA-seq or contamination of samples, low-quality cells were filtered out based on the number of unique detected genes (<200 genes per cell) and the total number of molecules (<1000 UMI per cell). We kept the datasets with ≥1000 high-quality cells, which were considered to be PASS datasets. In addition, PASS datasets without available expression matrix or with a large number of cells that consume high memory on our laptop were considered to be pending datasets. The rest of the scRNA-seq datasets were considered to be FAIL datasets (<1000 high-quality cells).

In order to eliminate the batch effect of different scRNA-seq datasets, the SCTransform function was applied to data normalization and data scaling. After data integration, the top 3000 variable features identified were employed for dimensional reduction using the principal component analysis (PCA) function. The first 50 principal components were chosen for cell clustering using K-Nearest Neighbours (KNN) and Louvain algorithm by FindNeighbors and FindClusters function. The differentially

expressed genes (DEGs) for each cluster were determined by false discovery rate (FDR <0.05) and the log-scale fold change (logFC > = 0.25) using the FindAllMarkers function.

Cell-type annotation

The cell types of clusters were determined by prior knowledge or supervised approaches. First, we automatically assigned the cell type identity to clusters based on signature files using the RNAAnnotateCelltype function from the MAESTRO R package (Wang *et al*., 2020). The signature files were constructed for plant species by collecting marker genes of each cell type of different tissues from the publicly available resources (Chen *et al*., 2021; Efroni *et al*., 2015). Second, we calculated an index of cell identity (ICI) score for each scRNA-seq dataset to identify cell types of clusters using an information-theory-based approach (Efroni *et al*., 2015). Third, we annotated the cell clusters based on a weighted marker-based index of cell identity (MICI) (Wang *et al*., 2021). Finally, we annotated cells against microarrays data (if available) using SingleR (Aran *et al*., 2019), an automatic annotation method for scRNA-seq datasets.

To achieve a more reliable performance of cell-type annotation, several other automatic cell identification methods were used to compare the performance of the classifiers (Data S8). Here, a total of 7 human peripheral blood mononuclear cells (PBMC) scRNA-seq datasets with known cell-type annotation were used to perform benchmarking analyses (Data S3). The 7 PBMC datasets (PbmcBench) with 7 different sequencing protocols were downloaded from Zenodo (https://doi.org/10.5281/zenodo.3357167) (Abdelaal *et al*., 2019). The performance of the classifiers was evaluated based on the F1-score and computation time.

The score of ligand-receptor interactions

To investigate the expression of genes on both transcript level and protein level, we reanalyzed the relationship between mRNA and protein data from four tissues of *Arabidopsis thaliana* (Data S9), namely flower, rosette leaf, silique, and seed (Mergner *et al*., 2020), using Spearman's correlation. We concluded that the majority of genes (26 828/37137, 72.24%, ¦Spearman's correlation¦ > 0.3) were highly correlated between transcript level and protein level for four tissue types, suggesting that transcript abundance could be a proxy for protein abundance (Figure S1c, d). Therefore, we scored ligand-receptor interactions based on the transcript level of genes.

Importantly, several strategies of communication scores were proposed to infer the PPI strength and showed advantages based on their different assumptions (Armingol *et al*., 2021), i.e. SingleCellSignalR and CellPhoneDB. In addition, an edge-score model was developed to model PPI by the law of mass (Altmann *et al*., 2020). PlantPhoneDB R package provides these optional approaches for users to score ligand-receptor interactions. It is worth noting that each ligand-receptor pair was expressed in at least 10% of cells of a given cell type.

For comparison, we used the 3 k and 8 k 10 × PBMCs datasets from https://support.10xgenomics.com/single-cell-gene-expression/datasets, to evaluate the performance of four scoring approaches (LRscore, WeightProduct, Average, and Product). We filtered the ligand-receptor pairs with LRscore <= 0.5 imposed by Cabello-Aguilar *et al* (Cabello-Aguilar *et al*., 2020). For the Product and Average scoring approaches, we calculated the *P*-values based on the interaction score distribution of randomly permuted cell types (100 times by default). *P*-values <0.05 were considered significant. As for the WeightProduct scoring approach, we considered that the mean expression level of ligand and receptor above 0.1 could be co-detected.

