https://doi.org/10.1093/plphys/kiad332 PLANT PHYSIOLOGY 2023: 193: 190–194

Plant Physiology®

Letter

RiceTFtarget: A rice transcription factor–target prediction server based on coexpression and machine learning

Baoyi Zhang,[1] Xueai Zhu,[1] Zixin Chen,[2] Hongsheng Zhang,[1] Junxian Huang[2,*] and Ji Huang[1,3,*]

1 State Key Laboratory of Crop Genetics & Germplasm Enhancement and Utilization, Jiangsu Province Engineering Research Center of Seed Industry Science and Technology, Nanjing Agricultural University, Nanjing 210095, China
2 College of Artificial Intelligence, Nanjing Agricultural University, Nanjing 210095, China
3 Jiangsu Key Laboratory for Information Agriculture, Nanjing Agricultural University, Nanjing 210095, China

*Author for correspondence: huangji@njau.edu.cn (Ji.H.), jim@njau.edu.cn (J.H.)

Dear editor,

Transcription factors (TFs), also known as trans-acting factors, usually recognize the DNA *cis*-regulatory elements in the promoter regions of target genes to activate or repress expression (Mitchell and Tjian 1989). Identifying target genes of TFs or the TFs binding to target genes is crucial to address the biological functions and regulatory networks of these TF–target modules. However, the TF–target interaction is time-consuming and laborious. Here, we constructed RiceTFtarget (https://cbi.njau.edu.cn/RiceTFtarget/), a website for robustly predicting TF–target pairs based on coexpression, pattern matching, and machine learning (Fig. 1A). Although some tools can be used for predicting *cis*-regulatory elements (Heinz et al. 2010; Grant et al. 2011; Mathelier et al. 2016), RiceTFtarget is a tool for retrieving specific TF–target interactions in plants.

RiceTFtarget provides a user-friendly and convenient operation interface and includes 6 main functions (Fig. 1B). The core function is mainly used to query which target genes can be regulated by inputting the gene ID of interested TFs or query which TFs can regulate the target gene of interest. The result comprises the coexpression correlation coefficient (Pearson correlation coefficient [PCC]), binding site information, and machine learning prediction results of binding sites (Fig. 1C), which can be further expanded (Fig. 1D). Also included is the entry for constructing the regulatory network diagram of TFs/targets by top PCCs (Fig. 1D). The other functions are detailed in Supplemental Text S1.

Coexpression and *cis*-regulatory element pattern matching

We collected 149 rice (*Oryza sativa* L.) RNA-seq datasets (Supplemental Table S1) without biological replicates from the NCBI GEO database (Barrett et al. 2013) to host a robust coexpression relationship for any TF–target pair by Pearson's correlation (ρ) (Supplemental Text S2). Many TF binding specificities have been summarized as position weight matrices (PWM), also known as TF motifs (Leporcq et al. 2020). 1,862 TF genes were obtained from PlantTFDB (Jin et al. 2017), 269 of which are manually curated, nonredundancy and high-quality TF binding motifs (Supplemental Table S2). High-quality TF motifs were scanned for rice gene promoters (<2 kb) to obtain putative transcription factor binding sites (TFBSs) by the default FIMO parameters (Grant et al. 2011). The putative TFBSs for each TF were obtained based on the sequence similarity analysis among TFs (Fig. 1A and Supplemental Text S3).

