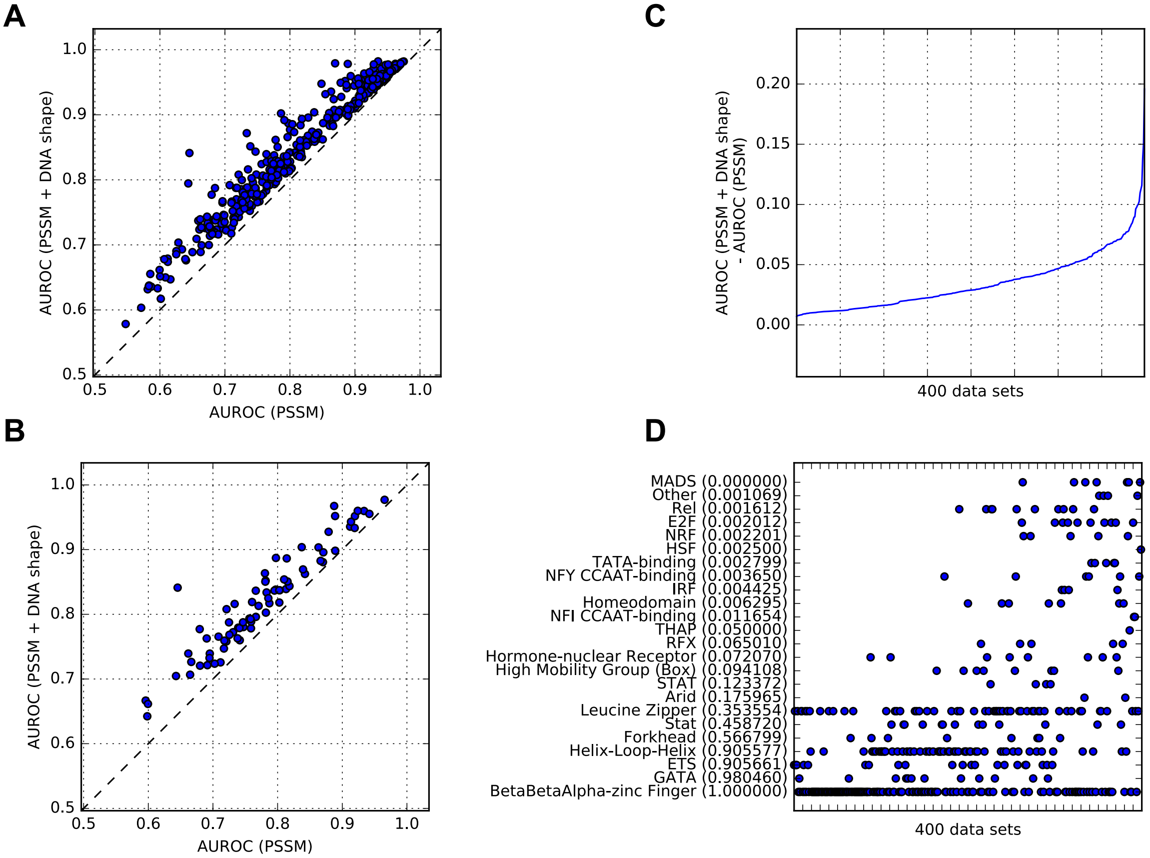
Data S2

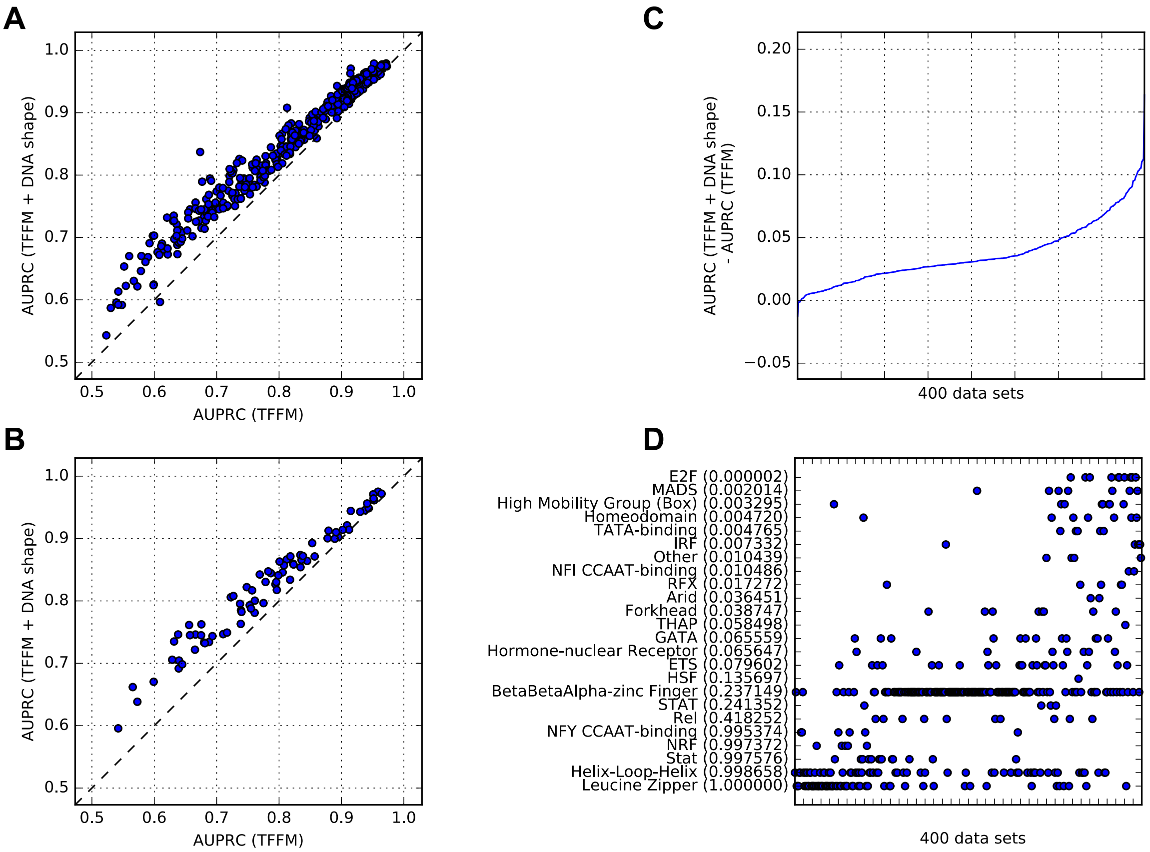
DNA shape features improve transcription factor binding site predictions *in vivo*

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

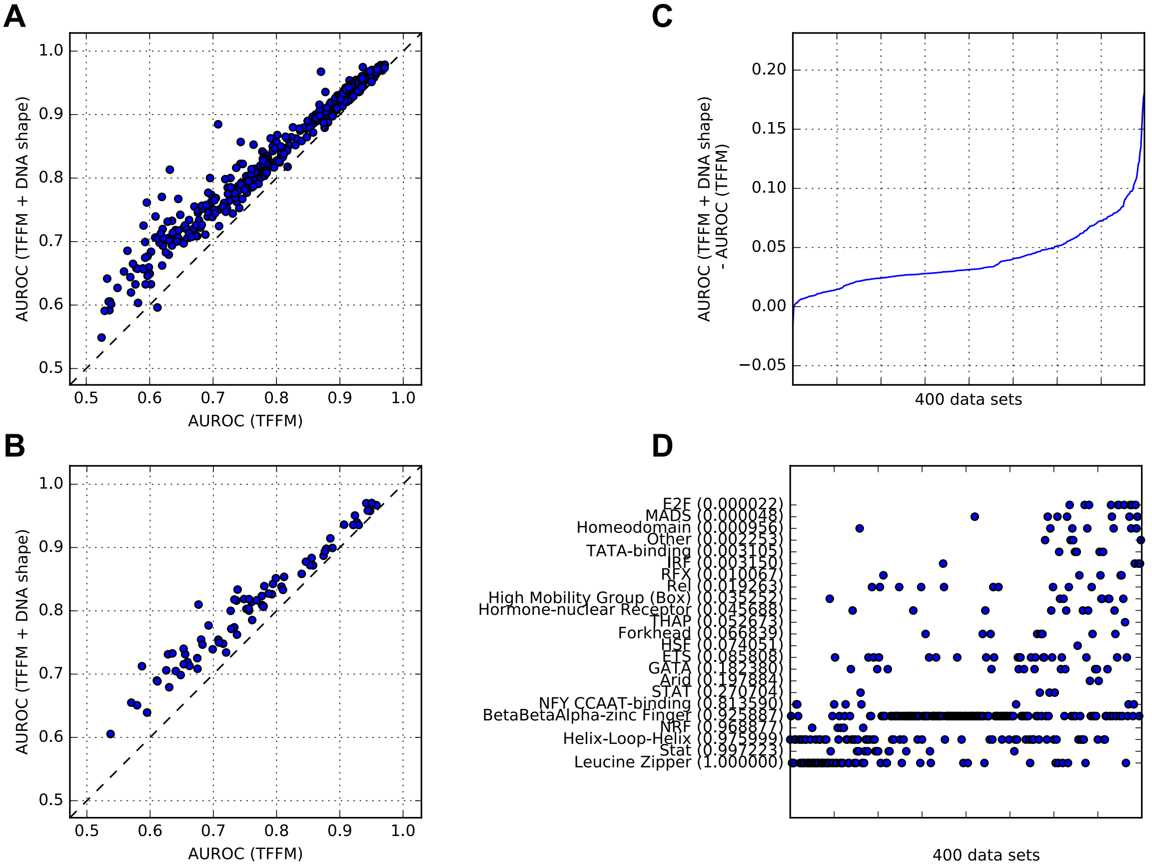
[1] Centre for Molecular Medicine at the Child and Family Research Institute,
Department of Medical Genetics, University of British Columbia,
980 West 28th Avenue, V5Z 4H4, Vancouver, BC, Canada
[2] Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership,
University of Oslo and Oslo University 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



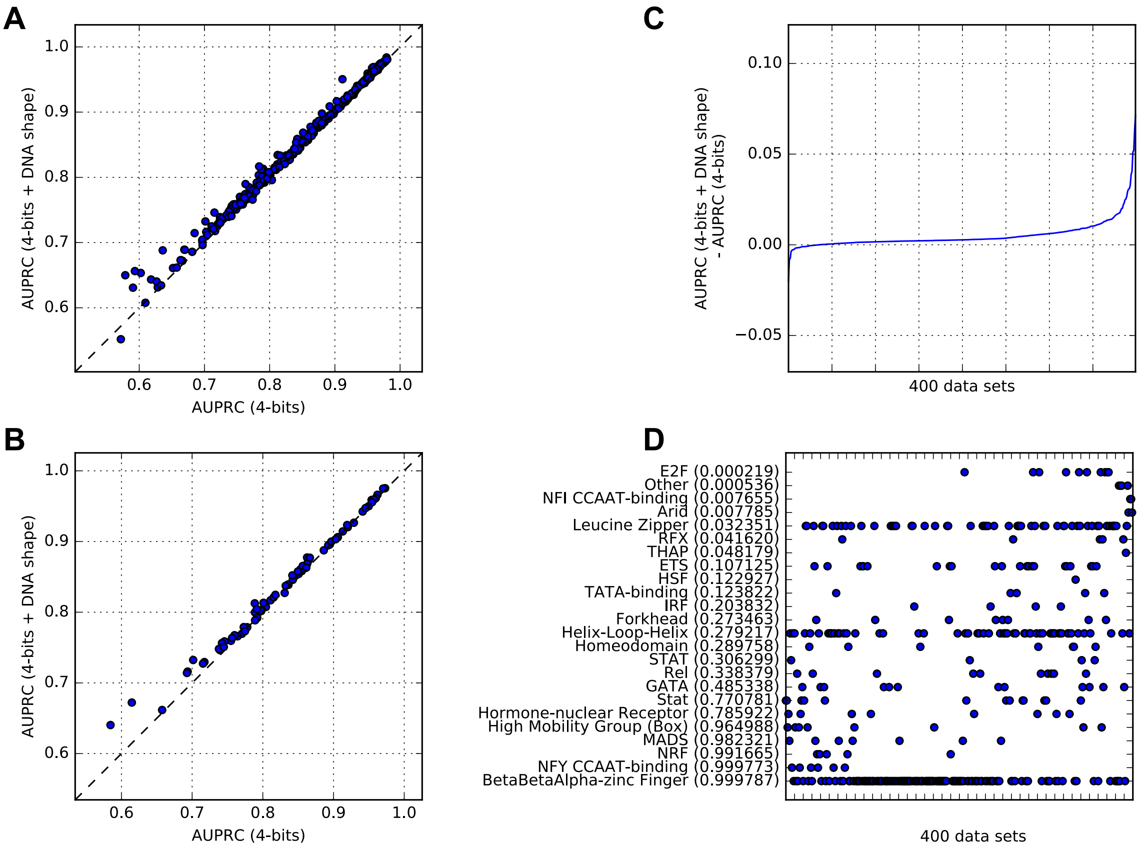
Related to Figure 2. Impact of DNA shape on predicting TFBSs. ***A.*** *Comparison of the AUROC obtained for the 400 ENCODE human ChIP-seq data sets when using either the PSSM scores (x-axis) or the classifiers combining PSSM scores and DNA shape features (y-axis). The dashed line represents equal AUROC values obtained with the two methods.* ***B.*** *Comparison of the median AUROC over all ChIP-seq data sets associated with each TF (one data point per TF) when using either the PSSM scores (x-axis) or the PSSM + DNA shape classifiers (y-axis). The dashed line represents equal AUROC values obtained with the two methods.* ***C.*** *Improvement in predictive power obtained when considering DNA shape features (y-axis) as the difference between the AUROC values obtained with the PSSM + DNA shape classifiers and PSSM scores. The higher the difference, the stronger the improvement when DNA shape information is incorporated. Note that the data sets (x-axis) were ranked by increasing difference values.