Construction of intracellular signalling pathway

We assumed the ligand-receptor pair as the seed node to transmit signalling from the surface of the cell membrane to downstream genes (Browaeys *et al*., 2020). In order to understand downstream biological pathways regulated by ligand-receptor pair, we identified highly variable genes as the primary downstream target genes of a specific ligand-receptor pair using SCTransform function from the Seurat R package (parameters by default), which could highlight biological signals in downstream analysis (Brennecke *et al*., 2013). Then, we further filtered downstream genes that meet the following two criteria: First, downstream genes were considered as differentially expressed genes (DEGs; FDR <0.05, logfc.threshold >0.25) identified by FindMarkers function from Seurat v.4.0.3 R package; second, Spearman's correlation greater than one threshold (0.013 by default) between mean expression of ligand-receptor pair and expression of each downstream gene.

Lastly, we constructed a weight adjacent matrix based on mutual information between all pairs of ligand-receptor pairs and downstream DEGs for each cell type using the infotheo R package (https://cran.r-project.org/web/packages/infotheo/index.html). The weak edges were removed to reconstruct an intracellular gene co-expression network using the parmigene R package (Sales and Romualdi, 2011). For a particular ligand-receptor pair, the 2nd-order neighbourhoods (default) were extracted to perform functional category enrichment analysis of signalling pathways using the Fisher's exact test. That is, we related ligand-receptor pair with all possible downstream pathways for each cell type. Pathway gene sets were obtained from PlantGSAD database (Ma *et al*., 2022). The statistical formula of the Fisher's exact test is defined as (1):

$$P = \frac{\binom{n}{k}\binom{N-n}{K-k}}{\binom{N}{K}} \quad (1)$$

where N is the number of unique detected genes from a scRNA-seq dataset, n is the number of the 2nd-order neighbourhoods, K is the number of one pathway gene sets and k is the number of overlapping genes. We adjusted enrichment *P*-values using the method of Benjamini & Hochberg (Korthauer *et al*., 2019).

Web interface of PlantPhoneDB

PlantPhoneDB is a ligand-receptor interactions database, which aims to infer cell–cell communications from scRNA-seq datasets in plants. We built the PlantPhoneDB web interface to present the results we analysed in a flexible way based on shiny (www.rstudio.com/shiny), a web application framework for R. All the processed scRNA-seq datasets can be searched and downloaded from the web interface, and saved as h5ad format file to meet requirements of cellxgene (https://chanzuckerberg.github.io/cellxgene/) for visualization. The website is freely available at https://jasonxu.shinyapps.io/PlantPhoneDB/ without login requirements. The documentation for PlantPhoneDB is available at https://plantphonedb.readthedocs.io/en/latest/index.html.

Acknowledgements

This work was financially supported by the Natural Science Foundation of China (NSFC) (32171740, 31870581, 31570586), and the National Key Research and Development Program of China (2017YFC0506102). We appreciate anonymous reviewers and the editor for the insightful comments and valuable suggestions. Also, we thank Dr. Jianming Zeng (University of Macau) and all the members of his bioinformatics team, biotrainee, for generously sharing their experience and codes.

Conflict of interests

The authors declare that they have no conflict of interest.

Author contributions

The project was conceived and directed by Ying Zhou and Hai-Lei Zheng. Dongna Ma and Qiansu Ding collected data, and data analysis was performed by Chaoqun Xu. Image processing was carried out by Chaoqun Xu and Dongna Ma. The manuscript was written by Chaoqun Xu, Ying Zhou, and Hai-Lei Zheng. All authors have read and approved the final manuscript.

References

Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M.J.T. and Mahfouz, A. (2019) A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.* **20**, 194.

Almagro Armenteros, J.J., Tsirigos, K.D., Sønderby, C.K., Petersen, T.N., Winther, O., Brunak, S., von Heijne, G. *et al*. (2019) SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **37**, 420–423.

Altmann, M., Altmann, S., Rodriguez, P.A., Weller, B., Elorduy Vergara, L., Palme, J., Marín-de la Rosa, N. *et al*. (2020) Extensive signal integration by the phytohormone protein network. *Nature* **583**, 271–276.

Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S. *et al*. (2019) Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172.

Armingol, E., Officer, A., Harismendy, O. and Lewis, N.E. (2021) Deciphering cell–cell interactions and communication from gene expression. *Nat. Rev. Genet.* **22**, 71–88.

