Cross-species single-cell annotation with orthologous marker gene groups

Tran N. Chau[1,2], Prakash Raj Timilsena[1], Sai Pavan Bathala[3], Bastiaan O. R. Bargmann[1], Song Li[1,2]

Abstract

The lack of known cell-type specific marker genes in most plants is a major hurdle for single-cell analysis. To address this challenge, we developed an approach to identify Orthologous Marker Gene groups (OMGs) across monocots and dicots and showed that they are capable of determining cell identities in tomato. This result indicates employing conserved OMGs from reference single-cell maps can annotate cell types in other plant species.

Single-cell RNA sequencing (scRNA-seq) technology has been widely used in identifying and characterizing various cell types in numerous tissues from different plant species[1–6]. Efforts to apply this technology in the model plant Arabidopsis have benefited from the extensive knowledge of cell-type identity markers[7,8]. Contrastingly, accurate labeling of cell types in other plant species remains a challenge due to the scarcity of known cell-type marker genes[9]. Various approaches have been explored to address this issue, such as using literature-derived marker gene sets[10], direct integration of data from diverse species[11], or comparing the gene expression profiles of single-copy orthologous genes between closely related species[12]. However, these studies also found many cell-type marker genes in Arabidopsis do not exhibit the same cell-type identity across diverse plant species[3,10,11,12].

Many cross-species integration methods have been developed in non-plant systems[13], which relied on matching gene expression matrices based on one-to-one orthologous genes. Although these methods were successful for non-plant systems, we have found direct integration of single-cell data from diverse plant species resulted in mixed clusters with variegated cell identities (Supplementary Fig. 1, Table S1). Because one-to-one orthologous genes were assigned solely based on sequence similarity, many such genes may have diverged in their cell-type specific expression levels in different plant species (e.g. Supplementary Fig. 2). To tackle this challenge in annotating cell clusters across species, we have developed a pipeline to identify OMGs (Fig. 1a), which include not only one-to-one but also one-to-many and many-to-many orthologous genes. Cell type assignments were based on counting overlapping OMGs between a reference single-cell map from a well-established model system and a species that lacks marker genes. This method eliminated the step to explicitly determine the best orthologous genes based on both expression and sequence similarities.

Using OMGs to assign cell types to single-cell clusters is fast and memory efficient because this method does not require complex cross-species data integration. First, we identified the top N marker genes (N=200) for each cell cluster in each species using Seurat (v4.0). This choice of N maintains a balance by ensuring sufficient overlapping OMGs between clusters across species while preserving marker gene specificity. Second, we employed OrthoFinder (v2.5.4) to generate orthologous gene

groups for M plant species (m=4, including Arabidopsis, rice, maize, and tomato). Third, we performed a pairwise comparison using an overlapping OMG heatmap followed by a statistical test (i.e., Fisher's exact test) to determine clusters with a significant number of shared OMGs (Fig. 1a). The scripts for the analysis can be found at our GitHub repository (https://github.com/LiLabAtVT/OrthoMarkerGenes) and shinyR website for visual comparisons (https://orthomarkergenes.shinyapps.io/omg_browser_version1/).

Using rice and Arabidopsis as one example, we found 11 clusters in Arabidopsis and 10 clusters in rice have corresponding cell types in the counterpart species (Fig. 1b, Table S2). From a total of 110 possible pairwise comparisons, we identified 15 statistically significant cell cluster pairs (FDR < 0.01). Among these 15 cluster pairs, 10 pairs are between identical or near identical orthologous cell types, including hair, non-hair, xylem, cortex, and phloem cells. We also observed mismatched clusters, and interestingly, four of these mismatched clusters are in the middle region of the Arabidopsis UMAP, suggesting these are relatively undifferentiated cell clusters (Supplementary Fig. 3). Similarly, this concordance of cell types was found between other pairwise comparisons (Supplementary Fig. 4, Table S2), with 7 out of 11 for maize vs. rice, and 4 out of 9 for Arabidopsis vs. maize. We further investigated the overlapping of marker genes between three species and found that xylem and cortex exhibited a higher number of OMGs across three species as compared to hair and nonhair cells (Fig. 1c, Table S3).

Despite the morphological and functional similarity of hair and non-hair cells in these species, we found only 1 OMG shared across three species for each of these cell types respectively (Fig. 1c). One way to address this problem is to increase the number of marker genes (i.e., N > 200) with the same statistical method but reduce the specificity threshold of these genes. To increase the number of markers without sacrificing the specificity, we have demonstrated that applying feature selection in machine learning enabled us to identify markers distinct from those identified by statistical approaches[7]. Therefore, we employed two additional machine-learning (ML) methods (SHAP-RF and SVM) to identify the top 200 marker genes for each cell-type cluster in each species. We found a high level of agreement between the ML-derived markers and fewer markers were shared between ML methods and the Seurat method (Fig. 1d, Table S6). With ML-derived markers, we found 2 additional OMGs for each of the hair/non-hair clusters in three species and new OMGs between stele clusters and phloem clusters (Fig. 1e, Table S4,5). These results suggest using different methods to identify cluster marker genes could lead to new markers that were not found with a single method. However, it also highlights the substantial reduction of conserved marker genes in certain cell types when trying to find them across multiple diverse species.

To assess the applicability of using OMGs to identify cell types in another plant species, we utilized tomato root data as our test dataset[4]. Tomato root is an ideal system because single-cell clusters were assigned based on comparison with bulk cell type-specific expression analysis with promoter-GFP marker lines and cell sorting. The cell type assignment in tomato provides a higher confidence in the cell-type identities[4,14] (Supplementary Fig. 5) as compared to species without such data. We found that

conserved OMGs across three species are also enriched in corresponding cell types in tomato root single-cell data (Fig. 2a). We further performed a pairwise comparison between tomato and Arabidopsis or rice respectively (Fig. 2b,c, Table S2), and we found 8 out of 11 and 11 out of 12 significant pairwise comparisons are between orthologous or similar cell types in these species. Finally, we extended our analysis to tissue types beyond plant roots, where established marker genes are rare. We utilized shoot apical meristem tissues from Arabidopsis and tomato as our model system[6] and we observed 13 out of 17 significant overlapping clusters in these pairwise comparisons from identical or near-identical cell types between two species (Fig. 2d, Table S2).

