Supplemental Items

DNA shape features improve transcription factor binding site predictions *in vivo*

Anthony Mathelier[1,2], Beibei Xin[3], Tsu-Pei Chiu[3],
Lin Yang[3], Remo Rohs[3,*], and Wyeth W. Wasserman[1,*]

[1] Centre for Molecular Medicine at the Child and Family Research Institute,
Department of Medical Genetics, University of British Columbia,
980 West 28th Avenue, V5Z 4H4, Vancouver, BC, Canada
[2] Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership,
University of Oslo and Oslo University Hospital, Norway
[3] Molecular and Computational Biology Program, Departments of Biological Sciences,
Chemistry, Physics, and Computer Science,
University of Southern California, Los Angeles, CA 90089, USA
[*] Co-corresponding authors

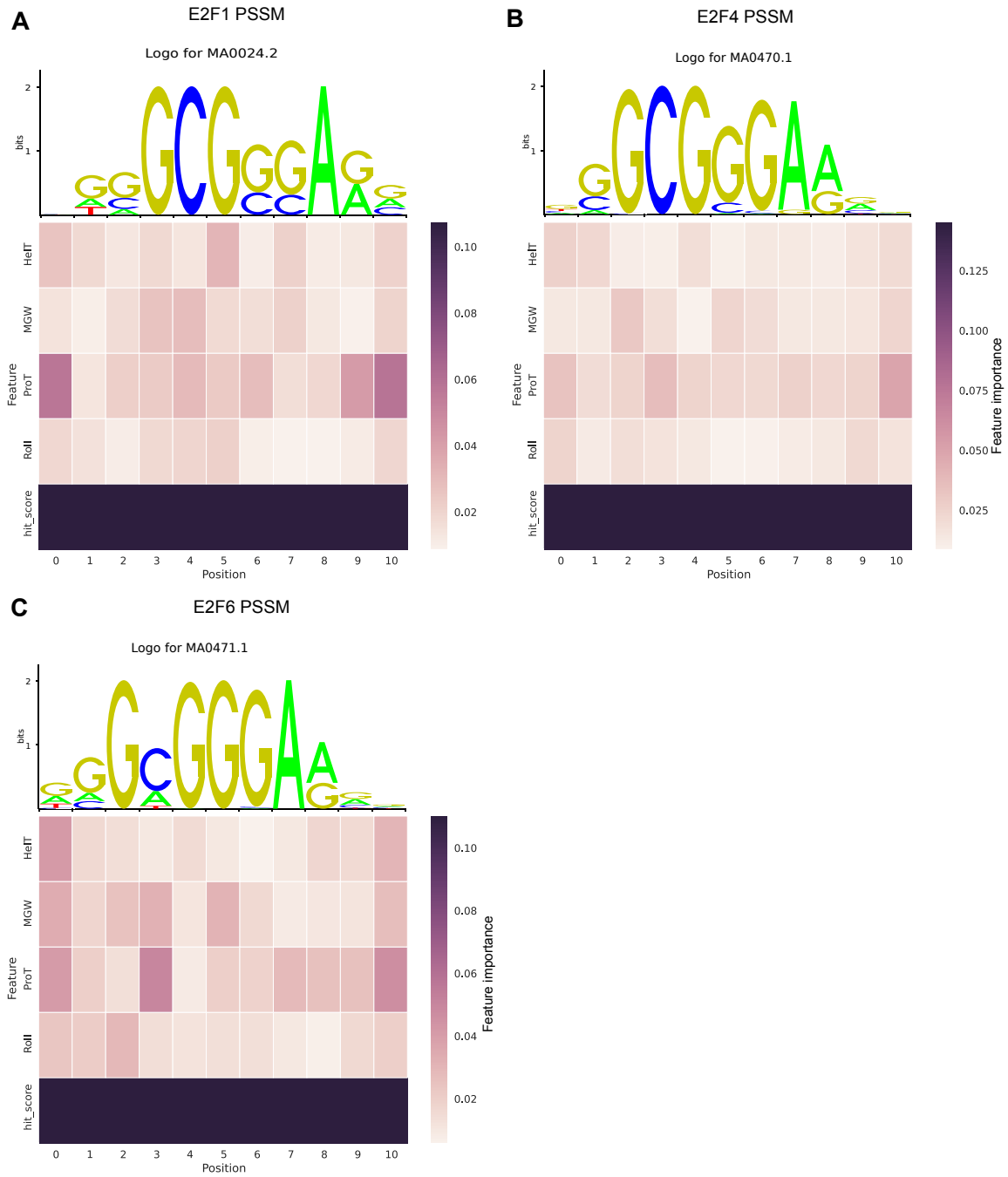

Figure S1 *Related to Figure 7. Feature importance measures for human E2F TFBS recognition in ChIP-seq. Weblogos of the E2F TF profiles for E2F1 (**A**), E2F4 (**B**), and E2F6 (**C**) from JASPAR are provided at the top of the panels. Heat maps providing the average level of feature importance (y-axis) 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.*

