Title: Identification of cell-type marker genes from plant single-cell RNA-seq data using machine learning

Haidong Yan[1], **Qi Song**[1,2,4], **Jiyoung Lee**[1,2], **John Schiefelbein**[3], **Song Li**[1,2*]

[1]School of Plant and Environmental Sciences (SPES). [2]Graduate program in Genetics, Bioinformatics and Computational Biology (GBCB), [3]Department of Molecular, Cellular, and Developmental Biology, University of Michigan. Ann Arbor, MI 48109. *Corresponding author.

Abstract

An essential step of single-cell RNA sequencing analysis is to classify specific cell types with marker genes in order to dissect the biological functions of each individual cell. In this study, we integrated five published scRNA-seq datasets from the *Arabidopsis* root containing over 25,000 cells and 17 cell clusters. We have compared the performance of seven machine learning methods in classifying these cell types, and determined that the random forest and support vector machine methods performed best. Using feature selection with these two methods and a correlation method, we have identified 600 new marker genes for 10 root cell types, and more than 70% of these machine learning-derived marker genes were not identified before. We found that these new markers not only can assign cell types consistently as the previously known cell markers, but also performed better than existing markers in several evaluation metrics including accuracy and sensitivity. Markers derived by the random forest method, in particular, were expressed in 89-98% of cells in endodermis, trichoblast, and cortex clusters, which is a 29-67% improvement over known markers. Finally, we have found 111 new orthologous marker genes for the trichoblast in five plant species, which expands the number of marker genes by 58-170% in non-Arabidopsis plants. Our results represent a new approach to identify cell-type marker genes from scRNA-seq data and pave the way for cross-species mapping of scRNA-seq data in plants.

Introduction

Single cell RNA sequencing (scRNA-seq) has recently emerged as a powerful approach to investigate gene expression in complex multi-cellular organisms. Compared to bulk RNA-seq, scRNA-seq can identify rare cell populations and reveal transitions of cell states at different developmental stages, which are difficult to capture using traditional methods (Butler et al., 2018; Trapnell, 2015; Wang and Navin, 2015). As a transformative technology, scRNA-seq is particularly important for plant research because traditional methods for determining gene expression in individual cell types rely on transgenic lines expressing cell-type-specific fluorescent markers, which are not available in most non-model species. Because of the advantages of using scRNA-seq in plant, this approach has been applied in a number of studies to profile transcriptomes of *Arabidopsis* and maize (Bezrutczyk et al., 2020; Satterlee et al., 2020). The *Arabidopsis* root is an ideal system to address important questions in plant biology using scRNA-seq, including analysis of expression pattern of rare cell types (Denyer et al., 2019; Ryu et al., 2019), determination of developmental trajectories of root cells (Denyer et al., 2019; Jean-Baptiste et al., 2019; Ryu et al., 2019; Zhang et al., 2019b), and characterization of stress responsive genes at the single cell level (Jean-Baptiste et al., 2019; Ryu et al., 2019; Shulse et al., 2019).

Identification of cell types is a key step in the analysis and interpretation of scRNA-seq data (Luecken and Theis, 2019). Currently, approaches to define *Arabidopsis* root cell types fall into three major categories: (1) Calculate index of cell identity (ICI) using selected marker genes based on information theoretic scores from a published cell expression profiles (Efroni et al., 2015; Shulse et al., 2019; Turco et al., 2019); (2) Assign cell types by visualizing expression patterns using known marker genes (Jean-Baptiste et al., 2019; Ryu et al., 2019; Zhang et al., 2019b). (3) Compute correlation coefficient with published gene expression data (Jean-Baptiste et al., 2019; Shulse et al., 2019). All these strategies rely on some knowledge of marker genes: the first two methods require knowledge of expression patterns of fewer than 200 genes whereas the third method needs expression profiles of thousands of genes from known cell types.

In addition to these methods, several other computational approaches have been developed to identify novel marker genes from scRNA-seq data. The Seurat software first identifies highly variable genes across different clusters and then defines marker genes based on a statistical test between a cluster of cells against other clusters of cells (Butler et al., 2018). SCMarker is a statistical approach where marker genes are selected as mutual exclusively expressed with some other genes based on a mixture distribution model (Wang et al., 2019). A database called CellMarker was developed to provide a comprehensive overview of cell markers in both human and mouse (Zhang et al., 2019c). In plants, ~1500 cell-type marker genes have been determined from FACS-based gene expression data in the root cells of *Arabidopsis* (Birnbaum et al., 2005; Brady et al., 2007; Bruex et al., 2012; Efroni et al., 2015; Li et al., 2016). Recently, scRNA-seq data has been used to identify novel cell-type markers in plants by identifying genes with predominant expression in particular cell clusters with known identities (Jean-Baptiste et al., 2019; Shulse et al., 2019). However, complex heterogeneity of cell populations is sometimes characterized by multiple sub-populations of cells within one cluster, which limits the accurate identification of novel markers.

Machine learning (ML) has been widely applied to solve classification problems in genomics (Libbrecht and Noble, 2015). With regard to scRNA-seq data, supervised ML algorithms have been used to build cell-type classifiers (Alquicira-Hernandez et al., 2019; Pliner et al., 2019; Zhang et al., 2019a), which have outperformed traditional correlation-based approaches. However, none of these ML methods addresses the question of selecting marker genes in scRNAseq data. Feature selection is a key component of modern machine learning methods because it provides interpretability to the ML models (Azodi et al., 2020). Feature selection refers to a class of techniques that assign scores to the input features to indicate how much each feature contributes to the performance of a predictive ML model (Cai et al., 2018). For example, a Support Vector Machine (SVM)-based recursive feature elimination was used to identify marker genes to differentiate developing neocortical cells from neural progenitor cells (Hu et al., 2016). These novel marker genes not only perform better than traditional gene sets, but also uncover hidden regulatory networks with novel interactions (Hu et al., 2016). We have also developed a feature selection-based approach to determine key regulators of transcription regulatory networks (Song et al., 2020). In this work, we have demonstrated that feature selection significantly outperformed traditional statistical methods such as enrichment tests in selecting regulatory transcription factors in scRNA-seq data.

In the present study, we have developed machine learning based approaches to identify novel cell-type marker genes in the *Arabidopsis* root. Our studies evaluated seven machine learning methods for classification of ten different root cell types in *Arabidopsis*. We discovered that

Random forest (RF) and SVM exhibited relatively better performance and were used to identify marker genes. In the RF model, the markers were identified based on the feature importance derived from the SHapley Additive exPlanations method (SHAP markers). Comparing with traditional variable importance algorithms that only display results across all samples, the SHAP method allows to calculate SHAP value for each observation, which greatly increases the model transparency (Lundberg et al., 2020). The SVM model (SVMM) markers were identified based on feature selection using the weight of each gene in the trained model. We also derived a list of new markers that correlated to the ICIM markers (CORR markers). We compared these three new types markers (SHAP, SVMM and CORR) to other existing markers including 1) ICIM, Index of Cell Identity Method-based markers, 2) BULR, markers identified from a Bulk RNA-Seq study (Li et al., 2016), and 3) KNOW, markers derived from microarray analysis (Brady et al., 2007). We found that the SHAP and CORR markers have similar or stronger preferentially expression than the other markers in multiple cell types. Further, the SHAP, SVMM and CORR markers performed better for cell-type classification than other markers using RF models. Interestingly, we found that the majority of SHAP and SVMM maker genes are novel markers that were not identified by previous approaches. It is important to note that our method can also detect novel "one-off" marker genes that may have unique biological functions. Finally, our approach identified 111 new orthologous marker genes for trichoblast cells in five plant species, which suggest our approach can identify novel marker genes to facilitate cell type identification in scRNA-seq data from other plant species.

Results

Overview of SPmarker. The function of Single cell Predictive marker (SPmarker) is to identify novel marker genes in different cell types of the *Arabidopsis* root (**Figure 1**). This approach includes two major steps. In the first step, the expression data of cells from different datasets were normalized and integrated using an established approach (Butler et al., 2018). The identities of these cells were assigned by the ICI method (Efroni et al., 2015). The genes with highly variable expression were used as features for downstream analysis (**Figure 1A**). In the second step, several machine learning methods were trained and compared with five-fold cross validation. The goal of machine learning was to determine a function that best predict cell types based on gene expression features in each cell. Only the best performing machine learning methods were used for follow up analysis to determine marker genes using feature selection. For example, the SHAP method was used to calculate the SHAP value which represents feature importance for each gene (**Figure 1B**). The top 20 (default setting) genes with the highest SHAP value in each cell type were selected as the marker genes.

Training dataset preparation

Five scRNA-seq datasets of *Arabidopsis* roots were downloaded from the GEO database (accessions: GSE123013, GSE121619, GSE122687 and GSE123818) and a webserver (PRJNA517021; http://wanglab.sippe.ac.cn/rootatlas/) (Denyer et al., 2019; Jean-Baptiste et al., 2019; Ryu et al., 2019; Shulse et al., 2019; Zhang et al., 2019b). We excluded cells under treatment or in a mutant background and only used the cells derived from control samples in each dataset. The five data sets were merged into a single data set with 57,333 cells where 29,929 genes were detected in this combined data set (**Figure S1**). ICI was used to label the cell type for each cell (Efroni et al., 2015). The ICI score (0 <= score <= 1) of each cell represent a similarity of each cell to one of the 15 known cell types in Arabidopsis root. Cells in each cell type were ranked by the ICI scores and 32,146 cells with the scores over 0.5 were selected. Before the training process, two steps were conducted to balance the number of cells for each cell type: 1) Five cell types (Phloem; Late_PPP; Pericycle; LRM; Late_XPP) with low cell number (<300) were removed; 2) Randomly selected 5000 cells from Atrichoblast and Cortex were used for model training. Finally, 25,618 cells with 29,929 genes from ten cell types (Trichoblast; Atrichoblast; Cortex; Endodermis; Phloem_CC; Protophloem; Protoxylem; Meri_Xylem; Columella; QC) were kept to compare different machine learning models for cell-type classification performance (**Table S1**).

