4800–4812 *Nucleic Acids Research, 2014, Vol. 42, No. 8* *Published online 12 February 2014*
doi:10.1093/nar/gku132

An improved predictive recognition model for Cys_2-His_2 zinc finger proteins

Ankit Gupta[1,2,†], Ryan G. Christensen[3,†], Heather A. Bell[4], Mathew Goodwin[5], Ronak Y. Patel[3], Manishi Pandey[3], Metewo Selase Enuameh[1], Amy L. Rayla[1], Cong Zhu[1], Stacey Thibodeau-Beganny[5], Michael H. Brodsky[1,6], J. Keith Joung[5,7], Scot A. Wolfe[1,2,*] and Gary D. Stormo[3,*]

[1]Program in Gene Function and Expression, University of Massachusetts Medical School, Worcester, MA 01605, USA, [2]Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA 01605, USA, [3]Department of Genetics, Washington University School of Medicine, St Louis, MO 63108, USA, [4]Department of Biochemistry and Biology and Biotechnology, Worcester Polytechnic Institute, Worcester, MA 01609, USA, [5]Molecular Pathology Unit, Center for Computational and Integrative Biology, and Center for Cancer Research, Massachusetts General Hospital, Charlestown, MA 02129, USA, [6]Department of Molecular Medicine, University of Massachusetts Medical School, Worcester, MA 01605, USA and [7]Department of Pathology, Harvard Medical School, Boston, MA 02115, USA

Received December 5, 2013; Revised January 21, 2014; Accepted January 22, 2014

ABSTRACT

Cys_2-His_2 zinc finger proteins (ZFPs) are the largest family of transcription factors in higher metazoans. They also represent the most diverse family with regards to the composition of their recognition sequences. Although there are a number of ZFPs with characterized DNA-binding preferences, the specificity of the vast majority of ZFPs is unknown and cannot be directly inferred by homology due to the diversity of recognition residues present within individual fingers. Given the large number of unique zinc fingers and assemblies present across eukaryotes, a comprehensive predictive recognition model that could accurately estimate the DNA-binding specificity of any ZFP based on its amino acid sequence would have great utility. Toward this goal, we have used the DNA-binding specificities of 678 two-finger modules from both natural and artificial sources to construct a random forest-based predictive model for ZFP recognition. We find that our recognition model outperforms previously described determinant-based recognition models for ZFPs, and can successfully estimate the specificity of naturally occurring ZFPs with previously defined specificities.

INTRODUCTION

Defining the grammar underlying the transcriptional regulatory elements within the human genome remains a critical step in understanding both developmental and disease processes (1). The advent of high-throughput sequencing technology has fueled the development of methodologies for the genome-wide characterization of regulatory features, such as global histone modifications (1–10). These data coupled with global analysis of RNA transcript levels (6,11), chromatin immunoprecipitation (ChIP)-based occupancy data for sequence-specific transcription factors (TFs) (7,12–14) and chromatin conformational capture techniques (15) provide a framework for deconvoluting regulatory networks directing gene expression patterns (16,17). Currently, only a small subset of human TFs has been characterized by ChIP-based approaches in any given cell line (7,13,14), although some sequence occupancy can be inferred from DNaseI (12,17) and MNase (18) data. In the absence of genome-wide binding data, knowledge of the DNA-binding specificities of the TFs within regulatory networks in concert with data sets on sequence conservation, chromatin accessibility and histone modifications can be exploited by computational algorithms to predict TF genomic occupancy, and thereby construct more elaborate transcriptional regulatory models (1,9,17,19–24). Given the difficulty in characterizing the diverse binding

*To whom correspondence should be addressed. Tel: +1 508 856 3953; Fax: +1 508 856 5460; Email: scot.wolfe@umassmed.edu
Correspondence may also be addressed to Gary D. Stormo. Tel: +1 314 747 5534; Fax: +1 314 362 2156; Email: stormo@wustl.edu

†These authors contributed equally to the paper as first authors.

patterns of all expressed TFs in all possible temporal and spatial expression patterns in vertebrates, the ability to estimate the specificity of the constellation of TFs expressed at any given time in a given cell type provides a critical data set for constructing these regulatory models.

Cys_2-His_2 zinc finger proteins (ZFPs) are the largest class of TFs within most metazoans (25), with an estimated 675 members in the human genome (26) harboring an average of 8.5 finger units per gene (27). The majority of these ZFPs are believed to be involved in DNA-recognition, as many of the neighboring fingers are connected by a Krüppel-type TGE(K/R)P linker, which is a hallmark of DNA-binding fingers (28). The canonical DNA-recognition model for an individual finger is based on the ZFP-DNA co-crystal structure of Zif268 (29,30) and other naturally occurring and engineered ZFPs (31–35), wherein each finger potentially recognizes a 4-bp subsite that overlaps the recognition site of the neighboring N- and C-terminal fingers by 1 bp (Figure 1A). Amino acid residues at positions −1, +2, +3 and +6 of the recognition helix typically mediate the recognition preference of a finger within its subsite. The target site preference of a tandem array of fingers reflects a complex interaction between the individual finger modules, as the recognition properties of an individual finger can be influenced by its position within an array and the recognition determinants displayed by its immediate neighbors (36–41).

DNA-binding specificities have been determined for only a small fraction of ZFPs in metazoan genomes (13,17,26,47–50). Unlike other TF families where the majority of the resident factors in diverse species share a high degree of homology (26,51–54), evolutionary analysis of ZFPs indicates that a substantial fraction of resident members do not have highly conserved homologs across metazoans. Instead, the number and composition of fingers within these ZFPs is dynamic between species (27,55,56) and can even vary within a species [e.g. the variation in human PRDM9 isoforms (57,58)]. The specificity determinants within these ZFPs are under strong positive selection, implying the rapid diversification of their recognition potential (27). Consequently, naturally occurring ZFPs can specify a wide variety of different DNA sequences based on both the number and composition of fingers within the array.

