DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo

Anthony Mathelier,1,2,3 Beibei Xin,2 Tsu-Pei Chiu,4 Lin Yang,4 Remo Rohs,4,* and Wyeth W. Wasserman1,*
1Centre for Molecular Medicine and Therapeutics at the Child and Family Research Institute, Department of Medical Genetics, University of British Columbia, 980 West 28th Avenue, Vancouver, BC V5Z 4H4, Canada
2Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo and Oslo University Hospital, 0318 Oslo, Norway
3Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, 0372 Oslo, Norway
4Molecular and Computational Biology Program, Departments of Biological Sciences, Chemistry, Physics, and Computer Science, University of Southern California, Los Angeles, CA 90089, USA
*Correspondence: rohs@usc.edu (R.R.), wyeth@cmmt.ubc.ca (W.W.W.)
http://dx.doi.org/10.1016/j.cels.2016.07.001

SUMMARY

Interactions of transcription factors (TFs) with DNA comprise a complex interplay between base-specific amino acid contacts and readout of DNA structure. Recent studies have highlighted the complementarity of DNA sequence and shape in modeling TF binding in vitro. Here, we have provided a comprehensive evaluation of in vivo datasets to assess the predictive power obtained by augmenting various DNA sequence-based models of TF binding sites (TFBSs) with DNA shape features (helix twist, minor groove width, propeller twist, and roll). Results from 400 human ChIP-seq datasets for 76 TFs show that combining DNA shape features with position-specific scoring matrix (PSSM) scores improves TFBS predictions. Improvement has also been observed using TF flexible models and a machine-learning approach using a binary encoding of nucleotides in lieu of PSSMs. Incorporating DNA shape information is most beneficial for E2F and MADS-domain TF families. Our findings indicate that incorporating DNA sequence and shape information benefits the modeling of TF binding under complex in vivo conditions.

INTRODUCTION

One of many mechanisms that control gene expression, transcriptional regulation involves transcription factors (TFs) as key proteins (Jacob and Monod, 1961; Ptashne and Gann, 1997). Most TFs are sequence-specific DNA binding proteins that recognize specific genome positions through a complex interplay between nucleotide-amino-acid contacts (base readout) and readout of DNA structure (shape readout) (Slattery et al., 2014). Deciphering how TFs identify and bind specific target sequences—the TF binding sites (TFBSs)—is a key challenge in understanding transcriptional gene regulation (Dror et al., 2016; Wasserman and Sandelin, 2004; Zambelli et al., 2013).

TFBSs are short and often degenerate sequence motifs. These characteristics make it computationally difficult to model and predict TFBSs at the genomic scale (Badis et al., 2009). Moving beyond initial consensus sequence methods, the classical computational model to describe TFBSs is the position-specific scoring matrix (PSSM), which uses an additive method to summarize frequencies of every nucleotide at each position of the TFBS (Stormo, 2013). These second-generation models, however, do not capture position interdependencies or variable spacing. Therefore, several experimental assays have been designed to unravel characteristics of TF-DNA interactions at the large scale. In vitro high-throughput (HT) binding assays, such as protein binding microarrays (PBMs) (Berger et al., 2006), HT SELEX (Jolma et al., 2010; Zhao et al., 2009), and SELEX-seq (Slattery et al., 2011), expose DNA sequences selected by TFs and reveal their binding preferences. Chromatin immunoprecipitation sequencing (ChIP-seq) represents the in vivo counterpart of these in vitro assays, allowing for the identification of DNA regions bound by a targeted TF at the genomic scale (Johnson et al., 2007).

Large-scale data derived from HT experiments highlight higher-order positional interaction features of TFBSs that cannot be captured by classical PSSMs, even though the methods based on these traditional models perform quite well (Weirauch et al., 2013). Recently, computational advances have used experimental assays to construct sophisticated models that capture the broad range of TFBS representations. For instance, PSSMs have been extended to dinucleotides to capture interrelationships within TFBSs (Siddharthan, 2010). Using PBM data, binding energy models include energy parameters to describe contributions of dinucleotides to binding affinity (Zhao et al., 2012). These models describe TF-DNA binding specificity well in cases in which PSSMs have performed insufficiently. Utilizing ChIP-seq data, we developed the TF flexible model (TFFM) framework to improve in vivo prediction of TFBSs (Mathelier and Wasserman, 2013). TFFMs capture interdependencies of successive nucleotides within TFBSs and the flexible length of TFBSs within a single hidden Markov model framework.

