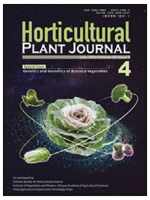
July 2024. Horticultural Plant Journal, 10 (4): 971–982

Horticultural Plant Journal

Available online at www.sciencedirect.com
The journal's homepage: http://www.keaipublishing.com/en/journals/horticultural-plant-journal



Research Paper

Gene expression, transcription factor binding and histone modification predict leaf adaxial–abaxial polarity related genes

Wei Sun [a,b,c,d], *Zhicheng Zhang* [a,b], *Guusje Bonnema* [d], *Xiaowu Wang* [a,b,*], and *Aalt Dirk Jan van Dijk* [e,*]

[a] *State Key Laboratory of Vegetable Biobreeding, Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China*
[b] *Sino-Dutch Joint Lab of Vegetable Genomics, Beijing 100081, China*
[c] *Bioinformatics Group, Wageningen University and Research, 6708 PB Wageningen, the Netherlands*
[d] *Plant Breeding, Wageningen University and Research, 6708 PB Wageningen, the Netherlands*
[e] *Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, 1090 GE Amsterdam, the Netherlands*

Received 20 February 2024; Received in revised form 1 May 2024; Accepted 4 June 2024

Available online 29 June 2024

ABSTRACT

Leaf adaxial–abaxial (ad–abaxial) polarity is crucial for leaf morphology and function, but the genetic machinery governing this process remains unclear. To uncover critical genes involved in leaf ad–abaxial patterning, we applied a combination of *in silico* prediction using machine learning (ML) and experimental analysis. A Random Forest model was trained using genes known to influence ad–abaxial polarity as ground truth. Gene expression data from various tissues and conditions as well as promoter regulation data derived from transcription factor chromatin immunoprecipitation sequencing (ChIP-seq) was used as input, enabling the prediction of novel ad–abaxial polarity-related genes and additional transcription factors. Parallel to this, available and newly-obtained transcriptome data enabled us to identify genes differentially expressed across leaf ad–abaxial sides. Based on these analyses, we obtained a set of 111 novel genes which are involved in leaf ad–abaxial specialization. To explore implications for vegetable crop breeding, we examined the conservation of expression patterns between *Arabidopsis* and *Brassica rapa* using single-cell transcriptomics. The results demonstrated the utility of our computational approach for predicting candidate genes in crop species. Our findings expand the understanding of the genetic networks governing leaf ad–abaxial differentiation in agriculturally important vegetables, enhancing comprehension of natural variation impacting leaf morphology and development, with demonstrable breeding applications.

Keywords: Machine learning; Leaf polarity; *Arabidopsis thaliana*; *Brassica rapa*; Transcription factor

1. Introduction

Leaves constitute the primary site of photosynthesis, and their structure contributes to this essential task. Leaf primordia develop into fully expanded leaves by growth and development among three axes: lateral defining leaf width, longitudinal defining leaf length and dorsi-ventral defining leaf thickness and differentiation of ad–abaxial sides (Machida et al., 2015). The acquisition of a flat lamina with correct ad–abaxial polarity optimizes the primary leaf function of photosynthesis and transpiration and is a key innovation in leaf evolution. The adaxial side of the leaf originates closest to the meristem and develops a cuticle, an epidermal cell layer and a layer of palisade mesophyll cells that are closely packed to maximize light absorption, while

[*] Corresponding authors.
E-mail addresses: wangxiaowu@caas.cn; a.d.j.vandijk@uva.nl

Peer review under responsibility of Chinese Society of Horticultural Science (CSHS) and Institute of Vegetables and Flowers (IVF), Chinese Academy of Agricultural Sciences (CAAS)

https://doi.org/10.1016/j.hpj.2024.06.002

the abaxial side contains a loosely packed spongy mesophyll layer and a higher density of stomatal pores in the epidermal cell layer that facilitates gas exchange and regulates transpiration (Braybrook and Kuhlemeier, 2010; Yamaguchi et al., 2012). The initiation and appropriate differentiation of ad–abaxial cells is essential for photosynthesis and transpiration, and consequently, for the survival of a plant. Leaf polarity along the ad–abaxial axis underpins key agricultural traits across vegetable crops. In Chinese cabbage for example, the change from flat, extended leaves to incurving inward leaves revealing abaxial surfaces is crucial for head development (Gao et al., 2020). As a result, modifying genetic variables that regulate the formation of adaxial and abaxial leaf identities can influence key morphogenic phases such as the early rosette period, with significant implications for agricultural productivity.

Considerable efforts have been made to investigate the developmental mechanisms underlying leaf polarity specification. Three members of the HD-ZIPIII gene family, namely *PHABULOSA* (*PHB*), *PHAVOLUTA* (*PHV*) and *REVOLUTA* (*REV*), are essential for adaxial determination. Mutations in a presumed microRNA regulatory target of the *PHB* and *PHV* genes resulted in leaf polarity defects and the formation of meristems in ectopic positions, whereas similar mutations in the REV gene resulted in vascular bundle and leaf polarity defects (McConnell et al., 2001; Yamaguchi et al., 2012). Another gene family that plays a key role in adaxial fate specification is the ARP family, which includes *ASYMMETRIC LEAVES1 (AS1)* and *ASYMMETRIC LEAVES2 (AS2)*. *AS1* and *AS2* in *Arabidopsis thaliana* (hereafter *Arabidopsis*) form a protein complex and positively regulate *PHB, PHV* and *REV*. *AS2* encodes a leucine-zipper motif-containing AS2/LOB domain protein that is expressed on the adaxial surface of leaf primordia. Overexpressing *AS2* causes adaxialized leaf development even though neither *as1* or *as2* single mutants nor the *as1/as2* double mutant exhibit clear ad–abaxial polarity abnormalities. This suggests that the *AS1/AS2* pathway functions redundantly with other routes in adaxial destiny specification (Iwakawa et al., 2007). In addition, a variety of mutations, including those in the ribosomal protein genes, 26S proteasome components, the trans-acting short-interfering RNA (ta-siRNA) biogenesis genes, and chromatin modification genes raise the severity of ad–abaxial polarity deficits in as1 and as2 (Huang et al., 2006; Yao et al., 2008). The determination of abaxial destiny is facilitated by the KANADIs and *AUXIN RESPONSE FACTOR 3/4 (ARF3/4)* (Kerstetter et al., 2001; Pekker et al., 2005; Maugarny-Calès and Laufs, 2018). The dramatic adaxialization of leaves with loss of lamina flattening seen in *kan1 kan2* double mutants and *kan1 kan2 kan3* triple mutants of the *Arabidopsis* plant is caused by the progressive loss of KANs activity. While *ARF3/ARF4* work downstream of KANs, *ARF3* loss of function can partly revert the near-radialized and abaxialized phenotype of genotypes expressing *KAN1* ectopically, illustrating that *ARF3* is an integral part of the KAN abaxial promoting pathway (Pekker et al., 2005). Additionally, combined loss of *ARF3* and *ARF4* also results in adaxialized leaves, which are extremely similar to double *KAN1* and *KAN2* mutant leaves (Pekker et al., 2005). In addition, YABBY genes are important for maintaining polarity, but not for the initial development of leaf polarity. Polar YABBY expression becomes localized to the boundary between the ad–abaxial domains at the leaf margins, where lamina outgrowth occurs (Eshed et al., 2004; Juarez et al., 2004). Moreover, research on ad–abaxial patterning in vegetable crops has explored links to agricultural traits. For instance, it was demonstrated that in lettuce, the KNOX transcription factor *LsKN1* suppresses expression of the adaxial gene *LsAS1* (Yu et al., 2020). This polarized control of morphogenesis along the ad–abaxial axis explains development of multi-layered lettuce heads impacting yield and taste attributes. Four genes were identified in Chinese cabbage, namely *BrARF3.1*, *BrARF4.1*, *BrKAN2.1* and *BrKAN2.3* located in genomic regions under selection (Cheng et al., 2016). *Arabidopsis* mutants of these orthologs exhibit comparable leaf curling phenotypes, suggesting ad–abaxial patterning has been a target of selection for the heading trait in Chinese cabbage. Together, these findings in lettuce and Chinese cabbage demonstrate that genes controlling ad–abaxial polarity can regulate morphology changes like heading that have been prioritized in vegetable crop breeding. While our fundamental understanding of the influence ad–abaxial patterning has been expanded, a thorough characterization of the genetic elements involved in this specific patterning pathway has not yet been performed.

