

APPENDIX

Transcription factor family-specific DNA shape readout revealed by quantitative specificity models

Lin Yang^{1,†}, Yaron Orenstein^{2,†,‡}, Arttu Jolma³, Yimeng Yin³, Jussi Taipale³, Ron Shamir^{2,*} & Remo Rohs^{1,**}

¹ Molecular and Computational Biology Program, Departments of Biological Sciences, Chemistry, Physics, and Computer Science, University of Southern California, Los Angeles, CA 90089, USA

² Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel

³ Division of Functional Genomics and Systems Biology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm SE 141 83, Sweden

[†] These authors contributed equally to this work.

[‡] Present address: Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

* Corresponding author. Tel: +972-3-640-5383; E-mail: rshamir@tau.ac.il

** Corresponding author. Tel: +1-213-740-0552; E-mail: rohs@usc.edu

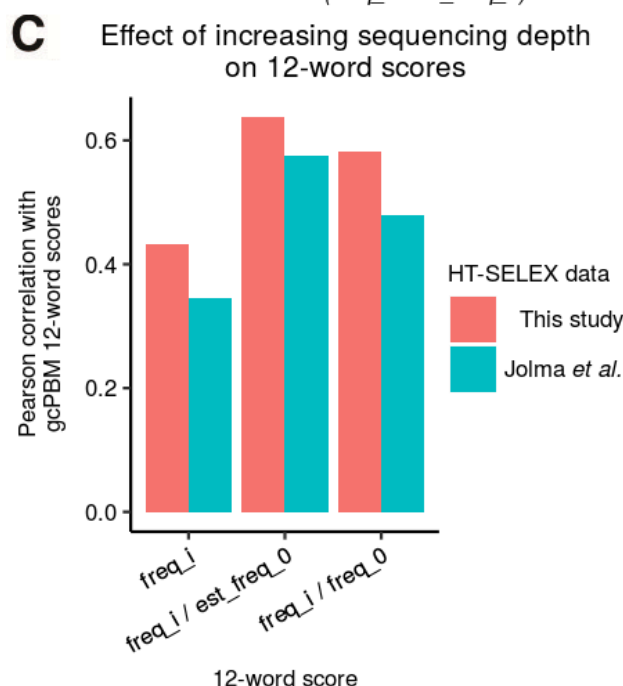
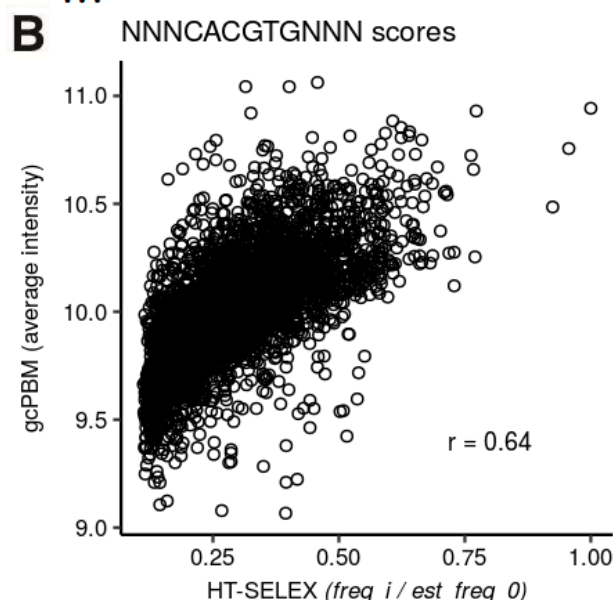
TABLE OF CONTENTS

Appendix Figures

<i>Appendix Figure S1. Processed data compared to gcPBM binding score.</i>	2
<i>Appendix Figure S2. Principal component analysis (PCA) using randomly generated features.</i>	3
<i>Appendix Figure S3. Performance comparison between different models.</i>	4
<i>Appendix Figure S4. Positional DNA shape importance revealed by feature selection for different TF families.</i>	6
<i>Appendix Figure S5. Feature-specific positional DNA shape importance revealed by feature selection for different TF families.</i>	11
<i>Appendix Figure S6. Structure analysis of a nuclear receptor protein in complex with DNA revealed distinct characteristics of DNA shape features in the central region which is also highlighted in the heat map analysis.</i>	15

A

K-mer	Score	Count
AATATTGCGCAATATT	1.0	533
ACTATTACGCAATATG	0.872	786
GCAATCGAGTTATTGA	0.860	573
GCAATGGTACAAATCA	0.800	633
CCGATTGCGCAATAGT	0.800	465
CCGATTGCGCAATCGT	0.799	432
CCTATTGCGCAATGTG	0.797	492
ATGATTACGCAATCAG	0.795	516
ACGATTACGCAATCGG	0.793	325
CGCAATACGAAATTTA	0.785	470
CGCAATTTCGTACTGCT	0.784	432
CCGATTGCGTAATAGA	0.782	379
CGCAATACATAATCAA	0.778	663
ACTATTGCGCAACAGG	0.771	561

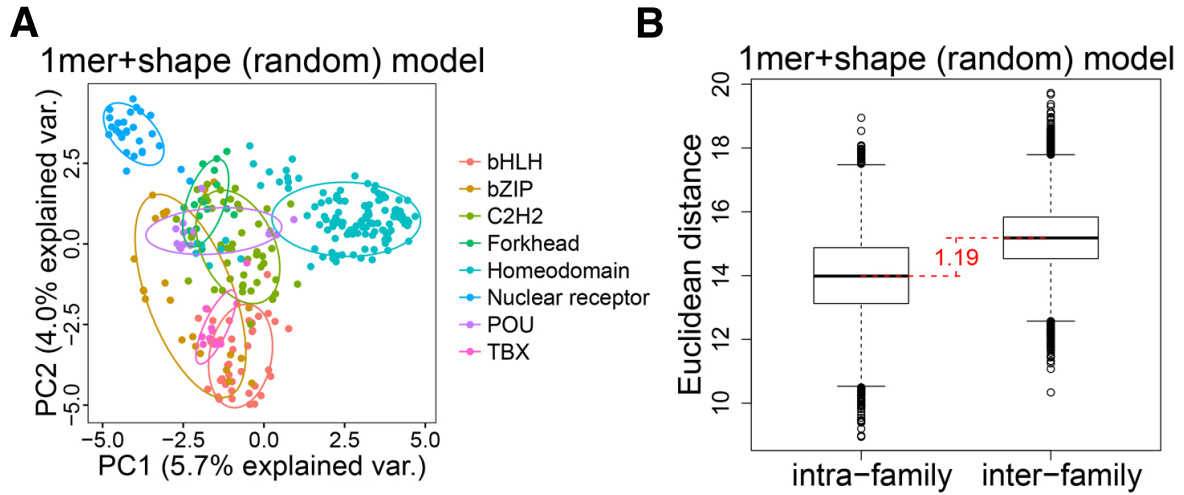


Appendix Figure S1. Processed data compared to gcPBM binding score.