The above-mentioned third-generation methods enable TFBS prediction by representing sequence properties. A parallel approach utilizes the 3D DNA structure, or DNA shape, to capture, at least in part, the interdependencies between nucleotide
positions within TFBSs (Gordan et al., 2013; Tsai et al., 2015; Yang and Ramsey, 2015; Zhou et al., 2015). Large-scale DNA structural information can be computed by the DNAshape method (Zhou et al., 2013), which computes four DNA shape features: helix twist (HeIT), minor groove width (MGW), propeller twist (ProT), and Roll. Recent studies have demonstrated the complementary role of DNA sequence and shape information in determining protein-DNA binding specificity in vitro (Joshi et al., 2007; Rohs et al., 2009; Slattery et al., 2011). For example, the binding specificity of Hox proteins was analyzed using SELEX-seq data to show the direct role of DNA shape features in protein-DNA readout (Abe et al., 2015). Using PBM and SELEX-seq data, we showed that complementing DNA sequence with shape information enhanced the prediction of TF binding affinities (Zhou et al., 2015). DNA shape information at regions flanking core binding sites was highly predictive of differential binding derived from BunDLE-seq assays (Levo et al., 2015).

While previous works have demonstrated that models combining DNA sequence and shape improve quantitative models of TF binding in vitro, we have addressed here three key questions:

1. Do more complex in vivo protein-DNA interactions exhibit similar properties?
2. When DNA shape properties are integrated with sequence-based TFBS prediction methods, do we observe an improvement in performance?
3. Do specific TF families benefit more than others from the integration of DNA shape features in TF binding models?

Here, we have capitalized on the availability of DNA shape information extracted from GBshape (Chiu et al., 2015), our genome browser database of DNA shape features computed from our DNAshape prediction tool (Zhou et al., 2013), at TF-bound regions derived from ChIP-seq experiments to address the three aforementioned questions.

RESULTS

Machine Learning Models Combining DNA Sequence and Shape Features
To assess the effects of including DNA structural information in predictions of TFBSs in ChIP-seq datasets, we developed a computational framework combining DNA sequence and shape information to model and predict TFBSs. The availability of numerous ChIP-seq regions enables the application of a discriminative supervised machine learning approach (Libbrecht and Noble, 2015). Specifically, a DNA sequence that is considered as a potential TFBS was represented by a (feature) vector that combined 1 to 4n features that encode sequence information and 8n features that capture DNA shape information, where n is DNA sequence length. We encoded DNA sequence information of the putative TFBS by using either the PSSM or TFFM score computed from the sequence, or a binary encoding using 4 bits per nucleotide (Zhou et al., 2015). DNA shape-related features are the predicted values of HeIT, MGW, ProT, and Roll at each position of the TFBS, extracted from GBshape (Chiu et al., 2015). The vector was further augmented with four second-order shape features that capture structural dependencies at adjacent nucleotide positions (Zhou et al., 2015) (Figure 1). Assuming that each ChIP-seq region contains a TFBS, we constructed a feature vector for the best hit per ChIP-seq peak and background region predicted by a TF binding profile (PSSM or TFFM) to train a classifier. To discriminate between TF bound (ChIP-seq) and unbound (background) regions, we used a gradient boosting classifier, which is an ensemble machine learning classifier that combines multiple weak learners to improve predictive power (Friedman et al., 2001). The gradient boosting classifier was based on decision trees that, given an input feature vector, output the probability of the feature vector to be associated with a ChIP-seq peak or a background region. This approach naturally handles heterogenous features (e.g., DNA sequence and shape information), is robust to outliers, and is able to manage irrelevant input, such as noise from ChIP-seq experiments (Friedman et al., 2001).