In conclusion, the OMG framework is a straightforward and effective method to assign cell types across multiple plant species using single-cell reference maps. This method is also inherently "explainable" because the lists of overlapping marker genes can be inspected by the user. Most cell clusters in our analysis only showed one or two orthologous cell clusters that are statistically significant, suggesting a high specificity of this method. However, limitations of this method include that it can only assign cell types to each cluster, and not to individual cells. Different pre-processing methods and the choice of the number of clusters can potentially result in different markers from the same dataset, which create a hurdle for determining the optimal number of clusters and clustering methods. This work can be further improved with community-based agreement on meta-data publication standards such that it is easier to reproduce the published clusters, extract marker genes automatically, and to enable faster cross-species comparison in many plant species.

Fig. 1: Common OMGs detection.
A. the OMG pipeline. **B**. pairwise comparison of cell clusters from rice and Arabidopsis. Each block displays the number of conserved marker genes between the two cell types being compared, with darker colors indicating a higher frequency of conserved OMGs and lighter colors representing fewer conserved OMGs. The red boxes highlight statistically significant overlaps. **C**. conserved Seurat-OMGs across three species, where darker colors correspond to a larger number of OMGs. **D.** comparison of the top 200 marker genes found by Seurat and machine learning methods. The height of the square and the thickness of the connecting lines represent the number of common marker genes found by the two methods, with each color representing a distinct cell-type cluster. **E.** conserved OMGs found by machine learning approaches.

Fig. 2: Cell type identification in another plant species.
A. using conserved OMGs as markers for identifying different cell types in tomato roots, with the color intensity indicating the expression level. **B.** pairwise comparison between Arabidopsis and tomato root data with marker genes found by the Seurat method. **C.** pairwise comparison of rice and tomato root cell types, highlighting the number of conserved OMGs using marker genes identified by the SHAP method. **D.** comparison of shoot OMGs between Arabidopsis and tomato with marker genes found by Seurat method.

Methods:

Cell type assignment for published scRNA-seq datasets of three plant species

In this study, gene expression data from three different plant species, namely *Arabidopsis thaliana*, *Zea mays*, and *Oryza sativa* were analyzed. The data consisted of 7522 Arabidopsis cells, 14,733 *Zea mays* cells, and 27,426 *Oryza sativa* cells, which were published by Ryu et al., Ortiz-Ramírez et al., and Zhang et al., respectively (GSE123013, GSE172302, PRJNA706435). To process the gene expression matrix, a standard workflow that included the NormalizeData, FindVariable, and ScaleData functions, was used with default parameters in the Seurat package (verseion 4.1.1). Principal component analysis (PCA) was used to reduce the high-dimensional data into 30, 15, and 32 principal components (PCs) for Arabidopsis, maize, and rice, respectively. The FindCluster function with resolutions of 0.5, 0.55, and 0.4 was applied to these PCs using the Louvain algorithm to identify the clusters, and the uniform manifold approximation and projection (UMAP) method was used to visualize all cells in two dimensions. These parameters were selected to reproduce UMAP plots that best resemble those in the published results. Finally, lists of known marker genes from these published papers were used to label all the clusters of cells. If a cluster did not include any marker genes from the publication, the cluster is labeled as “unknown”. The cell-type assignment scripts can be found in our GitHub repository (https://github.com/LiLabAtVT/OrthoMarkerGenes). These analyses produced species-specific UMAPs.

Integration of three plant species datasets together

To integrate scRNA-seq datasets from three species, the Seurat integration pipeline was used. First, genes from each species were converted into orthologous gene IDs, and 5000 high variable genes were identified and ranked from 1 to 5000 using the FindVariable function and nfeature=5000. Since one orthologous gene can consist of one or more species of genes, ranking the list of high-variable genes supported the selection of the top genes in each orthologous group for integration. Second, top 2000 (out of 5000) unique high-variable orthologous genes corresponding to the top high variable genes for each species were selected to ensure that the most informative genes were used in the integration process. To ensure the quality of each dataset, the merged Seurat object of three datasets was split into its individual components, and each was independently normalized and scaled using the SCTransform method. The SelectIntegrationFeatures function was then used to identify the top informative features for integrating multiple datasets, and the PreSCTIntegration function was employed to ensure that the Pearson residuals for the selected features were present in each Seurat object. To integrate the three datasets, the FindIntegrationAnchors function was used to identify a set of anchors between all Seurat objects in the list, and these anchors were then used to integrate the three objects together with IntegrateData. Finally, the standard workflow for processing scRNA-seq data was applied, including RunPCA, RunUMAP, FindNeighbors, and FindClusters. The clustering was performed using 30 dimensions and a resolution of 0.3. This analysis produced an integrated UMAP for Arabidopsis, rice, and maize.

The integrated metadata was annotated based on the cell-type annotation obtained from the original analyses where each species was analyzed separately. This step avoided the potential overfitting in the integration process that could lead to loss of biologically important variations across species. To investigate whether the number of homogenous clusters would increase, the resolution parameter of FindClusters was modified to 0.9, resulting in the number of clusters being increased to 30. In addition, to examine the potential impact of heterogenous clusters caused by different integration methods, the harmony integration was employed by applying the RunHarmony function after conducting PCA.

Marker gene identification by differential gene expression analysis

The FindAllMarkers function in the Seurat package was used to detect marker genes in species-specific UMAPs. A default differential expression test of expression level in a single cell type versus the expression in all other cell types was performed based on the Wilcoxon rank-sum test. To detect marker genes, we included only.pos = TRUE, which selects only positive log-fold changes, min.pct = 0.3, which sets a minimum percentage of cells that express the gene, and logfc.threshold = 0.45, which sets the threshold for log-fold change to select significant genes. After running the FindAllMarkers function, the list of differentially expressed genes was sorted based on their log-fold change values, and the top 200 most significant marker genes were selected for each cluster.

Marker gene identification by SHAP and SVM methods

We used the SPMarker Python package to identify marker genes in our Arabidopsis, rice, maize, and tomato datasets (Table S7). The SPMarker package employed two feature selection techniques, namely Shapley additive explanations (SHAP) based on random forest and support vector machine, to develop a pipeline for calculating the contribution of each gene on the cell type assignment model. This approach allowed us to handle imbalanced data caused by major and minor clusters in scRNA-seq data and identify unique marker genes in each cell type cluster. To generate the cell type and gene expression matrix files, a custom R script was used with Seurat's FindVariable function to select 5000 high-variable genes and the SCTransform function to normalize the selected dataset. Using the SHAP value and SVM coefficients, cell type-specific markers were identified in the species-specific UMAPs. Similar to the default Seurat method, the top 200 most significant marker genes were selected for each cluster.