Athar, A., Füllgrabe, A., George, N., Iqbal, H., Huerta, L., Ali, A., Snow, C. *et al*. (2019) ArrayExpress update - from bulk to single-cell expression data. *Nucleic Acids Res.* **47**, D711–D715.

Berardini, T.Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E. and Huala, E. (2015) The arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. *Genesis* **53**, 474–485.

Boos, F., Mühlhaus, T. and Herrmann, J.M. (2018) Detection of internal matrix targeting signal-like sequences (iMTS-Ls) in mitochondrial precursor proteins using the TargetP prediction tool. *Bio Protoc* **8**, e2474.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bansal, P., Bridge, A.J. *et al*. (2016) UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. *Methods Mol. Biol.* **1374**, 23–54.

Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B. *et al*. (2013) Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093–1095.

Browaeys, R., Saelens, W. and Saeys, Y. (2020) NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat. Methods* **17**, 159–162.

Busch, W. and Benfey, P.N. (2010) Information processing without brains – the power of intercellular regulators in plants. *Development* **137**, 1215–1226.

Cabello-Aguilar, S., Alame, M., Kon-Sun-Tack, F., Fau, C., Lacroix, M. and Colinge, J. (2020) SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. *Nucleic Acids Res.* **48**, e55.

Chakraborty, S., Nguyen, B., Wasti, S.D. and Xu, G. (2019) Plant leucine-rich repeat receptor kinase (LRR-RK): structure, ligand perception, and activation mechanism. *Molecules* **24**, 3081.

Chen, H., Yin, X., Guo, L., Yao, J., Ding, Y., Xu, X., Liu, L. *et al*. (2021) PlantscRNAdb: a database for plant single-cell RNA analysis. *Mol. Plant* **14**, 855–857.

Efremova, M., Vento-Tormo, M., Teichmann, S.A. and Vento-Tormo, R. (2020) CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat. Protoc.* **15**, 1484–1506.

Efroni, I., Ip, P.-L., Nawy, T., Mello, A. and Birnbaum, K.D. (2015) Quantification of cell identity from single-cell gene expression profiles. *Genome Biol.* **16**, 9.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A. *et al*. (2021) Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29.

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S. *et al*. (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res.* **32**, D452–D455.

Hou, R., Denisenko, E., Ong, H.T., Ramilowski, J.A. and Forrest, A.R.R. (2020) Predicting cell-to-cell communication networks using NATMI. *Nat. Commun.* **11**, 1–11.

Huang, X., Chen, S., Peng, X., Bae, E.-K., Dai, X., Liu, G., Qu, G. *et al*. (2021) An improved draft genome sequence of hybrid Populus alba × Populus glandulosa. *J. For. Res.* **32**, 1663–1672.

Jean-Baptiste, K., McFaline-Figueroa, J.L., Alexandre, C.M., Dorrity, M.W., Saunders, L., Bubb, K.L., Trapnell, C. *et al*. (2019) Dynamics of gene expression in single root cells of Arabidopsis thaliana. *Plant Cell* **31**, 993–1011.

Jeon, B.W., Kim, M.-J., Pandey, S.K., Oh, E., Seo, P.J. and Kim, J. (2021) Recent advances in peptide signaling during Arabidopsis root development. *J. Exp. Bot.* **72**, 2889–2902.

Käll, L., Krogh, A. and Sonnhammer, E.L.L. (2007) Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res.* **35**, W429–W432.

Kim, Y.-D. and Choi, S. (2007) *Nonnegative tucker decomposition*. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8.

Korthauer, K., Kimes, P.K., Duvallet, C., Reyes, A., Subramanian, A., Teng, M., Shukla, C. *et al*. (2019) A practical guide to methods controlling false discoveries in computational biology. *Genome Biol.* **20**, 118.

Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580.

Kutschmar, A., Rzewuski, G., Stührwohldt, N., Beemster, G.T.S., Inzé, D. and Sauter, M. (2009) PSK-α promotes root growth in Arabidopsis. *New Phytol.* **181**, 820–831.

Lease, K.A. and Walker, J.C. (2006) The Arabidopsis unannotated secreted peptide database, a resource for plant peptidomics. *Plant Physiol.* **142**, 831–838.

Lei, Y., Li, S., Liu, Z., Wan, F., Tian, T., Li, S., Zhao, D. *et al*. (2021) A deep-learning framework for multi-level peptide–protein interaction prediction. *Nat. Commun.* **12**, 5465.