Classifiers combining PSSM score, TFFM score, or 4-bits nucleotide encoding with DNA shape features are referred to as PSSM + DNA shape, TFFM + DNA shape, or 4-bits + DNA shape classifiers, respectively. Open-source Python software for generating and using these classifiers is provided at https://github.com/amathelier/DNAshapedTFBS.

Incorporating DNA Shape Features Improves TFBS Prediction in Human In Vivo Datasets
We compiled a set of 400 uniformly processed human ENCODE ChIP-seq datasets for which a JASPAR TF binding profile (Mathelier et al., 2014) was available for the corresponding immunoprecipitated (ChIPed) TF (Data S1). These datasets, covering 76 TFs, were used to compare the predictive powers of three computational models that consider DNA sequence information alone with their DNA shape-augmented classifiers. The first two DNA sequence-based models are PSSMs and TFFMs, which are widely used to score TFBSs in ChIP-seq...
datasets. The third model, the 4-bits classifier, is a discriminative model that uses a binary encoding of DNA sequence information (Zhou et al., 2015).

Here, the predictive power of a model refers to its ability to discriminate ChIP-seq regions (defined as the 50-bp region surrounding each side of the ChIP-seq peak maximum) from matched background sequences. The 50-bp regions were selected because they are enriched for TFBSs (Worsley Hunt et al., 2014; Wilbanks and Facciotti, 2010). To avoid sequence composition biases, we selected each set of background sequences to match either the G+C (%GC) content or dinucleotide composition of the ChIP-seq regions. Unless otherwise indicated, background sequences matching the %GC content of ChIP-seq regions were used in the following results.

Predictive powers of PSSM scores and PSSM + DNA shape classifiers were assessed through 10-fold cross-validation (CV). We optimized the PSSMs derived from JASPAR TF binding profiles with the perceptron algorithm using the DIOMO tool (Patel and Stormo, 2014) on the constructed foreground and background training sets (range: 495–83,123, median: 15,171, mean: 21,098, SD: 17,220 sequences). Parameters of PSSM + DNA shape classifiers were learned from the same training sets. Vectors used by the classifiers for a ChIP-seq region correspond to the combination of the best PSSM score in the region and the 8n DNA shape feature values computed for this hit. To assess predictive power, we varied the threshold for scores to compute the recall (sensitivity), specificity, and precision values. Areas under the precision and recall curve (AUPRC) were computed for each model on each ChIP-seq dataset to evaluate predictive power. Unless otherwise noted, we provide the AUPRC values and the p values for significance calculated by a Wilcoxon signed-rank test.

By comparing AUPRC values derived from the PSSM scores or PSSM + DNA shape classifiers, we found that shape-augmented classifiers performed better for all 400 ChIP-seq datasets ($p = 2.7 \times 10^{-67}$; Figure 2A). By considering the median AUPRC values per TF over all ChIP-seq datasets associated with the TF, we observed consistent improvement for all TFs when DNA shape features were incorporated ($p = 3.6 \times 10^{-14}$; Figure 2B).

We computed the difference of discriminative power between the two models (Figure 2C) to assess the improvement obtained by using the PSSM + DNA shape classifiers. Predictive power improvement obtained when considering DNA shape features (y axis) as the difference between AUPRC values obtained with PSSM + DNA shape classifiers and PSSM scores. Larger difference corresponds to stronger improvement. Datasets (x axis) are ranked by increasing difference values.

Finally, we compared the 4-bits and 4-bits + DNA shape classifiers, which were trained and tested on sequences of the highest-scoring hit per ChIP-seq region derived from the PSSMs. DNA shape-augmented classifiers performed consistently better than 4-bits classifiers for 365/400 ChIP-seq datasets ($p = 2.7 \times 10^{-57}$) and 70/76 TFs ($p = 1.3 \times 10^{-12}$) when considering the median AUPRC values (Data S2).

We confirmed the improvement in discriminative power of the models incorporating DNA shape features by considering background sequences matching the dinucleotide composition of ChIP-seq regions (Data S3) and TF-bound regions recurrently found in multiple ChIP-seq datasets for the same TF (Data S4).