Construction of a single-TF model

To build rice TFBS machine learning prediction models and verify the accuracy of RiceTFtarget, we downloaded the 53 ChIP/DAP-seq data of 13 TFs in rice from the NCBI GEO database (Supplemental Table S3) and obtained the corresponding motifs (Supplemental Fig. S1). Subsequently, element scanning on gene promoters identified putative TFBSs of each TF by FIMO (Grant et al. 2011). Thirty TFBS features

Received February 24, 2023. Accepted May 26, 2023. Advance access publication June 9, 2023

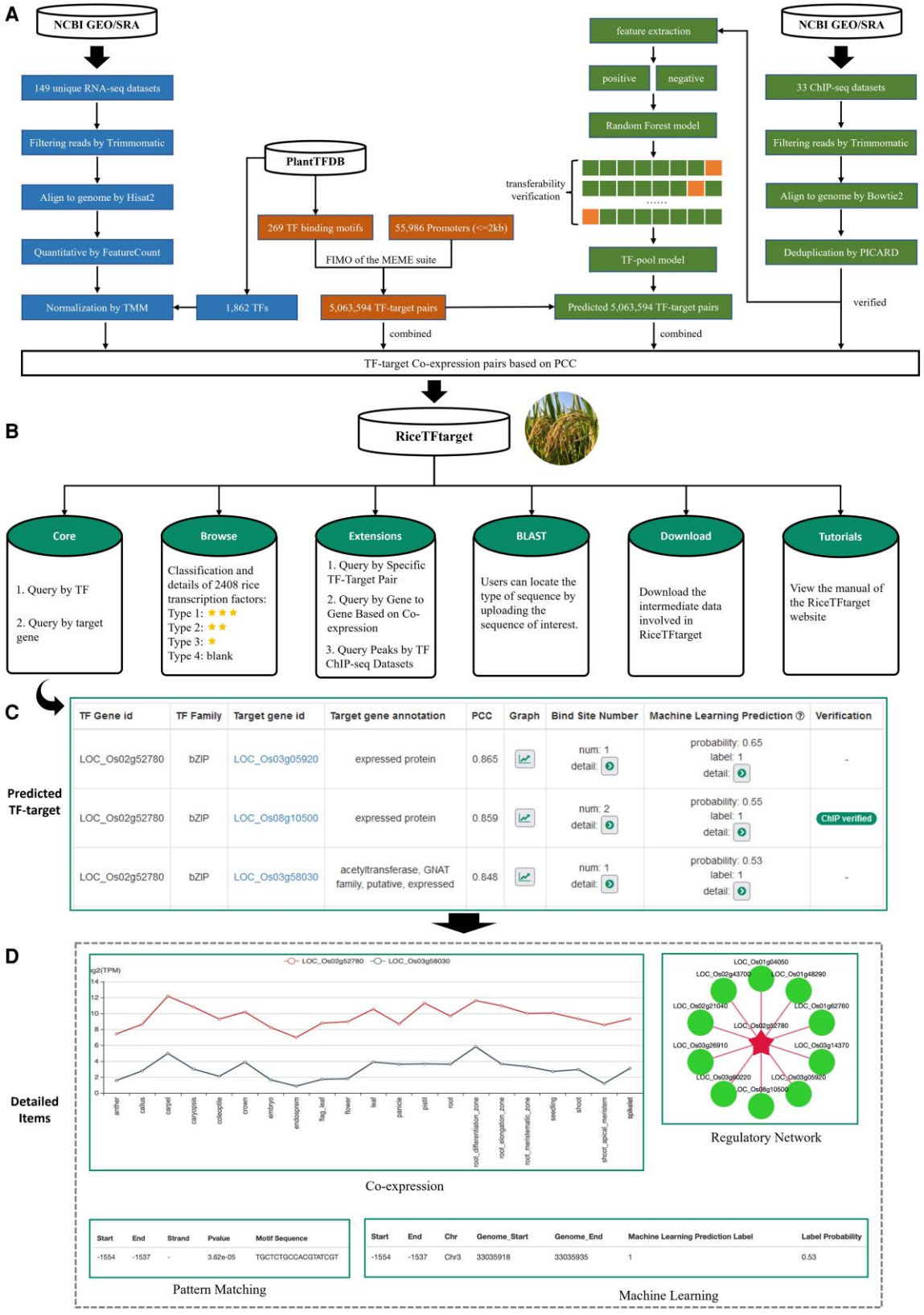

Figure 1. Overview of RiceTFtarget pipeline, structure, functions, and applications. **A**) RiceTFtarget analysis pipeline. **B)** RiceTFtarget website structure. **C)** RiceTFtarget functions. **D)** Detailed items in RiceTFtarget prediction results, including coexpression, pattern matching, machine learning, and regulatory network.