* ***D.*** *For each TF family (y-axis), an associated data set is represented at the corresponding x-coordinate where the data set appears in B. The name of the TF families are given on the y-axis along with the Mann-Whitney U test p-values of enrichment for significant improvement in parenthesis. Note that the given p-values are not corrected for multiple hypothesis testing.*



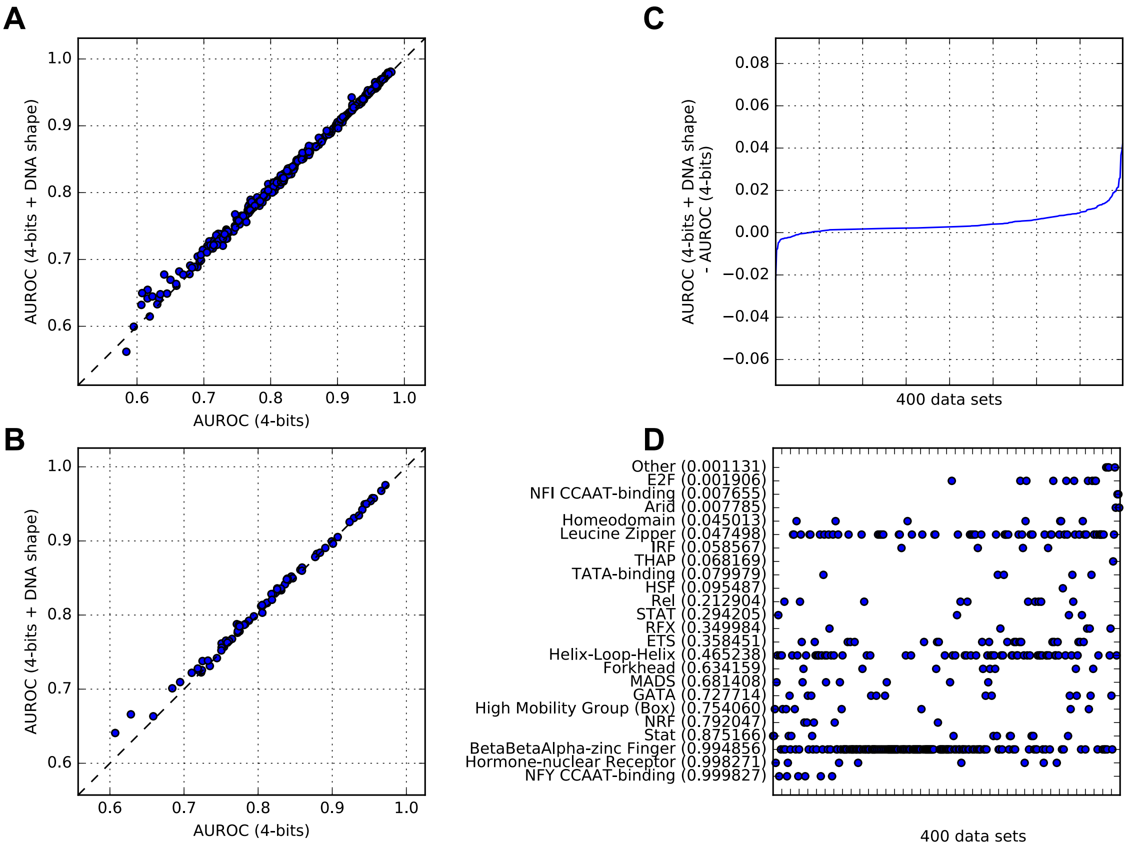
Related to Figure 2. Impact of DNA shape on predicting TFBSs. ***A.*** *Comparison of the AUPRC obtained for the 400 ENCODE human ChIP-seq data sets when using either the TFFM scores (x-axis) or the classifiers combining TFFM scores and DNA shape features (y-axis). The dashed line represents equal AUPRC values obtained with the two methods.* ***B.*** *Comparison of the median AUPRC over all ChIP-seq data sets associated with each TF (one point per TF) when using either the TFFM scores (x-axis) or the TFFM + DNA shape classifiers (y-axis). The dashed line represents equal AUPRC values obtained with the two methods.* ***C.*** *Improvent in predictive power obtained when considering DNA shape features (y-axis) as the difference between the AUPRC values obtained with the TFFM + DNA shape classifiers and TFFM scores. The higher the difference, the stronger the improvement when DNA shape information is incorporated. Note that the data sets (x-axis) were ranked by increasing difference values.