Although some principles that govern the recognition properties of zinc fingers have been established, the accurate prediction of their DNA-binding specificity remains challenging. Specificity determinants at individual recognition helix positions with defined base preferences have been extracted from the biochemical and structural characterization of naturally occurring ZFPs (42,47,49,50,59–61) and the selection and characterization of artificial ZFPs that recognize novel target sequences (37,38,41,44,62–74). These data provide a foundation for the construction of predictive recognition models that estimate DNA-binding specificity based on the sequence of the recognition helix of each incorporated finger. Initial models focused on using the amino acid identity at key determinant positions (−1, +2, +3 and +6) to estimate the base preference at their primary DNA contact positions

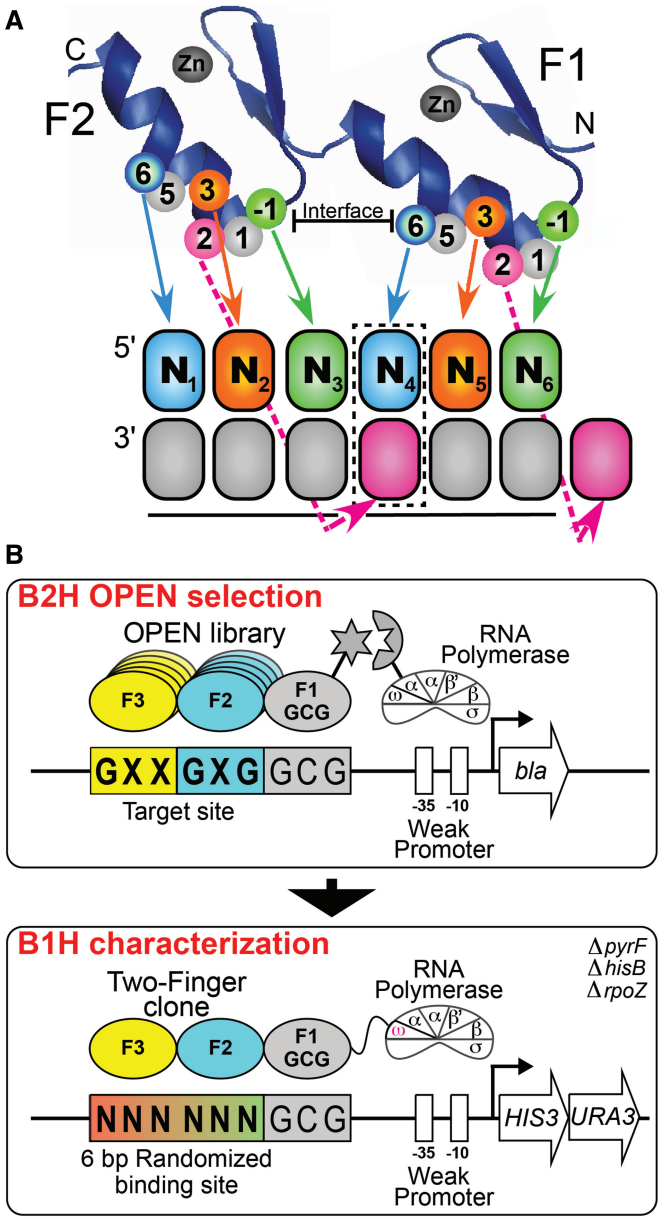

Figure 1. (**A**) Schematic representation of the canonical recognition pattern of two zinc fingers recognizing a hexamer sequence. Each zinc finger unit spans ~30 amino acids and folds into a ββα-motif around a tetrahedrally coordinated zinc ion (42,43). DNA-binding specificity is typically mediated by residues at positions −1, +2, +3 and +6 of the recognition helix, where the numbering scheme refers to the position of each residue relative to the start of the α-helix. The boxed base pair (N_4) represents the position of potential recognition overlap in the canonical recognition model. (**B**) Schematic representation of the two-stage process used to identify two-finger modules with the desired sequence preference. In Stage 1, the B2H system is used to select two-finger modules from an OPEN-based library, where the finger pools used correspond to the finger 2 (F2) and finger 3 (F3) subsites in each target site (44,45). These two-finger libraries are selected in the context of a constant finger 1 (F1) module that recognizes GCG in the neighboring subsite. The DNA-binding specificity of active clones recovered from the B2H selection was determined using the B1H system using a 6-bp randomized library adjacent to the constant GCG F1 binding site. The recovered binding sites are determined by Illumina sequencing and then a binding site motif is calculated from these sequences (46).

within the DNA subsite bound by each individual finger (75–77). Recently, more advanced predictive models have been constructed with improved performance that incorporate context-dependent recognition, which allows determinants to influence more binding site positions than prescribed by the standard recognition model (76–82). However, the construction of these models has been hampered by the limited amount of existing quantitative specificity data for ZFPs that links individual fingers with recognition of particular subsites.

A comprehensive recognition model for canonically binding ZFPs should be achievable using the growing archive of quantitative specificity data from recent bacterial one-hybrid (B1H) analysis of a large number of artificial (41,62,71) and naturally occurring ZFPs (49,50), where the position of each finger within the recognition sequence is defined or can be inferred. This data set spans 678 two-finger modules, including the characterization of 95 two-finger modules generated using the Oligomerized Pool ENgineering (OPEN) system (44,45) described herein. A sizeable fraction of these data explicitly examines the impact of recognition residues at the finger–finger interface on the preferred specificity at the junction of the finger binding sites, which remains the most challenging recognition feature to model. These data permit an improved estimation of context-dependent effects requiring the use of predictive models [such as support vector machine (83) or random forests (RFs) (84)] that implicitly capture these complex properties. Building on our previous efforts using RF models to estimate the specificity of homeodomains (85), we have constructed an RF predictive model for ZFPs using our B1H data that are superior to existing predictive models and that can effectively estimate the DNA-binding specificity of a number of naturally occurring ZFPs.

MATERIALS AND METHODS

OPEN finger selections

OPEN selections were performed to generate a set of two-finger modules that recognize all 64 possible GNNGNG-type sequences in the context of an N-terminal 'GCG' binding anchor zinc finger (recognition helix: RSDTLAR). All target sites used in the selection of novel recognition fingers were of the form GNNGNGGCG. Zinc finger libraries for each target site were assembled from the corresponding Finger 2 and Finger 3 OPEN pools as previously described but with a fixed Finger 1 module (44,45). OPEN selections were performed essentially as previously described (44,45) but using a beta-lactamase (*bla*) antibiotic-resistance gene instead of the *HIS3* gene (70). For each of the 64 selections, we assayed the abilities of up to five clones to activate expression of a *lacZ* reporter gene in a bacterial two-hybrid (B2H) system as previously described (45) and determined the amino acid sequences of these clones. Fifty-eight of the 64 selections displayed active clones, from which we chose 95 clones that could activate expression of *lacZ* in the B2H system by ~2.5-fold or more for further evaluation via B1H binding site selections (Supplementary Table S1).