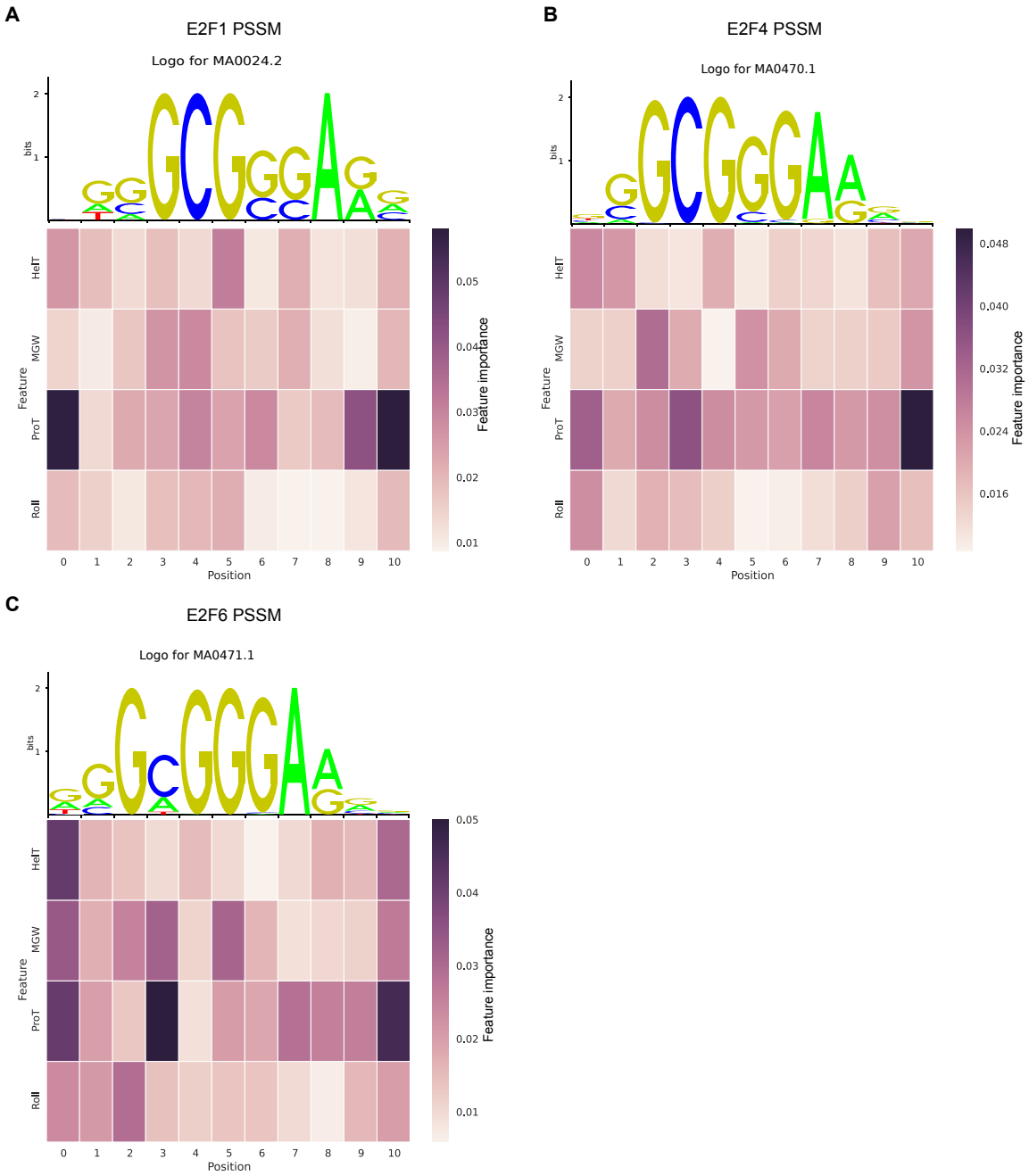

Figure S2 *Related to Figure 7. DNA shape only feature importance measures for human E2F TFBS recognition in ChIP-seq. Weblogos of the E2F TF profiles for E2F1 (**A**), E2F4 (**B**), and E2F6 (**C**) from JASPAR are provided at the top of the panels. Heat maps providing the average level of DNA shape 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.*