Liu, Q., Liang, Z., Feng, D., Jiang, S., Wang, Y., Du, Z. *et al*. (2021) Transcriptional landscape of rice roots at the single-cell resolution. *Mol. Plant* **14**, 384–394.

Liu, Y., Li, J.S.S., Rodiger, J., Comjean, A., Attrill, H., Antonazzo, G., Brown, N.H. *et al*. (2022) FlyPhoneDB: an integrated web-based resource for cell-cell communication prediction in Drosophila. *Genetics* **220**, iyab235.

Ma, X., Yan, H., Yang, J., Liu, Y., Li, Z., Sheng, M., Cao, Y. *et al*. (2022) PlantGSAD: a comprehensive gene set annotation database for plant species. *Nucleic Acids Res.* **50**, D1456–D1467.

McWhite, C.D., Papoulas, O., Drew, K., Cox, R.M., June, V., Dong, O.X., Kwon, T. *et al*. (2020) A pan-plant protein complex map reveals deep conservation and novel assemblies. *Cell* **181**, 460–474.e14.

Mergner, J., Frejno, M., Messerer, M., Lang, D., Samaras, P., Wilhelm, M., Mayer, K.F.X. *et al*. (2020) Proteomic and transcriptomic profiling of aerial organ development in Arabidopsis. *Sci Data* **7**, 334.

Merret, R., Nagarajan, V.K., Carpentier, M.-C., Park, S., Favory, J.-J., Descombin, J., Picart, C. *et al*. (2015) Heat-induced ribosome pausing triggers mRNA co-translational decay in Arabidopsis thaliana. *Nucleic Acids Res.* **43**, 4121–4132.

Murphy, E., Smith, S. and De Smet, I. (2012) Small signaling peptides in Arabidopsis development: how cells communicate over a short distance. *Plant Cell* **24**, 3198–3217.

Nakaminami, K., Okamoto, M., Higuchi-Takeuchi, M., Yoshizumi, T., Yamaguchi, Y., Fukao, Y., Shimizu, M. *et al*. (2018) AtPep3 is a hormone-like peptide that plays a role in the salinity stress tolerance of plants. *Proc. Natl. Acad. Sci. USA* **115**, 5810–5815.

NCBI Resource Coordinators (2018) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **46**, D8–D13.

Nemhauser, J.L., Hong, F. and Chory, J. (2006) Different plant hormones regulate similar processes through largely nonoverlapping transcriptional responses. *Cell* **126**, 467–475.

Oh, E., Seo, P.J. and Kim, J. (2018) Signaling peptides and receptors coordinating plant root development. *Trends Plant Sci.* **23**, 337–351.

Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.-J., Stark, C., Willems, A., Boucher, L. *et al*. (2021) The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.* **30**, 187–200.

Ren, X., Zhong, G., Zhang, Q., Zhang, L., Sun, Y. and Zhang, Z. (2020) Reconstruction of cell spatial organization from single-cell RNA sequencing data based on ligand-receptor mediated self-assembly. *Cell Res.* **30**, 763–778.

Restrepo-Montoya, D., Brueggeman, R., McClean, P.E. and Osorno, J.M. (2020) Computational identification of receptor-like kinases "RLK" and receptor-like proteins "RLP" in legumes. *BMC Genomics* **21**, 459.

Sales, G. and Romualdi, C. (2011) parmigene—a parallel R package for mutual information estimation and gene network reconstruction. *Bioinformatics* **27**, 1876–1877.

Shao, X., Liao, J., Li, C., Lu, X., Cheng, J. and Fan, X. (2021) CellTalkDB: a manually curated database of ligand–receptor interactions in humans and mice. *Brief. Bioinform.* **22**, bbaa269.

Shiu, S.-H. and Bleecker, A.B. (2001) Receptor-like kinases from Arabidopsis form a monophyletic gene family related to animal receptor kinases. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 10763–10768.

Shiu, S.-H., Karlowski, W.M., Pan, R., Tzeng, Y.-H., Mayer, K.F.X. and Li, W.-H. (2004) Comparative analysis of the receptor-like kinase family in Arabidopsis and rice. *Plant Cell* **16**, 1220–1234.