CV-B1H method

To determine binding site specificities of OPEN-selected and other 2F-modules, the CV-B1H (Constrained Variation Bacterial one-Hybrid) assay was performed essentially as described previously (46). Two-finger modules were evaluated as fusions to the GCG anchor finger. Following transformation into the selection strain, 1×10^6 cells containing the zinc finger plasmid (1352-omega-UV2-ZFP) and the 6-bp randomized binding site library (in pH3U3) were plated on selective NM minimal medium plates (100 × 15 mm) containing 50 μM IPTG and 1 or 2 mM 3-AT and grown at 37°C for 22–30 h. All cells on the plate were pooled, and the pH3U3 plasmids containing the compatible binding sites were isolated for identification of the functional DNA sequences. The binding site region was PCR amplified, barcoded and sequenced via Illumina sequencing, and then binding specificities were determined from these data using GRaMS modeling and the log-odds method (46,71,86).

Construction of the RF ZFP regression model

Based on a pilot study and previous work with homeodomain recognition modeling (85), we developed a recognition modeler based on a RF regression approach (84) using the 'randomForest' module from the *R* package [http://www.r-project.org/(87)]. Two different ZFP RF regression models were trained based on the B1H specificity data: one-finger and two-finger models. The training data for the two-finger model consisted of 678 protein sequences for two fingers of ZFPs and the position frequency matrices (PFMs) obtained from the B1H experiments described above. The one-finger model was trained on the same set but contained 1209 individual fingers (redundancy removed, Supplementary Table S2). Preliminary analysis showed that including additional protein positions beyond the canonical −1, +2, +3 and +6 recognition positions in each finger did not improve the accuracy of the model, so all further training used only those positions. Of the 678 two-finger examples, there are 530 unique combinations of residues at positions −1, +2, +3 and +6; all of them are kept in the data set because the PFMs, while similar between repeats, are not identical and this maintains the inherent variability in the data. These models use the RF regression engine that was previously described (85). The modeler predicts the PFM for a zinc finger protein based on its sequence at the recognition positions, and the RF regression minimizes the mean-squared error (MSE) between the predicted and observed PFMs. MSE values for a single position can range from 0, if the two PFMs are identical, to 0.5 if they contain probabilities of 1.0 for different bases. A random position (probability of 0.25 for each base) would have a maximum MSE of 0.1875 compared with a position with probability of 1.0 for any base. This has the effect of generating PFMs that tend toward random at some positions instead of making high probability predictions that are frequently incorrect.

We used the default value of 500 trees while training the RF model. In this model, a single tree picks predictive variables, specific amino acids at specific positions, randomly and then applies regression to estimate their contribution to each PFM parameter. The set of individual trees are then weighted by regression to minimize the overall MSE between the observed and predicted PFMs. Accuracies were determined by 10-fold cross-validation, where the total data set was divided into 10 subsets and training was based on nine of them and the accuracy measured on the remaining subset. Each of the subsets was left out in turn, and the testing accuracy is reported as the means and medians on the test sets.

We chose to minimize MSE because we are specifically trying to find optimal PFMs that fit the entire distribution of binding site affinities. However, other objectives could be used instead. There have been a large number of different methods proposed to compare motifs with each other and determine a quantitative measure of similarity (88–94). The MSE that we use is closely related to maximizing the Pearson correlation and is often a highly ranked method, particularly when trying to assign a motif to a specific class of transcription factors. In other approaches more emphasis is put on matching high information content positions in the binding sites and low information content positions are scored similar to mismatches. For example, the recently published zinc finger predictor from the Princeton group (82) specifically maximizes the number of correctly predicted positions with high information content, which has advantages for some purposes (see later in the text).

Construction of ZFP recognition motif predictions

We established a Web site that will predict the binding motif for an input ZFP containing any number of fingers (http://stormo.wustl.edu/ZFModels/). ZFP sequences can be submitted in two forms as follows: a concatenation of the four critical recognition residues of each finger (−1, +2, +3 and +6) or the entire protein sequence. In the latter case, the Web site will determine the locations of the recognition residues in each finger based on a HMMER analysis (95) of zinc finger motifs present within the sequence. Three different ZFP motif generation methods are available based on the trained RF regression models: one-finger model, multi-finger model and the average of these models. In the one-finger model, the predictions are based on training of single fingers, and the complete motif is predicted by concatenating the individual predictions. In the multi-finger model, the predictions are based on the two-finger training data, and the complete motif is stitched together from the overlapping two-finger predictions, where the positions of overlap between the motifs are averaged (Supplementary Figure S1). The third method averages together the prediction from the one-finger and two-finger models to generate the final prediction. Generally, the different predictions are in close agreement but sometimes there is a divergence and the most accurate may depend on the specific zinc finger protein; therefore, we advocate testing with each model to examine the inherent variation.

Evaluation of Bcl6 predictive motif for predicting ChIP-seq peaks

The predicted DNA-binding specificity of Bcl6 was estimated using the multi-finger model through the ZFModels interface. The top 100 ChIP-seq peaks for Bcl6 (96) were extracted using Galaxy (97), and a motif for Bcl6 was extracted from these peaks using MEME (zoops mode) (98). MSE was calculated from this PFM against different motifs as described above. FIMO (99) was used to determine the number of the top 100 ChIP peaks containing favorable Bcl6 binding sites ($P < 10^{-4}$) based on each motif.

RESULTS

Selection and characterization of two-finger modules recognizing GNNGNG target sites

We used OPEN selections (44,45) to identify two-finger modules recognizing 64 different 6-bp target sites of the form GNNGNG (Figure 1B). This set of target sites was chosen to include a focused set of sequences that were available in the OPEN system to explore the quality of the B2H-generated fingers. In addition, for the defined target positions (constant guanines), there are strong expectations about the complementary recognition determinants that would be selected. Deviations from the expected residues in the recovered sequences would be indicative of context-dependent effects. These two-finger modules were selected via the B2H system in the context of a three-finger array harboring a fixed N-terminal anchor finger that recognizes a GCG subsite. Fifty-eight of these selections yielded zinc finger arrays that bound their target site as evidenced by their ability to activate transcription in a B2H *lacZ* reporter assay (Supplementary Table S1).

We determined the DNA-binding specificity of a representative set of the B2H-selected two-finger modules using the B1H system (49,71). Each two-finger module was characterized using a reporter system containing a 6-bp randomized binding site library adjoining the finger 1 recognition element—GCG (46,71) (Figure 1B). After selection, surviving colonies carrying the functional DNA-sequences for each two-finger module recovered from this library were pooled and characterized by Illumina sequencing from which a preferred recognition motif was determined (46). This analysis yielded motifs for 95 OPEN-selected two-finger modules (Supplementary Figure S2). For 64 of these two-finger modules, the preferred recognition sequence matched the expected target site. The remaining modules are complementary to their target sequence, but actually prefer a related binding site. These modules expand the population of characterized two-finger modules for the construction of artificial zinc finger arrays, and the coupled specificity data provide additional information on the recognition potential of specific determinant combinations for the construction of improved predictive models.

