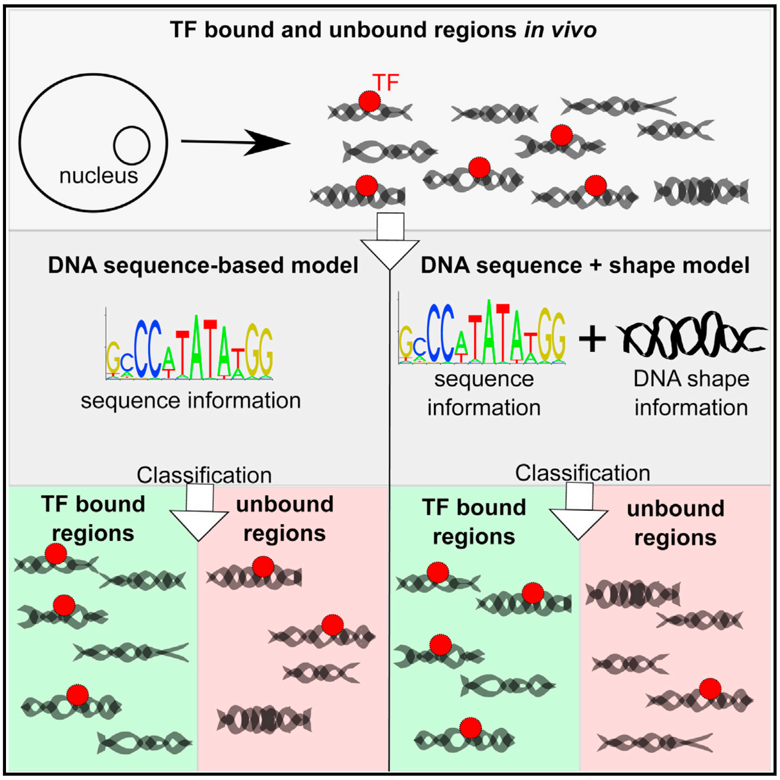
Cell Systems

Article

DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo

Graphical Abstract



Highlights

- Considering DNA shape features improved the prediction of TF binding in vivo
- DNA shape at flanking regions of binding sites refined the prediction of TF binding
- Larger improvements were observed for the E2F and MADS-domain TF families
- Propeller twist at specific nucleotide positions of the MADS-box contributed most

Authors

Anthony Mathelier, Beibei Xin, Tsu-Pei Chiu, Lin Yang, Remo Rohs, Wyeth W. Wasserman

Correspondence

rohs@usc.edu (R.R.),
wyeth@cmmt.ubc.ca (W.W.W.)

In Brief

The study confirms the importance of considering DNA shape features when modeling TF binding profiles in in vivo studies. For the available TF families, DNA shape features are most critical for the E2F and MADS-domain TF binding in a position-specific manner.

Mathelier et al., 2016, Cell Systems *3*, 278–286
September 28, 2016 © 2016 The Author(s). Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.cels.2016.07.001

Cell Systems, Volume *3*

Supplemental Information

DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo

Anthony Mathelier, Beibei Xin, Tsu-Pei Chiu, Lin Yang, Remo Rohs, and Wyeth W. Wasserman