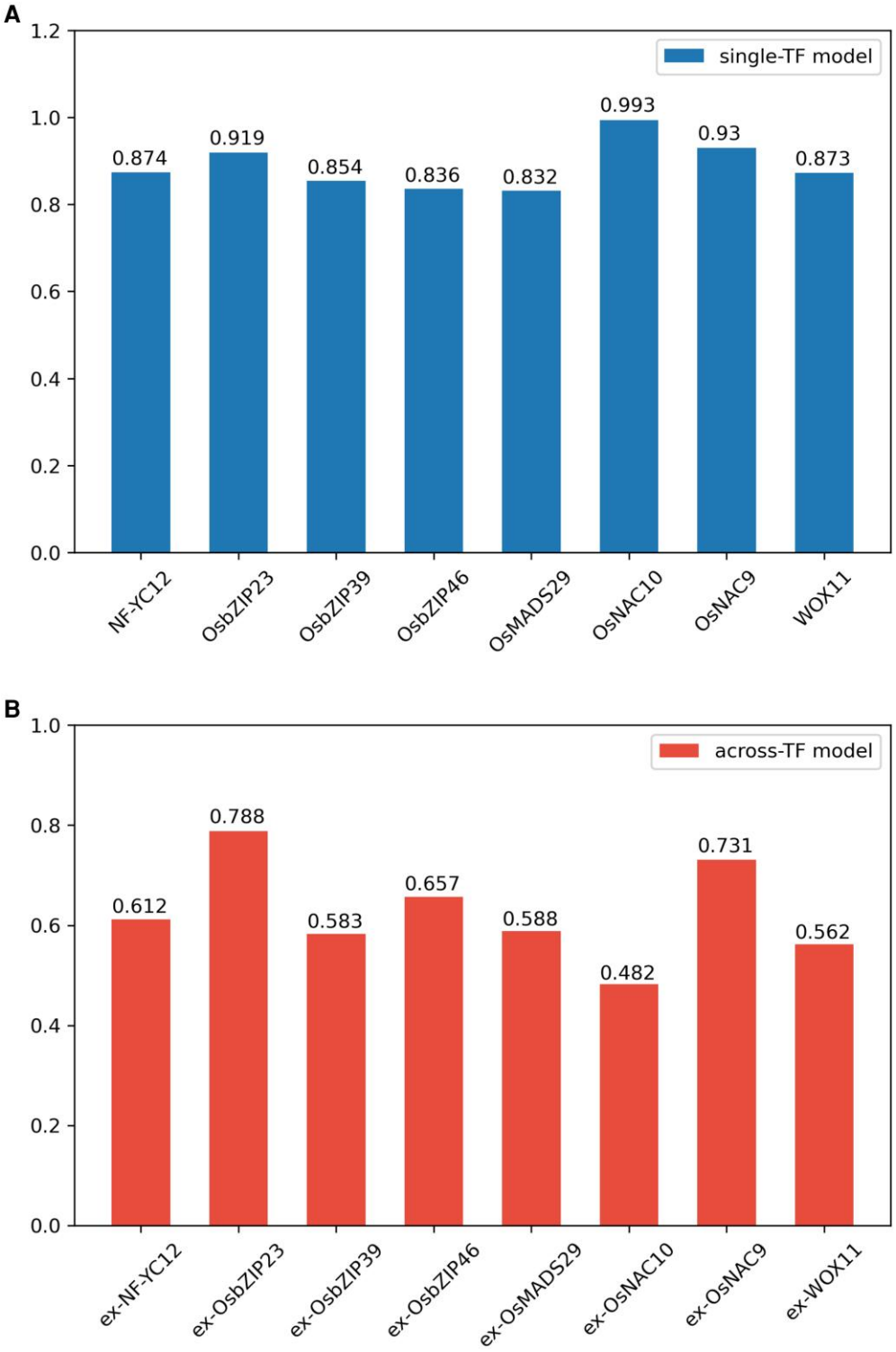

Figure 2. The performances of the 8 single/across-TF models based on 46 TFBS features and random forest algorithm. The *y* axis is the AUC value. **A)** The performances of single-TF models. Training sets: The genome location and 46 features of the peaks obtained from the analysis of the corresponding TF ChIP-seq data. **B)** The performances of across-TF models. Training sets: The genome location and 46 features of the peaks obtained from all but 1 of the TFs of the ChIP-seq data (7 of 8 TFs). Taking NF-YC12 as an example, the ChIP-seq data of 7 TFs, OsbZIP23, OsbZIP39, OsbZIP46, OsMADS29, OsNAC10, OsNAC9, and WOX11, are combined into an integrative dataset to train the ex-NF-YC12 model. The prediction accuracy of the model is evaluated by ChIP-seq data of NF-YC12. The prefix "ex-" of the TF label on the *x* axis means "excluding".

were integrated into the prediction model (Supplemental Table S4). We further implemented a supervised machine learning strategy to construct 8 single-TF prediction models (Supplemental Fig. S2). The remaining 5 TFs were filtered out due to the lack of training set data (Supplemental Table S5). The detailed analysis pipeline is shown in Supplemental Text S4.

Different machine learning algorithms exhibit different performances on the same training data. We selected 8 machine learning algorithms, including 4 traditional algorithms (SVM, logistic regression [LR], decision tree [DR], and K-nearest neighbor [KNN]) and 4 ensemble learning classification algorithms (XGBoost, LightGBM, CatBoost, and random forest [RF]), to construct TFBS prediction models with 8 different TFs based on the 30 features of TFBS. The AUC value was utilized to evaluate the performance of the model. The results showed that RF, CatBoost, LightGBM, and XGBoost performances were similar, with an average AUC of 0.873, 0.854, 0.854, and 0.861, respectively (Supplemental Table S6). The performance