The relative improvement obtained when incorporating DNA shape information varied depending on the baseline DNA sequence-based approach. Unsurprisingly, the 4-bits + DNA shape classifiers exhibited a smaller improvement over the 4-bits classifiers compared to the shape-based improvements.
obtained with PSSMs and TFFMs. The higher baseline performance of the 4-bits method is consistent with the superiority of discriminative over generative models to distinguish bound from unbound regions in ChIP-seq (Libbrecht and Noble, 2015) (Figure 3A; Data S5). Nonetheless, PSSM + DNA shape classifiers performed consistently better than 4-bits + DNA shape classifiers for 344/400 datasets ($p = 7.7 \times 10^{-43}$, Figure 3B) and 64/76 TFs ($p = 1.0 \times 10^{-8}$; Data S5).

Although 4-bits classifiers outperformed PSSM scores, the higher discriminative power of PSSM + DNA shape compared to 4-bits + DNA shape classifiers reinforces the capacity of DNA shape features to improve TFBS predictions in ChIP-seq datasets. Importantly, the combination of sequence information (captured by PSSMs, TFFMs, or 4-bits classifiers) with DNA shape properties performed better than generative (PSSM and TFFM) and discriminative (4-bits classifier) approaches modeling DNA sequence, indicating that DNA shape provides additional information.

Although the utility of DNA shape to predict TFBSs was reported before (Abe et al., 2015; Yang and Ramsey, 2015; Yang et al., 2014; Zhou et al., 2015), we provide evidence, from an extensive collection of 400 human in vivo datasets for 76 TFs, that this observation is generalizable and relevant to noisy environments and data (Fan and Struhl, 2009; Worsley Hunt et al., 2014; Worsley Hunt and Wasserman, 2014; Jain et al., 2015; Park et al., 2013; Teytelman et al., 2013).

**Figure 3. Predictive Power of PSSM and 4-Bit Approaches for TFBSs in ChIP-Seq Regions**

(A and B) AUPRC values obtained for 400 human ENCODE ChIP-seq datasets, obtained by considering PSSM scores (x axis) and 4-bits classifiers (y axis) (A) or PSSM + DNA shape (x axis) and 4-bits + DNA shape (y axis) classifiers (B). See also Data S5.

**DNA Shape at Genomic Flanking Regions Improves TFBS Predictions In Vivo**

Sequences immediately flanking TFBSs have been previously shown to contribute to TF binding specificity (Gordân et al., 2013), which is determined, in part, by DNA shape outside the core binding sites (Afek et al., 2014; Barozzi et al., 2014; Dror et al., 2015). We extended our DNA shape-augmented models to consider eight DNA shape features at 15-bp-long regions 5’ and 3’ of the TFBSs, as in Barozzi et al. (2014).

Augmenting DNA shape-based classifiers with additional DNA shape information from flanking sequences has improved the discriminatory power of classifiers trained using 10-fold CV for 378 (~94%), 373 (~93%), and 375 (~94%) datasets compared to PSSM + DNA shape, TFFM + DNA shape, and 4-bits + DNA shape classifiers, respectively (Figure 4; Data S6). Our findings agree with results from in vitro studies of the role of flanking regions in TF-DNA binding (Dror et al., 2016; Gordân et al., 2013; Levo et al., 2015).

**E2F and MADS-Domain TF Families Benefit Most from DNA Shape Information**

Next, motivated by the observation that DNA structural information improves the prediction of TFBSs for some ChIP-seq datasets more than others (Figure 2C; Data S2 and S3), we investigated whether predictions for certain TF families with similar DNA-binding domains specifically benefit from incorporating DNA shape information.

By using JASPAR (Mathelier et al., 2014), we extracted TF family information of DNA binding domains associated with the

**Figure 4. Predictive Power of DNA Shape Features at TFBS Flanking Regions**

(A) AUPRC values obtained for 400 human ENCODE ChIP-seq datasets when using classifiers combining PSSM scores and DNA shape features at core TFBSs (x axis), or classifiers combining PSSM scores and DNA shape features at both core TFBSs and surrounding 15 bp on each side (y axis). Dashed line represents equal AUPRC values for both methods.