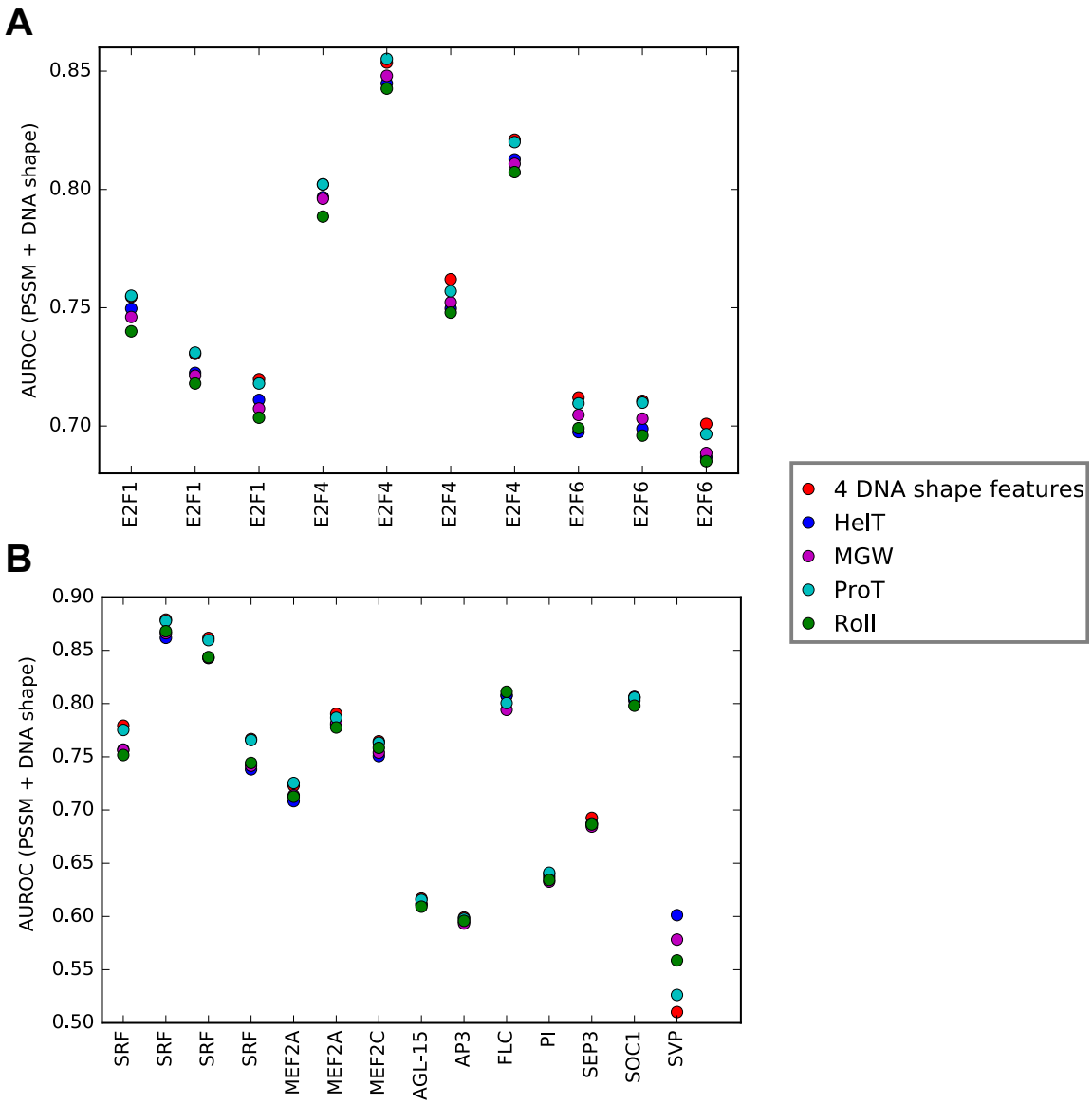

Figure S3 *Related to Figure 5. Using a single DNA shape feature for E2F and MADS-box TFBS recognition in ChIP-seq. Comparison of AUROC (y-axis) for the E2F (**A**) and the MADS-domain (**B**) TF data sets (x-axis) when using 4 DNA shape features or a single feature along with the PSSM scores in the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