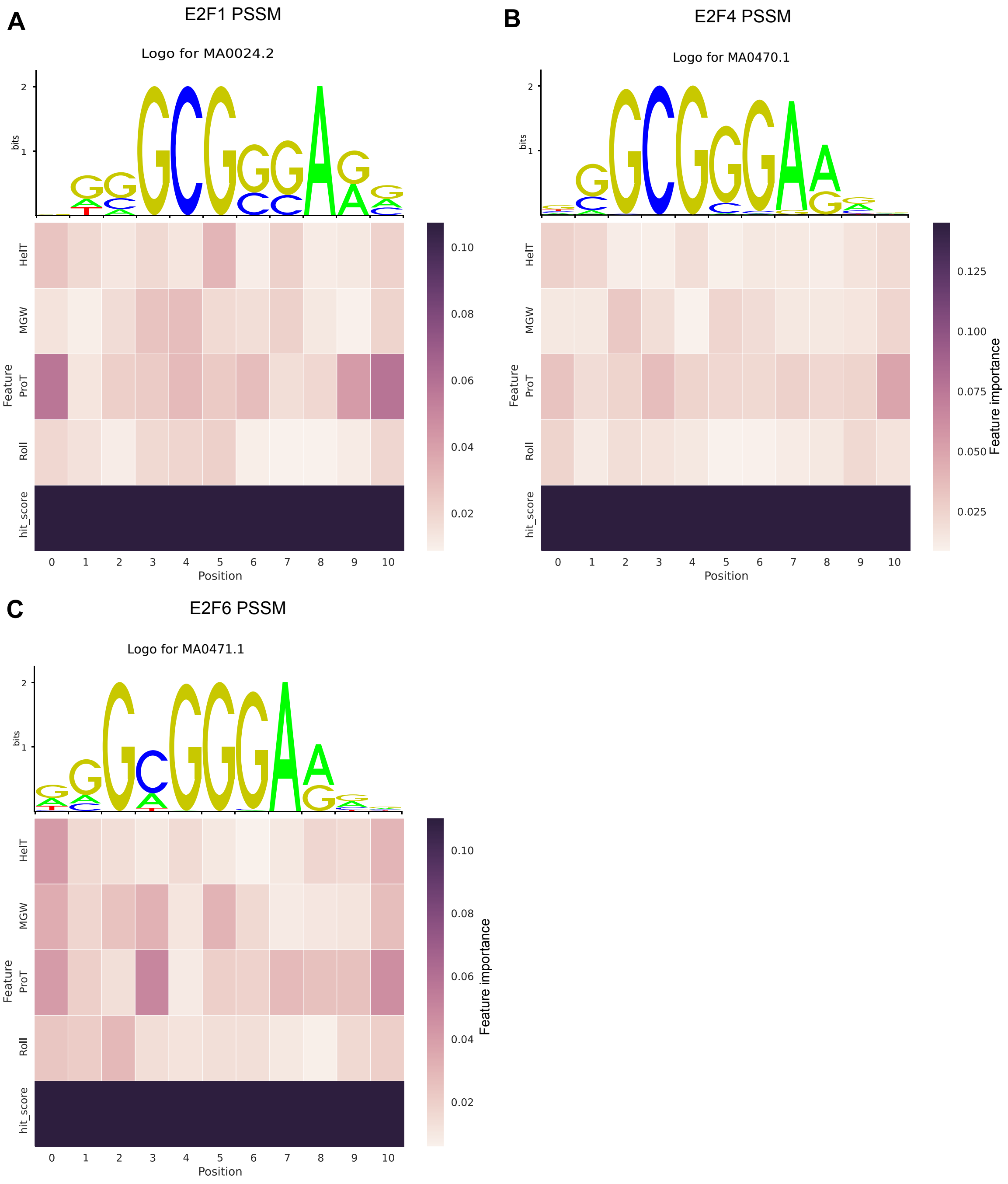


Figure S1 *Related to Figure 7. Feature importance measures for human E2F TFBS recognition in ChIP-seq. Weblogos of the E2F TF profiles for E2F1 (**A**), E2F4 (**B**), and E2F6 (**C**) from JASPAR are provided at the top of the panels. Heat maps providing the average level of feature importance (y-axis) at each position (x-axis) of the TFBSs in the PSSM + DNA shape classifiers are provided at the bottom of the panels. First-order DNA shape features have been considered in the classifiers. Note that the 'hit score' feature corresponds to the PSSM scores used in the classifiers. The 'hit score' feature spans all the positions for graphical representation but a single hit score is provided per vector in the classifiers.*