(B) AUPRC value differences (y axis) between flank-augmented classifiers and PSSM + DNA shape classifiers (x axis). Datasets (x axis) are ranked by increasing difference values. See also Data S6.
400 human ChIP-seq experiments. In aggregate, we analyzed profiles derived from DNA binding domains associated with 24 TF families, which were associated with TFs in the JASPAR database using a classification scheme (Fulton et al., 2009) (Data S1). By comparing the predictive powers of DNA shape-augmented classifiers to those of DNA sequence-based approaches, we assessed the enrichment of a TF family for larger AUPRC or AUROC difference values using the one-sided Mann-Whitney U test. Predictive power comparisons were performed considering %GC- and dinucleotide-matched background sets. Depending on the DNA sequence-based approach (PSSM, TFFM, or 4-bits), background type (%GC- or dinucleotide-matched), and assessment method (AUPRC or AUROC), we observed enrichment (p < 4.17 × 10^-14, with Bonferroni correction for desired α < 0.01) for different TF families (Figure 2D; Data S1 and S2). Taken together, the results across all three DNA sequence-based approaches suggest that the E2F and MADS-domain TF datasets benefited the most from inclusion of DNA shape information (Figure 2D; Data S1 and S2). It is noteworthy that results for the E2F and MADS-domain TF datasets were not derived from a single TF, but were consistent over several TFs from the same family. Namely, the ten E2F-associated ChIP-seq datasets were derived from E2F1 (three datasets), E2F4 (four datasets), and E2F6 (three datasets), and seven MADS-domain-associated ChIP-seq datasets were derived from MEF2A (two datasets), MEF2C (one dataset), and SRF (four datasets).

To confirm the results obtained when incorporating DNA shape features for MADS-domain TFs, we considered an independent set of seven Arabidopsis thaliana TFs (Heyndrickx et al., 2014), for which we had JASPAR TF binding profiles (Data S1). Similar to human MADS-domain TF results (Data S7), we observed improved discriminative power when DNA structural information was considered for plant MADS-domain TFs with PSSM + DNA shape and TFFM + DNA shape models (Data S7). Only the two smallest datasets (94 sequences for FLC and 54 for SVP in training sets) showed decreased discriminative power with DNA shape-augmented classifiers. For one of the three approaches, the 4-bits + DNA shape classifiers, we found no improvement in predictive power for plant MADS-domain TF datasets compared to the 4-bits classifiers (Data S7).

Taken together, we observed that, among all studied protein families, TFs from the E2F and MADS-domain families most benefited from the inclusion of DNA shape information in the classifiers when compared to the three DNA-sequence-based models.

**Structural Analyses of E2F and MADS-Domain TF Binding Specificities**

**DNA Shape Readout Contributes to E2F-DNA Binding**

Given our observed improvement in predictive power based on DNA shape information, we next characterized the specific DNA shape features that contributed the most to the improvement in predictive accuracy for the family of E2F TFs. We extracted the importance of each feature learned by the DNA shape-augmented classifiers by combining DNA sequence and shape information for the E2F TFs. To consider the same DNA sequence-based model per TF for all associated ChIP-seq datasets, we selected PSSMs derived from the corresponding JASPAR TF binding profiles. To simplify interpretation, we considered classifiers based on sequence and first-order DNA shape features (Zhou et al., 2015). Figures S1 and S2 plot the average feature importance measures obtained over the 10-fold CV training for all ChIP-seq datasets associated with the TFs.

Although the PSSM score was consistently the most important feature (Figure S1), several DNA shape features at different positions were important for TFBS predictions. A commonality
among the three E2Fs was the contribution of ProT (and, to lesser extents, HelT and MGW) at proximal flanks of the TFBSs among the three E2Fs was the contribution of ProT (and, to lesser extents, HelT and MGW) at proximal flanks of the TFBSs. This intricate system of eight hydrogen bonds enables bidentate hydrogen bonds with guanine bases of the core binding sites. This intricate system of eight hydrogen bonds enables bidentate hydrogen bonds with guanine bases of the core binding sites. The angular orientation of the arginine side chains stabilized by other amino acids selects for rotational parameters of the contacted G/C bp through hydrogen bond geometries. This observation is reflected by the importance given to rotational parameters, such as ProT, in our models. Similar results were reported in a recent study of diverse protein families (Dror et al., 2015).