A. HT-SELEX *M*-word scores. For each *M*-word with the core consensus in the center, we produced an *M*-word score by the ratio of the frequency in cycle 3 (or later cycles) over estimated frequency in the initial cycle. The core consensus is highlighted in red.

B. The gcPBM and HT-SELEX 12-word scores for the Max homodimer. The gcPBM scores were the average of log-normalized binding intensities. HT-SELEX scores were the log of the ratio of the frequency in cycle 3 over the estimated frequency of the initial cycle.

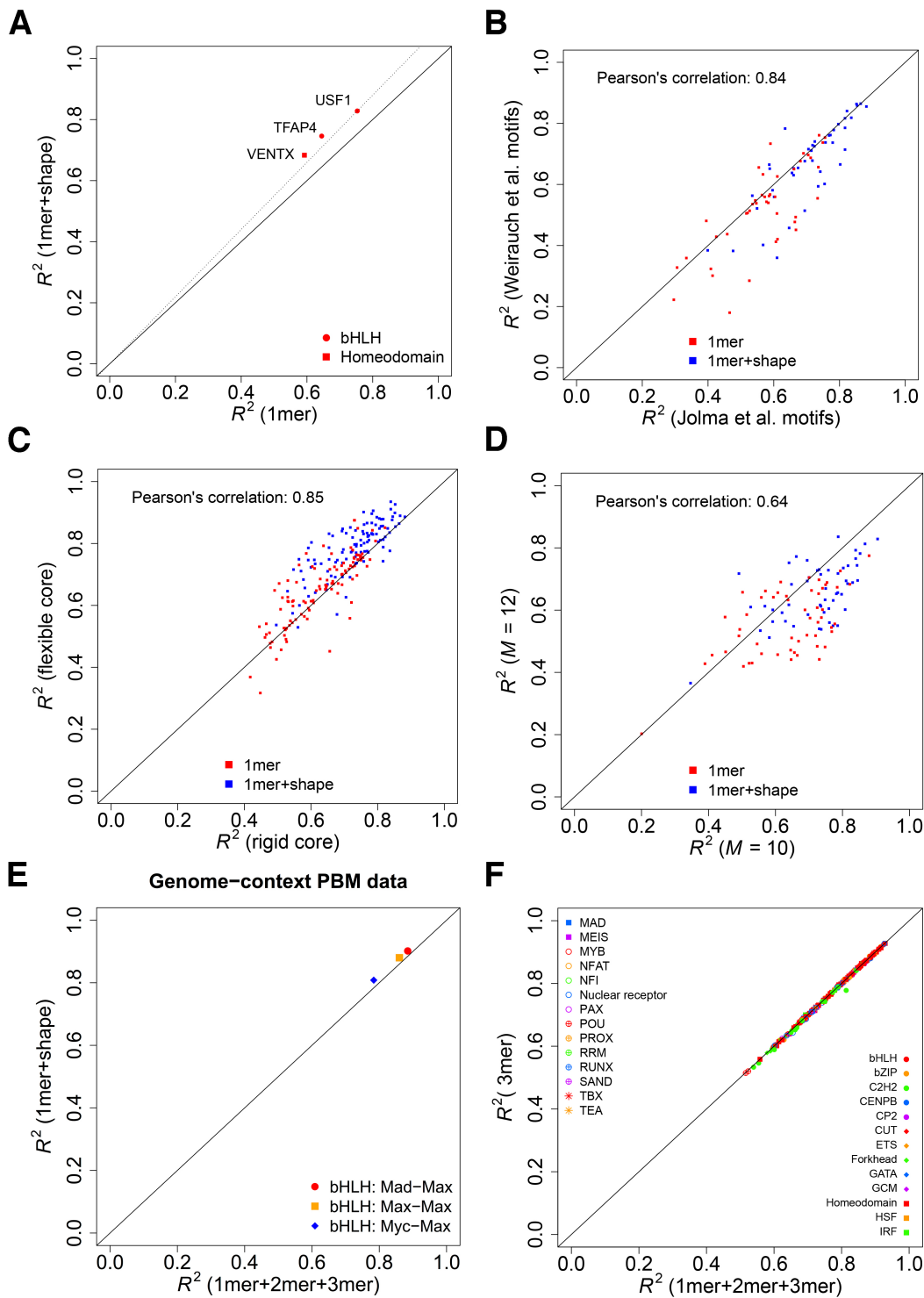
C. Comparison of *M*-word scores applied to previously published data (Jolma et al, 2013) and new augmented data (this study) in correlation to gcPBM 12-word scores for the Max homodimer. Freq_i is the frequency at cycle *i*; est_freq_0 is the estimated frequency at the initial cycle based on a fifth-order Markov model



Appendix Figure S2. Principal component analysis (PCA) using randomly generated features.

A. PCA using 1-mer and randomly generated shape features. Each dot represents a transcription factor (TF). Dots of the same color belong to the same TF family. An ellipse was drawn for each TF family. The ellipse is a contour of a fitted two-variate normal distribution that encloses 0.68 (R package default) probability.

B. Boxplots of inter- and intra-family TF distances derived from A. Difference between medians of inter- and intra-family distances is 1.19 (red).

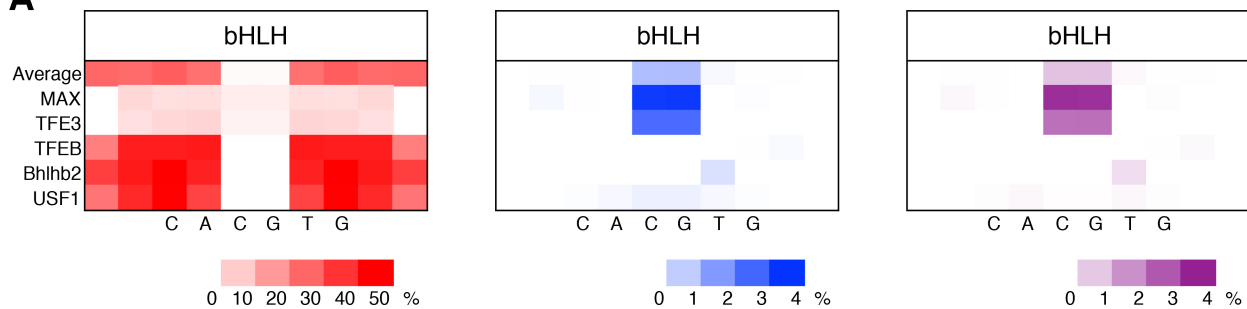


Appendix Figure S3. Performance comparison between different models.