High-throughput technologies have facilitated data collection in diverse areas including genomics, transcriptomics, epigenomics, proteomics, and metabolomics, making it possible to study gene function in terms of expression levels, post-translational modifications, etc. Compared to single-omics data, integration of multi-omics data can offer a more thorough understanding of the regulatory processes governing a trait at different levels. Researchers are increasingly aiming to integrate numerous omics data sources in order to capture diverse elements or acquire higher cellular resolution of the biological process under inquiry (Depuydt et al., 2022). However, since the outcomes of multi-omics data are typically acquired from different treatments and distinct samples, integration of these data remains a challenge. Machine learning (ML) refers to a general set of methodologies to find patterns in high-dimensional data and use these patterns in predictive models. In the analysis of a wide range of -omics data, ML can integrate various data sources in order to address relevant biological questions. This includes for example predicting gene regulatory networks, or interpreting the relationships between genotypes and phenotypes (Acharjee et al., 2011; Reel et al., 2021). ML has lately sparked improvements in a variety of scientific disciplines, and it is anticipated that it will do the same for the plant sciences (van Dijk et al., 2021).

Here, we applied a combination of *in silico* prediction using ML and experimental analysis to obtain a more comprehensive view of which genes are involved in leaf ad–abaxial specialization. Given the availability of substantial-omics data resources, our study first focused on *Arabidopsis*. As a well-studied model organism, *Arabidopsis* offers extensive genomics information and sufficient datasets that enable prediction and validation of novel genes. After developing a machine learning model trained on *Arabidopsis* data that accurately predicts ad–abaxial identity genes, we next sought to apply this model to the crop species *Brassica rapa (B. rapa)*. Translating findings from *Arabidopsis* to crops is challenging due to their greater genetic complexity, but by identifying *B. rapa* homologs with conserved expression patterns to already-validated *Arabidopsis* ad–abaxial regulators, we demonstrated that our computational approach facilitates discovery of biologically relevant patterning genes in less-studied species. Our strategy of coupling computational predictions between model and crop species with experimental validation expands knowledge of leaf polarity genetic components and their transferability across the crop plants phylogeny.

2. Materials and methods

2.1. *Expression data sources*

A total of 99 raw transcriptome datasets of *Arabidopsis* across different biological conditions were selected from the EMBL-EBI Expression Atlas (Cantelli et al., 2022) including plant development, biotic stress, abiotic stress, hormone treatment, and diurnal expression (Table S1). Sequence artifacts, including reads containing adapter contamination, low-quality nucleotides and unrecognizable nucleotide (N) were trimmed for subsequent bioinformatics analysis using Trimmomatic v.0.38 (Bolger et al., 2014). Post trimming reads were aligned to the *A. thaliana* TAIR10 genome using STAR (Dobin et al., 2013) with default stringencies and parameters. We mapped the read loci to RNA features using the featureCounts function of Subread (Liao et al., 2014). This resulted in 2 173 expression values which were used as numerical features for the genes. Twenty-one specific expression datasets biased towards leaf polarization in leaf primordia were also collected (Tian et al., 2019). In total, this meant 2 194 expression features were available.

2.2. *TF binding data and histone modification data sources*

Arabidopsis transcription factor (TF) binding sites and histone modification sites were collected from ReMap2022 (Hammal et al., 2022). Peaks located in regions 3 kb upstream or 3 kb downstream of the transcription start site were annotated with ChIPseeker (Yu et al., 2015). For each TF binding/modification dataset, annotated genes obtained the feature value 1 while the feature value for unannotated genes was set to 0. This introduced 423 TF binding site features and 33 histone modification features, leading to a total of 456 binary features.

2.3. *Machine learning classification*

For *Arabidopsis*, prediction models were built based on the above defined 2 194 continuous features and 456 binary features using the Random Forest (RF) algorithm implemented using the randomForest (Liaw and Wiener, 2002) R package. The number of trees was set to 500 and mtry (the number of features to consider at each split point) was set to 51.

A total of 61 genes related to leaf polarity development reported in the literature were treated as "CLASS ad/ab" genes (Table S2). Given that there is no definite knowledge on which genes are *not* involved in leaf polarity development, we randomly selected genes for "CLASS non-ad/ab". To obtain a balanced set of labeled genes for training the model, a total of 61 genes was therefore randomly selected from the rest of the genes to use as "CLASS non-ad/ab". Since some of these genes might actually be genes related to leaf polarity development, we repeated the sampling three times and compared the predictions obtained with these different sets of randomly-selected CLASS non-ad/ab genes.

In this way, the genes were divided into a modeling set (i.e. 61 "CLASS ad/ab" genes and the 61 "CLASS non-ad/ab" genes) and a prediction set (the remaining genes). To evaluate the performance of the model, 10-fold cross validation was performed on the modeling set. The CLASS ad/ab probability produced by the model was used as the prediction score; genes with probability greater than 0.5 were classified as CLASS ad/ab. The true positives (TP; genes predicted as CLASS ad/ab which indeed were CLASS ad/ab), false positives (FP; genes predicted as CLASS ad/ab which were not CLASS ad/ab), true negatives (TN; genes predicted as CLASS non-ad/ab which indeed were CLASS non-ad/ab) and false negatives (FN; genes predicted as CLASS non-ad/ab which were not CLASS non-ad/ab) were used to calculate the precision $(\frac{TP}{TP+FP})$, recall $(\frac{TP}{TP+FN})$ and F_1 score $(\frac{2TP}{2TP+FP+FN})$ for each method. Receiver Operating Characteristic curve (ROC curve) and the Area Under the ROC Curve (AUC) were computed using the R package ROCR (Sing et al., 2005). Partial dependence plots were made using the pdp R package (Greenwell, 2017) for visualization and to analyze how variation in the significant variables related to the prediction of CLASS non-ad/ab vs CLASS ad/ab.

To demonstrate the broader applicability of our approach, the same modeling strategy was used to predict leaf ad/abaxial genes in maize (*Zea mays*). A total of 22 genes were used as ground truth "CLASS ad/ab" genes, based on their association with leaf ad–abaxial polarization related Gene Ontology terms (GO: 2000011, GO: 0009943, GO: 0009944, GO: 0009955) according to MaizeGDB (Woodhouse et al., 2021). 22 genes were randomly selected one at a time from the rest of the genes to use as "CLASS non-ad/ab".

2.4. *Plant materials and RNA extraction*

The *A. thaliana* [L.] Heynh. ecotype Columbia (Col-0) plants were grown in the No.4 plastic tunnel on the experimental farm at The Institute of Vegetables and Flowers in Beijing, China in 2022. The voucher specimen of the wild *A. thaliana* has been deposited in the herbaria BNU (http://sweetgum.nybg.org/ih/herbarium.php) with the reference code 3152970. Double-sided tape was used to separate all the leaves of three to four-week-old seedlings to obtain the adaxial epidermis directly. Then a blade was used to gently scrape off the remaining leaf to obtain the epidermis abaxial side. Total RNA was extracted from both sides of the leaves with two replicates taken from each side, using the TransZol reagent (TransGen). RNA integrity was verified with an Agilent Fragment Analyzer 5400.

2.5. *Library preparation for transcriptome sequencing*

Libraries were prepared from total RNA using the NEBNext® UltraTM RNA Library Prep Kit for Illumina®, following the protocol and adding index codes for identification. Briefly, mRNA was isolated using poly-T magnetic beads, fragmented, and converted into cDNA using M-MuLV Reverse Transcriptase and DNA Polymerase I. After end repair and adapter ligation, cDNA fragments (250–300 bp) were purified (AMPure XP) and size-selected (USER Enzyme). Libraries were amplified (Phusion High-Fidelity DNA polymerase), purified, and assessed with an Agilent Bioanalyzer 2100.