**MADS-Domain TFs Recognize Position-Specific DNA Shape**

As described above for E2Fs, we individually considered DNA shape features in the DNA shape-augmented classifiers for the MADS-domain TF ChIP-seq datasets. Figure 5B highlights the importance of ProT for improving the discriminative power of models associated with human and plant MADS-domain TFs. The FLC and SVP datasets, which had small numbers of training sequences, were the only ones for which ProT was not the most important shape feature. Inclusion of a single DNA shape category ensured that DNA shape features did not compensate for each other when considering all four shape features due to dependencies among different features. Previous work showed the important role of MGW at the A-tract of the MADS-box for DNA-binding (Muiño et al., 2014). Our models captured this importance, although ProT remained the most important shape feature.

We extracted feature importance measures at each position within the TFBSs learned by the PSSM + DNA shape classifiers for human and plant MADS-domain TFs (Figures S4, S4, and S6). The most important DNA shape features for discriminating ChIP-seq bound sites from background genomic regions were ProT and Roll at specific positions within the MADS-box TFBSs (in agreement with Figure 5B). This observation was consistent across all of the human MADS-domain TFs, whereas the signal was more diffuse for plant MADS-domain TFs. DNA shape-augmented classifiers associated with human MADS-domain TF ChIP-seq datasets obtained the strongest discriminative improvements over sequence-based models (Figure 7A).

As an example, we plotted feature importance measures and sequence logos of the TF binding profiles associated with SRF and MEF2C (Figures 7B and 7C). These two TFs were associated with datasets that showed the strongest improvements in discriminative power when incorporating DNA shape features for the classification of bound vs. unbound sites (Figure 7A). ProT features seemed to contribute to the binding of the core CArG-box (CCW6GG) (red squares in Figures 7B, 7C, and S7), whereas Roll features were highlighted at the edges of the MADS-box core motif (blue squares in Figures 7B, 7C, and S7).

Structural analyses of various complexes of MADS-domain TFs with their DNA target sites suggested that protein residues recognize specific DNA conformations. Comparison of the feature importance measures for different structural features of the DNA binding sites of human MADS-domain proteins indicated a contribution of ProT to binding specificity (Figures 6, 7, S3B, and S7). Human SRF ChIP-seq datasets were consistently associated with strongest improvement in discriminative power when considering the DNA shape-augmented classifiers to predict TFBSs in TF-bound regions (Figure 7A). Hence, we analyzed the co-crystal structures available for complexes of human SRF and MEF2.

The SRF homodimer uses lysine residues to form base-specific hydrogen bonds with C/G bp in the CArG motif (Figures 6C and
DNA bends around the protein, and intrinsic structural features are likely responsible for this deformation. As our models indicate a contribution of ProT, we compared ProT observed in the co-crystal structure of the SRF-DNA complex (PDB: 1SRS) (Pellegrini et al., 1995) with the prediction of ProT for the unbound motif using our DNAshape method (Zhou et al., 2013). The ProT pattern in the unbound target site resembled the pattern in the protein-DNA complex (Figure 6E).

This observation suggests that ProT is intrinsic to the binding site and likely selected by the TF. A less negative ProT within the two adjacent G/C bp at each flank of the motif optimizes the geometry of hydrogen bonds between the lysine side chains and bases (Figure 6D). This preference is coupled to the sequence and selection of functional groups of the bases, whereas structural features enhance the energetics of these specific contacts.

**DISCUSSION**

Here, we used a machine learning classifier-based approach to demonstrate that combining DNA sequence information with DNA shape features improves discrimination between TF-bound sites in vivo and background genomic regions. These in vivo analyses complement our previous in vitro studies showing that inclusion of DNA shape properties can improve the accuracy of TF binding site prediction (Abe et al., 2015; Yang et al., 2014; Zhou et al., 2015).