To train the models, the dataset was divided into a training dataset of 23,056 cells (90%) and an independent testing dataset of 2,562 cells (10%). We performed five-fold cross validation in the training dataset and separated the training set to sub-training (80%) and validation (20%) sets. The independent testing dataset was used to compare the performance of the machine learning methods.

Comparison of different machine learning models for cell-type classification

Seven machine learning approaches were evaluated including three deep neural network (NN) approaches and four classical ML approaches. The NN approaches include triplet NN (Hoffer and Ailon, 2015), contrastive NN (Koch et al., 2015), and a baseline NN implemented using Keras and tensorflow (Gulli and Pal, 2017; Schmidhuber, 2015). The classical ML methods include SVM (which is method to find separating hyperplane), Random Forests (which is an ensemble tree-based classifier), K-Nearest Neighbors (KNN, which is a network-based approach), and Principal Component Analysis (PCA). These methods were selected because they represent approaches where each is based on a distinct underlying mechanism. The Area Under Precision-Recall Curve (AUPRC) value was used to demonstrate the prediction performances of all the methods. The SVM and RF showed relative higher AUPRC than the KNN and PCA models. In the deep learning-based models, the contrast NN, and triplet NN had similar performance but had higher AUPRC than the baseline NN. In general, traditional machine learning-based models showed relative higher AUPRC than the deep learning-based models. The SVM and RF achieved the best performance among the seven models (**Figure 2A**). Performance comparisons were also obtained using the Area Under the Receiver Operating Characteristics (AUROC), Geometric Mean (GM), Matthews Correlation Coefficient (MCC), Accuracy (AC), Precision (PR), Sensitivity (SE), and Specificity (SP) (**Figure S2 and S3**). The performances of SVM and RF are superior regardless of evaluation metrics used. Interestingly, machine learning models performed better in some cell types such as Trichoblast, Endodermis, Cortex, and Atrichoblast, while the QC had the lowest AUPRC (**Figure 2B**). This is not entirely due to the low number of cells in QC (537 cells) because other cell types such as Phloem_CC and Protophloem have similar or fewer cells (565 and 380 cells respectively) but the machine learning model performances on these cells are higher than QC (**Figure 2B and Table S1**).

Identification of marker genes

Due to their performances, RF and SVM were used to identify RF method-based (SHAP) and SVM method-based (SVMM) marker genes. The marker genes have higher feature importance in a cell-type classification model and may express variably among other cells. We first selected the top 20 percent of highly variable genes (5,986 genes) using the Seurat package and then identified SHAP and SVMM markers from these 5,986 genes.

SHAP markers

We used the RF model to extract feature importance based on TreeExplainer function from the SHAP package. The previous built independent testing, sub-training, and validation sets with the selected genes (5,986) were used to train a new RF model. Five models were generated from the five-fold-cross validation. The best performing model was selected to calculate the feature importance (*SHAP* value) (**Figure S4**). To understand the distribution of feature importance for SHAP markers, we plotted the cumulative distribution for SHAP values (**Figure 3A**). Interestingly, for each cell type, only approximately 100 to 500 genes account for 50% of all feature importance. Comparing Cortex and QC, 148 and 538 genes accounted for 50% of the feature importance for these two cell types respectively.

A higher *SHAP* value of a gene represents a greater contribution to classify the cell types. Since each gene has a *SHAP* value for each of the ten cell types, we assigned each gene to the cell type with the highest *SHAP* value among all ten cell types. By doing this, we identified SHAP marker genes that are unique to each cell type. We found 1,840 and 1,460 unique marker genes for Cortex and Atrichoblast respectively, while 63 and 37 unique marker genes were found for the QC and Protophloem (**Figure 3B**). In other words, there are almost 30 times more unique SHAP marker genes in Cortex as compared to QC cells. By overlapping the cumulative *SHAP* values from genes with unique SHAP marker, we found fewer than 50 unique genes accounted for 50% of the total *SHAP* values in each cell types. This suggests that QC has a small number (Figure 3B, 63 genes) of unique SHAP markers and more markers (Figure 3A, 538 – 63 genes) are shared with other cell types, whereas Cortex or Atrichoblast have large numbers of unique SHAP markers but a small fraction of these markers carry the most weight.

To study the identity of the SHAP marker genes, we focused on the top 20 genes in each cell type with the highest SHAP value. There are 146 genes out of 200 SHAP marker genes (73%) that were not identified before in a collection of 1,813 marker genes from other publications (**Table S2**). In particular, in the Protoxylem, all SHAP markers are new (**Figure 3C**) and 80% of SHAP markers from the Atrichoblast and Phloem_CC are new markers (**Figure S5**).

SVM markers (SVMM)

The SVMM genes were determined based on the feature importance estimated from the absolute coefficient values in the SVM model. Each gene has a coefficient for each of the ten cell types, and each gene is assigned to the cell type with the highest coefficient. The number of unique marker genes and novel marker genes determined by SVMM are similar to that determined by SHAP. For example, most unique SVMM genes were assigned to Atrichoblast (1,236), Cortex (1,180) and vascular tissue and QC cells have lower numbers of SVMM (**Figure S6A**). However, some cell types, such as trichoblast, have approximately twice as many SVMM markers as SHAP markers (1,098 SVMM vs 555 SHAP). In the top 20 genes with the highest coefficient of each cell type, on average, 69.5% of SVMM are novel marker genes (**Figure S6B**).

Because we found both SHAP and SVMM are mostly new marker genes, we compared the overlaps between marker genes across all known methods. The ICI method has identified 360 predefined marker genes based on ATH1 Affymetrix microarray (Efroni et al., 2015), and 232 of them were detected in our integrated scRNA-seq dataset (others were not detected because scRNAseq preferentially detect highly expressed genes). The expression patterns of the 29,929 genes in the integrated scRNA-seq data were correlated to the ICI markers, and the top 20 genes with highest correlations of each cell type were selected as the correlation (CORR) markers. We also collected 533 genes that have tissue-preferential expression in specific cell types in a bulk RNA-Seq study (Li et al., 2016), and these genes were defined as Bulk RNA-Seq (BULR) markers. The other published marker genes (1,532) that are not overlapped with the ICIM markers were defined as known (KNOW) markers (**Table S3**). These four methods, ICIM, CORR, BULR, and KNOW, are not machine learning based approaches and these markers were compared with the SHAP and SVMM markers (**Figure 3D**). Interestingly, there is little overlap between these marker genes such that 93.5% or 1027 marker genes are unique to a single method. This result raised the question of how to compare the performance of these marker genes determined in different studies.

SHAP markers extensively express in clusters with dominant cell types

One way to compare the performance of marker genes is to compare their representations in different cell clusters in the scRNA-seq data. In the following analysis, we compared the gene expression patterns and cell-classification ability among the six types of markers: SHAP, SVMM, CORR, ICIM, BULR, and KNOW. To make these methods comparable, a total of 200 SHAP markers consisting of the top 20 markers in each of ten cell types were used. The same methods were used to select 200 markers in SVMM and CORR. To select the top 20 BULR and KNOW markers of each cell type, we ranked them based on their expression specificity described in the method section. The top 180 BULR markers in the cell types were selected except for the Protophloem where no marker was detected. In the KNOW markers, the top 161 of them were selected since no Meri_xylem marker and only one Protoxylem marker were found. The 232 ICIM markers were all used in the comparison (**Table S4**).

An unsupervised clustering analysis was performed on the 25,618 single cells using Seurat package (v3.1). A total of 17 clusters were identified. Cell clusters were assigned an overall cell type identity based on the ICI score for each cell. According to the proportion of cell type in one cluster, these 17 clusters were classified into three groups (**Figure 4A**). The group one clusters were defined when the cells assigned by a single cell type accounted for over 80% cells in one cluster. The group two clusters were defined as when a dominant cell type accounted for between 50% and 80% of cells in a cluster. Group three clusters were defined when no cell type can account for over 50% of cells in a cluster. Group one clusters include more than half of clusters (9/17). Among these group one clusters, Trichoblast, Atrichoblast, Endodermis, and Cortex can be clearly assigned to one or more clusters. The group two clusters include clusters 6, 12, and 15. The Cortex and Protoxylem were dominant in these clusters. Other clusters belong to groups three, where these clusters were mixed with Meri_Xylem, Phloem_CC, Protophloem, and QC cell types (**Figure 4A and 4B; Table S5**).

We focused on five group one clusters (5, 8, 9, 13, and 14) with a dominant cell type that accounts for over 90% cells in each cluster (**Table S5; Figure 4B**). These clusters were selected because they are the most homogenous clusters, making the results easier to interpret. For each marker gene in each cluster, we calculated the proportion of cells in which this marker gene is expressed and we calculate the average "fraction of expressed cells" for all 20 markers in each cluster. For a more specific example, given 20 SHAP markers for cluster 8, on average each SHAP markers is detected in 95.9% of cells (**Figure 4C**, Cluster 8, Endodermis). These numbers were calculated for all marker types and compared in Figure 4C. In these homogenous cell clusters, the SHAP markers achieved higher or similar proportion of the expressed cells as compared to all the other markers. In the cluster 5 that represents Cortex, the SHAP markers had significantly (p < 0.05) higher expression rate (96.7%) than all other marker types (< 85%) (**Figure 4C**). BULR markers were detected in lowest number of cells and followed by ICIM markers. Because the ICI method does not require all ICIM to be detected in a given cell, it is not surprising that ICIM were only found in 50% of cells in some clusters. On average, the percent of cells expressing SHAP markers is 29% more than ICIM and 67% more than BULR markers. Because both BULR and ICIM were determined using data not from scRNA-seq experiments, these results might suggest BULR and ICIM include cell type-specific, but relatively low-expressed genes that cannot be detected by scRNA-seq.