Identification of conserved orthologous marker genes and analysis of the conserved orthologous marker genes between two species

OrthoFinder software (v2.5.4) was used to search homologous genes and cluster these genes into ortho-groups (Table S8). Protein sequences in FASTA format for each species were downloaded from the Ensembl plants database and used as input for OrthoFinder analysis. Species-specific marker gene names were mapped to the ortho-group name and these ortho-groups are named Orthologous Marker gene Groups (OMGs). These OMGs were used in downstream analyses, in place of the original marker genes, to enable comparison between species.

After obtaining lists of orthologous marker genes for all cell types in each of the species, the number of common orthologous marker genes between any two species was calculated. To evaluate the significance of the conserved marker gene list, Fisher's exact independent tests were performed, hypothesizing that the orthologous marker gene list for each cell-type cluster from one species was independent of those from another species. Fisher's exact test is a statistical method used to analyze the association between two categorical variables, and in this case, it was used to determine the significance of the overlap in marker genes between species. The p-value from the Fisher test could provide evidence to evaluate the significance of the conserved marker gene list. The p-value is obtained through the R code:

```
p_value_df[i,j] = fisher.test(rbind(c(a,b),c(c,d)), alternative = "greater")$p.value
```

df[i,j] is a matrix where i ∈ (1, M) where M is the number of clusters in species 1 and
j ∈ (1, N) where N is the number of clusters in species 2;

a = OMG[I,J]

b = $\sum_{j=1\ (j\ \neq J)}^{N} OMG[I,j]$

c = $\sum_{i=1\ (i\ \neq I)}^{M} OMG[i,J]$

d = $\sum_{j,i=1}^{N,M} OMG[i,j]$ -a -b -c

To address the issue of multiple testing, we applied an FDR correction to the p-values obtained from the Fisher exact test. A threshold of 0.01 was then set on the q-value to determine which tests were significant.

References:

1. Ryu, K. H., Huang, L., Kang, H. M. & Schiefelbein, J. Single-cell RNA sequencing resolves molecular relationships among individual plant cells. *Plant Physiol* **179**, 1444–1456 (2019).
2. Zhang, T. Q., Chen, Y., Liu, Y., Lin, W. H. & Wang, J. W. Single-cell transcriptome atlas and chromatin accessibility landscape reveal differentiation trajectories in the rice root. *Nat Commun* **12**, (2021).
3. Ortiz-Ramírez, C. *et al.* Ground tissue circuitry regulates organ complexity in maize and Setaria. *Science (1979)* **374**, 1247–1252 (2021).
4. Cantó-Pastor, A. *et al.* A suberized exodermis is required for tomato drought tolerance. doi:10.1101/2022.10.10.511665.
5. Zhang, T. Q., Chen, Y. & Wang, J. W. A single-cell analysis of the Arabidopsis vegetative shoot apex. *Dev Cell* **56**, 1056-1074.e8 (2021).
6. Tian, C., Du, Q., Xu, M., Du, F. & Jiao, Y. Single-nucleus RNA-seq resolves spatiotemporal developmental 1 trajectories in the tomato shoot apex 2. doi:10.1101/2020.09.20.305029.
7. Yan, H. *et al.* Identification of new marker genes from plant single-cell RNA-seq data using interpretable machine learning methods. *New Phytol* **234**, 1507–1520 (2022).

8. Shahan, R. *et al.* A single-cell Arabidopsis root atlas reveals developmental trajectories in wild-type and cell identity mutants. *Dev Cell* **57**, 543-560.e9 (2022).
9. Ryu, K. H., Zhu, Y. & Schiefelbein, J. Plant Cell Identity in the Era of Single-Cell Transcriptomics. *Annu Rev Genet* **55**, 479–496 (2021).
10. Marand, A. P., Chen, Z., Gallavotti, A. & Schmitz, R. J. A cis-regulatory atlas in maize at single-cell resolution. *Cell* **184**, 3041-3055.e21 (2021).
11. Liu, Q. *et al.* Transcriptional landscape of rice roots at the single-cell resolution. *Mol Plant* **14**, 384–394 (2021).
12. Guillotin, B. *et al.* A pan-grass transcriptome reveals patterns of cellular divergence in crops. *Nature* (2023) doi:10.1038/s41586-023-06053-0.
13. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* **19**, 41–50 (2022).
14. Kajala, K. *et al.* Innovation, conservation, and repurposing of gene function in root cell type development. *Cell* **184**, 3333-3348.e19 (2021).

Author information:

Authors and Affiliations

School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA, USA
Tran N. Chau, Prakash Raj Timilsena, Bastiaan O. R. Bargmann & Song Li

Graduate Program in Genetics, Bioinformatics, and Computational Biology, Virginia Tech, Blacksburg, VA, USA
Tran N. Chau & Song Li

Department of Computer Science, Virginia Tech, Blacksburg, VA, USA
Sai Pavan Bathala

Contributions

S.L. developed the project and supervised its execution. T.N.C. performed the analyses. S.P.B. created the web browser. P.R.T. provided tomato data. T.N.C & S.L. wrote the manuscript with contributions from B.O.R.B.

Corresponding authors

Correspondence to Song Li

Ethics declarations:

Competing interests

The authors declare no competing interests.

Peer review:

Peer review information

Supplementary tables:

Table S1: Seurat integration with 19 clusters, 30 clusters, and Harmony integration.

Table S2: Pairwise comparison of rice vs Arabidopsis, maize vs Arabidopsis, maize vs rice, tomato vs Arabidopsis, tomato vs rice with marker genes found by SHAP method, and tomato vs Arabidopsis shoot. Only the data corresponding to the highlighted red boxes are presented in the table.

Table S3: Conserved OMGs across three species using marker genes identified by the Seurat function.

Table S4: Conserved OMGs across three species using marker genes identified by the SHAP method.

Table S5: Conserved OMGs across three species using marker genes identified by the SVM method.

Table S6: The Sankey diagram displays the number of common marker genes between different methods.

Table S7: OMGs with top 200 marker genes found by Seurat, SHAP, and SVM for each cell-type cluster in Arabidopsis, rice, maize, and tomato.

Table S8: Orthologous genes are identified by OrthoFinder software.