2.6. *Clustering and sequencing*

The clustering of the index-coded samples was performed on a cBot Cluster Generation System using TruSeq PE Cluster Kit v3-

cBot-HS (Illumina) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illumina Novaseq 6000 platform and 150 bp paired-end reads were generated.

2.7. RNA sequencing analysis

The analysis to obtain expression values is the same as described above for the 99 transcriptome datasets. Differential expression was called using DESeq2 (Love et al., 2014) package. The Wald test was used to generate *P*-values and $\mathrm{Log_2FC}$. Genes with adjusted *P*-values < 0.05 and absolute $\log_2\mathrm{FC} > 1$ were called as differentially expressed genes for each comparison. The raw RNA sequencing reads generated in this study have been deposited in NCBI under the accession number PRJNA957749 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA957749).

2.8. Identification of expression patterns in Arabidopsis based on single-cell data

To identify whether preferential expression was exhibited by the 111 selected genes (found from our ML model; differential expression in mature leaves; and differential expression in leaf primordia) and 61 known CLASS ad/ab genes in upper epidermal cells, lower epidermal cells, palisade mesophyll cells or sponge mesophyll cells in *Arabidopsis*, we directly utilized available spatial transcriptome data to calculate the average expression level in the upper epidermis cells, lower epidermis cells, sponge mesophyll cells and palisade mesophyll cells (Xia et al., 2022). A statistical approach based on standard deviation from the mean was developed to determine if one expression value was significantly higher than the others within each gene sample. The mean and standard deviation of the expression values were calculated for each gene across the four tissue types. Subsequently, the z-score was computed for each value as the difference between the individual value and the mean, divided by the standard deviation. A z-score exceeding an absolute value of 1 corresponds to a value being more than one standard deviation above or below the mean. Therefore, any expression value with a z-score over 1 was considered significantly higher than the other values for that gene, as it represented an outlier over 1 standard deviation from the central tendency of the data. By implementing this standard deviation threshold, genes displaying markedly different expression in one tissue relative to the others can be identified in a statistically manner (Tables S7, S8).

2.9. Identification of expression patterns in B. rapa based on single-cell data

For the *B. rapa* genes, we utilized the published leaf scRNA-seq dataset from NGDC (PRJCA009630). The FASTQ files were processed by Cell Ranger 3.0.2. After quality control, a total of 16 055 cells were used for downstream analyses. The Seurat workflow with similar parameters was performed as the published article (Guo et al., 2022). We developed a strategy to identify the upper epidermis and lower epidermis which had not been distinguished in previous studies. Upper and lower epidermis were distinguished, respectively, on the basis that *BrAS2* and *BrFIL1.1* are exclusively expressed in these cell types (Guo et al., 2022; Fig. S1). Therefore, we defined the cell population expressing the *BrAS2* gene as upper epidermal cell and the cell population expressing the *BrFIL1.1* gene as lower epidermal cell. The average expression values of the genes including upper epidermal cells, lower epidermal cells, palisade mesophyll cells and sponge mesophyll cells were finally obtained. For each *Arabidopsis* gene, the *B. rapa* ortholog with the highest aggregate expression across four cell types was selected as the representative homolog for subsequent analyses. The same statistical approach based on standard deviation from the mean, as described for *Arabidopsis*, was applied to determine if one expression value was significantly higher than the others within each gene sample for *B. rapa*.

3. Results

To uncover genes involved in leaf ad–abaxial polarity of *Arabidopsis*, we applied a combination of *in silico* prediction using ML and experimental analysis (Fig. 1). Known ad–abaxial polarity-influencing genes were used to train a ML model. This model learned to recognize features which make a gene more or less likely to be involved in ad–abaxial polarity. Two major types of features were included: (i) information on where and when the genes are expressed (based on transcriptome data from diverse tissues and treatments); and (ii) information on regulation of the promoter region of the genes (based on transcription factor (TF) and histone binding as measured using ChIP-seq data). Subsequently, a dedicated experimental approach was used to obtain genes differentially expressed between the adaxial and abaxial sides of leaves. Integration of the resulting set of genes with the ML model predictions validated the model and led to our final set of predicted leaf ad–abaxial polarity influencing genes.

3.1. Model construction

Sixty-one known genes underlying ad–abaxial leaf polarity were utilized as ground truth, needed to train the ML model (Table S2). To investigate the contribution of the different types of data sources to the prediction, we combined the data in different ways. Therefore, four Random Forest models were constructed based on a distinct mix of data categories from Table 1 to describe genes. The performances by 10-fold cross-validation of different models are shown in Table 2. All models performed quite well, as characterized using the Area under the ROC Curve (AUC), where the ROC Curve shows the performance of a classification model based on True Positive Rate and False Positive Rate. AUC values ranged from 0.67 to 0.82 (Table 2). Surprisingly, the model which used a set of pre-selected experimental conditions (model 1) had a somewhat worse prediction performance (AUC: 0.67 ± 0.15) compared to the model which used a larger set of experimental conditions (model 2; AUC: 0.74 ± 0.13), even though the expression data in model 1 were biased towards conditions which are relevant for leaf polarity. We also wanted to know the impact of including ChIP-seq data as features. To do so, we compared the prediction performance of two versions of the model: one incorporating only expression data, vs. one incorporating both expression data and ChIP-seq data as features. The results revealed that including ChIP-seq data improves model performance (model 1 vs. model 3, AUC increased by 0.14; model 2 vs.

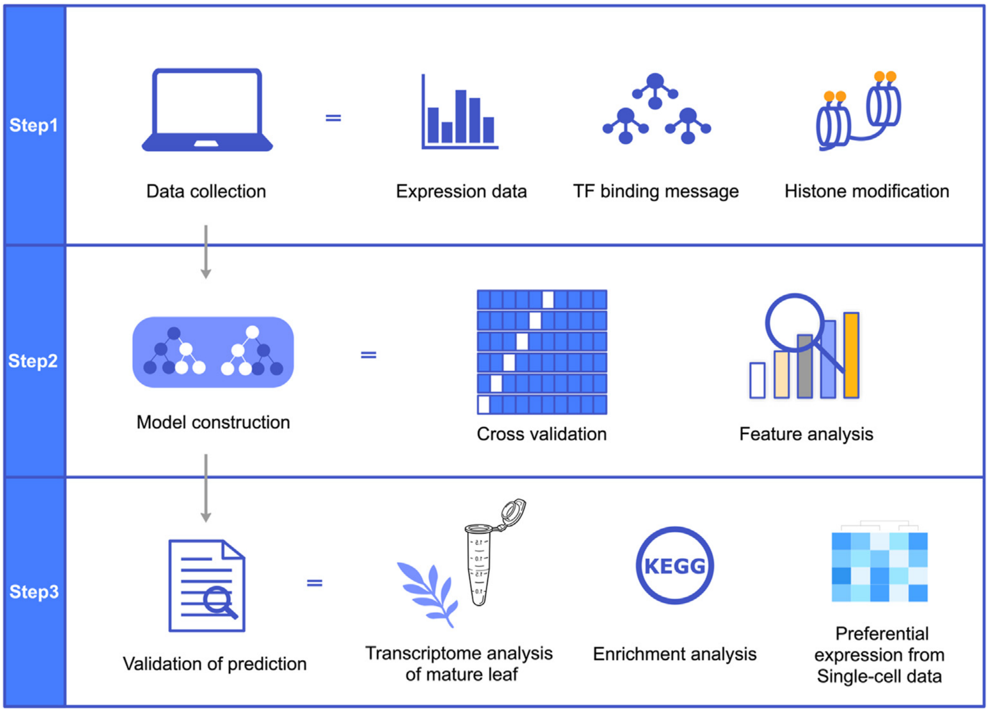


Fig. 1 Workflow of approach to predict ad–abaxial polarization genes
Three steps in this workflow include (i) collection of knowledge base and gene features to use as input for machine learning; (ii) construction of a Random Forest model to predict whether genes are involved in ad–abaxial polarization; and, (iii) validation of predicted genes.