SHAP, SVMM and CORR markers perform better on cell-type classification than KNOW, and BULR markers using machine learning models

To evaluate performance of the six marker types on cell-type classification, we performed five-fold cross validation using the same sub-training, validation, and independent testing datasets as the previous evaluation of different machine learning methods (**Figure 2**). The difference in this analysis as compared to those in Figure 2 is that only 20 marker genes for each cell type were used for each method respectively (with the exception of BULR, 161 makers and ICIM, 232 markers, see previous section.). The marker genes were used to train a RF model and an SVM model. As expected, the ICIM markers have the best performance in auROC and auPRC among the six marker types, since ICIM markers were used to assign cluster identities before training the models. The SHAP markers achieved the 2nd and 3rd highest auPRC and auROC performance in the RF model and the SVM model respectively (**Table 1**). SHAP, SVMM and CORR markers perform better than KNOW and BULR markers. This performance was consistently achieved when using other evaluation scores (MCC, GM, MAP, AC, PR, SE, and SP) (**Figure S13**).

Using newly developed markers to assign cell identity

To assign cell identities for clusters in single cell experiments, we use the following rules. First, expression values for each marker genes were normalized across all cells and clusters using AverageExpression function in the Seurat package. Second, for each of the marker genes, the normalized expressions were ranked from high to low (from 1 to N with N equals the number of clusters, N=17 in our dataset). Finally, the average rankings of all marker genes were used to determine the cluster identity. Among 17 clusters identified by our analysis, we found 15 clusters were assigned consistently to the same cell types by three or more marker types (black and grey color marked the clusters in **Figure 5A**). There are 11 clusters that were assigned to specific cell types by four or more marker types. For example, clusters 3 and 6 were assigned to endodermis and cortex respectively by all marker types. Among all cell types, tricoblast, atricoblast, cortex, endodermis and protoxylem all have clusters assigned consistently by 4 or more types of marker genes. These results showed that, although most marker genes from different approaches do not overlap, the results of cluster assignment are consistent.

Because only 20 marker genes were selected for each marker type for each cell type, we compared the expression patterns for these six types of markers using heatmap (**Figure S7** to **S12**). We found that different markers showed different levels of specificity from the heatmaps. Markers identified by SHAP, CORR and ICIM have better specificity than KNOW, SVMM and BULR. For example, we plotted the top three most specific markers from the 20 markers in SHAP and KNOW marker types (**Figure 5B** and **C**). We also plotted the expression of the bottom 3 markers (ranked 18, 19, and 20 by marker specificity) from the 20 markers (**Figure 5D** and **E**). We found that the SHAP markers showed high cell-type specificity in both cases whereas the specificity of KNOW markers is lower for those ranked at 18 to 20.

To further quantify the specificity of the SHAP makers and other marker types, we calculated their cumulative correlation with specific cell types in the atrichoblast, trichoblast, and endodermis (**Figure 5F-H**). These three cell types were selected because all of these cell types have a higher number of cells than other cell types and these cell types are consistently identified by the majority of marker types. Consistent with our observation in the heatmaps (**Figure S7-12**), the SHAP and CORR markers had relatively higher cumulative correlation rate in all three cell types, suggesting stronger preferentially expression for these two markers. The BULR and SVMM markers have the lowest correlation rate in the three cell types.

To confirm the analysis of cumulative correlations (**Figure 5F-H**), we also generated the distribution of correlation for six types of markers. The results are consistent with the results from cumulative correlation curves where SHAP, CORR and ICIM show consistently higher correlation values as compared to other marker types (**Figure S14A-C**). Interestingly, one SHAP marker (*AT2G3683*) for atrichoblast cells showed negative correlation with other genes expressed in the same cell type (**Figure S14A**, red circle). The expression pattern of this marker showed it had relatively lower expression levels on cluster 2, 10, 7, 11 than other clusters (**Figure S14D** and **E**). This result shows that the SHAP method is not only able to identify markers that are positively correlated with the cell type of interest, but also genes that are only repressed in the cell type of interest. Such negatively correlated markers can also be found by SVMM method.

Biological function of novel cell marker genes

Among the three new marker types (SHAP, SVMM and CORR), we are most interested in the function of SHAP markers because each cell can have a different SHAP marker value to determine the contribution of each gene to the cell identity. To further characterize specific functions of the SHAP markers, the Gene Ontology (GO) annotation from the SHAP markers was compared to the KNOW and ICIM markers which represent the majorities of published markers (**Table S6**). In the biological process, a total of 48 GO terms were found specifically to the SHAP markers (**Figure 6A**). Nearly one third (30.5%; 61/200) of the SHAP markers were involved in the responsiveness of stimulus in the biological process (**Figure 6B**). In contrast, the KNOW markers had 39 specific GO terms, and nearly half of the terms (46.2%; 18/39) are associated with root vegetative and reproduction development (**Figure 6A** and **6C**). In the ICIM markers, only two specific GO terms were identified (**Figure 6A**). One possible reason that KNOW markers were enriched in root developmental processes is that these KNOW marker genes might have been used to define GO annotations since these marker genes have been published more than 10 years ago.

The GO enrichment analysis also showed 19 SHAP markers are mainly enriched in water transport (GO:0006833), response to water (GO:0009415), and to water deprivation (GO:0009414) processes, and participate water transmembrane transporter activity (GO:0005372) and water channel activity (GO:0015250) (**Figure 6D**; **Table S7**). Among these markers, more than half of them (11/19) were under Cortex, Endodermis, Protoxylem, and Meri_Xylem cells (**Table S7**). These four cell types are essential for water transportation and minerals assimilation (Qiao and Libault, 2013; Steudle and Peterson, 1998). A number of 18 SHAP marker genes were also enriched in the cell wall biosynthesis (**Figure 6D**). One third (6/18) of these SHAP markers were found for the Protoxylem and Meri_Xylem cell types (**Table S7**). The xylem cells are known to be important for cell wall formation (Oda and Fukuda, 2012). In contrast to the SHAP markers, the KNOW markers had enriched GO terms on root system development, cell maturation, cell wall development, root hair, and epidermal cell differentiation (**Figure 6E**). No enriched gene was identified in the ICIM markers. In summary, the SHAP markers are enriched with environmental response functions, especially water responsiveness and cell wall formations, which suggest that SHAP marker genes not only can serve as cell identity markers, but also may play important cell-specific biological functions in roots.

As scRNA-seq experiments expand to include non-model plant species that lack cell-type markers, it becomes challenging to accurately determine cell types in these species. Therefore, a major goal of identifying new marker genes is to expand the list of candidate marker genes in other species. This is complicated by the likelihood that some marker genes in Arabidopsis may have altered their functions in other species or may be absent from genomes of other species. To evaluate whether our newly identified marker genes (SHAP, SVMM, CORR) in Arabidopsis can be used to define new marker genes in other species, we analyzed root hair cell expression from five other plant species where root hair specific expression data are available (Huang et al., 2017). To consider the most specific markers, we tested the top 20 markers from each marker type and excluded BULR markers. In these five species (**Figure 6F**; **Table S8**), we found a total of 161 genes that are expressed in root hair cells and also are orthologous to ICIM or KNOW marker genes, and 98 of these orthologous genes are significantly differentially expressed in root hair cells relative to all cells in the root differentiation zone (**Table S9**). We also found 184 genes in these five species that are expressed in root hair cells and are orthologous to SHAP, SVMM or CORR markers. Among these genes, 111 new marker genes are differentially expressed in root hair cells relative to the root differentiation zone (**Table S9**). In summary, new marker genes increase the number of candidate marker genes by 58% to 170% in these five species. This result demonstrates that our newly identified marker genes can substantially increase the number of marker genes that can be used to assign cell types in other plant species.

Discussion

The scRNA-seq technology provides a novel platform to analyze the transcriptomic profile of individual cells in plants to characterize heterogenous cell populations in detail. In plants, this process is heavily reliant on the use of marker genes that are preferentially expressed in specific cell types. Here we introduce a machine learning based approach to identify marker genes by analyzing their feature importance. One important distinction between machine learning-based marker detection and traditional methods is that most traditional methods select one gene at a time by calculating gene specificity. In contrast, the machine learning methods used in this work characterize the marker genes by analyzing combinations of many marker genes. Machine learning methods also provide a number of principled approaches to evaluate marker performance including cross-validation, leave-out testing sets, and evaluation metrics such as auROC, auPRC and F1 scores. These evaluations methods allow us to compare different marker genes in a more rigorous and unbiased fashion. Also, compared to the correlation markers that detect similar expression patterns as the known markers, the SHAP and SVMM markers can identify markers with novel expression patterns such as only repressed in a specific cell type. We have shown that these markers can yield high performance using machine learning-based evaluation metrics. Why certain genes are selectively repressed in a specific cell type and what unique biological functions these genes play in these cell types require further experimental testing. More importantly, because these machine-learning derived markers are not based on prior knowledge of gene functions, these markers may have new biological functions that are not characterized before.

In the evaluation of different machine learning methods, the SVM and RF methods outperform the three deep learning models (**Figure 2**). One possible reason is SVM and RF are effective for relatively small datasets or fewer outliers (Ali et al., 2012; Ben-Hur and Weston, 2010). The deep learning algorithms usually require relatively large dataset to work well and achieve good performance for solving more complex problems such as image classification (Zou et al., 2019). A previous study utilized the contrastive NN and triplet NN to successfully classify cells in mouse by using more than 100,000 cells to train these two models (Alavi et al., 2018), while our study used less than 30,000 cells. If more cells with accurate cell identity were available in *Arabidopsis*, the performance of the deep learning model in our study may be improved (Eraslan et al., 2019).