Our computational analyses of 400 human ChIP-seq datasets revealed that when DNA shape features were incorporated in the models, the E2F and MADS-domain TF families showed the largest improvement in TFBS prediction accuracies. TF families with the strongest predictive power improvements varied depending on the baseline model, background type, and assessment measure considered. These results highlight the importance of considering multiple background types and assessment measures when comparing models.

Whereas most bioinformatics tools rely on PSSMs for TFBS predictions, our findings imply that the field should consider more sophisticated modeling methods. For TFs for which only a small number of experimentally derived TF-bound regions are available, traditional PSSMs represent a reasonable alternative, as exemplified by the FLC and SVP plant TFs in our study. We envision that future tools relying on TFBS predictions will incorporate the most appropriate model for each TF or TF family.

With the increasing trove of whole-genome sequencing data, the identification of variants altering gene regulation through disruption of TF-DNA interactions has become an important challenge. Recent approaches have focused on allelic differences in PSSM scores to predict the functional impact of...
variants within TFBSs (Chen et al., 2016; Mathelier et al., 2015). Our study confirms that some TFs critically rely on DNA shape readout for TFBS recognition. Future work should assess how the incorporation of DNA structural properties could help to predict the impact of variants disrupting TFBSs.

**STAR METHODS**

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **CONTACT FOR REAGENT AND RESOURCE SHARING**
- **METHOD DETAILS**
  - ChiP-Seq Datasets and TF Binding Profiles
  - ChiP-Seq Peak Sequences
  - Discriminative Power Assessment
  - DNA Shape Features
  - Gradient Boosting Classifiers
  - Scanning DNA Sequences with TFFMs and PSSMs
  - TF Family Annotation
  - Feature Importance Measures
- **QUANTIFICATION AND STATISTICAL ANALYSIS**
  - Wilcoxon Signed-Rank Tests
  - Mann-Whitney U Tests
- **DATA AND SOFTWARE AVAILABILITY**
  - Software

**SUPPLEMENTAL INFORMATION**

Supplemental Information includes seven figures and seven data files and can be found with this article online at http://dx.doi.org/10.1016/j.cels.2016.07.001.

**AUTHOR CONTRIBUTIONS**

A.M., W.W.W., and R.R. conceived and designed the project. A.M. implemented and performed the experiments. A.M., W.W.W., and R.R. analyzed and interpreted the results. A.M. implemented the DNAshapedTFBS tool. B.X. performed experiments and analyzed data related to the 4-bits-based classifiers, with contributions from T.-P.C. T.-P.C. also generated tools for DNA shape analysis. L.Y. analyzed and interpreted structural data. A.M., W.W.W., and R.R. conceived and designed the project. A.M. implemented the DNAshapedTFBS tool. A.M., W.W.W., and R.R. analyzed and performed the experiments. A.M., W.W.W., and R.R. analyzed and interpreted the results. A.M. conceived and designed the project. A.M. implemented the DNAshapedTFBS tool. A.M., W.W.W., and R.R. conceived and designed the project.

**ACKNOWLEDGMENTS**

We thank Miroslav Hatas for systems support, Dora Pak for management support, François Parcy, Oriol Fornés Crespo, and Chih-Yu Chen for comments on the manuscript, and all members of the W.W.W. and R.R. labs for insightful discussions. A.M. and W.W.W. were supported by the Genome Canada Large scale Applied Research Grant (174CDE). Funding was provided by the Child and Family Research Institute, Vancouver, and the British Columbia Children’s Hospital Foundation (to A.M. and W.W.W.). A.M. was also supported by funding from the Norwegian Research Council, Helse Ser-Ost, and the University of Oslo through the Centre for Molecular Medicine Norway (NCMM), which is a part of the Nordic European Molecular Biology Laboratory partnership for Molecular Medicine. R.R. was supported by the NIH (grants R01GM106056 and U01GM103804) and an Alfred P. Sloan Research Fellowship. Open access charges were defrayed in part through the National Science Foundation (grant MCB-1413939 to R.R.).

Received: September 12, 2015
Revised: March 4, 2016
Accepted: June 30, 2016
Published: August 18, 2016

**REFERENCES**