Table 1 Features used to describe genes in order to predict role in ad–abaxial polarization

	Expression[a]	Biased_expression[b]	TF_binding[c]	Histone_modification[d]
Number of features	2 173	21	423	33
Resources	EMBL-EBI	Tian et al. (2019)	ReMap2022	ReMap2022

Note:
[a] The expression from 99 transcriptome datasets of *Arabidopsis* (2 173EXP).
[b] Twenty-one specific expression datasets biased towards leaf tissues of *Arabidopsis* (21EXP).
[c] TF binding features annotated as binary input (423TF).
[d] Histone modification features annotated as binary input (33HISTONE).

Table 2 Model performance of predicting ad–abaxial polarization genes using different features

	Precision	Recall	F1	AUC
Model 1 (21EXP)	0.67 ± 0.14	0.64 ± 0.23	0.64 ± 0.19	0.67 ± 0.15
Model 2 (2173EXP)	0.74 ± 0.13	0.69 ± 0.19	0.72 ± 0.14	0.74 ± 0.13
Model 3 (21EXP_423TF_33HISTONE)	0.81 ± 0.07	0.82 ± 0.18	0.81 ± 0.08	0.81 ± 0.08
Model 4 (2173EXP_423TF_33HISTONE)	0.83 ± 0.14	0.84 ± 0.13	0.84 ± 0.11	0.82 ± 0.11

Note: The standard deviation (SD) was obtained through 10-fold cross-validation.

model 4, AUC increased by 0.08). In addition, we compared two model versions to investigate the impact of using different sets of transcriptomics data: model 3, using pre-selected transcriptomics data, vs. model 4, using a larger set of transcriptomics data. Both model 3 and model 4 also used ChIP-seq data in addition to the transcriptomics data. Model 4 was taken as the final model for further investigation because of its largest AUC value. Moreover, by analyzing features underlying the predictive performance of this model, we can potentially learn new expression-based and ChIP-seq data-based features influencing leaf polarity (as below).

3.2. *Feature importance analysis*

The features with the greatest contribution to our final model were identified via feature importance analysis. The top 20

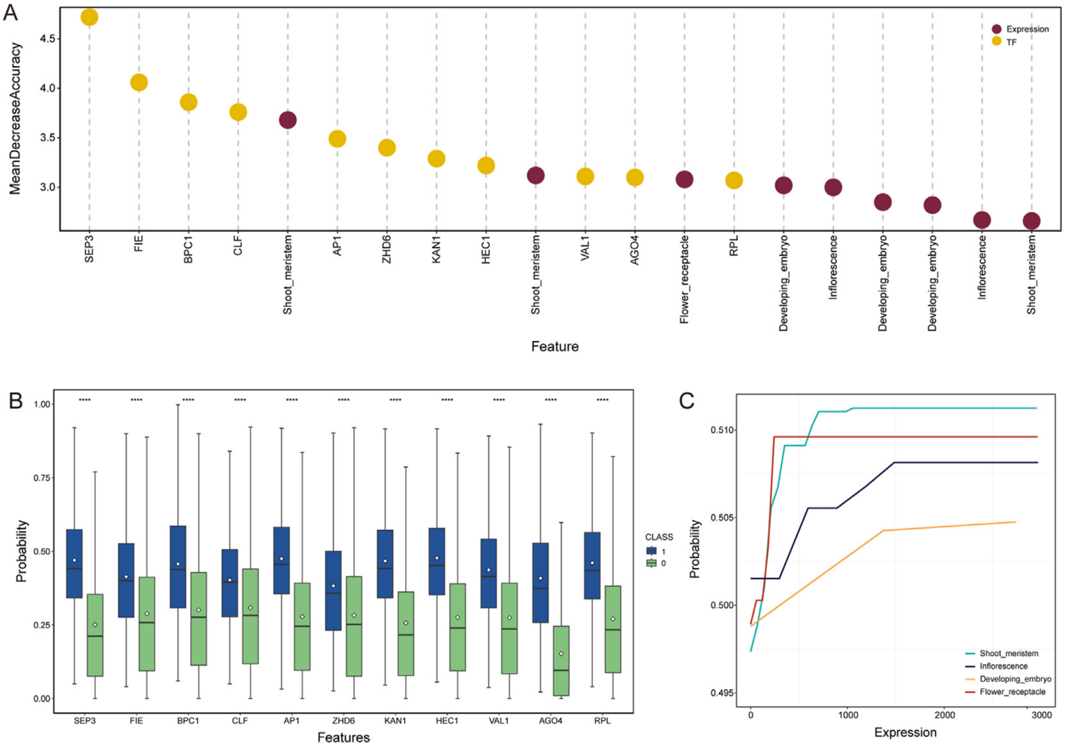


Fig. 2 Feature analysis for prediction of ad–abaxial polarization genes

(A) Feature importance of *Arabidopsis* final model; yellow points represent TF features; purple points represent expression features. (B) Partial dependence plot for top 11 binary TF features in the final model. Horizontal axis (Features) represents binding or non-binding of each TF to the promoter of a gene; for each feature, blue boxes show the probability distribution of binding genes and green boxes the probability distribution of non-binding genes. Vertical axis gives the predicted probability for genes to be involved in *Arabidopsis* ad/abaxial polarization. The white point inside of the box represents the mean value of each box. (C) Partial dependence plot for four typical expression features in the final model. The horizontal axis represents the value of gene expression in the different tissues; the vertical axis represents the predicted probability for a gene to be involved in ad/abaxial polarization.

features for mean decrease accuracy were listed; these are the features having the largest impact on the accuracy of the model, include 9 expression features and 11 TF features (Fig. 2, A). When using only these features as input, the AUC value of the model increased by 0.05, while the AUC value of the model was reduced by 0.06 when using all features except these 20, indicating the relevance of this set of features for predicting whether genes are involved in ad–abaxial polarization.

Among the most important features from the ChIP-seq data, some known leaf polarity regulators, such as CLF (CURLY LEAF) and KAN1 were found, indicating the relevance of the features selected by the model. For these as well as for other TFs obtained as important regulators (11 TF features in top 20 features), partial dependency analysis (Fig. 2, B) indicated that there was a positive relationship between these features and the predicted probability returned by the model. This means that the presence of a binding site for these TFs in the promoter of a gene makes it more likely that this gene is predicted to be involved in leaf polarity. For expression-based features, three of the top 20 important features related to expression in the shoot meristem and three related to expression in the developing embryo, aligning with their importance in shoot apical meristem (SAM) initiation and organization. In addition, one flower receptacle related feature and two inflorescences related features were also within the most influential predictors (Fig. 2, A).

Using partial dependency plots, we observed again a positive relationship between the expression of a gene in four tissues and the predicted probability for that gene to be involved in leaf polarity, among these, the shoot meristem and flower receptacle expression features exhibited a noticeably higher positive slope (Fig. 2, C).

3.3. *Functional analysis of predicted genes*

From our final model we identified 6 316 genes predicted to be involved in leaf polarity ("preGenes") with a probability larger than 0.5. Note that this large number probably reflects the variety of CLASS ad/ab genes used for training the model; we thus further analyzed this initial set of predictions in order to attain a smaller set of genes which would be most likely relevant. Specifically, there were 728 genes having a predicted probability greater than 0.9 (Fig. 3, A). To further interpret the predictions, KEGG pathway enrichment was performed on all the 6 316 preGenes (Fig. 3, B). Among the enriched pathways, ribosome biogenesis (ko: 03010) had the most significant enrichment. In particular, the two genes with the highest predicted probability (close to 1) were two ribosome protein genes (*UL13Z*, *EL27Y*). The preGenes also include 9 ribosome-related genes that have been proven to cause aberrant leaf shape (Table S3, Horiguchi et al., 2011). The proteasome pathway (ko: 03050), which is likewise known to influence leaf