Cell populations of *Arabidopsis* roots are characterized by a high level of heterogeneity. Results from animal systems have demonstrated that, even within a cell population, the cells are not homogeneous because sub-populations may exist (Liu and Trapnell, 2016). Furthermore, it is not clear whether all cell types have been discovered for the *Arabidopsis* root (Zhang et al., 2019b), in particular, for cells in a transition stage or regulated by periodical signals (Voß et al., 2015). This highlights the importance of identifying new marker genes which may be expressed at different levels in sub-populations as compared to traditional marker genes derived from bulk

RNA-seq and microarray experiments. However, traditional approaches for marker gene identification usually involve manual inspection of the cell population structure, which could be arbitrary (Luo et al., 2015; Usoskin et al., 2015).

Identification of new marker genes is particularly important for plant biology research because cell type markers are largely unknown from non-model species. We have demonstrated that our machine learning based approaches can substantially expand the number of known root hair marker genes and that orthologs of these marker genes can also be found in other plant species. One future direction is to define root cell types in non-model species from cross-species mapping of marker genes and their expression pattern in roots.

Materials and Methods

Data preprocessing

The scRNA-seq data of root cells from four publications were downloaded from the GEO website (GSE123013, GSE121619, GSE122687 and GSE123818) and a web server (http://wanglab.sippe.ac.cn/rootatlas/) (PRJNA517021) (Denyer et al., 2019; Jean-Baptiste et al., 2019; Ryu et al., 2019; Shulse et al., 2019; Zhang et al., 2019b). For each dataset, raw counts were used as input data, and the treatment groups were removed. A gene was retained if it was expressed in more than three cells, and each cell was required to have at least 200 expressed genes. The cells that have over 5% mitochondrial counts and have unique feature counts over 2,500 or less than 200 were removed out. A global-scaling normalization method ‘LogNormalize’ in the Seurat package (v3.1) is used to normalize the feature expression measurements for each cell by the total expression, and multiplies this by a scale factor (scale.factor = 10,000) (Butler et al., 2018). The five data sets were merged to obtain 57,333 cells and 29,929 genes. To correct for dataset-specific batch effects, the Seurat (v3.1) multicanonical correlation analysis was conducted on the merged dataset.

ICI method and cell type identification

To assign cell types to cells collected from the previous five datasets, Index of cell identity (ICI) score is computed using a R script from Efroni et al. (2015) study. The predefined marker genes and their specificity (Spec) scores (‘ath_root_marker_spec_score.csv’) for the 15 root cell types were provided as a supplementary file of this study. The 15 root cell types include

Trichoblast, Cortex, LRM (Lateral Root Meristem), Late_PPP (Late Phloem-Pole Pericycle), Protophloem, Meristematic_Xylem (Meri_Xylem), Phloem_CC (Companion Cell), Protoxylem, Phloem, Pericycle, Endodermis, Atrichoblast, Columella, QC (Quiescent Center), and Late_XPP (Xylem-Pole Pericycle).

The R script from the Efroni et al. (2015) study, computes the ICI score by averaging expression of all genes in the predefined set of marker genes and weights each gene by its Spec score for the specific cell type. The ICI score of each cell is calculated for all 15 cell types, and represent a similarity of each cell to each cell type. Cell type with the highest ICI score was assigned to the cell as final cell type label. We used the previous merged dataset before the Seurat integration to be input data in the ICI script, since negative value generated from the Seurat integration will allow this script to produce a negative ICI score that cannot be used to assign cell types to cells. The cells with ICI scores lower than 0.5 were filtered out. The Late_PPP, Late_XPP, Phloem, Pericycle, and LRM cell types with low number of cells (< 300) were removed. To balance the number of cells for each cell type, we randomly removed 3,074 and 2,923 cells in Arichoblast and Cortex, respectively. Finally, 25,618 cells, 29,929 genes, and ten cell types were remained and used for the further steps (**Table S1**).

Machine leaning classification

Several common machine learning approaches were evaluated for the task of cell type classification, including Support Vector Machine (SVM), K Nearest Neighbors (KNN), Random Forests (RF), Base Neural Network (Base NN), Siamese neural network with triple loss (triplet DNN), Siamese neural network with contrastive loss (contrastive NN). Python library scikit-learn was used to perform classification with KNN, SVM, and RF (v 0.23.1) (Pedregosa et al., 2011). Python library Keras (v 2.2.4) (Chollet, 2015) was used to perform classification with multi-task NN, Triplet NN, and Contrastive NN. All implementations of neural networks were modified based on the GitHub repository provided by the published study (Alavi et al., 2018).

The dataset was divided into training (90%) and independent testing (10%) sets for each cell type. We performed 5-fold cross validation for the training dataset that was separated to sub-training (80%) and validation (20%) sets. The machine learning models were trained with the sub-training dataset and were evaluated using the independent testing dataset. Each machine learning approach is briefly described below. Source code for our SPmarker pipeline is available at github (https://github.com/LiLabAtVT/SingleCellClassification/).

KNN. KNN is a commonly used simple classifier that does not have explicit training process. KNN makes new prediction by first computing Euclidean distance between the new input vector and every feature vector in the training dataset. Then the top K nearest neighbors are used for new prediction. In the last step, class label of the new input vector is determined by majority vote among the K nearest neighbors. The only hyperparameter for KNN is K, the number of top nearest neighbors. K is set to be 50 in the analysis. KNN is fast and simple, which makes it a first choice of machine learning classifier in many cases when computation resource is limited.

RF. RF is a tree-based machine learning approach built on a collection of decision trees. For each decision tree, a subset of training examples is randomly sampled as inputs and a subset of features are randomly sampled to split each tree node. The final class label is determined by majority vote. Number of trees (N) for RF is an important hyperparameter that can impact the model performance. Here, N was set as 50.

SVM. SVM is a machine learning classifier that maximizes the margin between different classes in a high dimensional space transformed by kernel function. Depending on kernel function, SVM can be a linear classifier (linear kernel) or a non-linear classifier (e.g., Gaussian kernel). To best capture the complex gene-gene relationships that characterize the cell type, Gaussian kernel was used to train SVM classifier.

Baseline NN. Baseline NN refers to a basic type of neural network that uses densely connected layer as input layer and hidden layers. The output layer has number of neurons equal to number of cell types (ten cell types). Architecture of the base NN is demonstrated in **Figure S15**. Briefly, input layer has number of neurons equal to number of genes used for classification (29,929 genes) and three hidden layers were used, of which each has 586, 256, and 100 neurons. The last layer is an output layer to which a softmax is applied to ensure output scores are summed to 1.

Triplet NN. Triplet NN is the implementation of Siamese neural network with triplet loss function. The use of triplet loss function was discussed in a published study (Alavi et al., 2018). Briefly, Siamese DNN consists of two subnetworks which have identical architecture and weights. The two neural networks connect to the same distance layer which computes a vector of distance between the last two hidden layers in the two subnetworks. The last two hidden layers are lower dimensional embeddings of original feature vectors. Architecture of Siamese NN is demonstrated in **Figure S15.** In this work, number of neurons in input layer is equal to number of genes used for classification (29,929). Numbers of neurons used in three hidden layers are 586, 256, and 100. In training dataset, each scRNA-seq expression profile is an "anchor" that can be paired with positive example and negative example. Positive examples are those labeled with the same cell type with anchor and negative examples are those with different cell type. For each anchor, it will be paired with a positive example and a negative example, which forms a group of triplets. Then for each group of triplets, anchor-positive and anchor-negative pairs will be respectively fed into triplet NN. Based on the discussion in (Schroff et al., 2015) and (Alavi et al., 2018), the loss function of triplet NN can be written as:

$$L(D)max\left\{ ,\left(\sum_{i=1}^{T} \left(D_{a,p}^{i}\right)^{2} - \left(D_{a,n}^{i}\right)^{2} + m \right)\right\}$$

Where T is the number of groups of triplets. $D_{a,p}^{i}$ is the Euclidean distance between anchor and positive samples and $D_{a,n}^{i}$ is the Euclidean distance between anchor and negative samples. m is a hyperparameter that represents the margin between $\left(D_{a,p}^{i}\right)^{2}$ and $\left(D_{a,n}^{i}\right)^{2}$.

To ensure that triplet NN can be effectively trained, the groups of triplets need to include anchor-positive pairs with large distances and anchor-negative pairs with small distances. These are the hard training examples that enforce the model to learn effectively. As discussed in Alavi's study (2018), batch hard loss function is used to generate hard training examples. In each iteration of optimization, M cell types which have K cells in each are sampled to generate a mini-batch. In this mini-batch, losses of hard training examples are selected and summed up as final loss value for the mini-batch. A slight modification of batch hard loss function was made in this study to include more training samples in each mini-batch. Instead of using one pair of hardest anchor-positive and anchor-negative respectively for each anchor, top k pairs of hardest pairs are selected for each anchor. The batch hard loss function therefore can be written as:

$$L'(D) = \left\{ 0, \sum_{i=1}^{M} \sum_{j=1}^{K} \left[topmax\left(k, P_{j}^{i}\right) - \ topmin\left(k, N_{j}^{i}\right) + m \ \right] \right\}$$

Where P_{j}^{i} is the set of distances between jth cell from ith cell type and all other cells in ith cell type (anchor-positive pairs) and N_{j}^{i} is the set of distances between jth cell from ith cell type and all other cells not from ith cell type (anchor-negative pairs). $topmax(k, P_j^i)$ selects the top k pairs with largest distances in P_j^i and sums the selected distances. $topmin(k, N_j^i)$ selects the top k pairs with smallest distances in N_j^i and sums the selected distances. This gives k pairs of anchor-positive sample pairs and k pairs of anchor-negative sample pairs for each anchor. In our analysis k was set as 10.