Assessing context dependence in our selected two-finger modules

As a basis set for constructing predictive recognition models for ZFPs, we have used quantitative B1H specificity data on a large group of naturally occurring (49,50) and artificial (41,62,71) zinc finger arrays. To facilitate the evaluation of DNA-recognition by these zinc fingers, we have parsed this data set into 1209 different one-finger modules or 678 different two-finger modules. For example, a characterized three-finger array is broken down into three one-finger modules or two overlapping two-finger modules with their associated subsite motifs (Supplementary Figure S1). Figure 2 shows the base preferences at base pair positions 1, 2 and 3 within the core subsite (contacted by specificity determinants at positions +6, +3 and −1, respectively; see Figure 1) for this data set of one-finger modules. In general, the observed amino acid to base correlations at each position are consistent with previous studies of recognition preferences for zinc finger proteins (42,43,50,76–78). The strongest correlations are observed at the central base; amino acid changes at position +3 in the recognition helix primarily influenced recognition at the middle base position of the altered finger subsite in our two-finger modules when examined over the data set (Supplementary Figure S3). The independence of recognition at this position was previously harnessed to expand the recognition diversity of our two-finger modules in a directed manner in many instances (71).

Weaker correlations at other positions highlight the role of context on specificity. The influence of context dependence on the DNA-binding specificity of individual fingers is apparent from a qualitative analysis of finger sets within our data set, particularly at the finger–finger interface for a subset of two-finger modules where residues on both sides of the interface were randomized to more effectively capture these effects (Figure 1A) (62,71). For many individual two-finger modules, the base at position 4 is highly specified. However, when the preferred specificity at this position is binned across the data set based on the type of residue at position +6 of the N-terminal finger (Figure 3A), some amino acids are associated with each of the four bases in different C-terminal finger contexts. Glutamate at position +6 provides a notable example, where two-finger modules containing this residue display distinct preferences for each of the four bases at position 4 (Figure 3B). The potential influence of residues within the C-terminal finger, in particular the residue at position +2, on recognition at base position 4 are well documented (29,31,38,100). Consistent with the potential influence of position +2 on recognition, changes in the residue at position +2 in the recognition helix in many instances appear to influence neighboring base preference, particularly at position 4 (Supplementary Figure S4). These data highlight the need for a predictive model that can capture the influence of each determinant position on multiple base positions within the zinc finger recognition sequence.

RF recognition models for ZFPs

Zinc fingers have been the focus of several studies on qualitative recognition codes [reviewed in (42,43)]. More recently, several groups have developed models that predict quantitative motifs for zinc finger proteins based on the residues present at canonical recognition positions

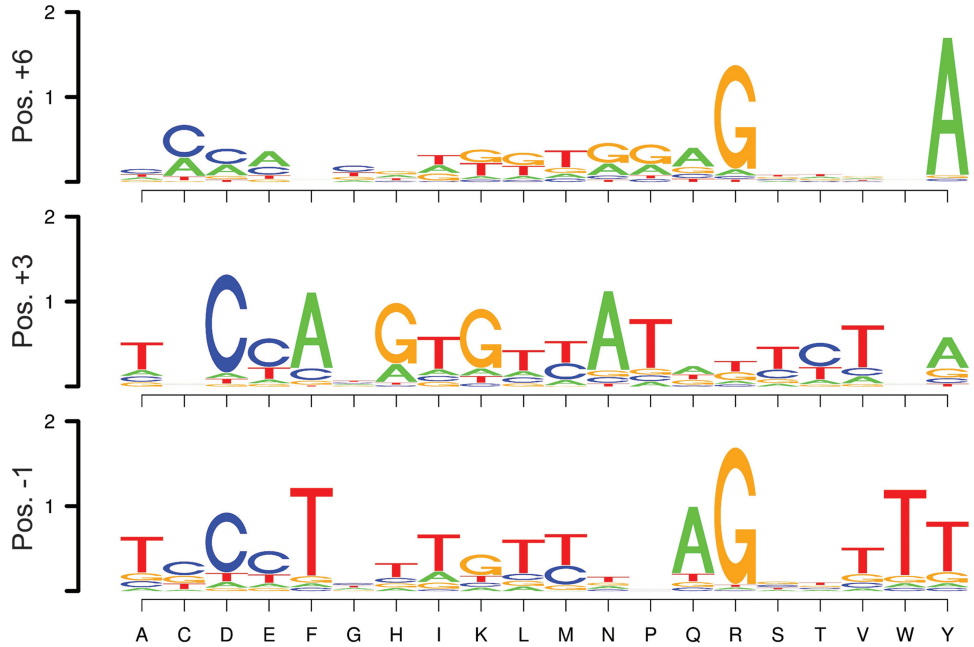

Figure 2. Base preferences observed across the data set for specificity determinants at each of the canonical recognition positions (+6, +3 and −1). For each amino acid (X-axis) at the finger positions +6 (top), +3 (middle) and −1 (bottom), the corresponding base preferences, averaged over all examples, are garnered from the B1H-determined recognition motifs. Base preferences at binding site position 1 are indicated for position +6 specificity determinants; base preferences at binding site position 2 are indicated for position +3 specificity determinants; base preferences at binding site position 3 are indicated for position −1 specificity determinants.