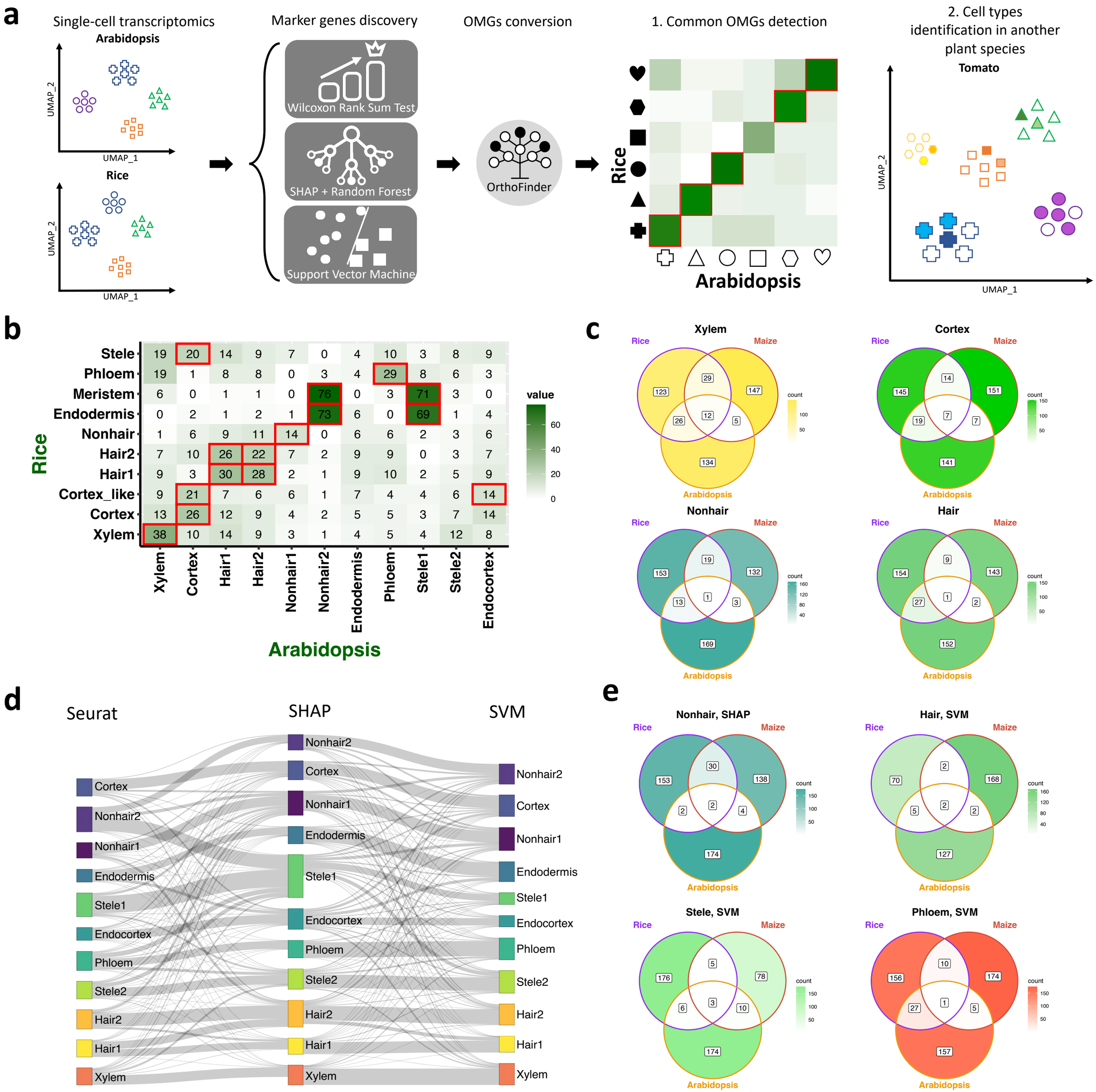

a
Single-cell transcriptomics
Arabidopsis
UMAP_2
UMAP_1
Rice
UMAP_2
UMAP_1
Marker genes discovery
Wilcoxon Rank Sum Test
SHAP + Random Forest
Support Vector Machine
OMGs conversion
OrthoFinder
1. Common OMGs detection
Rice
Arabidopsis
2. Cell types identification in another plant species
Tomato
UMAP_2
UMAP_1
b
Rice
Arabidopsis
Stele 19 20 14 9 7 0 4 10 3 8 9
Phloem 19 1 8 8 0 3 4 29 8 6 3
Meristem 6 0 1 1 0 76 0 3 71 3 0
Endodermis 0 2 1 2 1 73 6 0 69 1 4
Nonhair 1 6 9 11 14 0 6 6 2 3 6
Hair2 7 10 26 22 7 2 9 9 0 3 7
Hair1 9 3 30 28 2 1 9 10 2 5 9
Cortex_like 9 21 7 6 6 1 7 4 4 6 14
Cortex 13 26 12 9 4 2 5 5 3 7 14
Xylem 38 10 14 9 3 1 4 5 4 12 8
Xylem Cortex Hair1 Hair2 Nonhair1 Nonhair2 Endodermis Phloem Stele1 Stele2 Endocortex
value
60
40
20
0
c
Xylem
Rice
Maize
Arabidopsis
123 29 147 26 12 5 134
count
100
50
Cortex
Rice
Maize
Arabidopsis
145 14 151 19 7 7 141
count
150
100
50
Nonhair
Rice
Maize
Arabidopsis
153 19 132 13 1 3 169
count
160
120
80
40
Hair
Rice
Maize
Arabidopsis
154 9 143 27 1 2 152
count
150
100
50
d
Seurat
SHAP
SVM
Cortex
Nonhair2
Nonhair1
Endodermis
Stele1
Endocortex
Phloem
Stele2
Hair2
Hair1
Xylem
Nonhair2
Cortex
Nonhair1
Endodermis
Stele1
Endocortex
Phloem
Stele2
Hair2
Hair1
Xylem
Nonhair2
Cortex
Nonhair1
Endodermis
Stele1
Endocortex
Phloem
Stele2
Hair2
Hair1
Xylem
e
Nonhair, SHAP
Rice
Maize
Arabidopsis
153 30 138 2 2 4 174
count
150
100
50
Hair, SVM
Rice
Maize
Arabidopsis
70 2 168 5 2 2 127
count
160
120
80
40
Stele, SVM
Rice
Maize
Arabidopsis
176 5 78 6 3 10 174
count
150
100
50
Phloem, SVM
Rice
Maize
Arabidopsis
156 10 174 27 1 5 157
count
150
100
50

Fig. 1:

Fig. 1: a, the OMG pipeline. **b**, pairwise comparison of cell clusters from rice and Arabidopsis. Each block displays the number of conserved marker genes between the two cell types being compared, with darker colors indicating a higher frequency of conserved OMGs and lighter colors representing fewer conserved OMGs. The red highlighted cells indicate significant results based on Fisher's exact test ($p<0.01$). **c**, conserved OMGs across three species, employing the Seurat method, where darker colors correspond to a larger number of OMGs, and vice versa. **d**, comparison of the top 200 marker genes found by Seurat and machine learning methods. The height of the square and the thickness of the connecting lines represent the number of common marker genes found by the two methods, with each square's color representing a distinct cell type cluster. **e**, conserved OMGs found by machine learning approaches.