Contrastive NN. Contrastive NN is an implementation of Siamese neural network with contrastive loss function. Its use for cell type classification has been discussed in a published study (Alavi et al., 2018). In this work, contrastive NN was constructed using the same neural network architecture as triplet NN (**Figure S15**). The difference here is that contrastive NN uses paired samples which pair the cell assigned with same/different cell types. The idea is to penalize large distances between samples of same cell type and small distances between samples of different cell types. The loss function of Contrastive NN can be written as:

$$L(Y, D) = \sum_{i=1}^{P} (Y^i)\frac{1}{2}(D)^2 + (1 - Y^i)\frac{1}{2}(\{0, m - D\})^2$$

Where P represent number of pairs of training samples. $Y^i = 1$ if two samples in the ith pair are assigned with same cell type and $Y^i = 0$ if not. D is the Euclidean distance between the two samples in each pair, computed using the last hidden layers of the two sub-networks. m is a hyperparameter that represents the margin between two samples assigned with different cell types, usually set to 1.

Model evaluation

For the KNN, SVM, RF and multi-task NN, the sub-training dataset was used to train the models that can directly predict cell type label. The trained model was then used to predict cell type labels for the independent testing dataset. For triplet NN, contrastive NN, and PCA, the sub-training dataset was used to trained models that predict neural embeddings of the original feature vectors of in training dataset. For each new input vector from testing dataset, the trained model was first used to predict a neural embedding and this embedding was compared to all neural embeddings of the training dataset. The final cell type label was determined by majority vote of m nearest neighbors. Here we set $m = 50$.

To further evaluate performances of these seven models, we calculated true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The sensitivity (SE), accuracy (AC), specificity (SP), precision (PR), geometric mean (GM) of SE and SP, and matthews correlation coefficient (MCC) were used to evaluate these models. SE, SP, AC, PR, GM, MCC, and F1 were defined as follows:

SE = TP / (TP + FN);

SP = TN / (FP + TN);

PR = TP / (TP + FP);

GM = $\sqrt{SE \times SP}$;

AC = (TP + TN) / (TP + TN + FP + FN);

MCC = $(TP \times TN - FP \times FN)/\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}$

F1 = 2 × (PR × SE / (PR + SE))

Tukey's honestly significant difference test is used as a conservative statistical test to find significant differences in all pairwise comparisons and to control for family wise error rate (Abdi and Williams, 2010).

The Mean average precision (MAP) was also used as an evaluation metric for all classification approaches. The MAP works on ranked lists (e.g. a list of nearest neighbor cells in a retrieval database) by calculating the precision at exact-match cutoffs in the list, and then taking the mean of these. We follow the MAP calculation in Alavi's study (Alavi et al., 2018).

Identification of marker genes using the machine learning approaches

The TreeExplainer in SHAP package was used to calculate feature importance in the RF model (Lundberg et al., 2020). Due to its local interpretability that each observation can get its own set of SHAP values, we can calculate these values of each feature or gene under each cell type. The higher SHAP value suggests higher contribution of the feature to the classification. The novel marker genes assume to have higher SHAP values than other genes. The implementation of the SVM model is based on libsvm (Chang and Lin, 2011). The absolute size of the coefficient relative to the other ones gives an indication of how important the feature was for the separation. We assume the absolute coefficient values represent feature importance. The multiclass support of the SVM is handled according to a one-vs-one scheme. The attributes coefficients have the shape: (number_of_cell_type * (number_of_cell_type -1) / 2, number of features). To identify the feature importance of each feature on cell type, we calculated the average absolute coefficient values from all pairs for a specific cell type for each feature. Each feature has one coefficient of each cell type. The cell type with the highest coefficient was assigned to the feature.

Identification of correlation marker genes

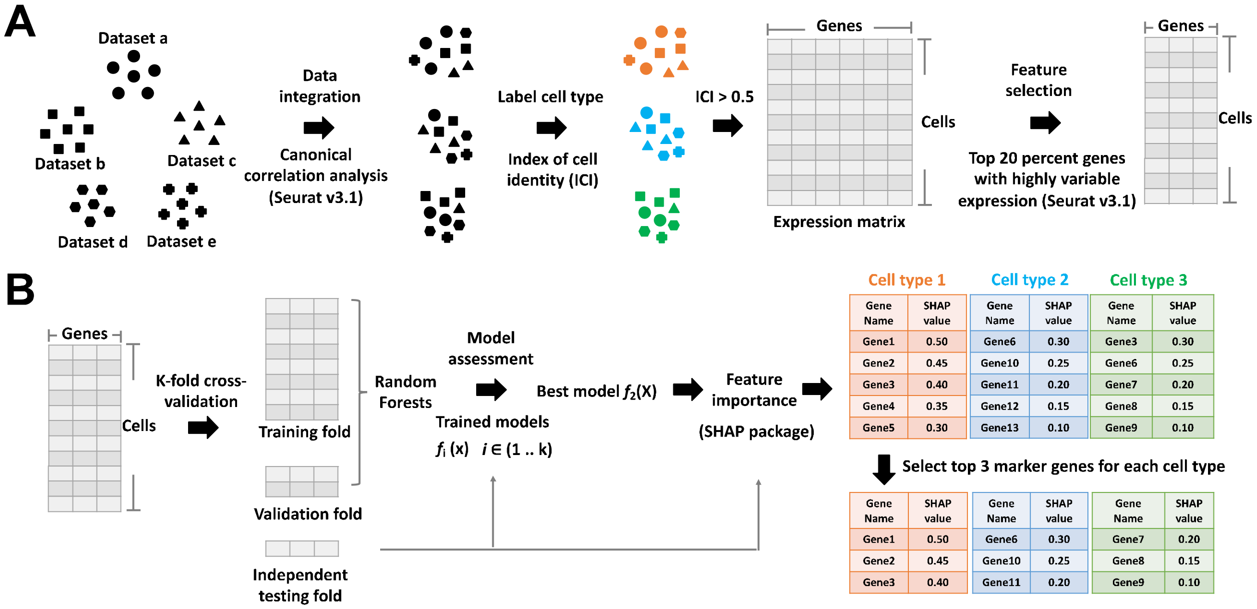
The Pearson correlation analysis was conducted between the cell expression of known and unknown marker genes (Benesty et al., 2009). Each of the unknown markers obtained a correlation rate for each cell type. We ranked the ten cell types for each marker based on the correlation rate. An unknown marker was assigned to a cell type where this marker achieved the highest *rho* in this cell type.

Cell clustering

The integrated dataset with 25,618 cells and 29,929 genes was used for clustering analysis. The top 30 aligned correlated components were used as input for tSNE dimension reduction and clustering analysis. Clusters were identified using Seurat FindClusters function with default settings. The DoHeatmap and DotPlot function in the Seurat was used to visualize expression patterns of the novel marker genes for the identified clusters.

Marker specificity

To calculate the marker specificity for a specific cell type, we generate a cell vector by labeling cells under this cell type to 1, and all the other cells to 0. The cells under the specific cell type was defined based on the ICI method. The marker expression was normalized across all cells using the formula: normalized expression = (Exp – Min)/(Max - Min) where Exp is original expression of the marker in each cell, Min is the smallest expression value of the marker in a cell, and Max is the largest expression value of the marker in a cell. The Pearson correlation analysis was used to calculate the correlation rate between the normalized marker expression and the dell vector developed before. The higher absolute correlation rate means higher marker specificity.



Cell type 1		Cell type 2		Cell type 3	
Gene Name	SHAP value	Gene Name	SHAP value	Gene Name	SHAP value
Gene1	0.50	Gene6	0.30	Gene3	0.30
Gene2	0.45	Gene10	0.25	Gene6	0.25
Gene3	0.40	Gene11	0.20	Gene7	0.20
Gene4	0.35	Gene12	0.15	Gene8	0.15
Gene5	0.30	Gene13	0.10	Gene9	0.10

Gene Name	SHAP value	Gene Name	SHAP value	Gene Name	SHAP value
Gene1	0.50	Gene6	0.30	Gene7	0.20
Gene2	0.45	Gene10	0.25	Gene8	0.15
Gene3	0.40	Gene11	0.20	Gene9	0.10

Figure 1 Summary of the SPmarker. A. Data processing pipeline. **B. Model training and identification of SHAP marker genes**. The integrated expression matrix was divided into the training dataset (90%) and the independent testing dataset (10%). The independent testing was used to evaluate the prediction performance a *fi*(x) model trained with the training dataset. The best model (*f_2*(x) in this case) was selected to identify the feature importance using the SHAP method. The top SHAP marker genes were selected from each cell type such that each cell type having its own marker genes that are not shared with others.