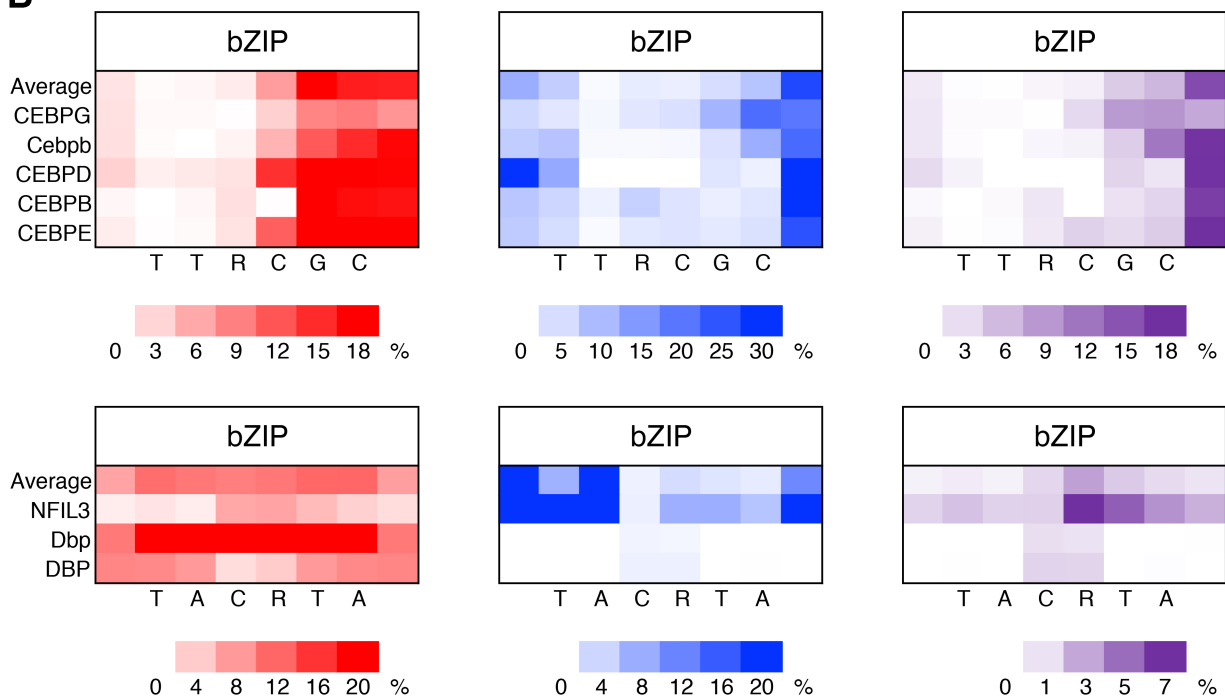
- A.** Performance comparison between 1mer and 1mer+shape models for replicate HT-SELEX experiments.
- B.** Comparison of model performance for preprocessing that used Jolma et al. (Jolma et al, 2013) seeds and Weirauch & Hughes (Weirauch & Hughes, 2011) seeds.
- C.** Comparison of model performance for preprocessing that allowed different numbers of mismatches at the core motif positions.
- D.** Comparison of model performance for preprocessing using different lengths of the flanking regions.
- E.** Performance comparison between 1mer+2mer+3mer and 1mer+shape models for gcPBM data.
- F.** Performance comparison between 1mer+2mer+3mer and 3mer models for HT-SELEX data.

Each dot represents one dataset. Coordinates of the dot are determined by the performance, measured in R^2 based on 10-fold cross-validation, of the corresponding models indicated in parentheses. Shape and color of the dots indicate the TF family. Dashed line in A has a slope of 1.1, indicating 10% performance increase.