of RF was slightly better, with an average AUC of 0.873 (0.806∼0.993) (Supplemental Fig. S3 and Table S6). Previous studies have shown that RF showed promising performance in TFBS prediction (Khamis et al. 2018). The comparison indicates that ensemble learning (RF, CatBoost, LightGBM, and XGBoost) is better than traditional algorithms (LR, DR, SVM, and KNN) in predicting TFBS (Supplemental Fig. S3). The TFBS is usually 6 to 20 bp long (Zeng et al. 2020). The dependence between sequence bases may be important evidence for determining TFBS. To verify this hypothesis, we added the *k*-mer (length-*k* substrings of a sequence) of TFBS to the 30 features of the RF model. Considering the performance and running speed, we determined the 2-mer on the 200-bp sequence of TFBS as the *k*-mer feature of the model (Supplemental Fig. S4). Compared with previous reports, the performance of most models was improved after adding *k*-mer (Supplemental Fig. S5). We finally applied 46 ($30 + 4^2$) TFBS features to build 8 single-TFBS prediction models based on RF, and the AUC was 0.832∼0.993 (Fig. 2A).

Construction of across-TF model for all TFs

Since a single-TF prediction model is only applicable to the specific TF, the territory of application is relatively narrow. Therefore, we expect to build a model that can be applied to the TFBS prediction of all TFs. We integrated the training data with all but 1 of the TFs of the dataset (7 of 8 TFs). The TF set aside was then used to evaluate the performance. We applied this strategy for each of the selected 8 TFs and compared the performance of the across-TF model and single-TF model for each TF (Fig. 2B). The prediction performance of all across-TF models was relatively lower than that of single-TF models, possibly due to the specificity between different TFs. Finally, we built an optimal across-TF model with an AUC of 0.788 that can be effectively applied to predicting the binding sites of all rice TFs.