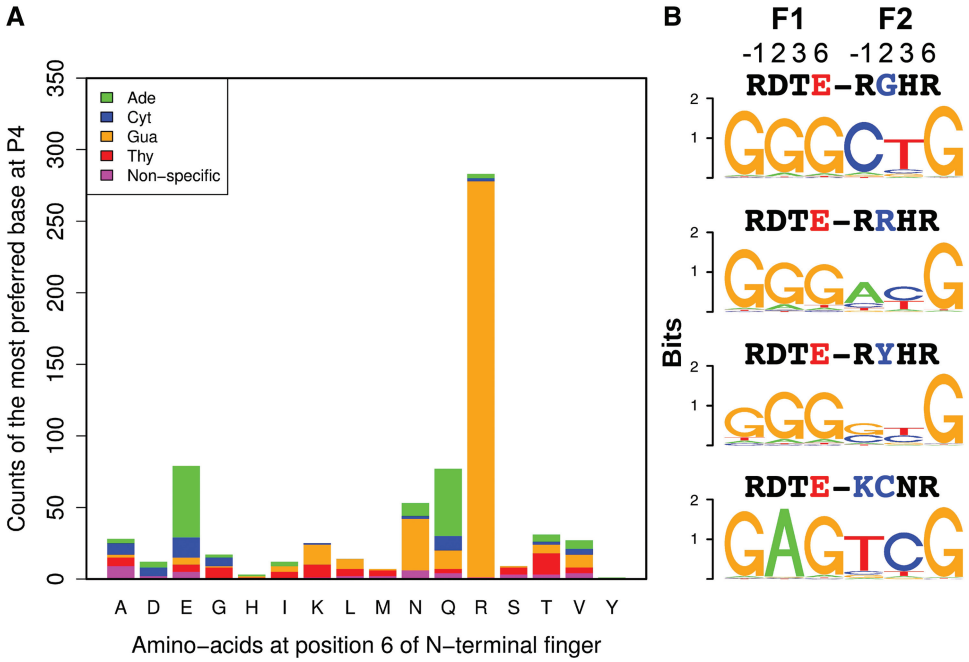

Figure 3. Context-dependent preferences observed for the base at position 4 (P4) recognized by the two-finger modules across the entire data set. (**A**) Stacked bar plot showing the distribution of base preferences dictated by each amino acid at position +6 of N-terminal finger in a two-finger module. The height of each bar corresponds to the number of zinc finger modules with the amino acid labeled on the X-axis. The height of each colored bar segment corresponds to number of modules preferring a particular base. Preference was defined as nonspecific if the information content at a position is <0.3. (**B**) Examples of context-dependent preference at position P4. Logos representing the specificity of four different two-finger modules with Glu at position +6 (red) of N-terminal finger with different base preferences at P4. Above each observed motif are the amino acids at the four canonical recognition positions (−1,+2,+3 and +6) for the N-terminal and C-terminal fingers.

within each finger (76–79). Although superior to purely qualitative recognition codes, their accuracies leave considerable room for improvement. These models were limited because they were trained primarily on qualitative data: collections of proteins and their binding sites with high binding affinity, but where the preference of each ZFP for its target site relative to other sequences was unknown. Our B1H-characterized zinc finger data provide a much larger training set with quantitative information about the preferences of different proteins for different DNA binding sites, which allows us to train new recognition models to obtain higher accuracy predictions. In pilot studies, we tested the feasibility of creating recognition models using several different machine learning algorithms, including neural networks (78), support vector machines (83), k-nearest neighbors (101), partial linear regression (102) and RF (84). We found that RF-based models performed as well or better than those of other methods and its implementation was computationally less demanding, so we used an RF regression algorithm to create a predictive model for ZFPs. The results of these preliminary studies were similar to those we previously reported for predicting the specificity of homeodomain proteins (85).

We trained RF predictive models on either one-finger or two-finger module specificity data, where the latter model is designed to capture context-dependent effects between neighboring fingers. Training the two-finger model takes as input the amino acids at the eight canonical recognition positions (−1, +2, +3 and +6 of each finger) and builds regression trees to predict recognition preference over the entire 6-bp binding site. (The one-finger model was similarly trained on individual fingers and each 3-bp binding site.) Importantly, these models are not restricted to the canonical interactions between particular finger recognition positions and bases within the binding site, unlike many previous recognition models (76,77). Because we have a much larger training set than was available for previous models, a wider range of potential interactions between these recognition positions and the binding site are allowed within the model to capture context-dependent effects observed within the data. Consequently, each recognition position within the two-finger module contributes to the overall predicted PFM, although the strongest contributions within the model will be between the most highly correlated amino acids and base pairs.

The objective during model training is to minimize the MSE between the observed and predicted PFM values for each two-finger module. Table 1 shows the average value (both the mean and median with standard deviations) obtained in a 10-fold cross-validation of our two-finger model. This was compared with predictions by each of four other published models that were readily available for testing (76–79). The MSE is greatly reduced with the new ZFModels predictions to less than half for means and less than one-third for medians when compared with other prior models. The prediction error is fairly evenly distributed across the positions of the binding sites

(Table 2). Figure 4 displays several examples that are near the median value of MSE to show the degree of similarity between observed and predicted PFMs. Many of the highest accuracy examples contain guanine at positions 1 and 6 because the training set was biased with fingers recognizing guanine at these positions. Figure 4 highlights examples deviating from this pattern, demonstrating that our ZFModels can generate accurate predictions for a wide variety of different types of motifs. As expected, the two-finger predictive model can capture the context dependence at the finger–finger junction observed in our data set, such as the motifs in Figure 3B, whereas the one-finger predictive model fails to capture this subtlety (Supplementary Figure S5).

Evaluating the utility of the RF-based zinc finger recognition model

Several published studies have determined specificity of ZFPs using SELEX (26,103–105). None of these examples were included in the training data and so they constitute an independent test set. Supplementary Figure S6 contains the logos from the published PFMs for a subset of these ZFPs and the logos predicted by ZFModels. In every case, the predictions match preferred binding sites from the experiments when we take into account the variable spacing between neighboring fingers due to noncanonical linkers in some instances. However, the quantitative models are less consistent than the average fits to zinc fingers within our data set via cross-validation analysis (Supplementary Table S3). This may be due to the SELEX data being evaluated after multiple rounds of selection where the resulting PFM is heavily weighted toward a subset of the highest affinity sites, leading to an over-specified motif. We also compared the ZFModels predictions on some of the same data sets with the predictions made by a recently published method (zf.princeton.edu) based on support vector machine training (83). ZFModels makes more accurate predictions as measured by MSE (Supplementary Table S4) on these independent test sets than the Princeton model, although the Princeton model often contains more matching positions with high information content (see Discussion).