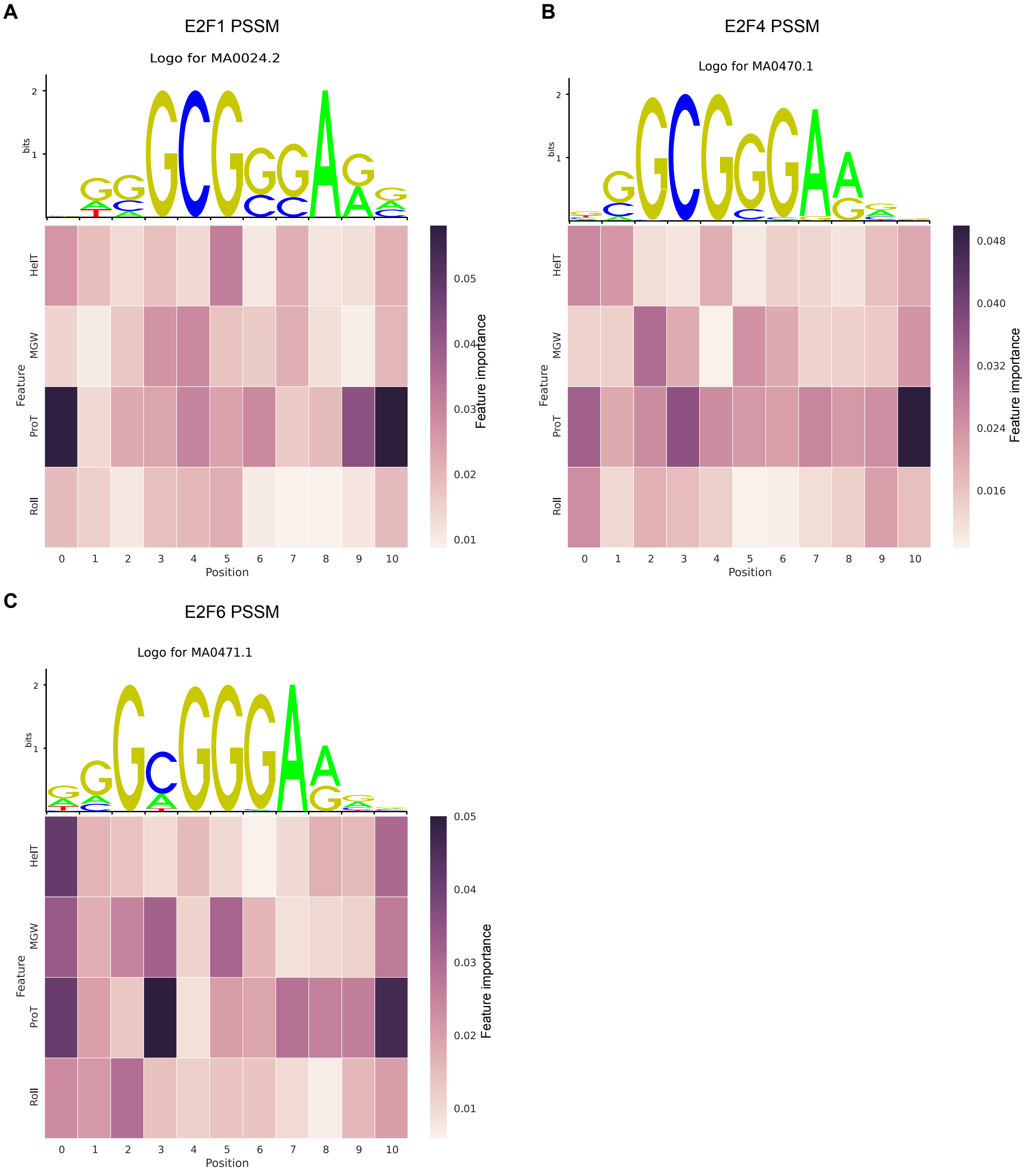


Figure S2 *Related to Figure 7. DNA shape only feature importance measures for human E2F TFBS recognition in ChIP-seq. Weblogos of the E2F TF profiles for E2F1 (**A**), E2F4 (**B**), and E2F6 (**C**) from JASPAR are provided at the top of the panels. Heat maps providing the average level of DNA shape feature importance (y-axis) at each position (x-axis) of the TFBSs in the PSSM + DNA shape classifiers are provided at the bottom of the panels. Note that the heat maps are zoomed in versions of the ones in Figure S1 when only DNA shape features are considered.*