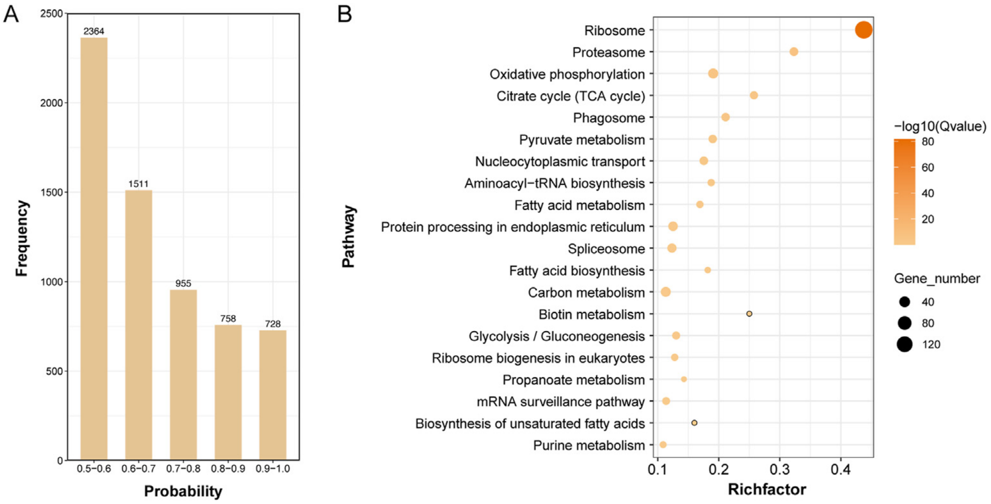


Fig. 3 Functional analysis of genes predicted to be involved in ad–abaxial polarization ("preGenes")

(A) Probability distribution of all preGenes in *Arabidopsis* model, indicating how likely genes are according to our model to be involved in ad–abaxial polarity. (B) KEGG enrichment of preGenes in *Arabidopsis* model. Richfactor for each pathway indicates the ratio between the gene numbers of significant differentially expressed genes and the total gene numbers of all annotated genes; dot color represents -$\log_{10}$ (Q value), for significance of the enrichment; dot size represents the number of genes that were enriched in each pathway.

ad–abaxial polarity (Huang et al., 2006), is also well-represented in second place, including 23 proteasome related genes with high predicted probability (Table S4).

3.4. *RNA sequencing of adaxial/abaxial leaves*

To further substantiate our set of predicted genes, we sequenced the adaxial side and abaxial side of leaves from twenty 4-week-old *Arabidopsis* plants. With the use of double-sided tape and a blade, the tissue on the abaxial and adaxial sides was collected separately. The irregular organization of leaf epidermal cells in both the adaxial and abaxial surfaces was revealed by cytological inspection, demonstrating the feasibility of taking samples after scraping the surface of leaves with a blade (Fig. 4, A). PCA analysis of RNA-seq data from the four samples revealed clear separation between adaxial and abaxial samples (Fig. 4, A). A total of 2 128 differentially expressed genes (DEGs) were obtained between the adaxial and abaxial side. Among Gene Ontology (GO) terms found to be enriched for the DEGs, the ribosome biosynthesis pathway (GO: 0042254) had the most significant enrichment. In addition, rRNA processing (GO: 0006364) and ribonucleoprotein complex biogenesis (GO: 0022613) were also identified as enriched GO terms (Fig. 4, B; Table S5).

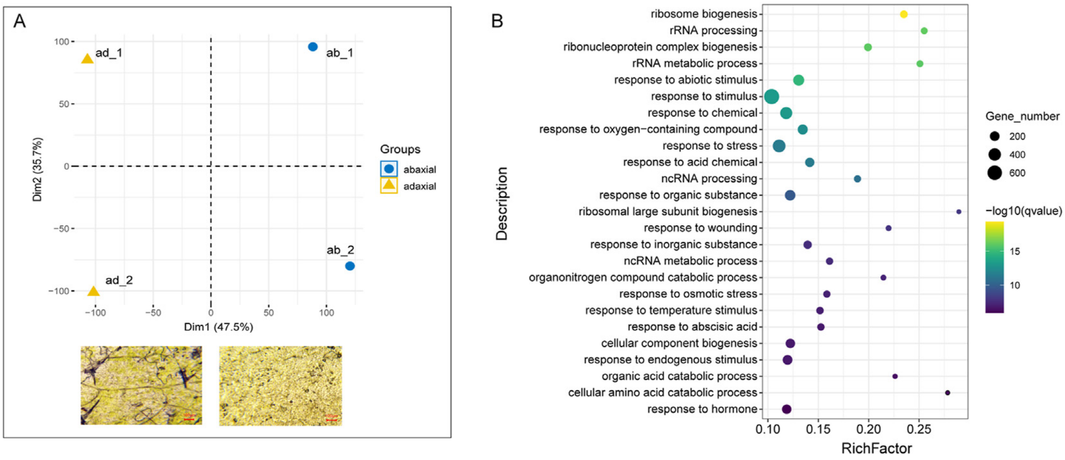


Fig. 4 Analysis of differential expressed genes between adaxial and abaxial mature leaves

(A) Principal Component Analysis of gene expression data of *Arabidopsis* adaxial and abaxial leaves and microscopic observation of both adaxial and abaxial leaf surface samples. (B) Top 25 GO terms of biological process class of 2 128 differentially expressed genes (DEGs) obtained between adaxial and abaxial mature leaves. RichFactor indicates the ratio between the number of differentially expressed genes assigned to the pathway and the total number of genes annotated to the pathway; dot size represents the gene number; dot color represents −$\log_{10}$(Q value), for significance of the enrichment.

3.5. *Intersection analysis of preGenes and DEGs*

The intersection of the genes predicted by the ML model and the genes obtained from the transcriptome analysis of mature leaves (DEGs between adaxial and abaxial leaf sides) consisted of 724 validated genes (*P*-value < 2.2e-16, indicating a significant overlap, Fig. 5, A). The distribution of prediction probabilities for these 724 genes (Fig. 5, B) showed that a substantial number exhibited predictive scores in the upper range of 0.8–1.0 by our model. Specifically, 69 genes had scores from 0.8 to 0.9, while 64 genes had scores from 0.9 to 1.0, indicating high confidence in the predictions for these genes. This upper range, containing around one-fifth of the overlap set, demonstrates that a sizable proportion of the validated genes were associated with the strongest predictive performance by the machine learning approach.

In addition, we also used a published DEG set between adaxial and abaxial side in leaf primordia to validate our preGenes (Tian et al., 2019). A total of 744 genes were found when comparing these "Tian" DEGs to our predicted genes (*P*-value < 2.2e-16, Fig. 5, A). The distribution of prediction probabilities for these 744 genes (Fig. 5, B) revealed that a considerable number of genes displayed high prediction probabilities. In this case, 70 genes had scores from 0.8 to 0.9, and 91 genes had scores from 0.9 to 1.0, suggesting that the machine learning model was able to predict a significant

