Data S6

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

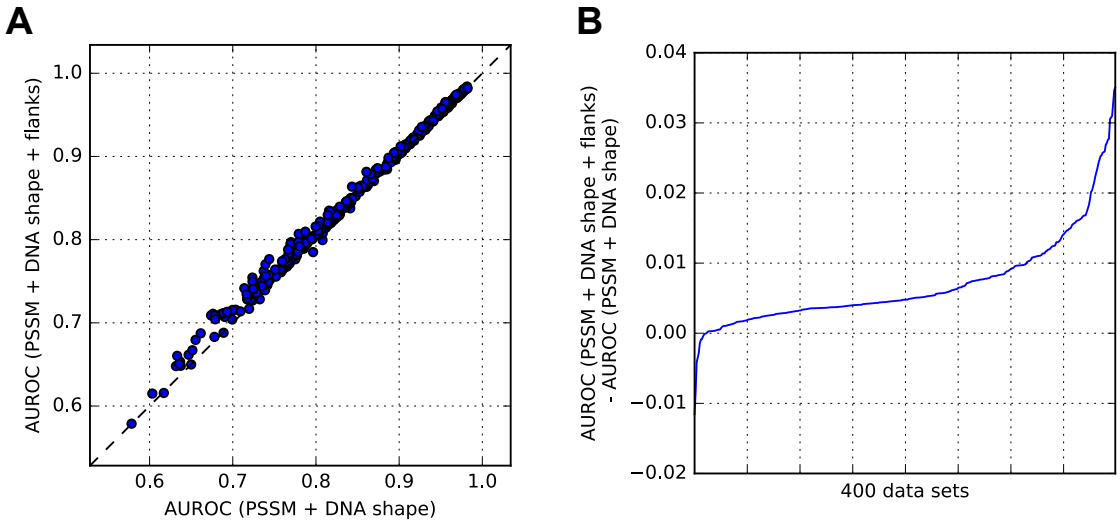
Anthony Mathelier[1,2], Beibei Xin[3], Tsu-Pei Chiu[3],
Lin Yang[3], Remo Rohs[3,*], and Wyeth W. Wasserman[1,*]

[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

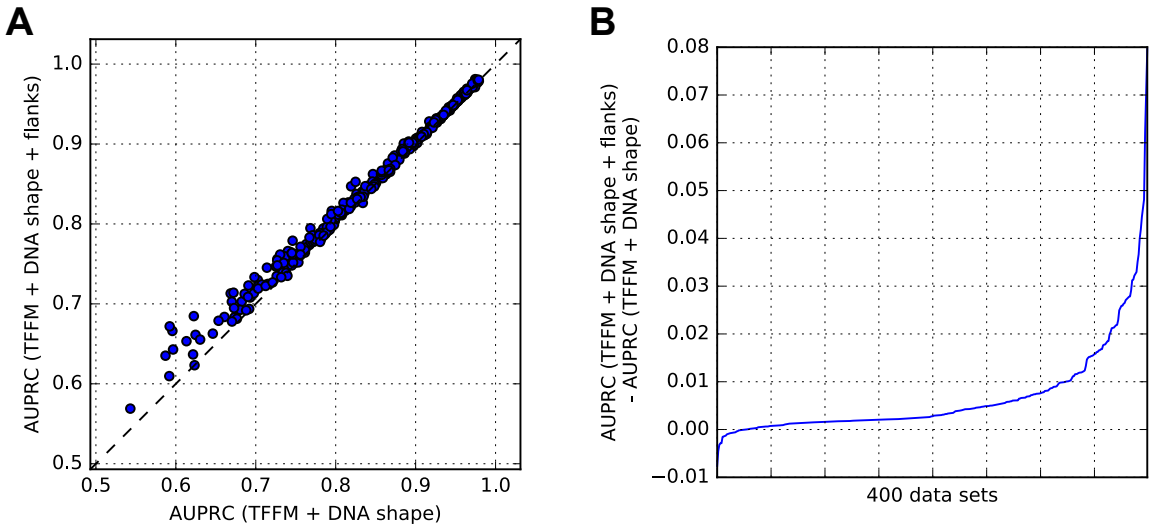
[2] Centre for Molecular Medicine Norway (NCMM), Nordic EMBL partnership,
University of Oslo and Oslo University Hospital, Norway