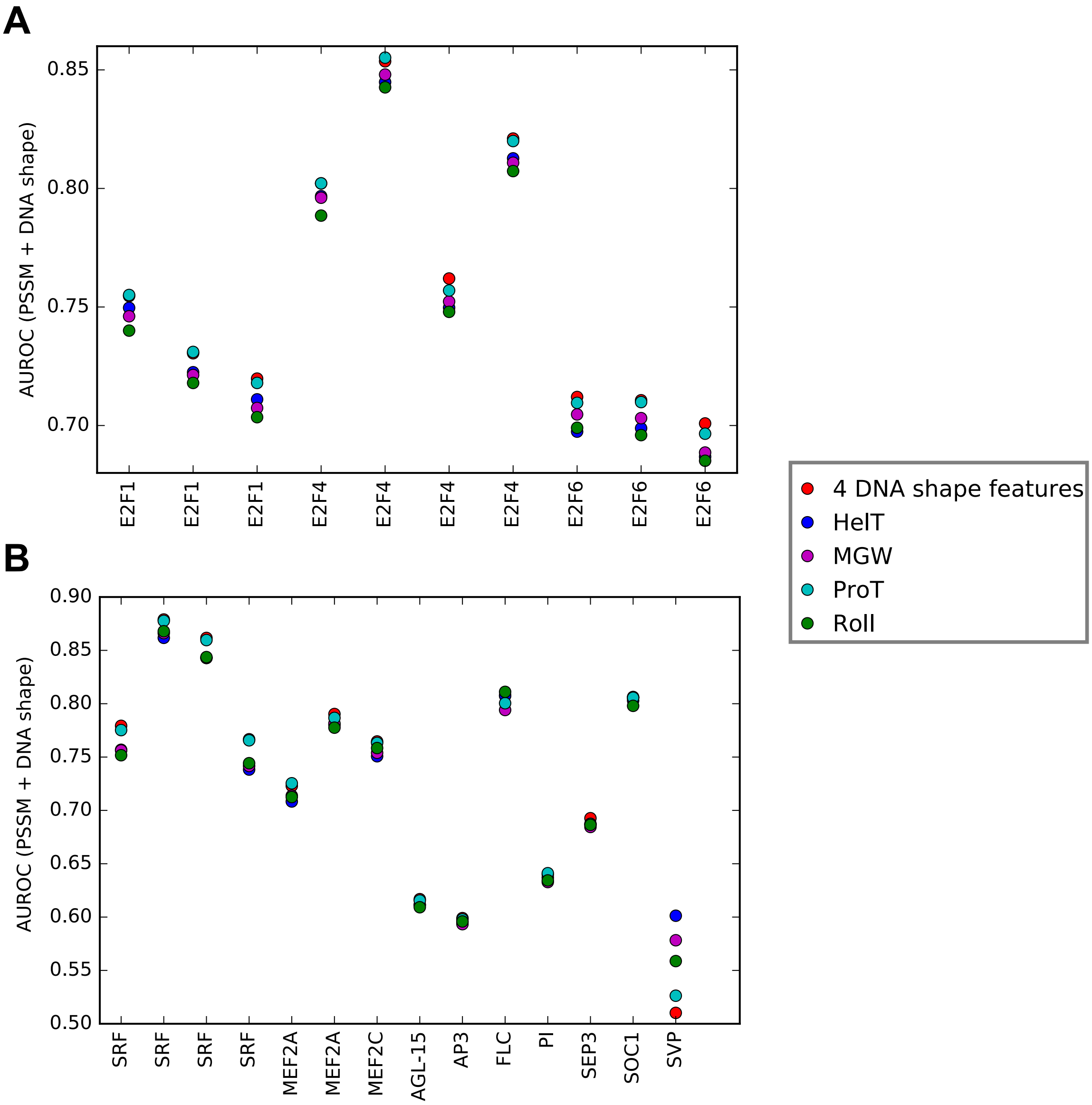


Figure S3 *Related to Figure 5. Using a single DNA shape feature for E2F and MADS-box TFBS recognition in ChIP-seq. Comparison of AUROC (y-axis) for the E2F (**A**) and the MADS-domain (**B**) TF data sets (x-axis) when using 4 DNA shape features or a single feature along with the PSSM scores in the PSSM + DNA shape classifiers.*