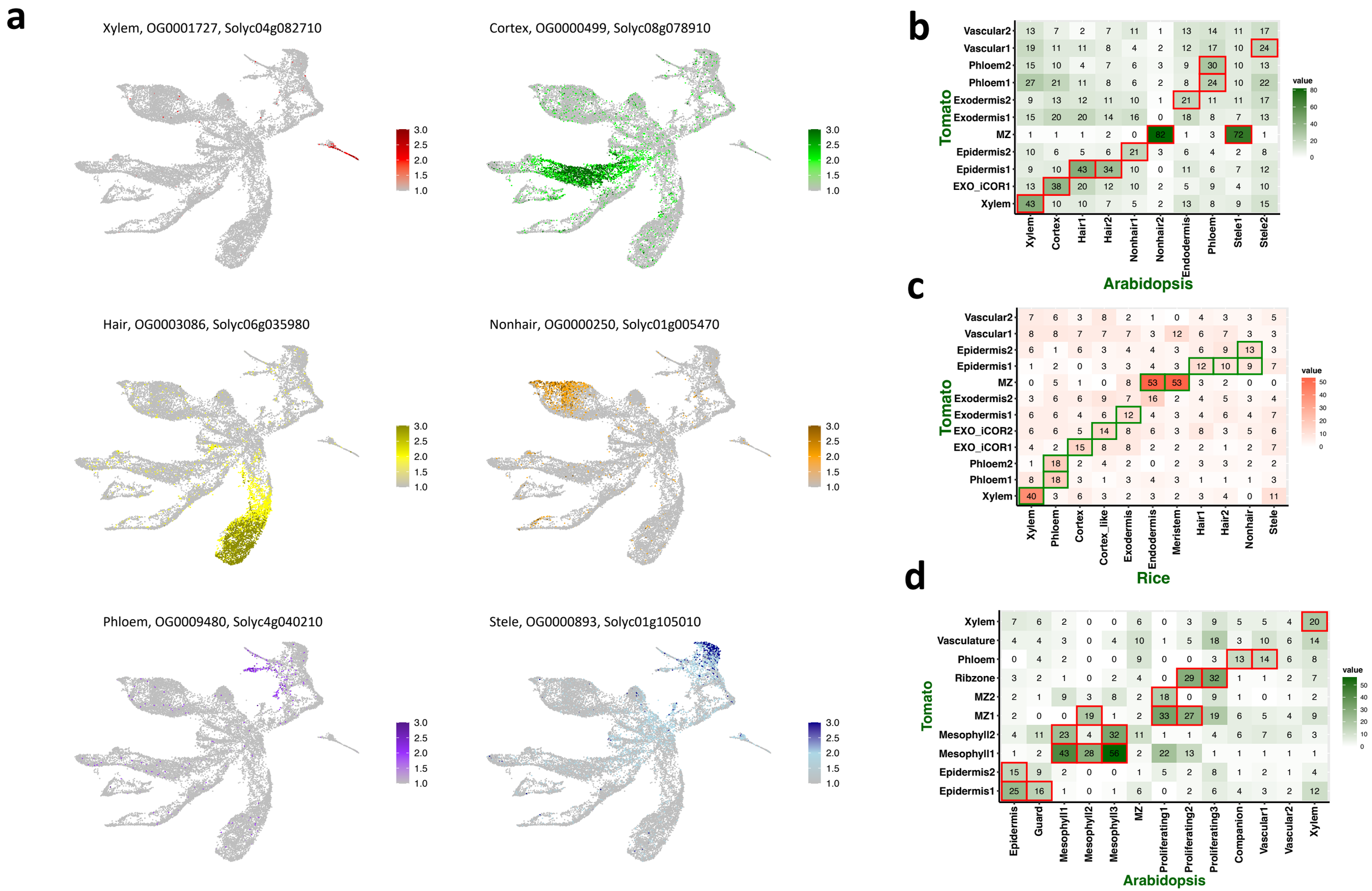


Fig. 2: a, conserved OMGs serve as markers for identifying different cell types in tomato roots, with the color intensity indicating the expression level. **b**, pairwise comparison between Arabidopsis and tomato with marker genes found by the Seurat method. The numbers within each cell denote the count of conserved marker genes between the corresponding cell types and the red boxes highlight significant results based on Fisher's exact test. **c**, pairwise comparison of rice and tomato root cell types, highlighting the number of conserved OMGs using marker genes identified by the SHAP method. **d**, comparison of shoot OMGs between Arabidopsis and tomato with marker genes found by Seurat function.