RiceTFtarget performance

We used OsbZIP23 (bZIP transcription factor 23) to verify the prediction accuracy of RiceTFtarget. RiceTFtarget predicted 41 targets (coexpression PCC > 0.6 and the machine learning prediction label is 1). Among 41 targets, 34 could be verified by ChIP-seq data (Supplemental Table S7 and Table S8) and the remaining 7 also might be the real targets missed by ChIP-seq. Moreover, RiceTFtarget successfully predicted the experimentally validated TF–target pairs, e.g. OsbZIP62-Osprx97 (Yang et al. 2019), OsMADS6-OsFDML1 (Tao et al. 2018), OsBZR1-OsPUB43 (Wu et al. 2022), and OsMYB30-Os4CL5 (Li et al. 2020), indicating that the TF–target pairs retrieved by RiceTFtarget are reliable and robust. The PCC and machine learning parameters can be flexibly adjusted to obtain more sensitive or accurate results predicted by RiceTFtarget. The ChIP-seq data for a TF or RNA-seq data for a TF mutant can also be compared to RiceTFtarget predictions to promote prediction accuracy. Compared to PlantRegMap (Tian et al. 2020), RiceTFtarget provides a more comprehensive search engine that not only includes TF binding motif scanning but also provides coexpression correlations of TF–targets and TFBS predicted by machine learning to obtain more reliable predictions of TF–target pairs. Additionally, RiceTFtarget implements bilateral search, which can predict the TFs binding to a gene of interest and target genes bound by a TF.

Overall, RiceTFtarget is a robust webserver for identifying TF–target pairs, which may substantially accelerate the study of the biological roles of TFs and TF regulatory networks in rice.

Acknowledgments

We are grateful for the public resources and their authors for providing rice RNA-seq and ChIP-seq libraries used in RiceTFtarget. We also thank Professor Yufeng Wu from the Center for Bioinformatics and the Information Center of Nanjing Agricultural University for their assistance in data storage and web maintenance.

Author contributions

J.H. and J.X.H. conceived the original research plan; B.Z. collected, processed, and analyzed RNA-seq and ChIP-seq data; B.Z. and X.Z. built the database and website; B.Z. and Z.C. structured the TFBS machine learning model; B.Z. and J.H. wrote the manuscript; J.H., J.X.H., and H.Z. supervised this work.

Supplemental data

The following materials are available in the online version of this article.

Supplemental Figure S1. The 11 corresponding motifs with the lowest *P*-value predicted by meme-chip.

Supplemental Figure S2. Pipeline of building a single-TF prediction model (taking OsbZIP23 as an example).

Supplemental Figure S3. Average performance of 8 different machine learning algorithms.

Supplemental Figure S4. Influence of different values of *k* and sequence length on model performance (taking OsbZIP23 as an example).

Supplemental Figure S5. Influence of k-mer on model performance.

Supplemental Table S1. Overall information of 149 unique RNA-seq samples.

Supplemental Table S2. 1,862 TF genes obtained from PlantTFDB, 269 of which are manually curated, nonredundancy and high-quality TF binding motifs.

Supplemental Table S3. Overall information of rice TF ChIP/DAP-seq experimental data.

Supplemental Table S4. 30 TFBS features used by RiceTFtarget to build the machine learning model.

Supplemental Table S5. Number of candidate TFBS for each TF model.

Supplemental Table S6. Prediction performance of 8 classification algorithms on 8 transcription factors, respectively.

Supplemental Table S7. ChIP-seq verification of OsbZIP23–target (coexpression PCC > 0.6 and the machine learning prediction label is 1)

Supplemental Table S8. Target genes of OsbZIP23 validated by ChIP-seq and RiceTFtarget.

Supplemental Text S1. Detailed functions of the RiceTFtarget website.