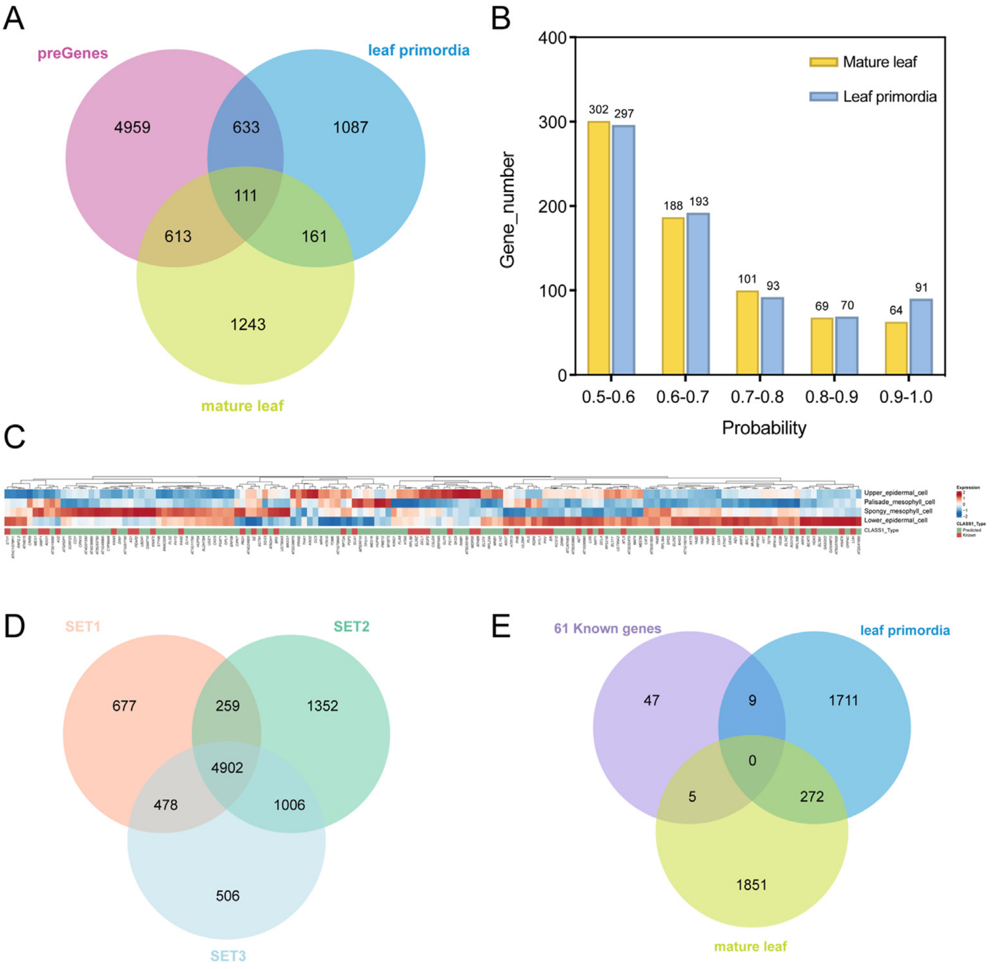


Fig. 5 Validation of ML predicted genes involved in *Arabidopsis* leaf ad/abaxial formation

(A) Venn diagram showing the number of predicted genes (preGenes), genes differentially expressed between adaxial and abaxial leaf primordia (Tian et al., 2019), and genes differentially expressed between adaxial and abaxial mature leaves. (B) Distribution of 724 validated genes overlapping between ML model predictions and mature leaf DEGs (yellow), and 744 validated genes overlapping between ML model predictions and leaf primordia DEGs (blue). The y-axis indicates the number of genes within each probability bin on the x-axis. Bins are arranged in 0.1 increments from 0.5 to 1.0 to visualize the spread of probabilities across genes. (C) Heatmap of 103 selected genes and 50 known CLASS ad/ab genes based on *Arabidopsis* single-cell spatial transcriptome profile. The genes are categorized into two groups: predicted CLASS1 genes (green) and known CLASS1 genes (red). (D) Venn diagram showing results obtained using three different selections of CLASS non-ad/ab genes. Numbers indicate total number of genes predicted to be CLASS ad/ab. (E) Venn diagram showing the number (61) of known CLASS ad/ab genes involved in leaf ad/abaxial polarity regulation, genes differentially expressed between adaxial and abaxial leaf primordia (Tian et al., 2019), and genes differentially expressed between adaxial and abaxial 20th mature leaves.

Table 3 Preferentially expressed genes among four leaf cell types in *A. thaliana* and *B. rapa*

	Highly expressed in upper epidermal cells	Highly expressed in palisade mesophyll cells	Highly expressed in spongy mesophyll cells	Highly expressed in lower epidermal cells
A. thaliana	15	6	24	35
B. rapa	30	23	10	41

portion of the genes differentially expressed in leaf primordia with high confidence.

Overall, 111 genes were found from all three approaches (our ML model; differential expression in mature leaves; and differential expression in leaf primordia; Fig. 5, A; Table S6). The presence of a substantial number of genes with high prediction scores in the overlap sets underscores the reliability of the predictions and the potential of the machine learning model to uncover novel genes related to leaf ad–abaxial polarity.

3.6. *Identification of expression pattern of selected genes based on single-cell data*

A single-cell spatial transcriptome profile of *Arabidopsis* leaves was used to identify ad/abaxial expression patterns of selected genes (Xia et al., 2022). Out of the set of 111 genes found above, 103 were present in this profile, and out of the 61 known CLASS ad/ab genes, 50 were present in this profile (the missing genes were not detected in the single-cell profile). The expression of our selected genes as well as known CLASS ad/ab genes in single cells showed different expression patterns between ab/adaxial sides (Fig. 5, C). For instance, the expression of *KAN1*, *KAN2*, *YAB1*, *YAB2* and *YAB3* was significantly higher in the abaxial side whereas some genes, such as *PGY1* and *DCL1* were highly expressed in the adaxial side. Interestingly, based on expression patterns of our selected genes and known CLASS ad/ab genes, we found that these genes shared various gene expression patterns among the four cell types, without distinct clusters for predicted or previously experimentally determined (CLASS ad/ab) genes. Some CLASS ad/ab genes are found showing similar expression pattern as their target genes. For example, the *KAN1* gene and its target gene, *AT1G18210*, are both highly expressed in spongy parenchyma and lower epidermal cells, which suggests that *KAN1* may modulate *AT1G18210* to control leaf ad/abaxial polarity in the abaxial side of the leaf.

3.7. *Identify expression preference of selected genes in Arabidopsis and B. rapa*

To determine whether the 111 selected gene and 61 known CLASS ad/ab genes showed preferential expression in upper epidermal cells, lower epidermal cells, palisade mesophyll cells or sponge mesophyll cells, we applied a statistical approach to identify genes with an expression value in one of the four cell types that was significantly higher than the expression values in the other three cell types. Based on single-cell spatial transcriptome profile of *Arabidopsis*, 35 genes showed preferential expression in lower epidermal cells, the highest among the four cell types. Expression of 15 genes was highest in upper epidermal cells, followed by 24 genes in spongy mesophyll cells and six genes in palisade mesophyll cells (Table 3, Gene ID can be found in Table S7). In order to identify whether homologs of genes predicted in *Arabidopsis* display preferential expression in *B. rapa*, we reanalyzed the published leaf scRNA-seq dataset from NGDC (PRJCA009630). A similar trend was observed in *B. rapa*, with 41 genes most highly expressed in lower epidermal cells. Preferential expression in upper epidermal cells was observed for 30 genes. Spongy mesophyll cells and palisade mesophyll cells had higher expression of 10 genes and 23 genes respectively (Table 3; gene names can be found in Table S8).

3.8. *Model validation on leaf ad/abaxial genes in maize*

To demonstrate our approach's wider applicability, we used it to predict maize leaf ad/abaxial genes. Expression data and ChIP-seq data were obtained from Zhou et al. (2020) and Tu et al. (2020). The same analysis steps were followed, using the B73 v4 reference genome, introducing 471 continuous features (expression) and 104 binary features (ChIP-seq). In line with the results obtained for *Arabidopsis*, the maize model which included the ChIP-seq data had a better performance (AUC increase of 0.06) compared to the model which used only gene expression data. When the model was applied to the maize genome, 2 830 genes were predicted to be involved in leaf ad/abaxial polarity with a probability larger than 0.9. Out of these, 93 genes had a predicted probability greater than 0.99 (Table S9), which constitute our final prediction for maize.

4. Discussion

Machine learning can predict faster and find more genes related to specific biological processes than traditional experimental methods. Importantly, the quality of ML-based predictions of gene function will depend on the quantity, the quality and the type of input features used. By comparing the results using only expression data based on specific leaf domains to those obtained using a large amount of data without leaf polarity bias, we discovered that the amount of data appeared to be an important determinant of the model's accuracy. When using the smaller set of 21 expression features, even though these are related to leaf polarity, predictions were less accurate. Although the standard deviation of the prediction performance in all the models was large, this is not unexpected given the small ground truth set. One concern could be that our random selection of negative genes (genes not involved in ad–abaxial specialization) could introduce bias to the results. To test this, we reselected CLASS non-ad/ab genes and followed the same procedure as with the original prediction. The predicted probability for genes found in these three tests showed a high Pearson correlation ($r = 0.85^{**}$, Fig. 5, D), demonstrating the robustness of our approach.

In addition, when TF binding and histone modification data were incorporated, the model's accuracy surpassed that of the model built exclusively on gene expression data. Many studies use gene-related features to predict gene functions, such as gene expression, GO terms, cis-regulatory elements, etc. (Hansen et al., 2018; Ng et al., 2022). However, our research highlights how in addition to these datasources, also ChIP-seq data can be used to improve the predictive ability of models.