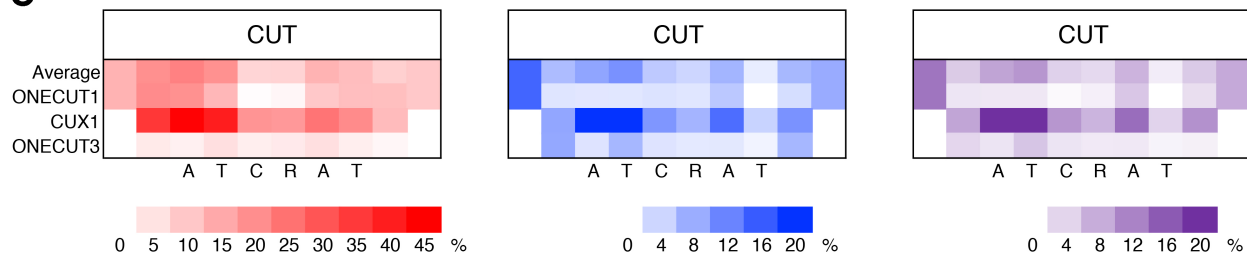
A



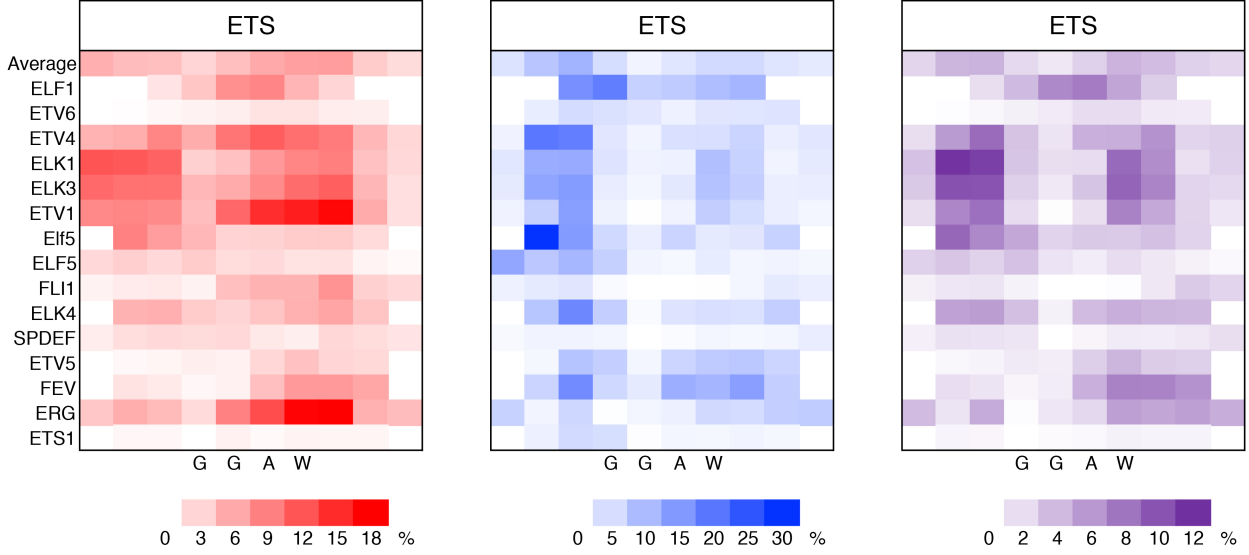
B



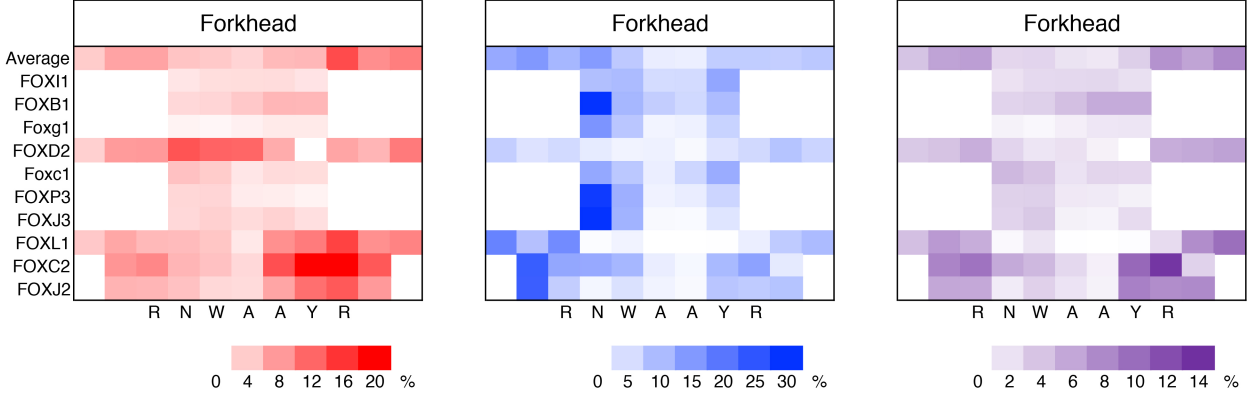
C



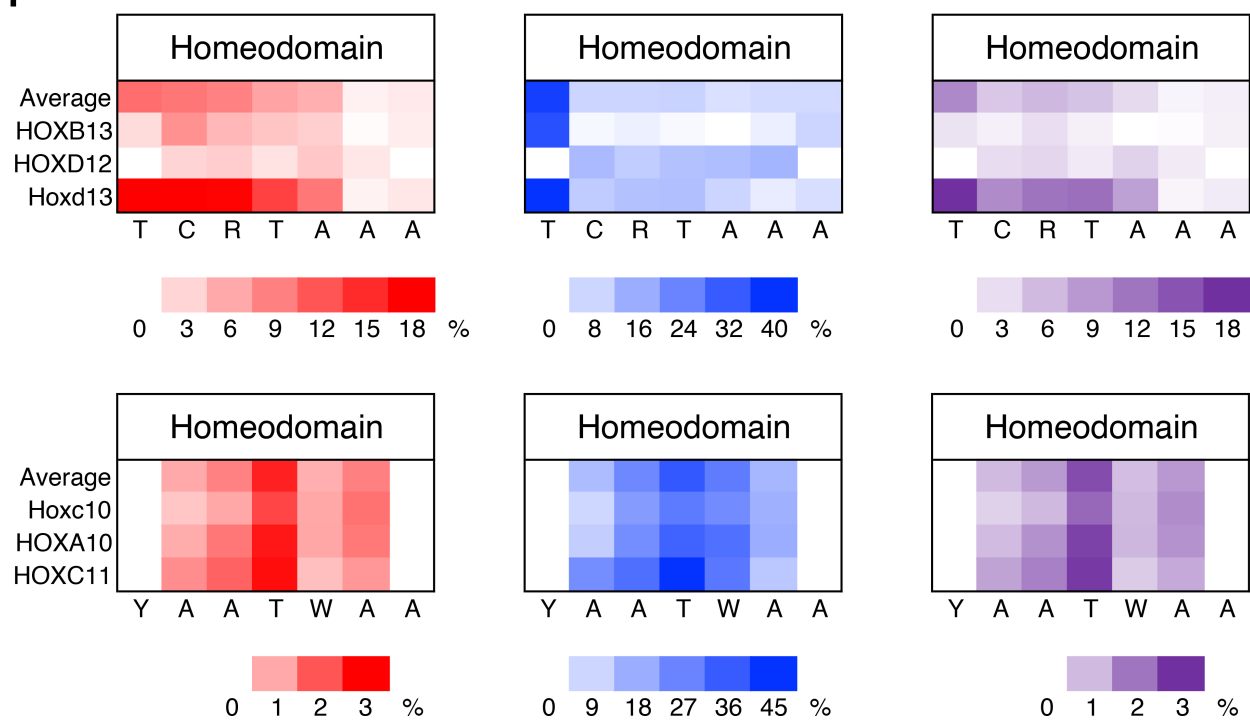
D



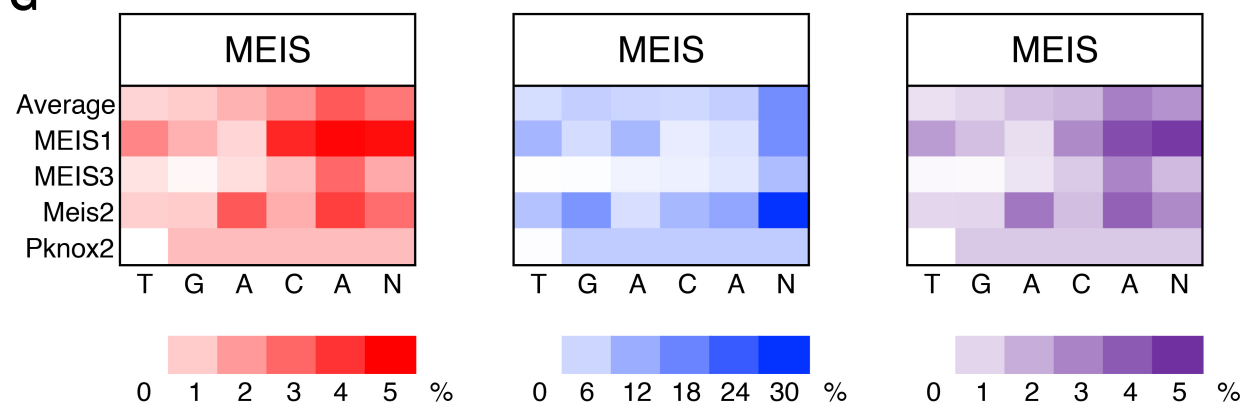
E



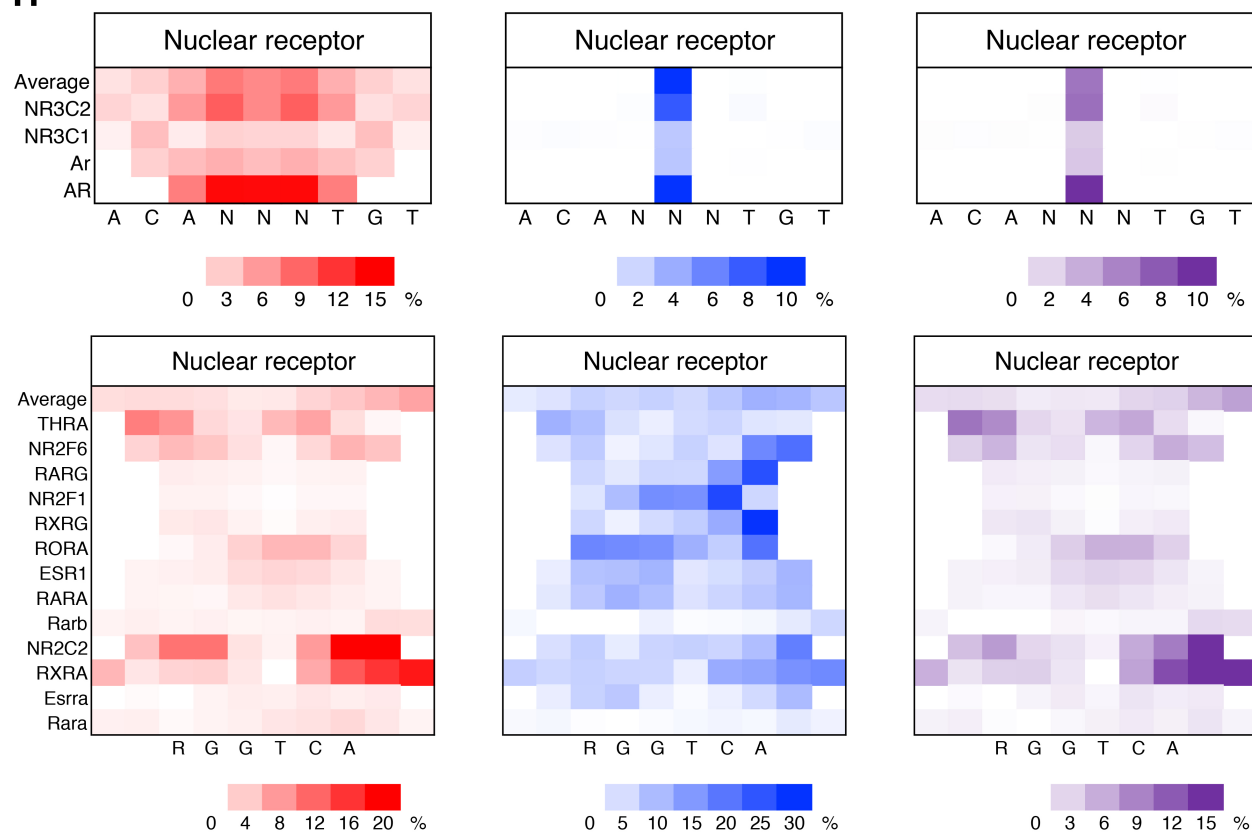
F



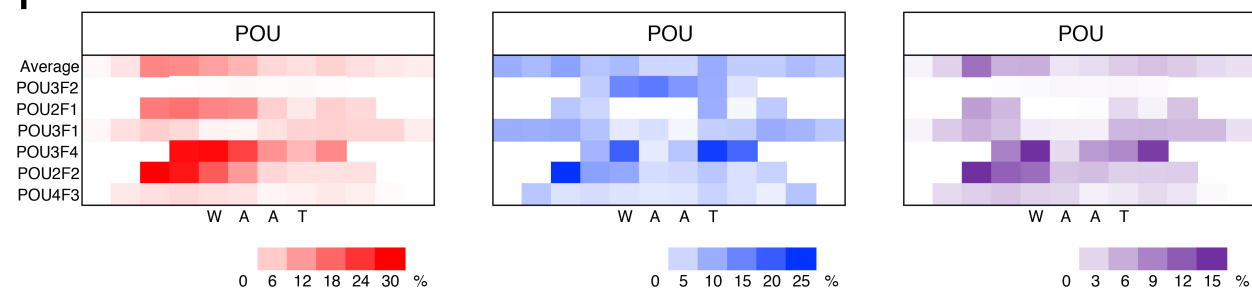
G

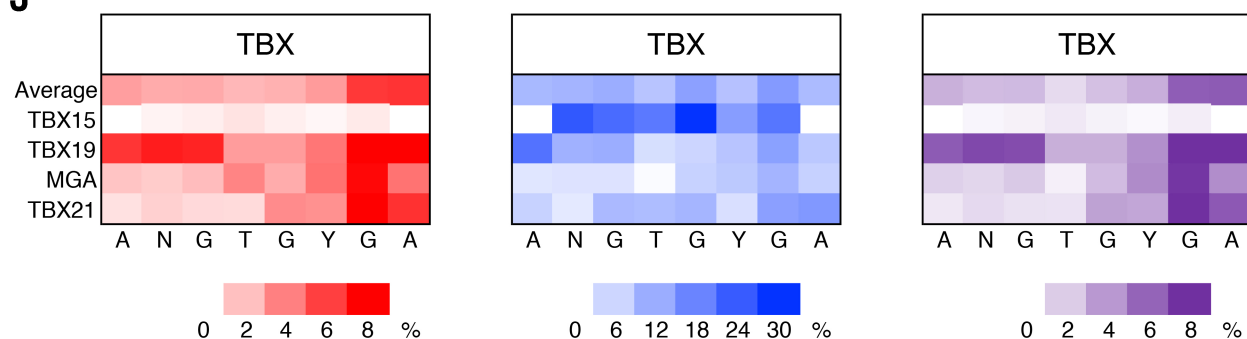


H



I