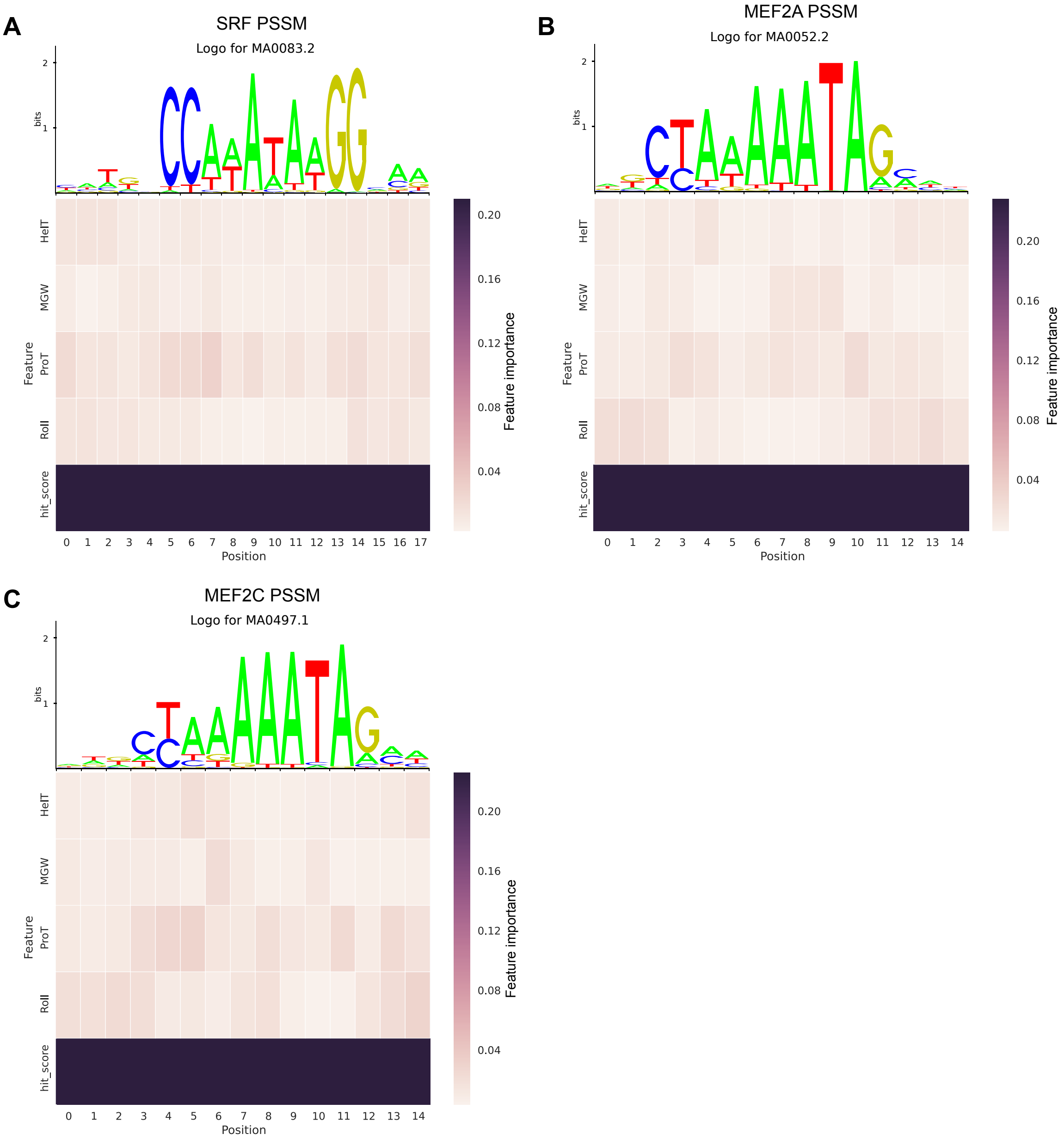


Figure S4 *Related to Figure 7. Feature importance measures for human MADS-box recognition in ChIP-seq. Weblogos of the MADS-domain TF profiles considered for SRF (**A**), MEF2A (**B**), and MEF2C (**C**) are provided at the top of the panels. Heat maps providing the average level of feature importance (y-axis) at each position (x-axis) of the TFBSs in the PSSM + DNA shape classifiers are provided at the bottom of the panels. Note that the 'hit score' feature corresponds to the PSSM scores used in the classifiers. The 'hit score' feature spans all the positions for graphical representation but a single hit score is provided per vector in the classifiers.*