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

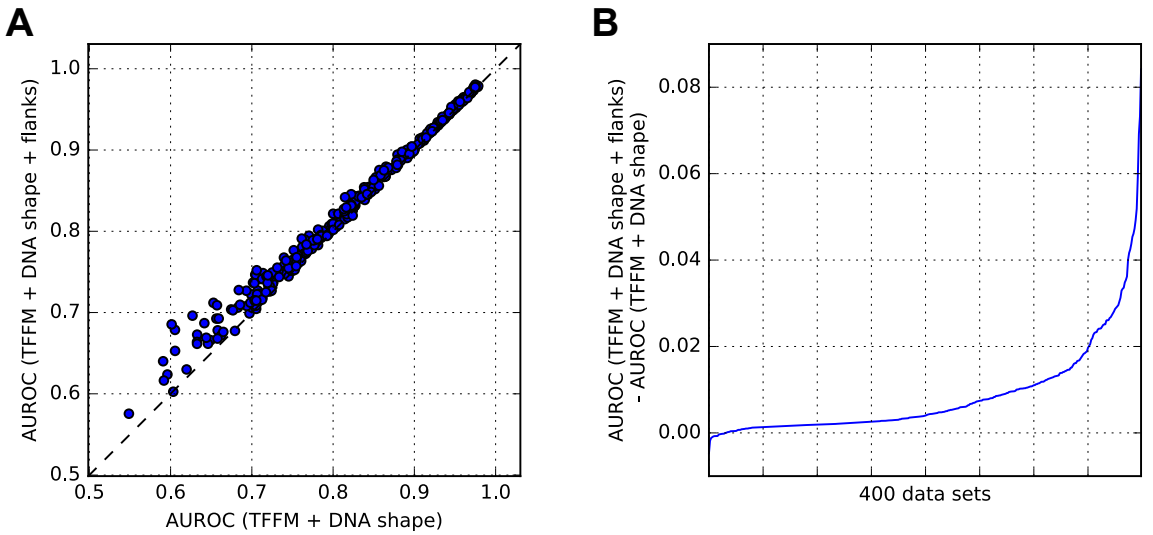
[*] Co-corresponding authors



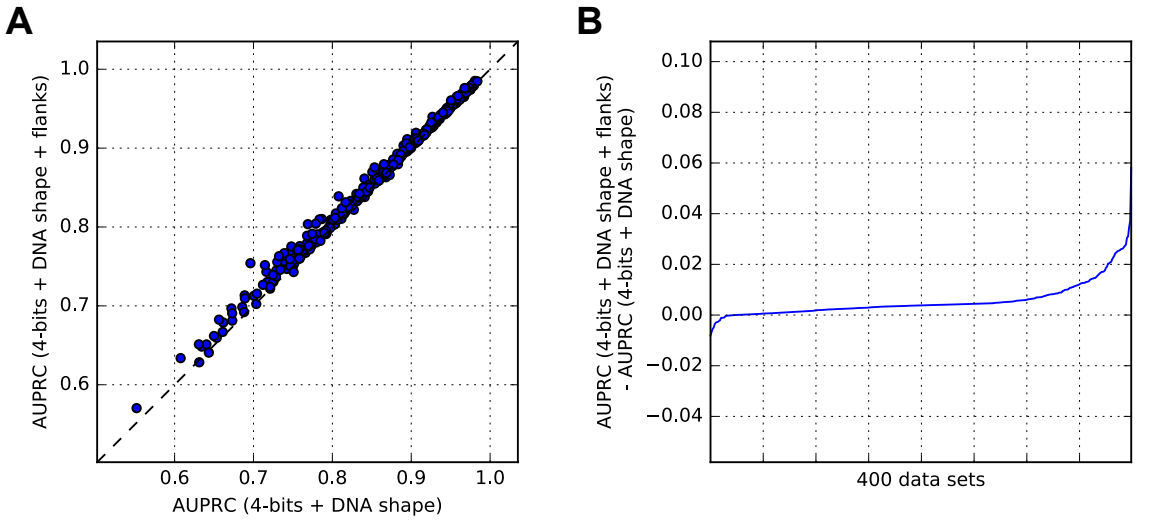
Related to Figure 4. Assessment of the predictive power of DNA shape features at TFBS flanking regions. ***A.*** *Comparison of the AUROC obtained for the 400 human ENCODE ChIP-seq data sets when using the classifiers combining PSSM scores and DNA shape features at the core TFBSs (x-axis) and the classifiers combining PSSM scores and DNA shape features at both the core TFBSs and the surrounding 15 bp on each side (y-axis). The dashed line represents equal AUROC values for the two methods.* ***B.*** *AUROC value differences (y-axis) between the flank-augmented classifiers and the PSSM + DNA shape classifiers. Data sets (x-axis) are ranked by increasing difference values.*