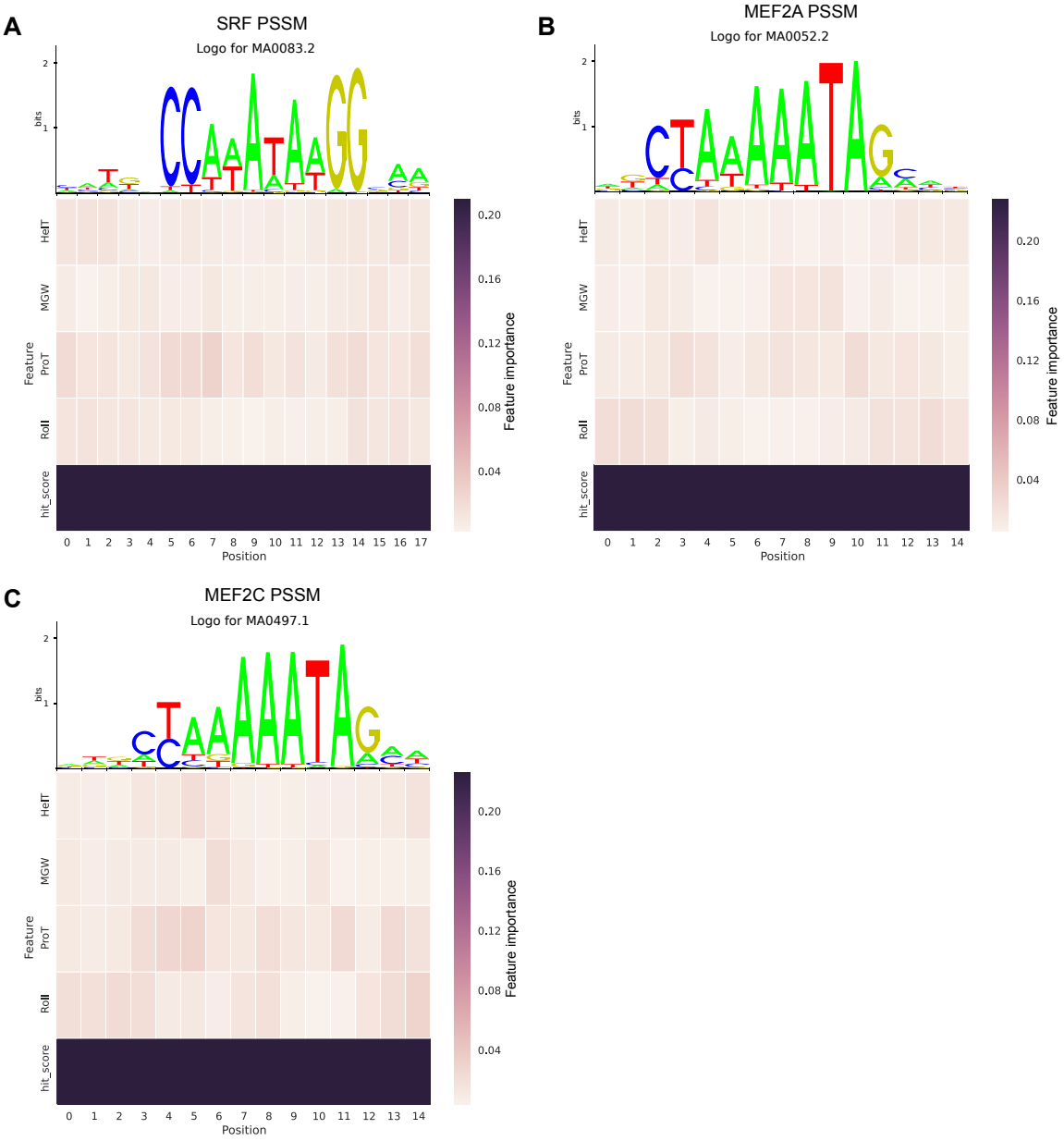

Figure S4 *Related to Figure 7. Feature importance measures for human MADS-box recognition in ChIP-seq. Weblogos of the MADS-domain TF profiles considered for SRF (**A**), MEF2A (**B**), and MEF2C (**C**) are provided at the top of the panels. Heat maps providing the average level of feature importance (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.*