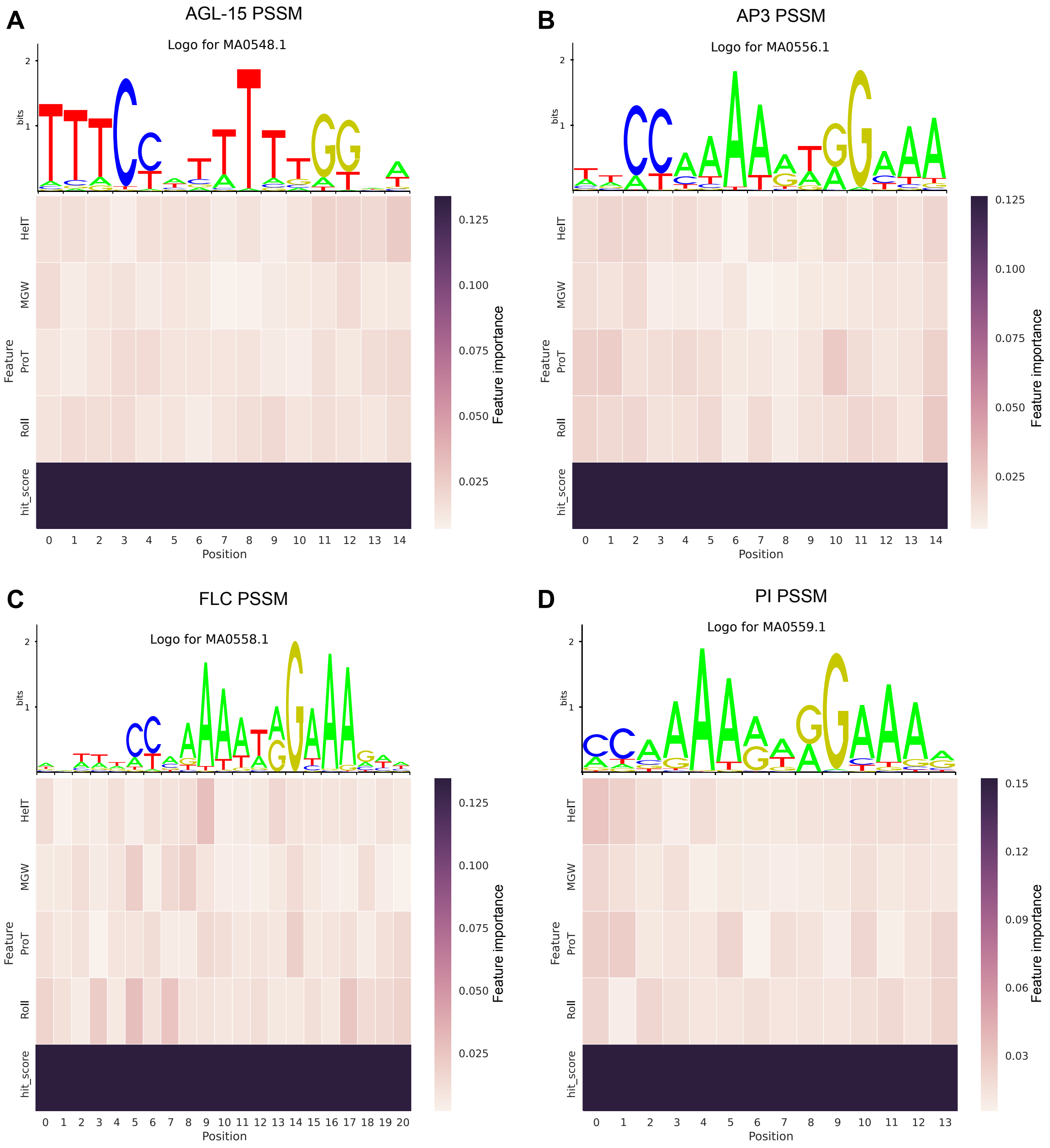


Figure S5 *Related to Figure 7. Feature importance measures for plant MADS-box recognition in ChIP-seq. Weblogos of the MADS-domain TF profiles considered for AGL-15 (**A**), AP3 (**B**, FLC (**C**), and PI (**D**) are provided at the top of the panels. Heat maps providing the average level of feature importance (y-axis) at each position (x-axis) of the TFBSs in the PSSM + DNA shape classifiers are provided at the bottom of the panels. Note that the 'hit score' feature corresponds to the PSSM scores used in the classifiers. The 'hit score' feature spans all the positions for graphical representation but a single hit score is provided per vector in the classifiers.*