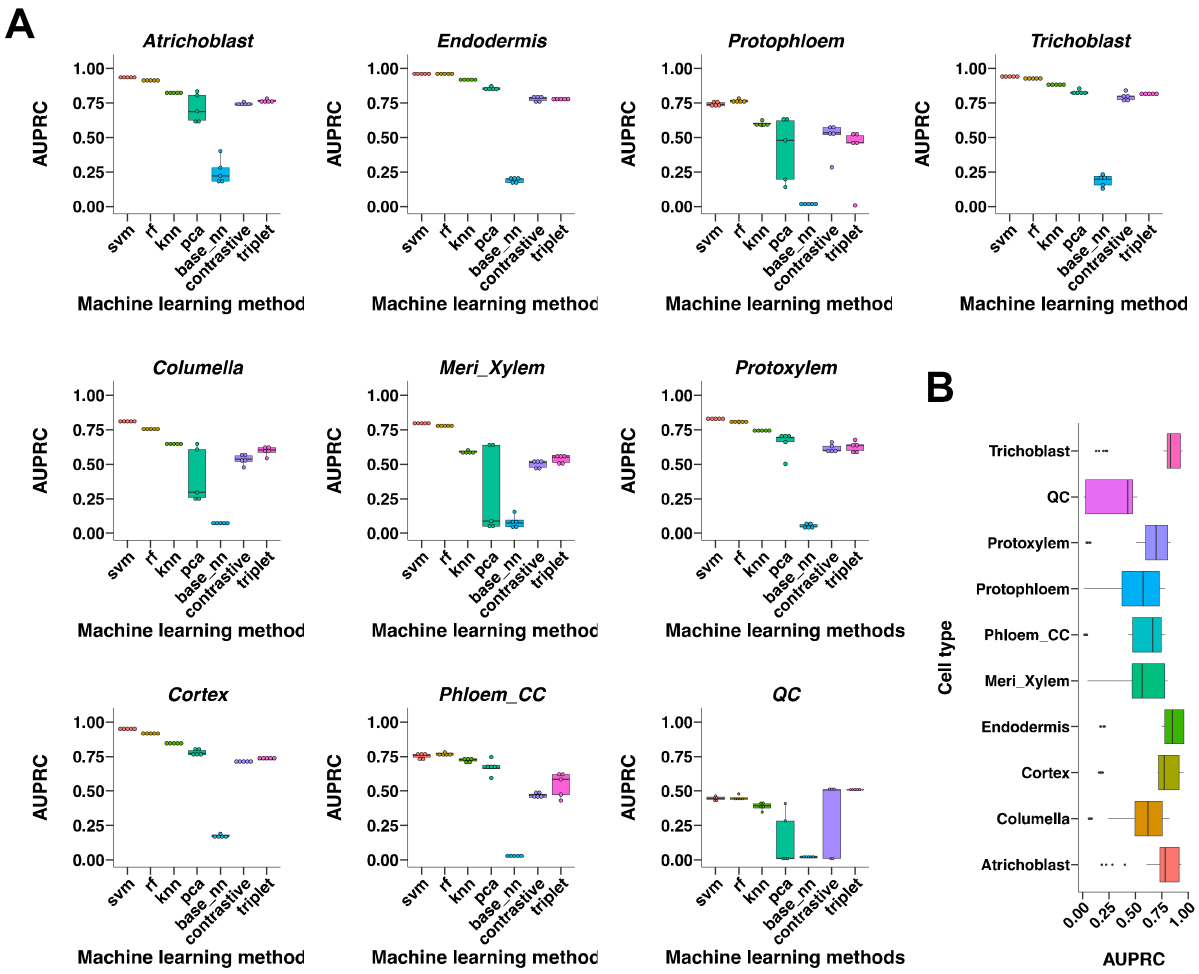


Figure 2 Classification performance of ten root cell types of *Arabidopsis*. A, comparison of seven machine learning models on cell type classification. B, comparison of classification performance of all the ten cell types. AUPRC means Area Under Precision-Recall Curve.

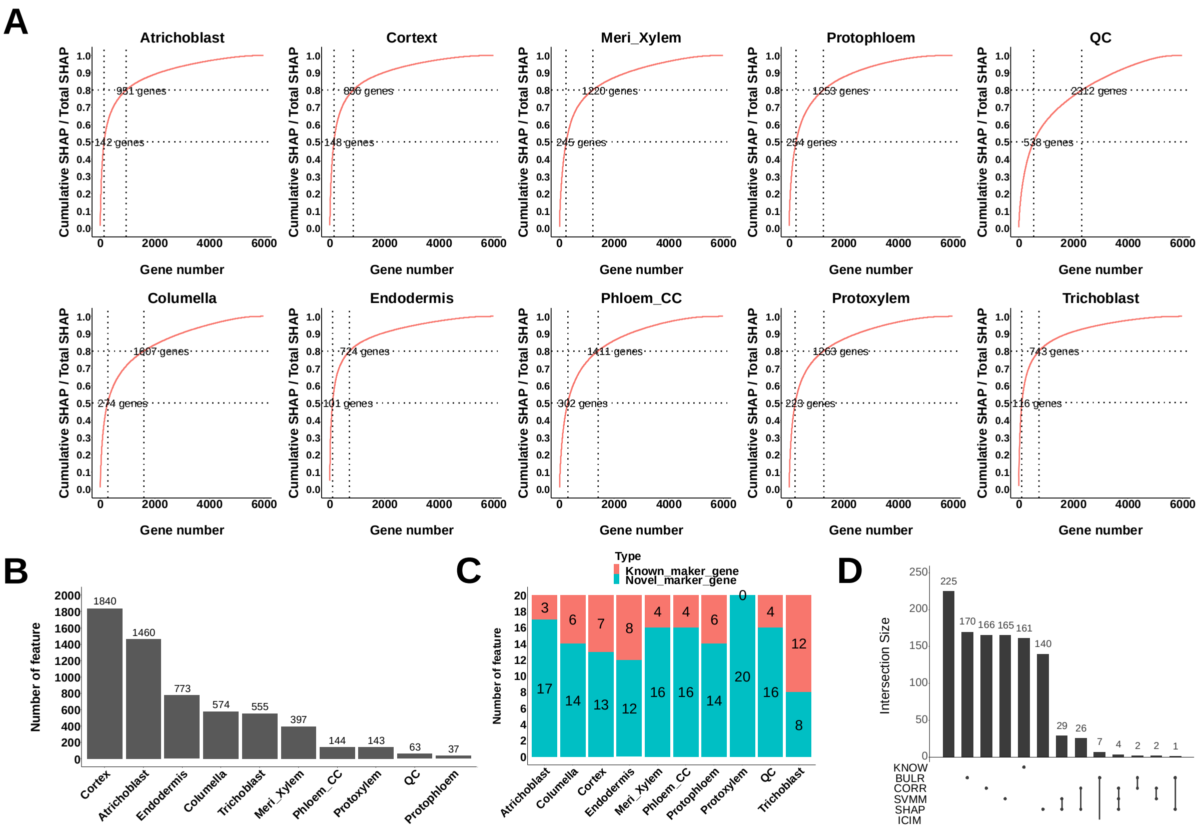


Figure 3 Summary of SHAP markers. A, the proportion of cumulative SHAP values to total SHAP value for each cell type. The two gene numbers in each plot are where cumulative SHAP value equal to 50% and 80%. B, the number of SHAP marker identified in each cell type. C, comparison of number of SHAP and known markers in the 20 genes with the highest SHAP value in each cell type. D, summary of gene counts from six marker types. Set size means gene count of different marker types. The dots under the bars mean the genes are specifically exist in the relative marker type. The line connected between two or more dots under the bars mean genes exist in two or more marker types. If two or more marker types do not have connection, it means these groups do not have shared genes.

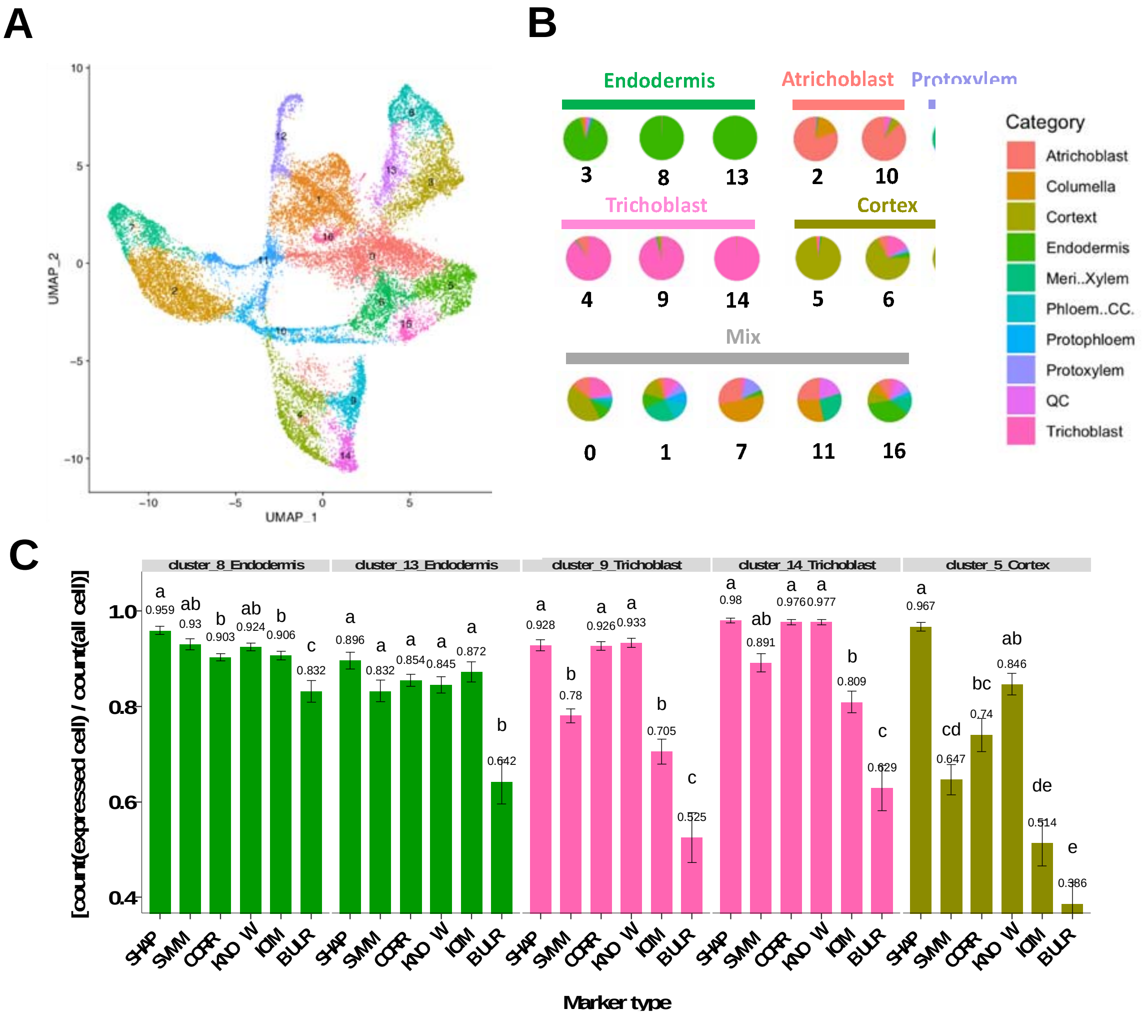


Figure 4 UMAP dimensional reduction of the root cells. A. A UMAP plot where cells were clustered into 17 populations using the Seurat package (Butler et al., 2018). Points suggest individual cells and are colored by the clusters. B. Pie plots of cell composition in each cluster. If a cell type with cells occupy over 50% in a cluster, this type is defined as the dominant cell type. The labels above the pies are names of the dominant cell types of the clusters. Otherwise, the label for the clusters is 'Mix'. C. Comparisons of proportion of expressed cells among the six marker types. All pair wise comparisons are statistically significant as indicated by different letters (a, b, c, d, and e). If two bars have the same letter, then they are not significantly different from each other.