Ideally, our recognition model would also allow prediction of ZFPs with uncharacterized DNA-binding specificity throughout the genome. We chose to evaluate its predictive utility for Bcl6, as this ZFP has been characterized by B1H (50), PBM (47) and SELEX-seq (26), which allows a comparison of our predictive motif against DNA-binding specificities determined via multiple methods, and against ChIP-seq data for this factor (96). The Bcl6 recognition motifs produced by B1H, PBM and SELEX-seq are all similar, although the SELEX-seq motif appears over-specified (Figure 5). We also generated a predicted recognition motif for Bcl6 using the Princeton SVM model for comparison with our model. The Princeton motif has greater information content than our ZFmodel motif, but at many positions, the Princeton motif predicts a particular base with absolute certainty, which much like the SELEX-seq motif suggests that it is over-specified. When judged against an independent source, a MEME (98) motif from the top 100 Bcl6 ChIP-seq peaks (96), the B1H and PBM motifs appear most similar. The ZFModels multi-finger predictive model also shows good similarity to the determined motifs (MSE values 0.04 from the MEME-ChIP motif, 0.05 from either the PBM- or B1H-based motifs, 0.05 from the Princeton motif and 0.08 from the SELEX-seq motif), but it is a bit worse than the average value of <0.01 in our cross validation studies. FIMO analysis (99) of these ChIP peaks using each motif confirms this assessment: the MEME-derived motif from the Bcl6 ChIP data discovers a good Bcl6 binding site ($P < 10^{-4}$) in 74 of 100 peaks, the B1H motif in 56 of 100 peaks, the PBM motif in 52 of 100, the SELEX-seq motif in 43 of 100, the ZFModels predicted motif in 25 of 100 and the Princeton motif in 9 of 100, where only four would be expected by chance. Thus, our predictive motif has value for the discrimination of binding sites within the genome, and in this example is superior to the Princeton motif, but it can still benefit from the incorporation of additional experimental data to improve its quality. Figure 5 displays logos in two formats, the original information-based method (106) and a PFM-based method where the height of each base is proportional to its frequency in the model (107). The frequency representation demonstrates that even in cases where our model does not make a confident (high probability and high information content)

Table 1. MSE for several prediction programs

Program	ZFModels[a]	Benos[b]	Kaplan[c]	Zifnet[d]	ZIFIBI[e]
Mean	0.017 ± 0.005	0.044	0.047	0.040	0.072
Median	0.009 ± 0.002	0.033	0.035	0.032	0.063

[a]This work. Values are mean and standard deviation from 10-fold cross-validation.
[b]Ref. (76).
[c]Ref. (77).
[d]Ref. (78).
[e]Ref. (79).

Table 2. MSE for each position, for one-finger and two-finger models (mean/median)

Nucleotide position	1	2	3	4	5	6
1 finger	0.016/0.004	0.015/0.005	0.008/0.001			
2 fingers	0.006/0.001	0.007/0.003	0.006/0.001	0.012/0.004	0.010/0.004	0.004/0.000

Note: The reported median values represent the bin the median value falls in, where the bins are 0.001 wide and labeled with the lower value. So if the median value is reported as 0.000 that means the median is in the bin between 0.000 and 0.001. These values come from training and testing on the complete data rather than from cross-validation, resulting in lower values than in Table 1.

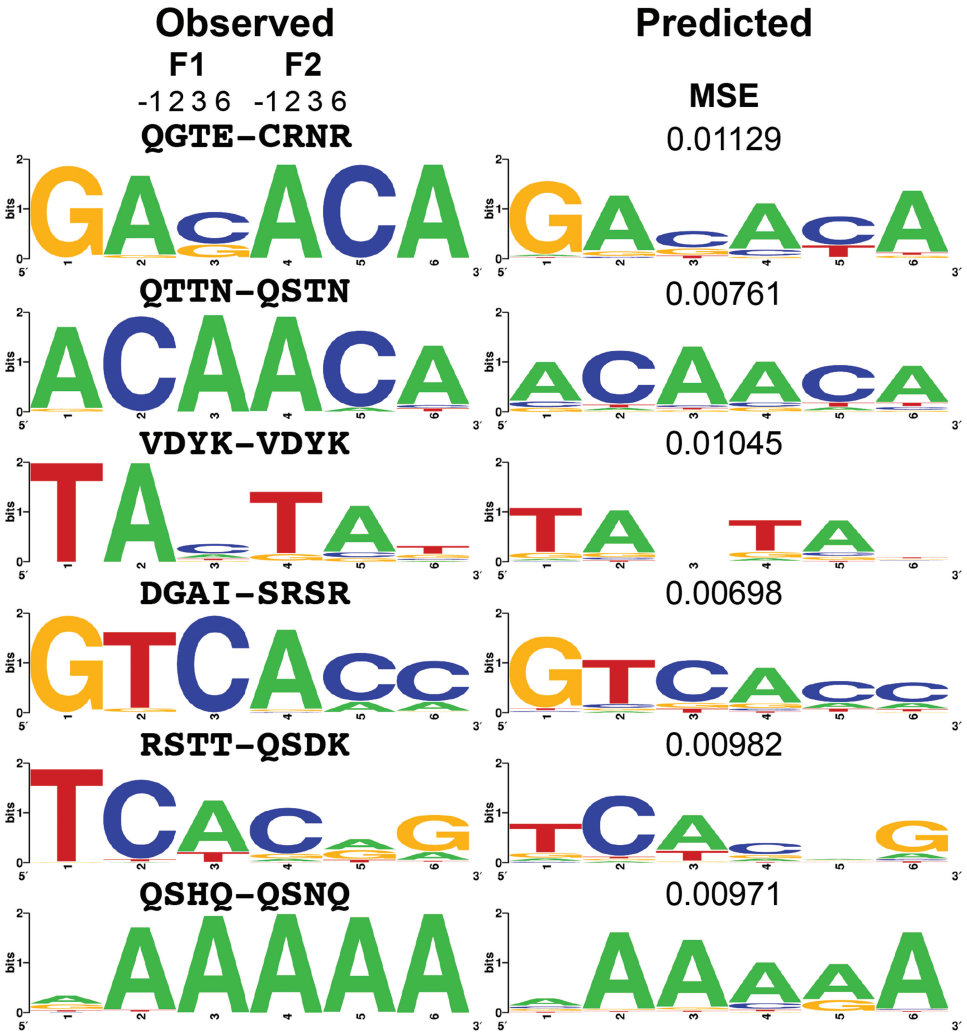

Figure 4. Examples of observed motifs for two-finger modules that are within our data set, and predicted motifs for these fingers using our final predictive model. Above each observed motif are the amino acids at the four canonical recognition positions (−1, +2, +3 and +6) for the N-terminal and C-terminal fingers. The MSE value between the observed and predicted PFMs is displayed above the predicted motif.

prediction, it generally gets the preferred base correct. Combining all of the experimental models with the MEME model from the ChIP-seq data, one finds a consensus sequence of TTCCTnGAAAG (positions 5–15 in the alignment). Our model agrees at every position except 13, where it prefers G slightly to A, but many of those predictions are low confidence. In contrast, the Princeton model has more high information content positions that match the consensus, but it also contains several positions where the preferred base is assigned a very low probability. Our model has an overall better fit to the other models, as evaluated by MSE and similarities to the rank distributions of all possible binding sites, but there are some purposes for which maximizing the number of high confidence, correct predictions is useful (see 'Discussion' section).