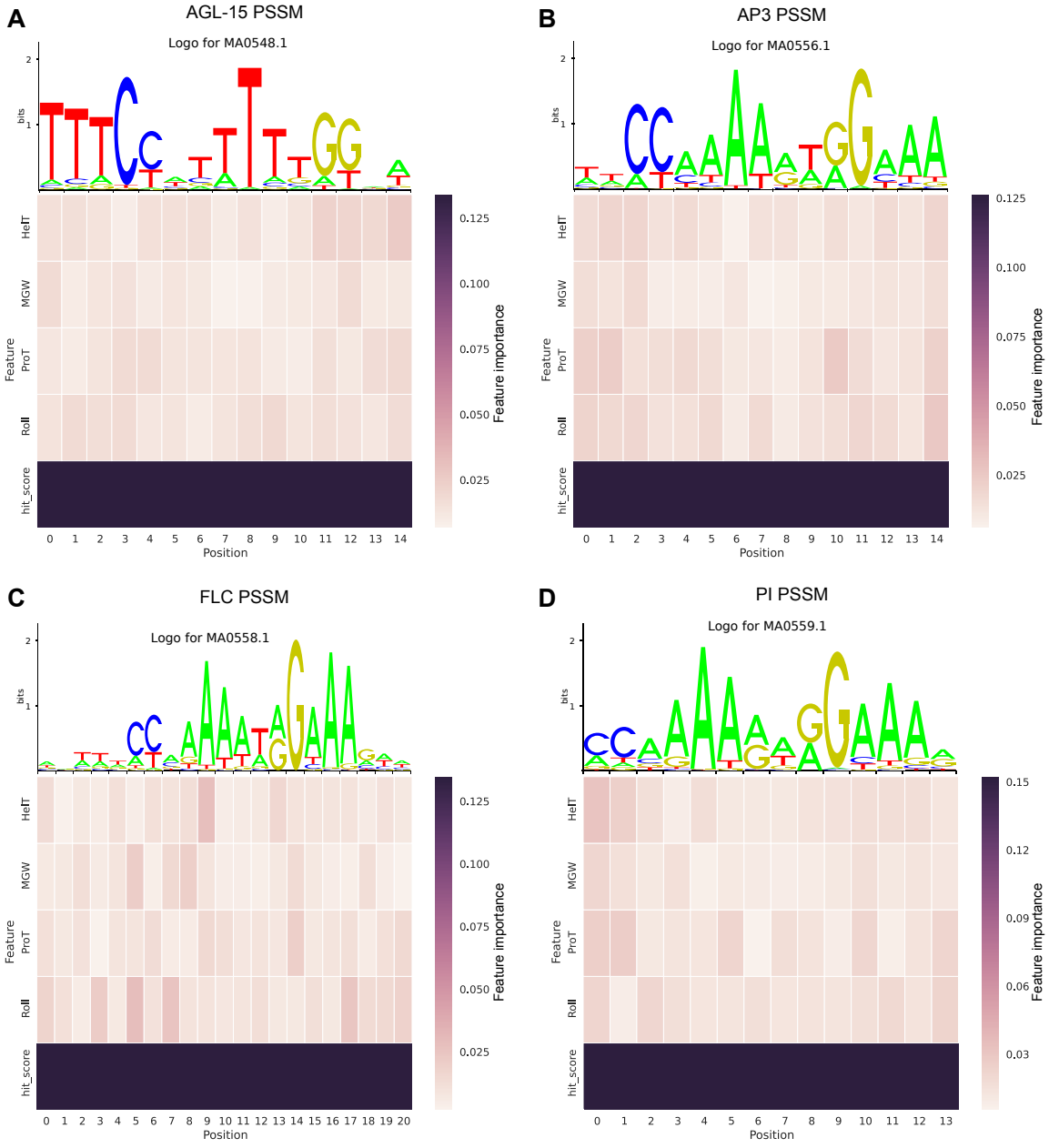

Figure S5 *Related to Figure 7. Feature importance measures for plant MADS-box recognition in ChIP-seq. Weblogos of the MADS-domain TF profiles considered for AGL-15 (**A**), AP3 (**B**, FLC (**C**), and PI (**D**) are provided at the top of the panels. Heat maps providing the average level of feature 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.*