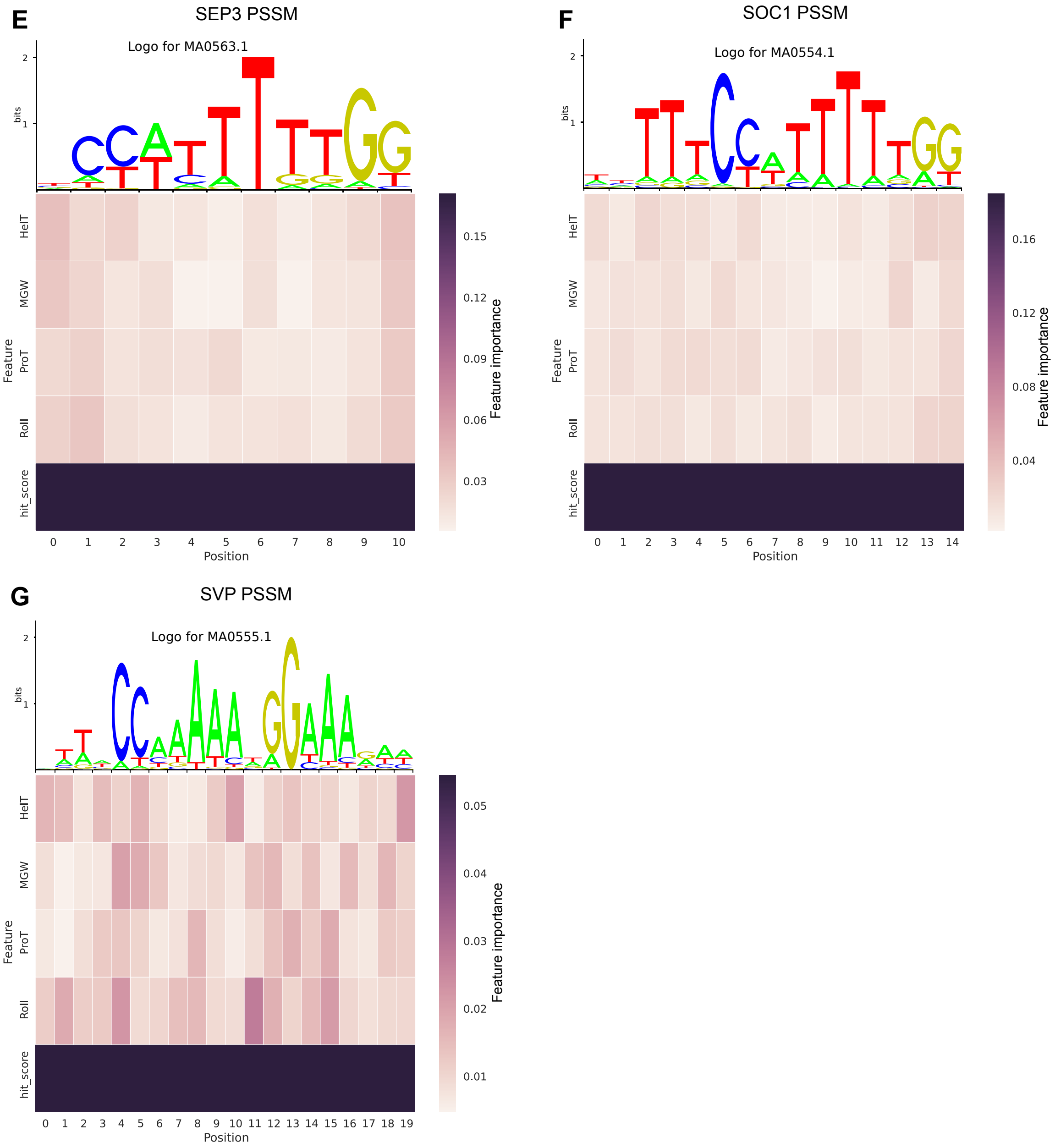


Figure S6 *Related to Figure 7. Feature importance measures for plant MADS-box recognition in ChIP-seq. Weblogos of the MADS-domain TF profiles considered for SEP3 (**E**), SOC1 (**F**), and SVP (**G**) are provided at the top of the panels. Heat maps providing the average level of feature importance (y-axis) at each position (x-axis) of the TFBSs in the PSSM + DNA shape classifiers are provided at the bottom of the panels. Note that the 'hit score' feature corresponds to the PSSM scores used in the classifiers. The 'hit score' feature spans all the positions for graphical representation but a single hit score is provided per vector in the classifiers.*