DISCUSSION

The development of platforms for rapidly characterizing the specificity of transcription factors has dramatically increased the amount of data that is available for all of the major TF families (108), but there are still barriers to generating data for all naturally occurring ZFPs. The average number of fingers in a human ZFP is 8.5 (27), and these polydactyl (i.e. many fingered) ZFPs may have complex binding modes due to the presence of independent DNA-recognition modules. For example, genome-wide ChIP analysis of NRSF (109,110), a 9-finger ZFP, recovered two different types of binding sites: a prominent motif that contains a juxtaposition of two subsites and a set of additional motifs with variable spacing between these subsites. Taipale and colleagues noted the difficulty in characterizing ZFPs by either SELEX-seq or PBM (26): they successfully characterized only 8% of ZFPs and only 3% with more than eight fingers (26). Similarly, our B1H motif set includes only seven naturally occurring ZFPs with ≥8 fingers with a success rate of ∼38% of the attempted *Drosophila* ZFP genes (50). With the possibility that polydactyl ZFPs use different finger sets to bind multiple distinct motifs, describing their recognition properties is critical to understanding their regulatory mechanisms.

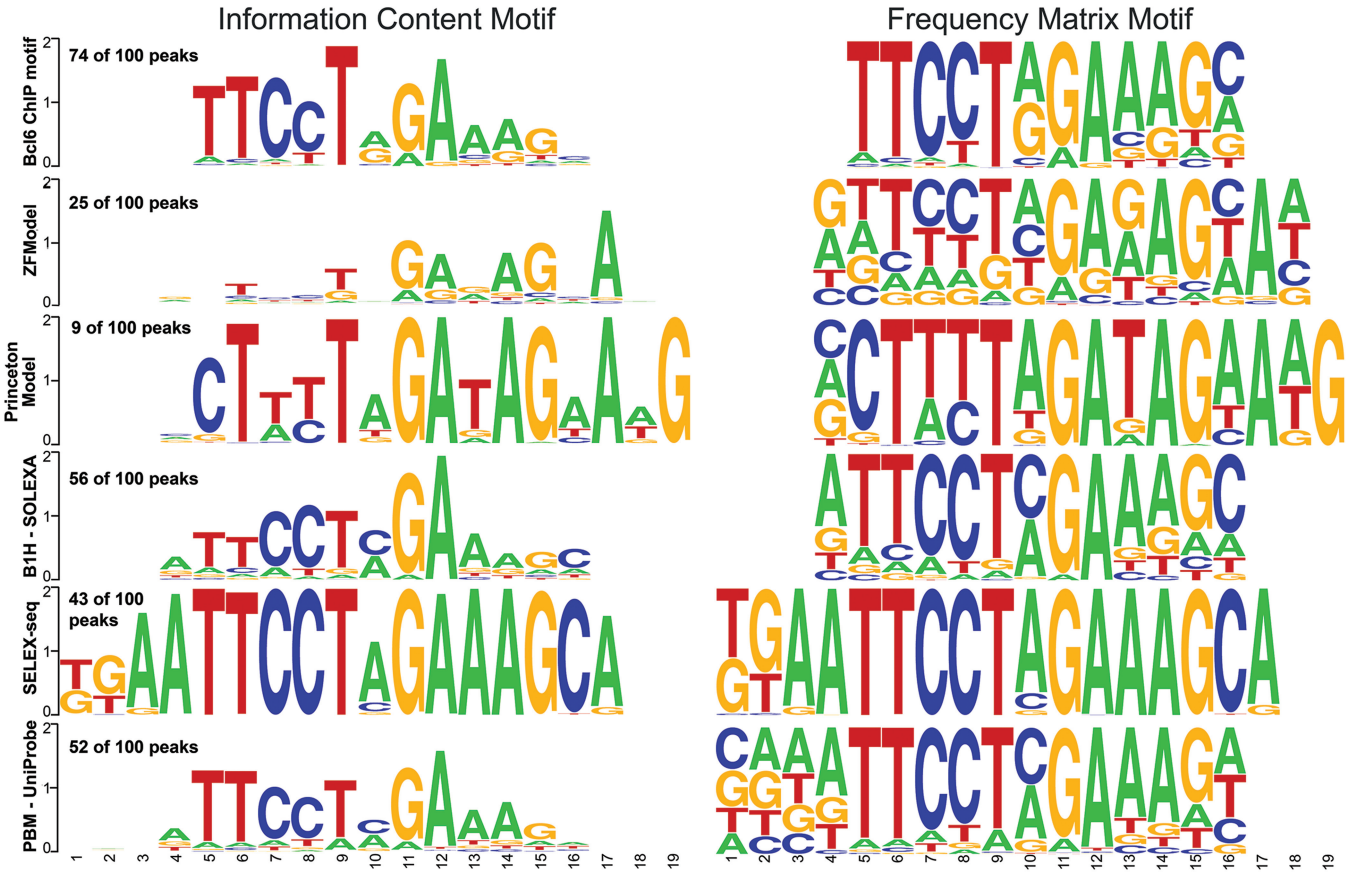

Figure 5. Comparison of the MEME motif from the top 100 Bcl6 ChIP peaks (96) with the motif predicted for the five canonically linked fingers by ZFModels and the Princeton SVM method (82) and the recognition motifs determined directly for Bcl6 by B1H (50), SELEX-seq (26) and PBM (47). The left column displays the motifs as information content, whereas the right column displays the motifs as position frequency plots. The frequency of a strong motif match ($P < 10^{-4}$) for each motif in the top 100 ChIP peaks as determined by FIMO is indicated above each motif.

The growing body of quantitative specificity data for naturally occurring and artificial ZFPs provides a foundation for the development of improved predictive models for this family to help facilitate a broader understanding of their function as regulators within the genome, where other direct analysis methods may be challenging to use.

Our efforts to construct an improved predictive model have focused on two aspects of the problem as follows: expanding the population of quantitatively characterized finger modules and using new methods for training improved recognition models. We have used OPEN-based ZFP selection methods (44,45) to expand our existing set of B1H-characterized artificial and naturally occurring fingers to 1209 one-finger modules and 678 two-finger modules. The latter group captures context-dependent effects that can occur at the finger–finger interface, allowing the construction of recognition models that span more than a single finger, thereby providing additional information on the recognition potential of specific determinant combinations for the construction of improved predictive models. These finger archives and the underlying data also have value in the design of artificial ZFPs to recognize specific sequences. Thus, the assembly of these modules can be data driven by applying 'rules' for recognition of particular sequences to estimate which assembled finger models are likely to provide the desired composite specificities.