Smakowska-Luzan, E., Mott, G.A., Parys, K., Stegmann, M., Howton, T.C., Layeghifard, M., Neuhold, J. *et al*. (2018) An extracellular network of Arabidopsis leucine-rich repeat receptor kinases. *Nature* **553**, 342–346.

Sonnhammer, E.L.L. and Östlund, G. (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.* **43**, D234–D239.

Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M. *et al*. (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613.

Takahashi, F., Suzuki, T., Osakabe, Y., Betsuyaku, S., Kondo, Y., Dohmae, N., Fukuda, H. *et al*. (2018) A small peptide modulates stomatal control via abscisic acid in long-distance signalling. *Nature* **556**, 235–238.

Thibivilliers, S. and Libault, M. (2021) Enhancing our understanding of plant cell-to-cell interactions using single-cell omics. *Front. Plant Sci.* **12**, 1585.

Tsuyuzaki, K., Ishii, M. and Nikaido, I. (2019) *Uncovering hypergraphs of cell-cell interaction from single cell RNA-sequencing data*, 566182.

Turco, G.M., Rodriguez-Medina, J., Siebert, S., Han, D., Valderrama-Gómez, M.Á., Vahldick, H., Shulse, C.N. *et al*. (2019) Molecular mechanisms driving switch behavior in xylem cell differentiation. *Cell Rep.* **28**, 342–351.e4.

Wang, C., Sun, D., Huang, X., Wan, C., Li, Z., Han, Y., Qin, Q. *et al*. (2020) Integrative analyses of single-cell transcriptome and regulome using MAESTRO. *Genome Biol.* **21**, 198.

Wang, Y., Huan, Q., Li, K. and Qian, W. (2021) Single-cell transcriptome atlas of the leaf and root of rice seedlings. *J. Genet. Genomics* **48**, 881–898.

Wendrich, J.R., Yang, B., Vandamme, N., Verstaen, K., Smet, W., Van de Velde, C. *et al*. (2020) Vascular transcription factors guide plant epidermal responses to limiting phosphate conditions. *Science* **370**, eaay4970.

Xu, X., Crow, M., Rice, B.R., Li, F., Harris, B., Liu, L., Demesa-Arevalo, E. *et al*. (2021) Single-cell RNA sequencing of developing maize ears facilitates functional analysis and trait candidate gene discovery. *Dev. Cell* **56**, 557–568.e6.

Zhang, H., Li, X., Wang, W., Li, H., Cui, Y., Zhu, Y., Kui, H. *et al*. (2022) SERKs regulate embryonic cuticle integrity through the TWS1-GSO1/2 signaling pathway in Arabidopsis. *New Phytol.* **233**, 313–328.

Zhong, S., Li, L., Wang, Z., Ge, Z., Li, Q., Bleckmann, A., Wang, J. *et al*. (2022) RALF peptide signaling controls the polytubey block in Arabidopsis. *Science* **375**, 290–296.

Zhou, G., Cichocki, A., Zhao, Q. and Xie, S. (2014) Nonnegative Matrix and Tensor Factorizations: an algorithmic perspective. *IEEE Sig. Proc. Magazine* **31**, 54–65.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1 Spearman correlation analysis and number of ligand-receptor pair comparison of PlantPhoneDB with scTensor.
Figure S2 Functions of PlantPhoneDB web interface.
Figure S3 Performance evaluation of each classifier on inter-datasets and intra-dataset models using the F1-score.
Figure S4 Comparison of databases and cell–cell communication network.
Figure S5 Application of ligand-receptor pairs in cell–cell communication.
Figure S6 Graphical representations.
Data S1 The detailed information of 29 scRNA-seq datasets, covering ~560 000 cells of 15 tissues from 5 plant species.
Data S2 The top 200 differentially expressed genes in lateral root cells.
Data S3 Overview of the datasets used in benchmark analysis.
Data S4 Compares the top communicating cell-type pairs using four scoring approaches on 3 k 10 × PBMCs dataset.
Data S5 Compares the top communicating cell-type pairs using four scoring approaches on 8 k 10 × PBMCs dataset.
Data S6 Identification of 1640 significant ligand-receptor pairs between pairwise cell types on GSE121619 dataset.
Data S7 Construction of intracellular signalling pathway.
Data S8 Automatic cell identification methods included in this study.
Data S9 Gene expression matrix from flower, rosette leaf, silique, and seed tissue of *Arabidopsis thaliana*.