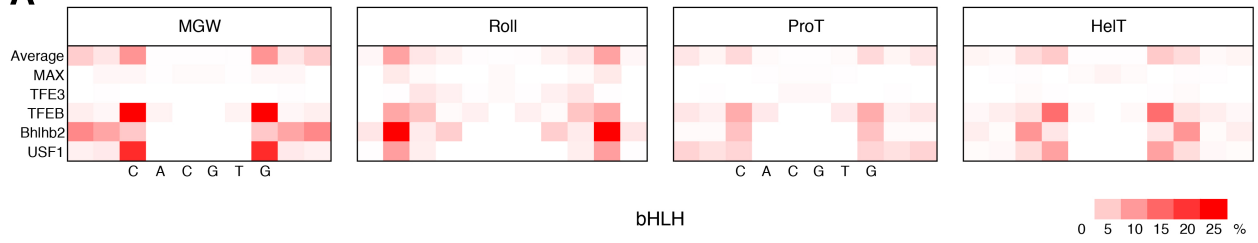
J

Appendix Figure S4. Positional DNA shape importance revealed by feature selection for different TF families.

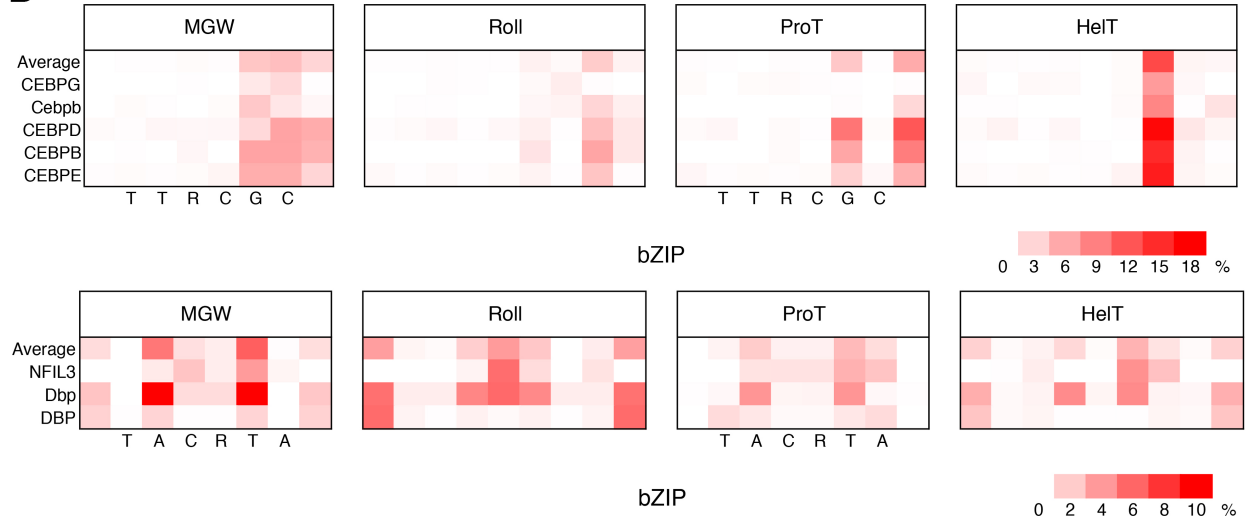
Heat maps derived from feature selection, where DNA shape features were added position by position to a baseline 1mer model (red heat maps), DNA shape features were removed position by position from a baseline shape-only model (blue heat maps), and the red and blue heat maps were combined by taking the minimum of these two heat maps (resulting in the combined purple heat maps), for the following TF families.

- A. bHLH (basic helix-loop-helix).
- B. bZIP (basic-leucine zipper).
- C. CUT.
- D. ETS.
- E. Forkhead.
- F. Homeodomain.
- G. MEIS.
- H. Nuclear receptor.
- I. POU.
- J. TBX.