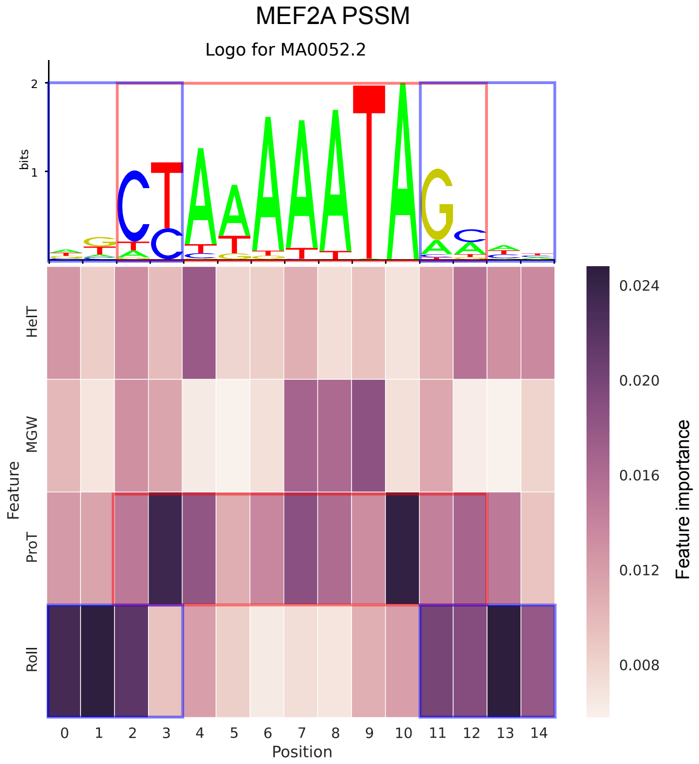


Figure S7 *Related to Figure 7. Feature importance measures for MADS-box recognition in ChIP-seq. The weblogo derived from the JASPAR TF binding profile associated with the MEF2A TF is provided at the top. The heat map providing the average feature importance (y-axis) at each position (x-axis) of the TFBSs in the classifiers trained for the 10-fold CV analysis of the ChIP-seq data sets are provided at the bottom. Note that only feature importances associated with DNA shape features are provided. The color scale used in the heat map is provided on the right of the heat map. The red box highlights the core MADS-box motif (CCW$_6$GG) while the blue boxes highlight the edges of the motif.*

Data S1 *Spreadsheets related to ChIP-seq datasets and discriminative improvements with DNA shape. Tables S1, S2, and S3 provide the list of ChIP-seq datasets used in this study. Table S4 summarizes the results for the TF families most benefitting from DNA shape information for TFBS prediction. Related to Figure 2.*

Data S2 *Impact of DNA shape on predicting TFBSs with genomic background sequences matching the %GC composition of ChIP-seq regions. Related to Figure 2.*

Data S3 *Impact of DNA shape on predicting TFBSs with background sequences matching the dinucleotide composition of ChIP-seq regions. Related to Figure 2.*

Data S4 *Impact of DNA shape on predicting TFBSs when considering recurrent ChIP-seq regions for each TF and genomic background sequences matching the %GC composition of ChIP-seq regions. Related to Figure 2.*

Data S5 *Comparison of the predictive powers between generative and discriminative approaches. Related to Figure 3.*

Data S6 *Assessment of the predictive power of DNA shape features at TFBS flanking regions with genomic background sequences matching the %GC composition of ChIP-seq regions. Related to Figure 4.*

Data S7 *Impact of DNA shape on predicting human and plant MADS-box TFBSs with background sequences matching the %GC or dinucleotide composition of the ChIP-seq regions. Related to Figures 2.*