By performing a feature significance analysis, 20 features with the high contribution to the model were discovered. The TF CLF, ranking fourth on the list of important features, was revealed to be critical in controlling the growth and development of leaves since deletion of this gene results in narrow leaf blades that curl upward along the longitudinal axis (Krizek et al., 2006). The TF KAN1 which has a well-known role in promoting abaxial organ identity was also in this list of important features. The partial dependency plot of TFs displays a positive relationship between the feature value (0/1) and the probability of predicting a gene as involved in ad/abaxial polarity. This is in line with the idea that targets of these TFs are likely candidates for being involved in ad–abaxial formation. For the expression features, the significant features often relate to embryo and meristem stage. Many genes, like *PIN-FORMED1 (PIN1)*, influence the regulation of leaf ad–abaxial polarity during embryogenesis (Izhaki and Bowman, 2007). These findings provide evidence that the cotyledons in embryos already exhibit ad/abaxial patterning.

In order to acquire DEGs on both sides of the leaf, we used Svozil's approach (Svozil et al., 2016) to separate the adaxial side of the leaf from the adaxial side and isolate mRNA from those cell populations. PCA analysis of the four samples revealed clear separation between adaxial and abaxial samples. The PCA results shows that the first principal component can explain 47.5% variance of the data which demonstrated a good difference between adaxial and abaxial tissues. However, there is also a great difference between the replicates of samples along the second principal component (35.7%). We propose that this is due to the fact that tissues from the same plant bring correlation between the ad–abaxial tissues.

Both machine learning and differential gene expression analysis resulted in candidate genes that highlight the critical role ribosomes have in ad–abaxial leaf polarity patterning in *Arabidopsis*. There have been studies on how ribosomes affect leaf polarity. Yao et al. (2008) hypothesized that ribosomes may promote the HD-ZIP III mediated pathway in the adaxial domain of leaves or genetically repress *ARF3/4, KAN* or their downstream genes (Yao et al., 2008). By creating mutants, 13 ribosomal genes were discovered that have various effects on leaf growth and development (Horiguchi et al., 2011); all but two of these genes, which are known to be involved in leaf polarity, are included in our preGenes set.

Differential gene expression analysis results between adaxial and abaxial leaf primordia were also used to validate our predictions (Tian et al., 2019). We discovered a significant intersection between our preGenes and the differentially expressed genes from leaf primordia. However, there were still many genes that were only identified as DEGs in either mature leaves of our study or in the existing data of leaf primordia. We thus set out to analyze the intersection of DEGs of mature and primordial leaves with the known 61 ad–abaxial patterning genes identified in *Arabidopsis*. We discovered that only nine CLASS ad/ab genes were present in the DEGs between adaxial and abaxial leaf primordia whereas five were present in the DEGs between adaxial and abaxial expanded leaves (Fig. 5, E). This demonstrates that some genes involved in the regulation of leaf ad–abaxial polarity are not significantly differentially expressed between the adaxial and abaxial leaf primordia, suggesting additional regulatory modes like post-translational control may also influence ad/abaxial patterning in these cases.

The single-cell spatial transcriptome profile of four cell subtypes in *Arabidopsis* leaves revealed that the CLASS ad/ab genes had different expression preferences. It also indicated that genes involved in the regulation of leaf ad/abaxial polarity are not only present in epidermal cells, but also in mesophyll cells and are highly expressed in these cell types. These results provide insight into the expression patterns of ad/abaxial polarity genes between different cell types. Analysis of the expression patterns of the selected genes and known CLASS ad/ab genes in single cells revealed that these genes exhibited various expression profiles among the four cell types. In particular, no clear separation of these genes into distinct groups was observed when clustered based on the expression of the predicted genes and previously experimentally-validated CLASS ad/ab genes. Furthermore, we analyzed preferential expression patterns in four leaf cell types between *Arabidopsis* and *B. rapa* using our statistical approach. In both species, the majority of genes showed highest expression in lower epidermal cells, followed by upper epidermal and then mesophyll cells. While the overall trends were similar, enrichment was observed for different numbers of genes in each cell type between species. This demonstrates both conserved and divergent mechanisms controlling ad/abaxial identity at cellular resolution in *Arabidopsis* and *B. rapa* leaves. Together, these results provide new insight into the transcriptional programs establishing leaf polarity across species. Although further study is still needed to determine the specific developmental functions of the putatively novel *B. rapa* polarity genes identified, the conserved polar expression signatures suggest that many uncharacterized orthologs likely possess polarity-associated activities worth exploring. This also allowed us to validate our computationally-predicted gene list across species and showed that our model facilitates targeted homolog gene discovery in other crops. Our integrative *in silico* prediction and experimental gene discovery framework have expanded the knowledge of the genetic components underlying leaf ad–abaxial polarity specification.

In addition to the study in *Arabidopsis*, we also predicted genes associated with leaf ad/abaxial polarity in maize using the same strategy. Since there are relatively few experimentally-validated genes available in maize, we used the genes annotated as leaf polarity-related in the Gene Ontology as input. Importantly, similar to what was observed in *Arabidopsis*, the results again illustrate that using ChIP-seq data as input helps to improve prediction accuracy. Based on the model prediction, we discovered 93 novel leaf ad/abaxial polarity maize candidate genes. This further exemplifies the adaptability of the modeling strategy described in this paper as well as the scope for its application in the prediction of candidate genes for phenotypes of interest in additional distant species.

5. Conclusions

In summary, we have substantiated a set of 111 genes likely contributing to the establishment of ad–abaxial leaf polarity through an integrated *in silico* prediction using ML and an experimental validation approach. By comparing the prediction accuracy of different versions of the ML model, we found that increasing the quantity of transcriptome data and application of ChIP-seq data can significantly improve model accuracy. Additionally, preferential expression of predicted genes in *Arabidopsis* and *B. rapa* were identified by analyzing single-cell RNA-seq data; these findings expand the molecular understanding of leaf polarity specification in crop plants.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (Grant No. 2022YFF1003003), the Central Public-interest Scientific Institution Basal Research Fund (Grant No. Y2023PT16) and the Agricultural Science and Technology Innovation Program (ASTIP). Wei Sun is supported by China Scholarship Council (Grant No. 202103250097).

Supplementary materials

Supplementary data to this article can be found online at https://doi.org/10.1016/j.hpj.2024.06.002.

REFERENCES

Acharjee, A., Kloosterman, B., de Vos, R.C.H., Werij, J.S., Bachem, C.W.B., Visser, R.G.F., Maliepaard, C., 2011. Data integration and network reconstruction with omics data using Random Forest regression in potato. Anal Chim Acta, 705: 56–63.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30: 2114–2120.

Braybrook, S.A., Kuhlemeier, C., 2010. How a plant builds leaves. Plant Cell, 22: 1006–1018.

Cantelli, G., Bateman, A., Brooksbank, C., Petrov, A.I., Malik-Sheriff, R.S., Ide-Smith, M., Hermjakob, H., Flicek, P., Apweiler, R., Birney, E., McEntyre, J., 2022. The European bioinformatics institute (EMBL-EBI) in 2021. Nucleic Acids Res, 50: D11–D19.

Cheng, F., Sun, R., Hou, X., Zheng, H., Zhang, F., Zhang, Y., Liu, B., Liang, J., Zhuang, M., Liu, Yunxia, Liu, D., Wang, X.B., Li, P., Liu, Y.M., Lin, K., Bucher, J., Zhang, N., Wang, Y., Wang, H., Deng, J., Liao, Y., Wei, K., Zhang, X., Fu, L., Hu, Y., Liu, J.S., Cai, C., Zhang, S.J., Zhang, S.F., Li, F., Zhang, H., Zhang, J., Guo, N., Liu, Z., Liu, J., Sun, C., Ma, Y., Zhang, H.J., Cui, Y., Freeling, M.R., Borm, T., Bonnema, G., Wu, J., Wang, X.W., 2016. Subgenome parallel selection is associated with morphotype diversification and convergent crop domestication in *Brassica rapa* and *Brassica oleracea*. Nat Genet, 48: 1218–1224.

Depuydt, T., De Rybel, B., Vandepoele, K., 2022. Charting plant gene functions in the multi-omics and single-cell era. Trends Plant Sci, 28: 283–296.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29: 15–21.