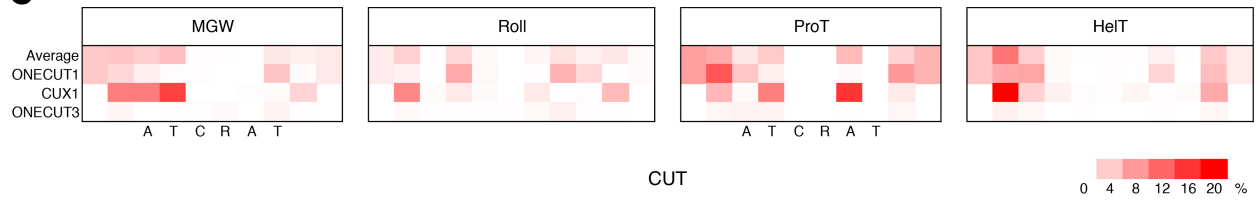
A



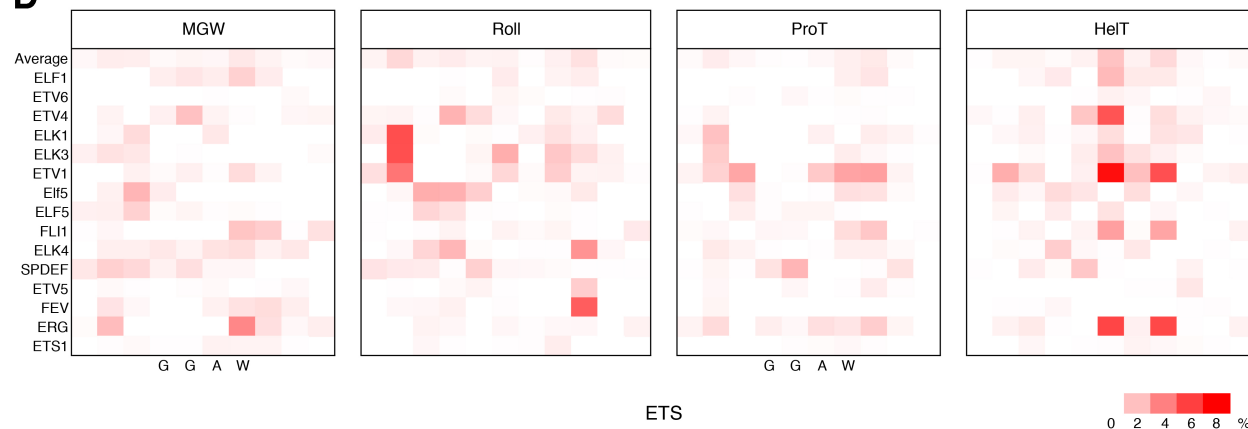
B



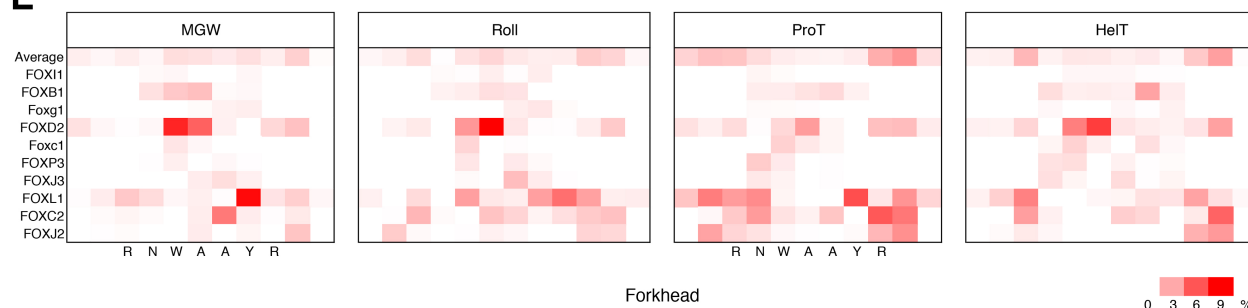
C



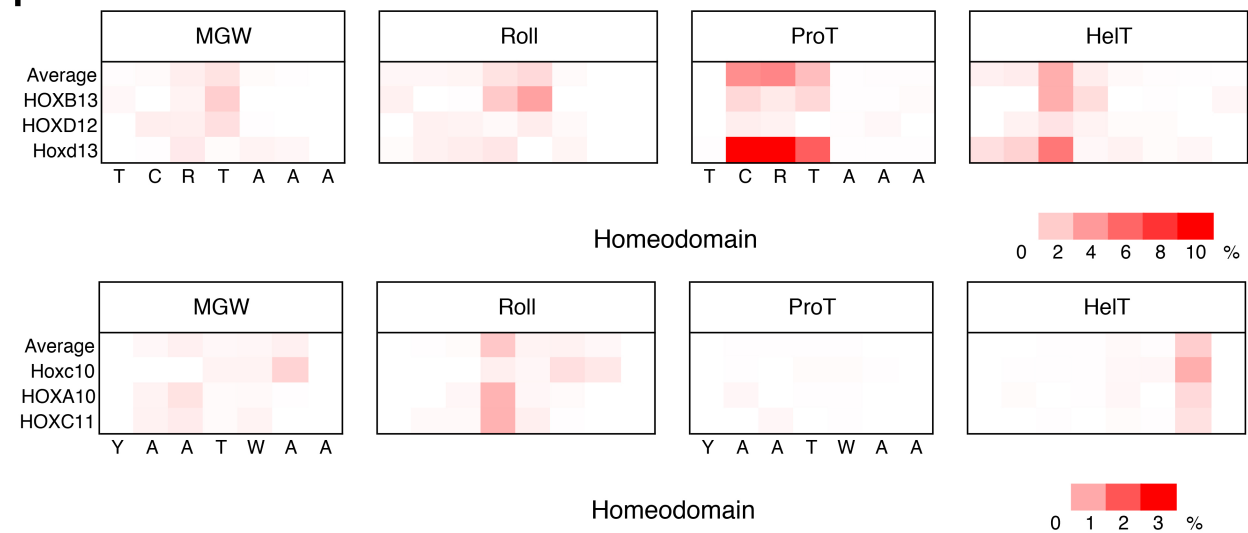
D



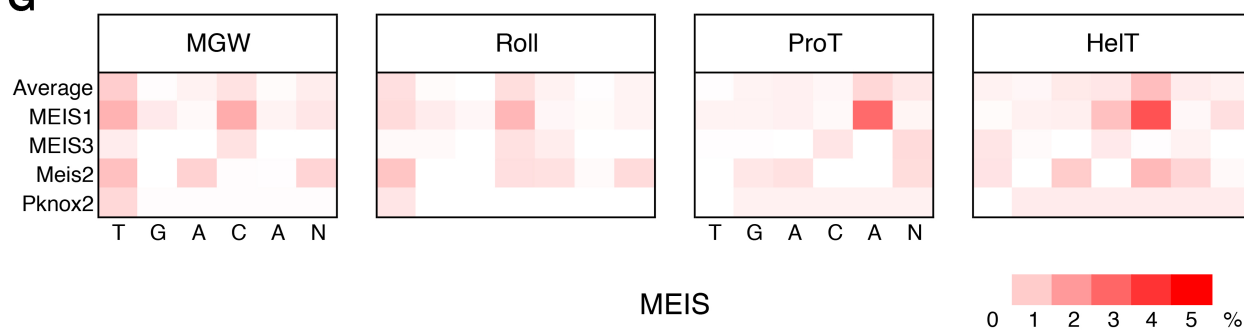
E



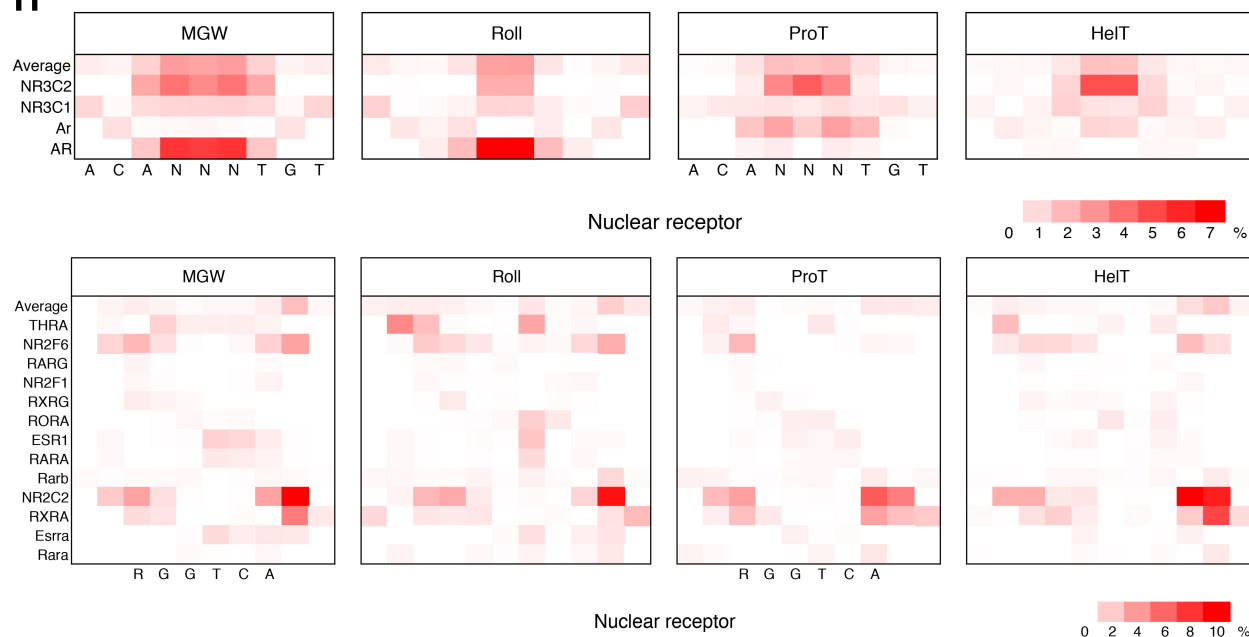
F



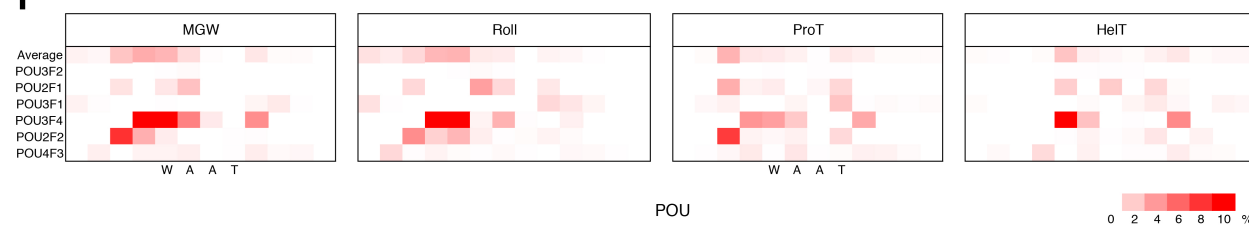
G



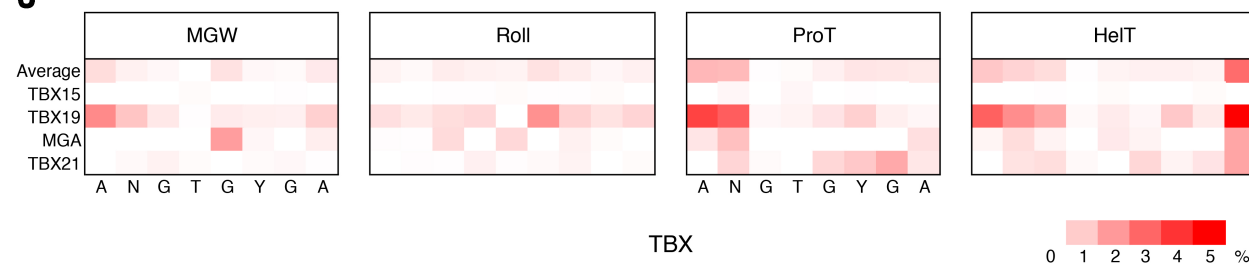
H



I



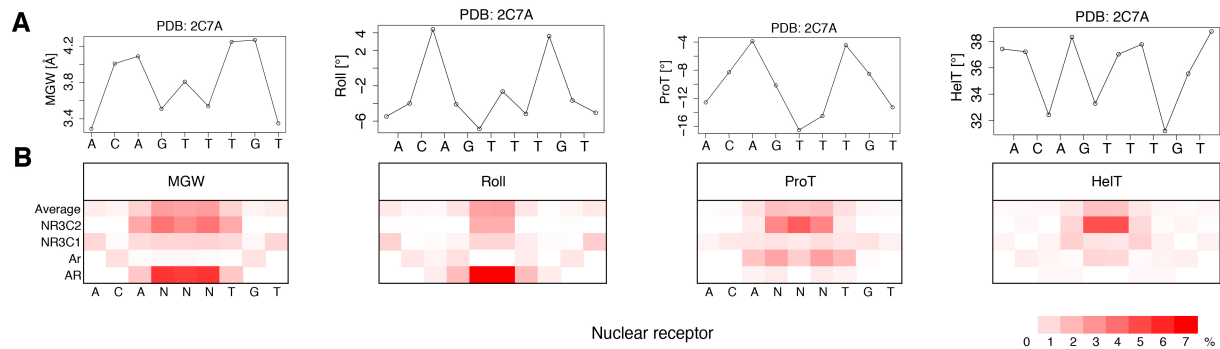
J



Appendix Figure S5. Feature-specific positional DNA shape importance revealed by feature selection for different TF families.

Heat maps derived from feature selection, where each category of DNA shape features, i.e. MGW, Roll, ProT and HelT, was added position by position to a baseline 1mer model, for the following TF families.

- A.** bHLH (basic helix-loop-helix).
- B.** bZIP (basic-leucine zipper).
- C.** CUT.
- D.** ETS.
- E.** Forkhead.
- F.** Homeodomain.
- G.** MEIS.
- H.** Nuclear receptor.
- I.** POU.
- J.** TBX.



Appendix Figure S6. Structure analysis of a nuclear receptor protein in complex with DNA revealed distinct characteristics of DNA shape features in the central region which is also highlighted in the heat map analysis.

- A.** CURVES (Lavery & Sklenar, 1989) derived plots for MGW, Roll, ProT, and HelT of the bound DNA in the progesterone receptor-DNA complex (PDB ID 2C7A).
- B.** Heat maps from Appendix Figure S5H aligned to the plots in A.

References:

Jolma A, Yan J, Whittington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, Palin K, Vaquerizas JM, Vincentelli R, Luscombe NM, Hughes TR, Lemaire P, Ukkonen E, Kivioja T, Taipale J (2013). DNA-binding specificities of human transcription factors. *Cell* 152: 327-339

Lavery R, Sklenar H (1989). Defining the structure of irregular nucleic acids: conventions and principles. *Journal of Biomolecular Structure and Dynamics* 6: 655-667

Weirauch MT, Hughes T (2011). A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. In *A Handbook of Transcription Factors*, pp 25-73. Springer Netherlands