Supplemental Text S2. Detailed process of gene quantitative analysis and coexpression analysis.

Supplemental Text S3. Identification of candidate TFBS by pattern matching.

Supplemental Text S4. Feature selection and model construction of machine learning models.

Funding

This work was supported by the Jiangsu Provincial Seed Industry Revitalization Project ([2021]009), the Fundamental Research Funds for the Central Universities and Jiangsu Collaborative Innovation Center for Modern Crop Production, and the Cyrus Tang Crop Seed Innovation Center.

Conflict of interest statement. None declared.

Data availability

All data can be accessed at https://cbi.njau.edu.cn/RiceTFtarget/.

References

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al. NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res. 2013:**41**(D1):D991–D995. https://doi.org/10.1093/nar/gks1193

Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011:**27**(7):1017–1018. https://doi.org/10.1093/bioinformatics/btr064

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010:**38**(4):576–589. https://doi.org/10.1016/j.molcel.2010.05.004

Jin J, Tian F, Yang DC, Meng YQ, Kong L, Luo J, Gao G. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res. 2017:**45**(D1):D1040–D1045. https://doi.org/10.1093/nar/gkw982

Khamis AM, Motwalli O, Oliva R, Jankovic BR, Medvedeva YA, Ashoor H, Essack M, Gao X, Bajic VB. A novel method for improved accuracy of transcription factor binding site prediction. Nucleic Acids Res. 2018:**46**(12):e72. https://doi.org/10.1093/nar/gky237

Leporcq C, Spill Y, Balaramane D, Toussaint C, Weber M, Bardet AF. TFmotifView: a webserver for the visualization of transcription factor motifs in genomic regions. Nucleic Acids Res. 2020:**48**(W1):W208–W217. https://doi.org/10.1093/nar/gkaa252

Li W, Wang K, Chern M, Liu Y, Zhu Z, Liu J, Zhu X, Yin J, Ran L, Xiong J, et al. Sclerenchyma cell thickening through enhanced lignification induced by OsMYB30 prevents fungal penetration of rice leaves. New Phytol. 2020:**226**(6):1850–1863. https://doi.org/10.1111/nph.16505

Mathelier A, Fornes O, Arenillas DJ, Chen CY, Denay G, Lee J, Shi W, Shyr C, Tan G, Worsley-Hunt R, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2016:**44**(D1):D110–115. https://doi.org/10.1093/nar/gkv1176

Mitchell PJ, Tjian R. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. Science. 1989:**245**(4916):371–378. https://doi.org/10.1126/science.2667136

Tao J, Liang W, An G, Zhang D. OsMADS6 controls flower development by activating rice FACTOR OF DNA METHYLATION LIKE1. Plant Physiol. 2018:**177**(2):713–727. https://doi.org/10.1104/pp.18.00017

Tian F, Yang DC, Meng YQ, Jin J, Gao G. PlantRegMap: charting functional regulatory maps in plants. Nucleic Acids Res. 2020:**48**(D1):D1104–D1113. https://doi.org/10.1093/nar/gkz1020

Wu Q, Liu Y, Huang J. CRISPR-cas9 mediated mutation in OsPUB43 improves grain length and weight in rice by promoting cell proliferation in spikelet hull. Int J Mol Sci. 2022:**23**(4):2347. https://doi.org/10.3390/ijms23042347

Yang S, Xu K, Chen S, Li T, Xia H, Chen L, Liu H, Luo L. A stress-responsive bZIP transcription factor OsbZIP62 improves drought and oxidative tolerance in rice. BMC Plant Biol. 2019:**19**(1):260. https://doi.org/10.1186/s12870-019-1872-1

Zeng YQ, Gong MQ, Lin M, Gao DR, Zhang YQ. A review about transcription factor binding sites prediction based on deep learning. Ieee Access. 2020:**8**(12):219256–219274. https://doi.org/10.1109/ACCESS.2020.3042903