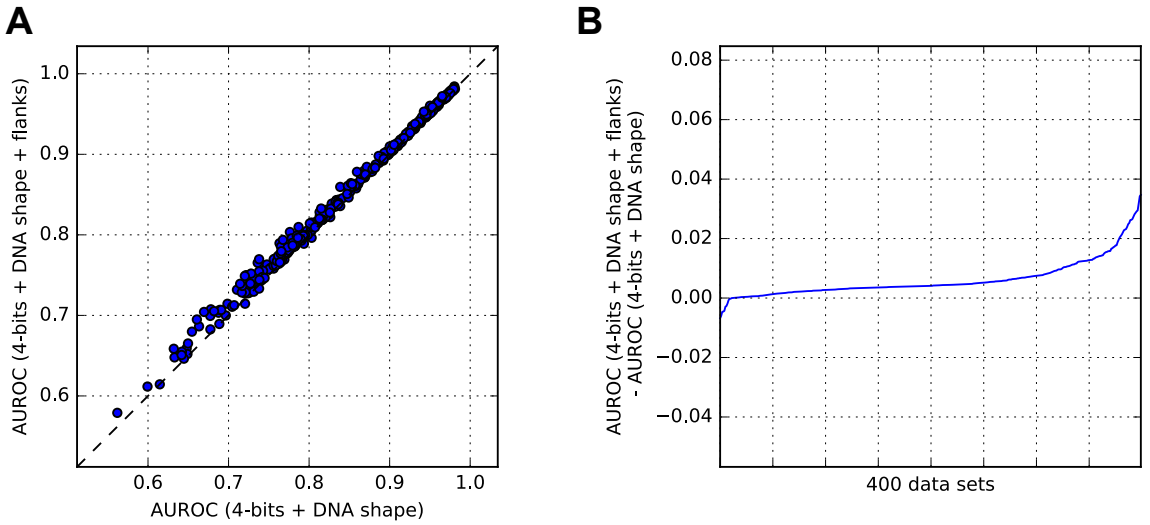
Related to Figure 4. Assessment of the predictive power of DNA shape features at TFBS flanking regions. ***A.*** *Comparison of the AUPRC obtained for the 400 human ENCODE ChIP-seq data sets when using the classifiers combining TFFM scores and DNA shape features at the core TFBSs (x-axis) and the classifiers combining TFFM scores and DNA shape features at both the core TFBSs and the surrounding 15 bp on each side (y-axis). The dashed line represents equal AUPRC values for the two methods.* ***B.*** *AUPRC value differences (y-axis) between the flank-augmented classifiers and the TFFM + DNA shape classifiers. Data sets (x-axis) are ranked by increasing difference values.*



Related to Figure 4. Assessment of the predictive power of DNA shape features at TFBS flanking regions. ***A.*** *Comparison of the AUROC obtained for the 400 human ENCODE ChIP-seq data sets when using the classifiers combining TFFM scores and DNA shape features at the core TFBSs (x-axis) and the classifiers combining TFFM scores and DNA shape features at both the core TFBSs and the surrounding 15 bp on each side (y-axis). The dashed line represents equal AUROC values for the two methods.* ***B.*** *AUROC value differences (y-axis) between the flank-augmented classifiers and the TFFM + DNA shape classifiers. Data sets (x-axis) are ranked by increasing difference values.*



Related to Figure 4. Assessment of the predictive power of DNA shape features at TFBS flanking regions. ***A.*** *Comparison of the AUPRC obtained for the 400 human ENCODE ChIP-seq data sets when using the 4-bits + DNA shape classifiers (considering DNA shape features at the core TFBSs; see (Zhou* et al.*, 2015) for the 4-bits encoding) (x-axis) and the same classifiers where DNA shape features at both the core TFBSs and the surrounding 15 bp on each side were considered (y-axis). The dashed line represents equal AUPRC values for the two methods.* ***B.*** *AUPRC value differences (y-axis) between the flank-augmented classifiers and the 4-bits + DNA shape classifiers. Data sets (x-axis) are ranked by increasing difference values.*



Related to Figure 4. Assessment of the predictive power of DNA shape features at TFBS flanking regions. ***A.*** *Comparison of the AUROC obtained for the 400 human ENCODE ChIP-seq data sets when using the 4-bits + DNA shape classifiers (considering DNA shape features at the core TFBSs; see (Zhou* et al.*, 2015) for the 4-bits encoding) (x-axis) and the same classifiers where DNA shape features at both the core TFBSs and the surrounding 15 bp on each side were considered (y-axis). The dashed line represents equal AUROC values for the two methods.* ***B.*** *AUROC value differences (y-axis) between the flank-augmented classifiers and the 4-bits + DNA shape classifiers. Data sets (x-axis) are ranked by increasing difference values.*