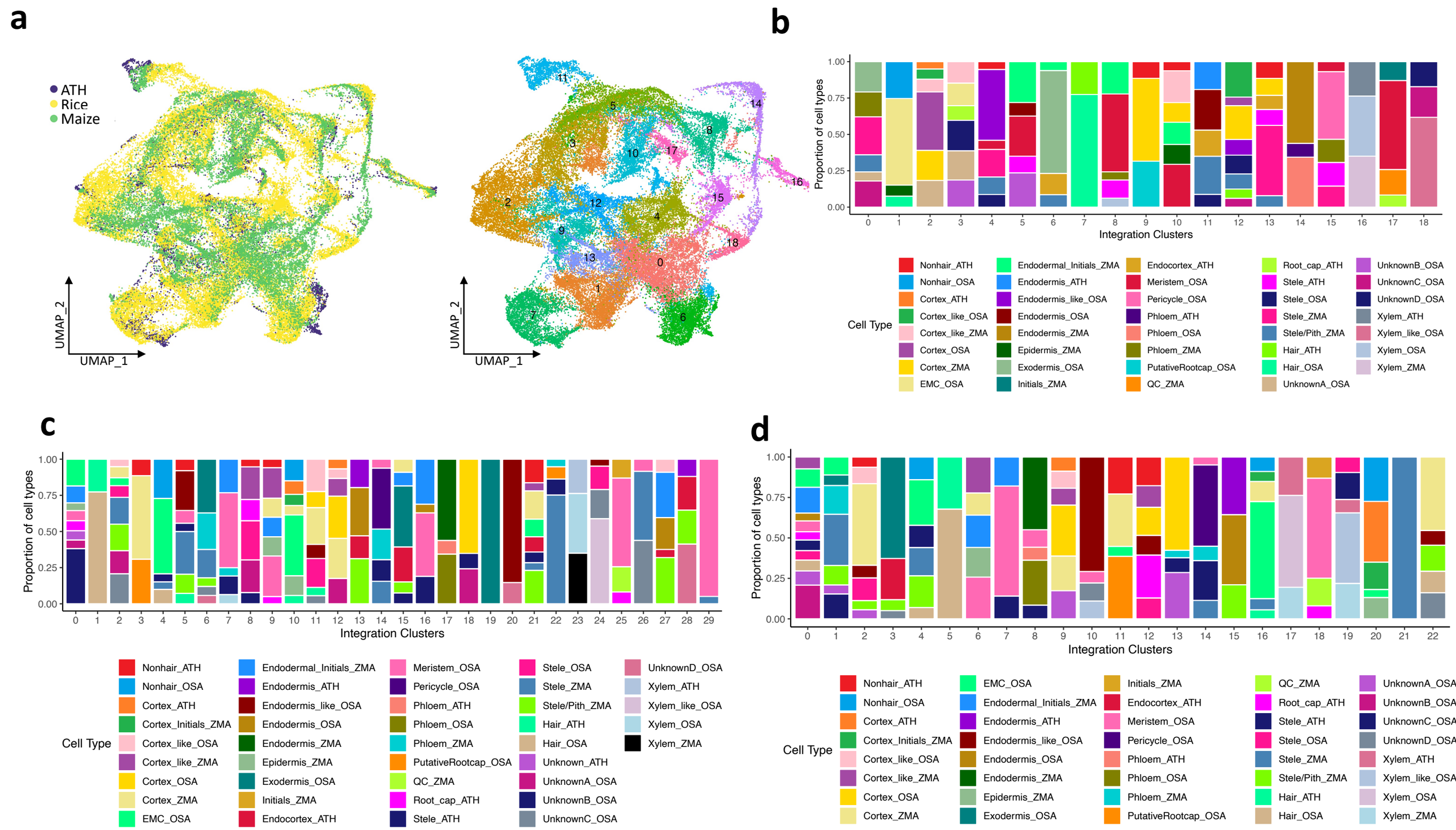


Fig. S1: a, integration of rice, maize, and Arabidopsis single-cell data. Left colors indicate different species. Right, colors indicate different clusters. **b**, bar plot shows all clusters are made of mixed cell types when using 19 clusters in the integrated map. Each color in the plot corresponds to a specific cell type cluster in a particular species, with the label indicating the cell type followed by the species ATH, OSA, or ZMA. **c**, Bar plot shows most clusters are made of mixed cell types when using 30 clusters. **d**, Bar plot shows most clusters are made of mixed cell types when applying Harmony integration.

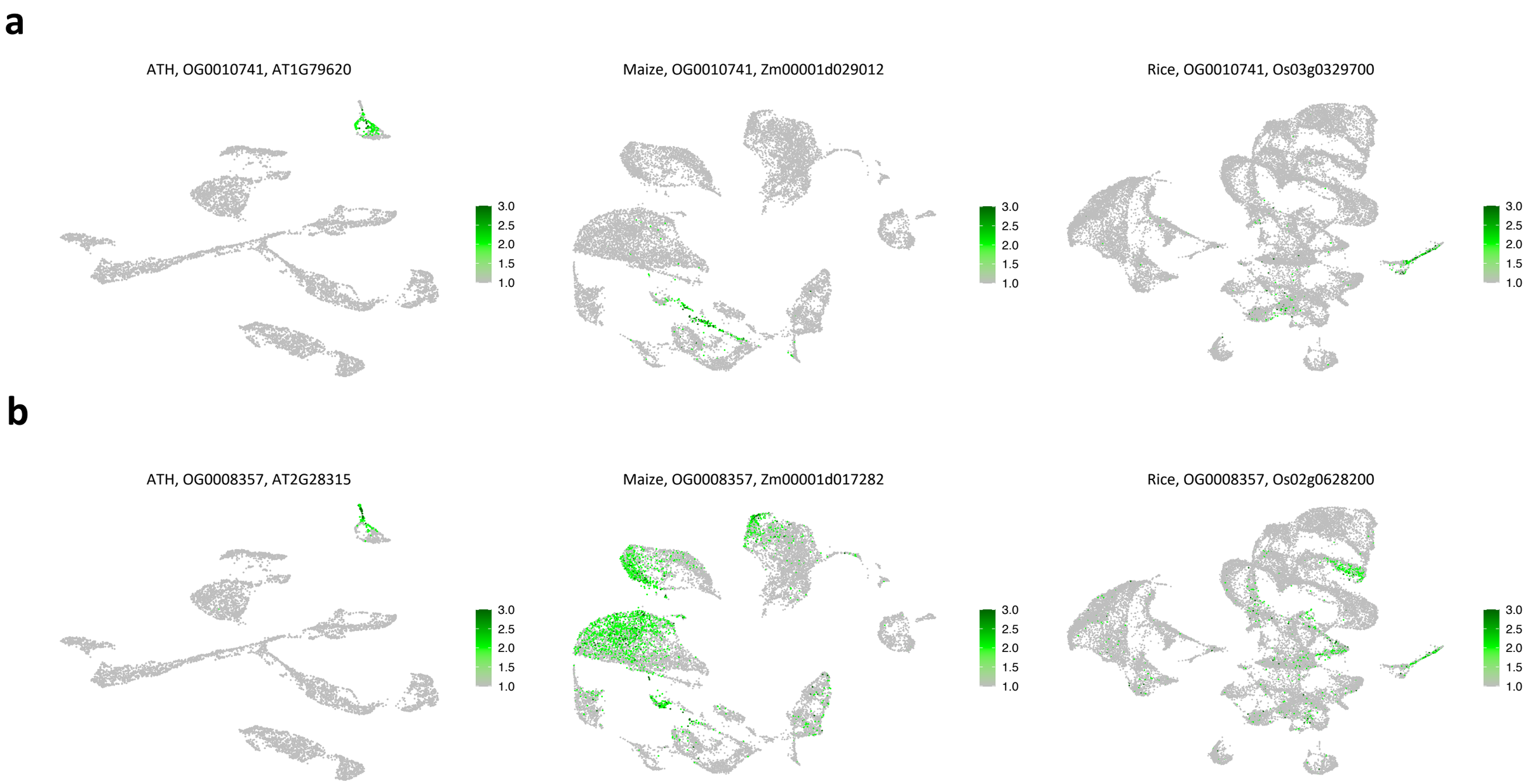


Fig. S2: a, example of one-to-one conserved markers for xylem. **b**, example of a one-to-one conserved gene. This genes is a marker for xylem in Arabidopsis, but this gene is not specific to any cell type in maize or rice. The color intensity corresponds to the level of gene expression.