Eshed, Y., Izhaki, A., Baum, S.F., Floyd, S.K., Bowman, J.L., 2004. Asymmetric leaf development and blade expansion in *Arabidopsis* are mediated by KANADI and YABBY activities. Development, 131: 2997–3006.

Gao, Y., Lu, Y., Li, X., Li, N., Zhang, X., Su, X., Feng, D., Liu, M., Xuan, S., Gu, A., Wang, Y., Chen, X., Zhao, J., Shen, S., 2020. Development and application of SSR markers related to genes involved in leaf adaxial-abaxial polarity establishment in Chinese cabbage (*Brassica rapa* L. ssp. *pekinensis*). Front Genet, 11: 773.

Greenwell, B.M., 2017. pdp: an R rackage for constructing partial dependence Plots. R J, 9: 421.

Guo, X., Liang, J., Lin, R., Zhang, L., Zhang, Z., Wu, J., Wang, X., 2022. Single-cell transcriptome reveals differentiation between adaxial and abaxial mesophyll cells in *Brassica rapa*. Plant Biotechnol J, 20: 2233–2235.

Hammal, F., de Langen, P., Bergon, A., Lopez, F., Ballester, B., 2022. ReMap 2022: a database of human, mouse, drosophila and *Arabidopsis* regulatory regions from an integrative analysis of DNA-binding sequencing experiments. Nucleic Acids Res, 50: D316–D325.

Hansen, B.O., Meyer, E.H., Ferrari, C., Vaid, N., Movahedi, S., Vandepoele, K., Nikoloski, Z., Mutwil, M., 2018. Ensemble gene function prediction database reveals genes important for complex I formation in *Arabidopsis thaliana*. New Phytol, 217: 1521–1534.

Horiguchi, G., Mollá-Morales, A., Pérez-Pérez, J.M., Kojima, K., Robles, P., Ponce, M.R., Micol, J.L., Tsukaya, H., 2011. Differential contributions of ribosomal protein genes to *Arabidopsis thaliana* leaf development: a comparative study of r-protein mutants. Plant J, 65: 724–736.

Huang, W., Pi, L., Liang, W., Xu, B., Wang, H., Cai, R., Huang, H., 2006. The proteolytic function of the *Arabidopsis* 26S proteasome is required for specifying leaf adaxial identity. Plant Cell, 18: 2479–2492.

Iwakawa, H., Iwasaki, M., Kojima, S., Ueno, Y., Soma, T., Tanaka, H., Semiarti, E., Machida, Y., Machida, C., 2007. Expression of the *ASYMMETRIC LEAVES2* gene in the adaxial domain of *Arabidopsis* leaves represses cell proliferation in this domain and is critical for the development of properly expanded leaves: AS2 represses adaxial cell proliferation of leaves. Plant J, 51: 173–184.

Izhaki, A., Bowman, J.L., 2007. KANADI and Class III HD-Zip gene families regulate embryo patterning and modulate auxin flow during embryogenesis in *Arabidopsis*. Plant Cell, 19: 495–508.

Juarez, M.T., Twigg, R.W., Timmermans, M.C.P., 2004. Specification of adaxial cell fate during maize leaf development. Development, 131: 4533–4544.

Kerstetter, R.A., Bollman, K., Taylor, R.A., Bomblies, K., Poethig, R.S., 2001. KANADI regulates organ polarity in *Arabidopsis*. Nature, 411: 706–709.

Krizek, B.A., Lewis, M.W., Fletcher, J.C., 2006. *RABBIT EARS* is a second-whorl repressor of *AGAMOUS* that maintains spatial boundaries in *Arabidopsis* flowers. Plant J, 45: 369–383.

Liao, Y., Smyth, G.K., Shi, W., 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30: 923–930.

Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R News, 2: 18–22.

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol, 15: 1–21.

Machida, C., Nakagawa, A., Kojima, S., Takahashi, H., Machida, Y., 2015. The complex of ASYMMETRIC LEAVES (AS) proteins plays a central role in antagonistic interactions of genes for leaf polarity specification in *Arabidopsis*. WIREs Dev Biol, 4: 655–671.

Maugarny-Calès, A., Laufs, P., 2018. Getting leaves into shape: a molecular, cellular, environmental and evolutionary view. Development, 145: dev161646.

McConnell, J.R., Emery, J., Eshed, Y., Bao, N., Bowman, J., Barton, M.K., 2001. Role of PHABULOSA and PHAVOLUTA in determining radial patterning in shoots. Nature, 411: 709–713.

Ng, J.W.X., Chua, S.K., Mutwil, M., 2022. Feature importance network reveals novel functional relationships between biological features in *Arabidopsis thaliana*. Front Plant Sci, 13: 944992.

Pekker, I., Alvarez, J.P., Eshed, Y., 2005. Auxin response factors mediate *Arabidopsis* organ asymmetry via modulation of KANADI activity. Plant Cell, 17: 2899–2910.

Reel, P.S., Reel, S., Pearson, E., Trucco, E., Jefferson, E., 2021. Using machine learning approaches for multi-omics data analysis: a review. Biotechnol Adv, 49: 107739.

Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T., 2005. ROCR: visualizing classifier performance in R. Bioinformatics, 21: 3940–3941.

Svozil, J., Gruissem, W., Baerenfaller, K., 2016. Meselect – a rapid and effective method for the separation of the main leaf tissue types. Front Plant Sci, 7: 220710.

Tian, C., Wang, Y., Yu, H., He, J., Wang, J., Shi, B., Du, Q., Provart, N.J., Meyerowitz, E.M., Jiao, Y., 2019. A gene expression map of shoot domains reveals regulatory mechanisms. Nat Commun, 10: 141.

Tu, X., Mejía-Guerra, M.K., Valdes Franco, J.A., Tzeng, D., Chu, P.Y., Shen, W., Wei, Y., Dai, X., Li, P., Buckler, E.S., Zhong, S., 2020. Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors. Nat Commun, 11: 5089.

van Dijk, A.D.J., Kootstra, G., Kruijer, W., de Ridder, D., 2021. Machine learning in plant science and plant breeding. iScience, 24: 101890.

Woodhouse, M.R., Cannon, E.K., Portwood, J.L., Harper, L.C., Gardiner, J.M., Schaeffer, M.L., Andorf, C.M., 2021. A pan-genomic approach to genome databases using maize as a model system. BMC Plant Biol, 21: 1–10.

Xia, K., Sun, H.X., Li, Jie, Li, Jiming, Zhao, Y., Chen, L., Qin, C., Chen, R., Chen, Z., Liu, G., Yin, R., Mu, B., Wang, X., Xu, M., Li, X., Yuan, P., Qiao, Y., Hao, S., Wang, Jing, Xie, Q., Xu, J., Liu, S., Li, Y., Chen, A., Liu, L., Yin, Y., Yang, H., Wang, Jian, Gu, Y., Xu, X., 2022. The single-cell stereo-seq reveals region-specific cell subtypes and transcriptome profiling in *Arabidopsis* leaves. Dev Cell, 57: 1299–1310.

Yamaguchi, T., Nukazuka, A., Tsukaya, H., 2012. Leaf adaxial-abaxial polarity specification and lamina outgrowth: evolution and development. Plant Cell Physiol, 53: 1180–1194.

Yao, Y., Ling, Q., Wang, H., Huang, H., 2008. Ribosomal proteins promote leaf adaxial identity. Development, 135: 1325–1334.

Yu, C., Yan, C., Liu, Yuling, Liu, Yali, Jia, Y., Lavelle, D., An, G., Zhang, W., Zhang, L., Han, R., Larkin, R.M., Chen, J., Michelmore, R.W., Kuang, H., 2020. Upregulation of a *KN1* homolog by transposon insertion promotes leafy head development in lettuce. Proc Natl Acad Sci, 117: 33668–33678.

Yu, G., Wang, L.G., He, Q.Y., 2015. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics, 31: 2382–2383.

Zhou, P., Li, Z., Magnusson, E., Gomez Cano, F., Crisp, P.A., Noshay, J.M., Grotewold, E., Hirsch, C.N., Briggs, S.P., Springer, N.M., 2020. Meta gene regulatory networks in maize highlight functionally relevant regulatory interactions. Plant Cell, 32: 1377–1396.