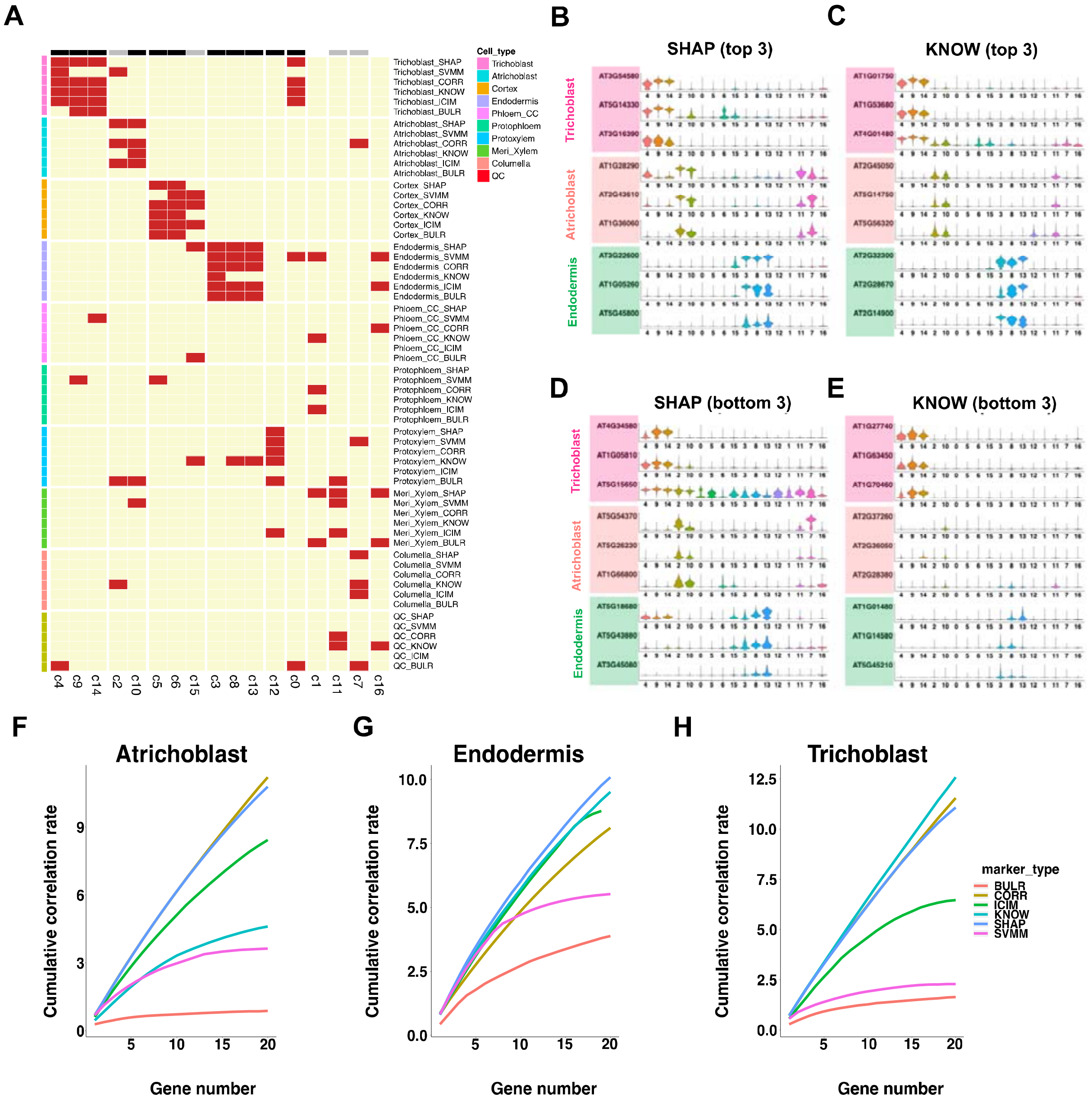


Figure 5 Comparison of expression patterns among different marker types. A. Heat map of cell types assigned to each clusters by different markers. B-E. Violin plots that show the expression of top three markers (B,C) and bottom three markers (D,E) in trichoblast, atrichoblast, and endodermis, respectively, across all clusters. F-H. Cumulative correlation plot for top 20 markers for six types of markers.

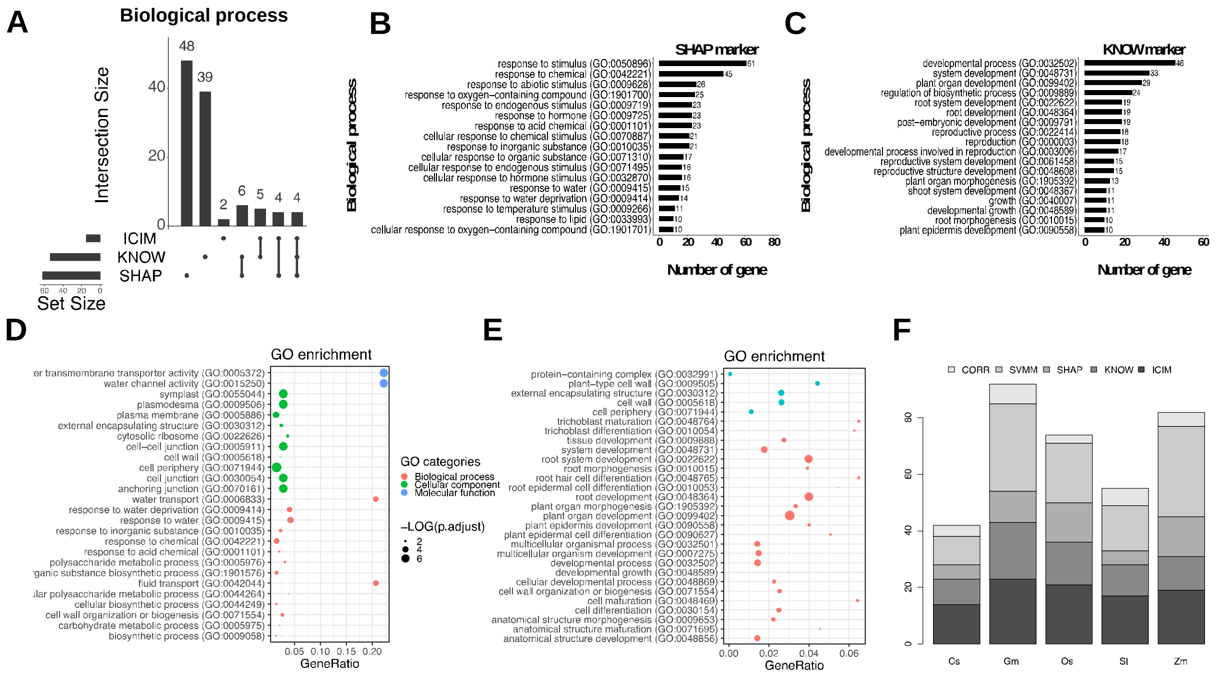


Figure 6 Comparison of GO annotation among SHAP, KNOW, and ICIM markers. A. Count summary of GO biological process items. Set size means count of GO terms under the SHAP, KNOW, and ICIM markers, respectively. The dots under the bars mean GO terms are specifically exist in the relative marker type. The line connected between two or three dots under the bars mean GO terms are overlapped in two or three marker types. B. GO terms of the SHAP markers involved in the responsiveness of stimulus in the biological process. C. GO terms of the KNOW markers associated with the root vegetative and reproduction developments

. The number beside the bar is marker gene number under the relative GO term. D. GO enrichment analysis on the SHAP markers. E. GO enrichment analysis on the KNOW markers. The GeneRatio in the bottom is proportion of gene number to the background gene number. F. Number of root hair marker genes identified using five marker types.

Table 1 Performance comparison among SHAP, SVMM, CORR, KNOW, ICIM, BULR markers using random forest model and SVM models

Marker_type	**Model**[a]	**Evaluation method**[b]	**Evaluation score**	**Standard deviation**	**Letters_mark_significant_difference**[c]	**Ranking**[d]
SHAP	RF	auPRC	0.775	0.001	b	2
SVMM	RF	auPRC	0.767	0.001	c	4
CORR	RF	auPRC	0.774	0.002	b	3
KNOW	RF	auPRC	0.732	0.003	d	5
ICIM	RF	auPRC	0.930	0.003	a	1
BULR	RF	auPRC	0.582	0.003	e	6
SHAP	RF	auROC	0.957	0.001	b	2
SVMM	RF	auROC	0.956	0.000	b	3
CORR	RF	auROC	0.956	0.001	b	4
KNOW	RF	auROC	0.946	0.001	c	5
ICIM	RF	auROC	0.993	0.001	a	1
BULR	RF	auROC	0.867	0.003	d	6
SHAP	SVM	auPRC	0.721	0.004	c	3
SVMM	SVM	auPRC	0.682	0.002	d	4
CORR	SVM	auPRC	0.732	0.004	b	2
KNOW	SVM	auPRC	0.682	0.003	d	5
ICIM	SVM	auPRC	0.974	0.002	a	1
BULR	SVM	auPRC	0.450	0.005	e	6
SHAP	SVM	auROC	0.936	0.001	c	3
SVMM	SVM	auROC	0.930	0.001	e	5
CORR	SVM	auROC	0.948	0.001	b	2
KNOW	SVM	auROC	0.933	0.001	d	4
ICIM	SVM	auROC	0.997	0.000	a	1
BULR	SVM	auROC	0.806	0.002	f	6

Note: a, RF: random forests; SVM: support vector machine. b, auPRC: area under the precision-recall curve; auROC: area under the receiver operating characteristic. c, all pair wise comparisons are statistically significant ($p < 0.05$) as indicated by different letters (a, b, c, d, e, and f). d, ranking of marker types according to their evaluation score.