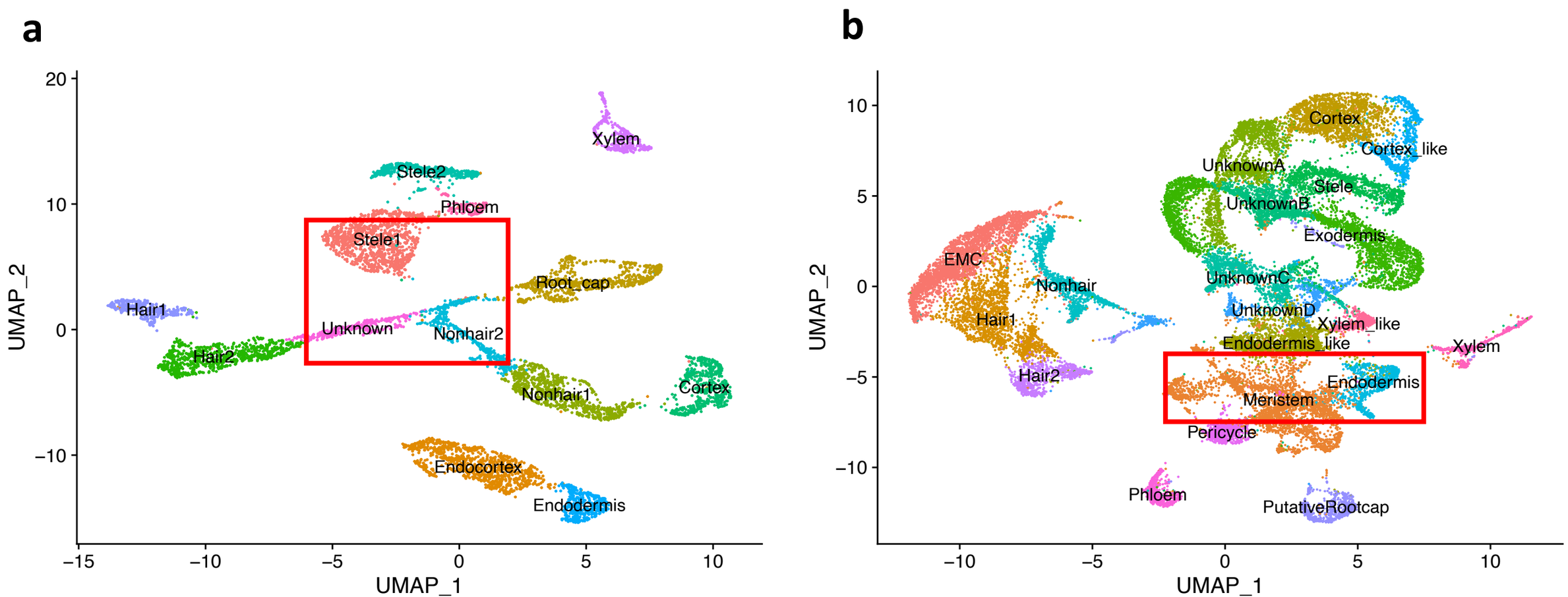
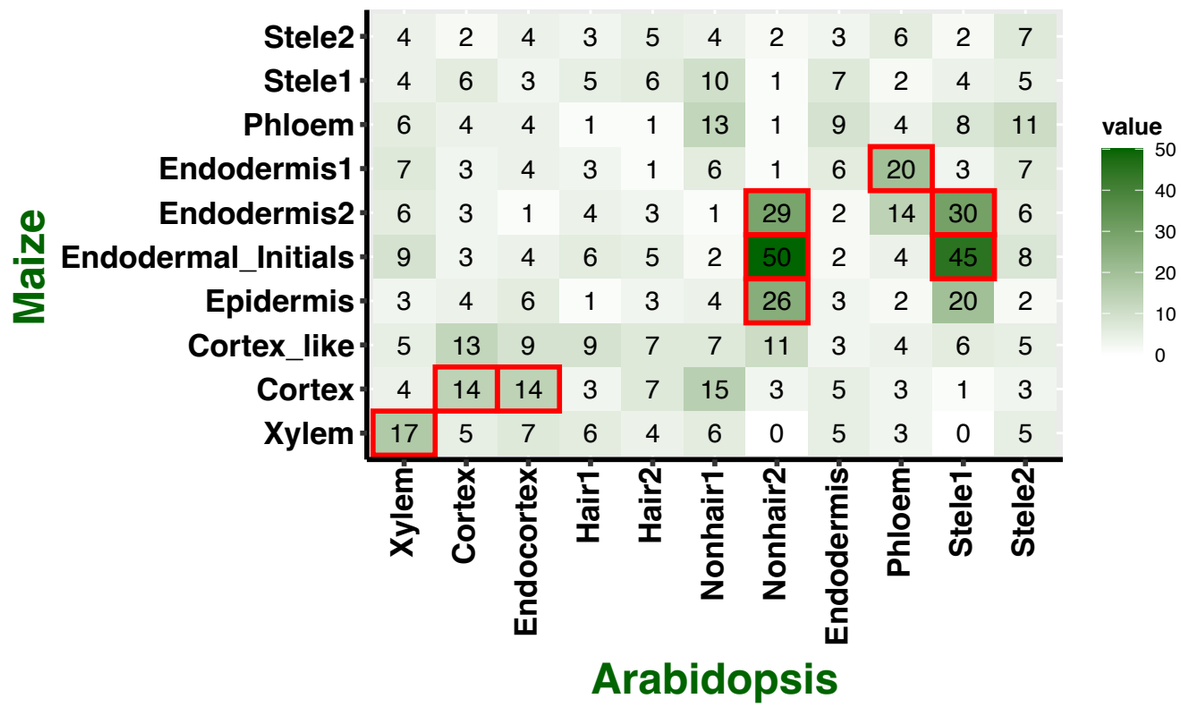


Fig. S3: The UMAP visualizations of the clustered cell groups of 7522 Arabidopsis cells **a**, and 27,426 Rice cells **b**, demonstrate that clusters such as the Stele1, Nonhair2, Meristem, and Endodermis, which are positioned at the center of the plots, exhibit a higher number of overlapping OMGs and this overlap can result in a mismatch between the cell type clusters of the two species (Fig. 1). Each dot represents an individual cell, with its color representing the respective cell type.