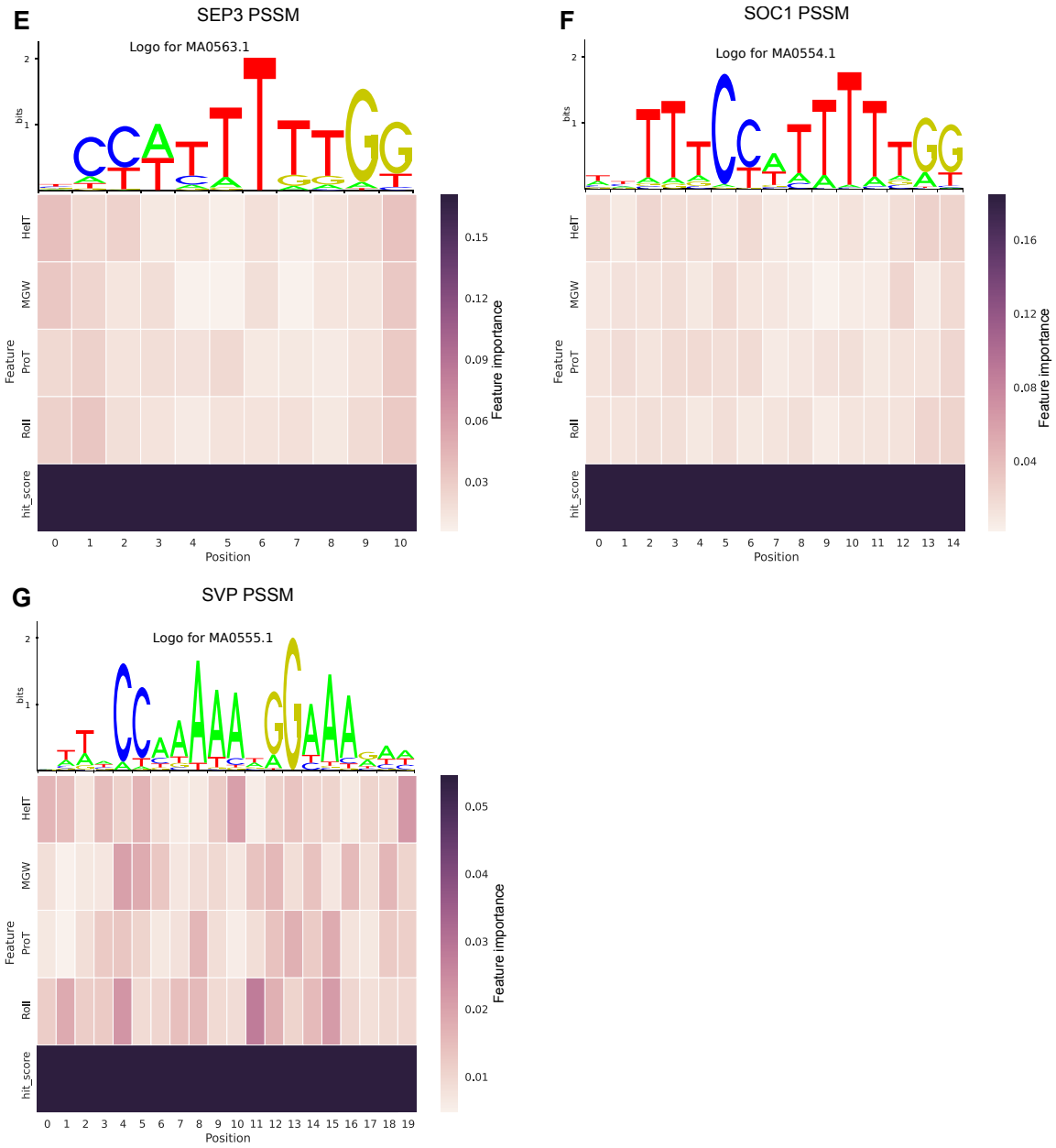

Figure S6 *Related to Figure 7. Feature importance measures for plant MADS-box recognition in ChIP-seq. Weblogos of the MADS-domain TF profiles considered for SEP3 (**E**), SOC1 (**F**), and SVP (**G**) are provided at the top of the panels. Heat maps providing the average level of feature importance (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.*

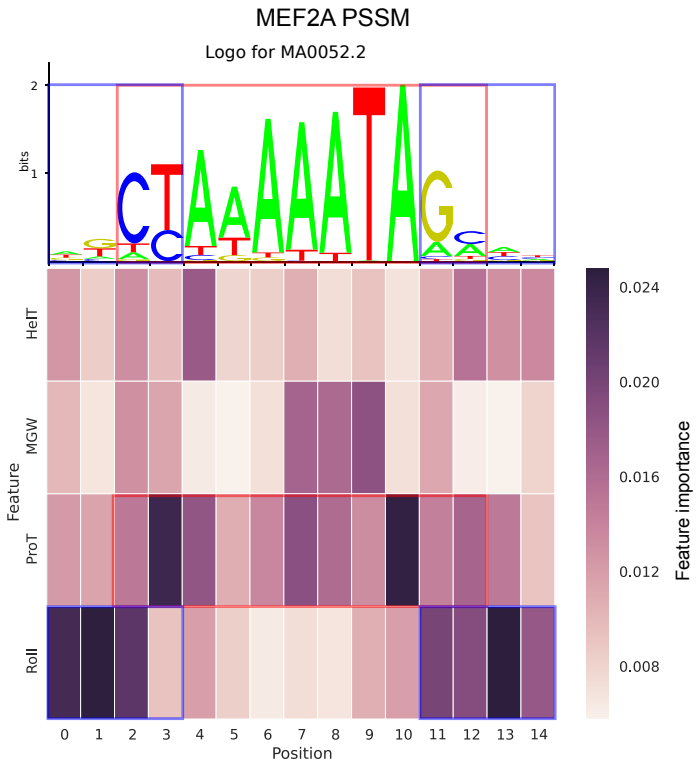

Figure S7 *Related to Figure 7. Feature importance measures for MADS-box recognition in ChIP-seq. The weblogo derived from the JASPAR TF binding profile associated with the MEF2A TF is provided at the top. The heat map providing the average feature importance (y-axis) at each position (x-axis) of the 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_6GG) while the blue boxes highlight the edges of the motif.*

Data S1 *Related to Experimental Procedures and Figure 2. Spreadsheets related to the ChIP-seq data sets used in this study (Tables S1-S3) and to the TF families benefiting most from DNA shape information (Table S4).*

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

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

Data S4 *Related to Figure 2. 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.*

Data S5 *Related to Figure 3. Comparison of the predictive powers between generative (TFFM- or PSSM-based models) and discriminative (4-bits-based models) approaches.*

Data S6 *Related to Figure 4. 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.*

Data S7 *Related to Figures 2. 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.*