Reference

Abdi, H., and Williams, L.J. (2010). Tukey's honestly significant difference (HSD) test. Encyclopedia of Research Design. Thousand Oaks, CA: Sage, 1-5.

Alavi, A., Ruffalo, M., Parvangada, A., Huang, Z., and Bar-Joseph, Z. (2018). A web server for comparative analysis of single-cell RNA-seq data. Nature communications *9*, 1-11.

Ali, J., Khan, R., Ahmad, N., and Maqsood, I. (2012). Random forests and decision trees. International Journal of Computer Science Issues (IJCSI) *9*, 272.

Alquicira-Hernandez, J., Sathe, A., Ji, H.P., Nguyen, Q., and Powell, J.E. (2019). scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. Genome biology *20*, 1-17.

Azodi, C.B., Tang, J., and Shiu, S.-H. (2020). Opening the Black Box: Interpretable machine learning for geneticists. Trends in Genetics.

Ben-Hur, A., and Weston, J. (2010). A user's guide to support vector machines. In Data mining techniques for the life sciences (Springer), pp. 223-239.

Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson correlation coefficient. In Noise reduction in speech processing (Springer), pp. 1-4.

Bezrutczyk, M., Zoellner, N., Kruse, C.P., Hartwig, T., Lautwein, T., Koehrer, K.-E., Frommer, W.B., and Kim, J.-Y. (2020). Phloem loading via the abaxial bundle sheath cells in maize leaves. bioRxiv.

Birnbaum, K., Jung, J.W., Wang, J.Y., Lambert, G.M., Hirst, J.A., Galbraith, D.W., and Benfey, P.N. (2005). Cell type–specific expression profiling in plants via cell sorting of protoplasts from fluorescent reporter lines. Nature methods *2*, 615-619.

Brady, S.M., Orlando, D.A., Lee, J.-Y., Wang, J.Y., Koch, J., Dinneny, J.R., Mace, D., Ohler, U., and Benfey, P.N. (2007). A high-resolution root spatiotemporal map reveals dominant expression patterns. Science *318*, 801-806.

Bruex, A., Kainkaryam, R.M., Wieckowski, Y., Kang, Y.H., Bernhardt, C., Xia, Y., Zheng, X., Wang, J.Y., Lee, M.M., and Benfey, P. (2012). A gene regulatory network for root epidermis cell differentiation in Arabidopsis. PLoS genetics *8*.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature biotechnology *36*, 411-420.

Cai, J., Luo, J., Wang, S., and Yang, S. (2018). Feature selection in machine learning: A new perspective. Neurocomputing *300*, 70-79.

Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST) *2*, 1-27.

Chollet, F. (2015). Keras: Deep learning library for theano and tensorflow. URL: https://keras. io/k *7*, T1.

Denyer, T., Ma, X., Klesen, S., Scacchi, E., Nieselt, K., and Timmermans, M.C. (2019). Spatiotemporal developmental trajectories in the Arabidopsis root revealed using high-throughput single-cell RNA sequencing. Developmental cell *48*, 840-852. e845.

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 biology *16*, 9.

Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F.J. (2019). Deep learning: new computational modelling techniques for genomics. Nature Reviews Genetics *20*, 389-403.

Gulli, A., and Pal, S. (2017). Deep learning with Keras (Packt Publishing Ltd).

Hoffer, E., and Ailon, N. (2015). Deep metric learning using triplet network. In International Workshop on Similarity-Based Pattern Recognition (Springer), pp. 84-92.
Hu, Y., Hase, T., Li, H.P., Prabhakar, S., Kitano, H., Ng, S.K., Ghosh, S., and Wee, L.J.K. (2016). A machine learning approach for the identification of key markers involved in brain development from single-cell transcriptomic data. BMC genomics *17*, 1025.
Huang, L., Shi, X., Wang, W., Ryu, K.H., and Schiefelbein, J. (2017). Diversification of root hair development genes in vascular plants. Plant Physiology *174*, 1697-1712.
Jean-Baptiste, K., McFaline-Figueroa, J.L., Alexandre, C.M., Dorrity, M.W., Saunders, L., Bubb, K.L., Trapnell, C., Fields, S., Queitsch, C., and Cuperus, J.T. (2019). Dynamics of gene expression in single root cells of Arabidopsis thaliana. The Plant Cell *31*, 993-1011.
Koch, G., Zemel, R., and Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. In ICML deep learning workshop (Lille).
Li, S., Yamada, M., Han, X., Ohler, U., and Benfey, P.N. (2016). High-resolution expression map of the Arabidopsis root reveals alternative splicing and lincRNA regulation. Developmental cell *39*, 508-522.
Libbrecht, M.W., and Noble, W.S. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics *16*, 321-332.
Liu, S., and Trapnell, C. (2016). Single-cell transcriptome sequencing: recent advances and remaining challenges. F1000Research *5*.
Luecken, M.D., and Theis, F.J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. Molecular systems biology *15*.
Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature machine intelligence *2*, 2522-5839.
Luo, Y., Coskun, V., Liang, A., Yu, J., Cheng, L., Ge, W., Shi, Z., Zhang, K., Li, C., and Cui, Y. (2015). Single-cell transcriptome analyses reveal signals to activate dormant neural stem cells. Cell *161*, 1175-1186.
Oda, Y., and Fukuda, H. (2012). Secondary cell wall patterning during xylem differentiation. Current opinion in plant biology *15*, 38-44.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research *12*, 2825-2830.
Pliner, H.A., Shendure, J., and Trapnell, C. (2019). Supervised classification enables rapid annotation of cell atlases. Nature methods *16*, 983-986.
Qiao, Z., and Libault, M. (2013). Unleashing the potential of the root hair cell as a single plant cell type model in root systems biology. Frontiers in plant science *4*, 484.
Ryu, K.H., Huang, L., Kang, H.M., and Schiefelbein, J. (2019). Single-cell RNA sequencing resolves molecular relationships among individual plant cells. Plant physiology *179*, 1444-1456.
Satterlee, J.W., Strable, J., and Scanlon, M.J. (2020). Plant stem cell organization and differentiation at single-cell resolution. bioRxiv.
Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks *61*, 85-117.
Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815-823.

Shulse, C.N., Cole, B.J., Ciobanu, D., Lin, J., Yoshinaga, Y., Gouran, M., Turco, G.M., Zhu, Y., O'Malley, R.C., and Brady, S.M. (2019). High-throughput single-cell transcriptome profiling of plant cell types. Cell reports *27*, 2241-2247. e2244.

Song, Q., Lee, J., Akter, S., Rogers, M., Grene, R., and Li, S. (2020). Prediction of condition-specific regulatory genes using machine learning. Nucleic Acids Research.

Steudle, E., and Peterson, C.A. (1998). How does water get through roots? Journal of experimental Botany *49*, 775-788.

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. Genome research *25*, 1491-1498.

Turco, G.M., Rodriguez-Medina, J., Siebert, S., Han, D., Vahldick, H., Shulse, C.N., Cole, B.J., Juliano, C., Dickel, D.E., and Savageau, M.A. (2019). Molecular Mechanisms Driving Bistable Switch Behavior in Xylem Cell Differentiation. bioRxiv, 543983.

Usoskin, D., Furlan, A., Islam, S., Abdo, H., Lönnerberg, P., Lou, D., Hjerling-Leffler, J., Haeggström, J., Kharchenko, O., and Kharchenko, P.V. (2015). Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. Nature neuroscience *18*, 145.

Voß, U., Wilson, M.H., Kenobi, K., Gould, P.D., Robertson, F.C., Peer, W.A., Lucas, M., Swarup, K., Casimiro, I., and Holman, T.J. (2015). The circadian clock rephases during lateral root organ initiation in Arabidopsis thaliana. Nature communications *6*, 1-9.

Wang, F., Liang, S., Kumar, T., Navin, N., and Chen, K. (2019). SCMarker: ab initio marker selection for single cell transcriptome profiling. PLoS computational biology *15*, e1007445.

Wang, Y., and Navin, N.E. (2015). Advances and applications of single-cell sequencing technologies. Molecular cell *58*, 598-609.

Zhang, A.W., O'Flanagan, C., Chavez, E.A., Lim, J.L., Ceglia, N., McPherson, A., Wiens, M., Walters, P., Chan, T., and Hewitson, B. (2019a). Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. Nature methods *16*, 1007-1015.

Zhang, T.-Q., Xu, Z.-G., Shang, G.-D., and Wang, J.-W. (2019b). A single-cell RNA sequencing profiles the developmental landscape of Arabidopsis root. Molecular plant *12*, 648-660.

Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., Luo, T., Xu, L., Liao, G., and Yan, M. (2019c). CellMarker: a manually curated resource of cell markers in human and mouse. Nucleic acids research *47*, D721-D728.

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., and Telenti, A. (2019). A primer on deep learning in genomics. Nature genetics *51*, 12-18.