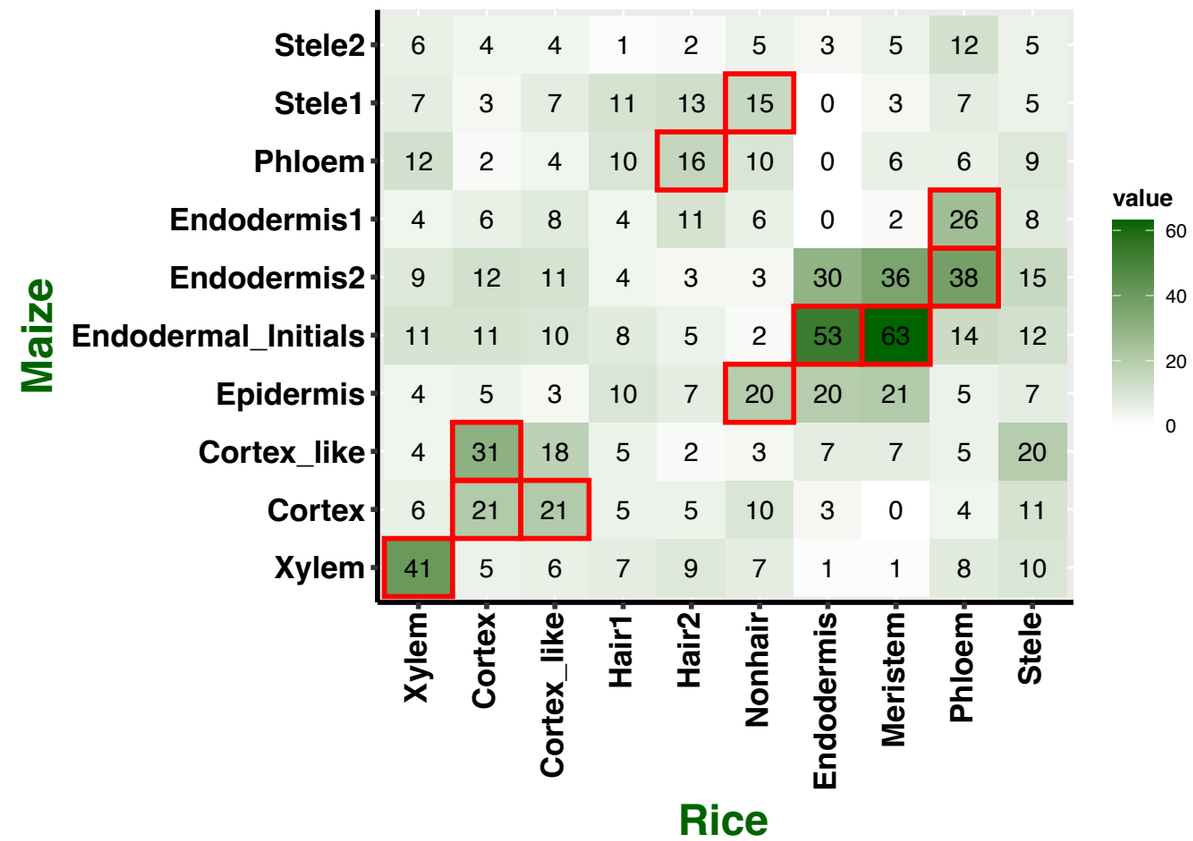


Fig. S4: Overlapping of OMGs between **a**, maize and Arabidopsis, and **b**, maize and rice. The numbers displayed in each cell indicate the count of conserved marker genes between the two cell-type clusters. Significance is denoted by the red boxes, which are based on the results of Fisher's exact test.

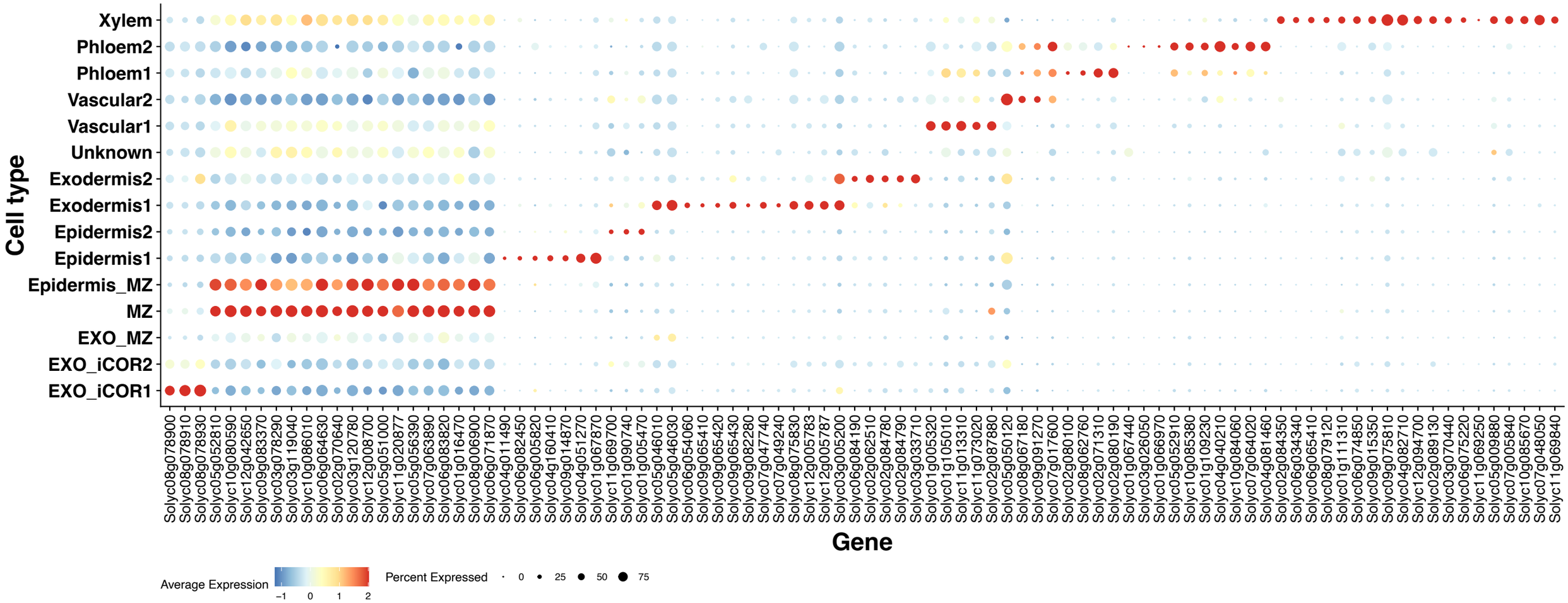


Fig. S5: Expression patterns of tomato cell-type marker genes which are derived from the conserved OMGs across species. The dot size represents the proportion of cells in a particular cluster expressing the marker gene, while the color of the dot indicates the level of expression. Red dots indicate high expression, while blue dots indicate low expression.