* ***D.*** *For each TF family (y-axis), an associated data set is represented at the corresponding x-coordinate where the data set appears in B. The name of the TF families are given on the y-axis along with the Mann-Whitney U test p-values of enrichment for significant improvement in parenthesis. Note that the given p-values are not corrected for multiple hypothesis testing.*



Related to Figure 2. Impact of DNA shape on predicting TFBSs. ***A.*** *Comparison of the AUROC obtained for the 400 ENCODE human ChIP-seq data sets when using either the TFFM scores (x-axis) or the classifiers combining TFFM scores and DNA shape features (y-axis). The dashed line represents equal AUROC values obtained with the two methods.* ***B.*** *Comparison of the median AUROC over all ChIP-seq data sets associated with each TF (one point per TF) when using either the TFFM scores (x-axis) or the TFFM + DNA shape classifiers (y-axis). The dashed line represents equal AUROC values obtained with the two methods.* ***C.*** *Improvement in predictive power obtained when considering DNA shape features (y-axis) as the difference between the AUROC values obtained with the TFFM + DNA shape classifiers and TFFM scores. The higher the difference, the stronger the improvement when DNA shape information is incorporated. Note that the data sets (x-axis) were ranked by increasing difference values.* ***D.*** *For each TF family (y-axis), an associated data set is represented at the corresponding x-coordinate where the data set appears in B. The name of the TF families are given on the y-axis along with the Mann-Whitney U test p-values of enrichment for significant improvement in parenthesis. Note that the given p-values are not corrected for multiple hypothesis testing.*



Related to Figure 2. Impact of DNA shape on predicting TFBSs. ***A.*** *Comparison of the AUPRC obtained for the 400 ENCODE human ChIP-seq data sets when using either 4-bits (x-axis) or 4-bits + DNA shape (y-axis) classifiers (see (Zhou* et al.*, 2015) for the 4-bits encoding). The dashed line represents equal AUPRC values obtained with the two methods.* ***B.*** *Comparison of the median AUPRC over all ChIP-seq data sets associated with each TF (one point per TF) when using either 4-bits (x-axis) or 4-bits + DNA shape (y-axis) classifiers. The dashed line represents equal AUPRC values obtained with the two methods.* ***C.*** *Improvent in predictive power obtained when considering DNA shape features (y-axis) as the difference between the AUPRC values obtained with the 4-bits + DNA shape and 4-bits classifiers. The higher the difference, the stronger the improvement when DNA shape information is incorporated. Note that the data sets (x-axis) were ranked by increasing difference values.* ***D.*** *For each TF family (y-axis), an associated data set is represented at the corresponding x-coordinate where the data set appears in B. The name of the TF families are given on the y-axis along with the Mann-Whitney U test p-values of enrichment for significant improvement in parenthesis. Note that the given p-values are not corrected for multiple hypothesis testing.*



Related to Figure 2. Impact of DNA shape on predicting TFBSs. ***A.*** *Comparison of the AUROC obtained for the 400 ENCODE human ChIP-seq data sets when using either 4-bits (x-axis) or 4-bits + DNA shape (y-axis) classifiers (see (Zhou* et al.*, 2015) for the 4-bits encoding). The dashed line represents equal AUROC values obtained with the two methods.* ***B.*** *Comparison of the median AUROC over all ChIP-seq data sets associated with each TF (one point per TF) when using either 4-bits (x-axis) or 4-bits + DNA shape (y-axis) classifiers. The dashed line represents equal AUROC values obtained with the two methods.* ***C.*** *Improvement in predictive power obtained when considering DNA shape features (y-axis) as the difference between the AUROC values obtained with the 4-bits + DNA shape and 4-bits classifiers. The higher the difference, the stronger the improvement when DNA shape information is incorporated. Note that the data sets (x-axis) were ranked by increasing difference values.* ***D.*** *For each TF family (y-axis), an associated data set is represented at the corresponding x-coordinate where the data set appears in B. The name of the TF families are given on the y-axis along with the Mann-Whitney U test p-values of enrichment for significant improvement in parenthesis. Note that the given p-values are not corrected for multiple hypothesis testing.*