Our assessment of ZFModels shows that the motif predictions obtained are superior to previously published predictors. This is likely due to our larger and better (i.e. quantitative) training sets that allow us to consider more interactions, not just the canonical ones that have been primarily used in the past. We have also leveraged our two-finger module data to extend the model construction beyond a one-finger to two-finger units, where the two-finger model constructs motifs by assembling interfaces via a stitching assembly (62) to try to minimize edge effects of the two-finger module data on the resulting motif. This model is accessible to the community though our Web site (http://stormo.wustl.edu/ZFModels/). Users can input a protein sequence and an HMM-based algorithm will extract the determinants in each finger for construction of a recognition motif. Users can use either the one-finger or multi-finger model, or a hybrid (average) of these two models for generating a motif for their factor. On an independent test set, the hybrid model performed slightly better (Supplementary Table S3), although the results from each method are similar.

There is still room for improvement in our predictive model, especially for some classes of C2H2 ZFs with

noncanonical linkers that may lead to alternate finger sequences or binding modes, but in nearly every case tested the predictions are at least partially correct and allow for the alignment of the individual fingers with the segments of the binding motifs that they interact with. A recently reported large compendium of zinc finger proteins selected for binding to specific DNA sequences (74), and then with their specificities determined by B1H, may provide additional, more diverse information to improve the predictive models further, but this has not been tested yet. Currently, predictions from our models are not accurate enough on their own to make reliable regulatory networks, but may be useful in conjunction with accessibility data and DNaseI footprinting data (12) to identify their regulatory sites. They can also aid in assigning ZF-TFs to particular motifs that are discovered through computational analysis of other genomic features, although for that particular problem, the alternative SVM approach of the Princeton group (82) will sometimes work better. Their approach trains their model to maximize the number of high information content positions that are correctly predicted. By then applying string matching methods, one can sometimes identify a ZF-TF that is likely to bind to a known motif [e.g. PRDM9 (58)] in cases where our model may yield a less definitive consensus because it may predict many low information content positions. In some cases, these approaches may also allow us to determine whether only a subset of ZFs are used to recognize DNA, or if different subsets are used to recognize different classes of binding sites, as when ZFPs use alternative modes of binding for interacting with different sequences. Given the rapid diversification of ZFPs during evolution and the technical challenges associated with experimental determination of their specificities, the continued refinement of predictive models will likely play an important role in understanding the roles of these proteins in transcriptional regulatory networks.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank members of the Brodsky, Joung, Stormo and Wolfe laboratories for their assistance with these studies.

FUNDING

U.S. National Institutes of Health (NIH) [GM068110 to S.A.W., HG000249 to G.D.S., HG004744 to M.H.B. and S.A.W., GM078369 to J.K.J., S.A.W., G.D.S.]. Funding for open access charge: U.S. National Institutes of Health (NIH).

Conflict of interest statement. J.K.J. has financial interests in Editas Medicine and Transposagen Biopharmaceuticals. J.K.J.'s interests were reviewed and are managed by Massachusetts General Hospital and Partners HealthCare in accordance with their conflict of interest policies.

REFERENCES

1. Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J., Kaul,R. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
2. Kundaje,A., Kyriazopoulou-Panagiotopoulou,S., Libbrecht,M., Smith,C.L., Raha,D., Winters,E.E., Johnson,S.M., Snyder,M., Batzoglou,S. and Sidow,A. (2012) Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Res.*, **22**, 1735–1747.
3. Song,L., Zhang,Z., Grasfeder,L.L., Boyle,A.P., Giresi,P.G., Lee,B.K., Sheffield,N.C., Graf,S., Huss,M., Keefe,D. *et al.* (2011) Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res.*, **21**, 1757–1767.
4. Wang,H., Maurano,M.T., Qu,H., Varley,K.E., Gertz,J., Pauli,F., Lee,K., Canfield,T., Weaver,M., Sandstrom,R. *et al.* (2012) Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.*, **22**, 1680–1688.
5. Thurman,R.E., Rynes,E., Humbert,R., Vierstra,J., Maurano,M.T., Haugen,E., Sheffield,N.C., Stergachis,A.B., Wang,H., Vernot,B. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.
6. Natarajan,A., Yardimci,G.G., Sheffield,N.C., Crawford,G.E. and Ohler,U. (2012) Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res.*, **22**, 1711–1722.
7. Arvey,A., Agius,P., Noble,W.S. and Leslie,C. (2012) Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res.*, **22**, 1723–1734.
8. Sanyal,A., Lajoie,B.R., Jain,G. and Dekker,J. (2012) The long-range interaction landscape of gene promoters. *Nature*, **489**, 109–113.
9. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.
10. Arnold,C.D., Gerlach,D., Stelzer,C., Boryn,L.M., Rath,M. and Stark,A. (2013) Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, **339**, 1074–1077.
11. Djebali,S., Davis,C.A., Merkel,A., Dobin,A., Lassmann,T., Mortazavi,A., Tanzer,A., Lagarde,J., Lin,W., Schlesinger,F. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.
12. Neph,S., Vierstra,J., Stergachis,A.B., Reynolds,A.P., Haugen,E., Vernot,B., Thurman,R.E., John,S., Sandstrom,R., Johnson,A.K. *et al.* (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, **489**, 83–90.
13. Wang,J., Zhuang,J., Iyer,S., Lin,X., Whitfield,T.W., Greven,M.C., Pierce,B.G., Dong,X., Kundaje,A., Cheng,Y. *et al.* (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.*, **22**, 1798–1812.
14. Yip,K.Y., Cheng,C., Bhardwaj,N., Brown,J.B., Leng,J., Kundaje,A., Rozowsky,J., Birney,E., Bickel,P., Snyder,M. *et al.* (2012) Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.*, **13**, R48.
15. Dekker,J., Marti-Renom,M.A. and Mirny,L.A. (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.*, **14**, 390–403.
16. Gerstein,M.B., Kundaje,A., Hariharan,M., Landt,S.G., Yan,K.K., Cheng,C., Mu,X.J., Khurana,E., Rozowsky,J., Alexander,R. *et al.* (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature*, **489**, 91–100.
17. Neph,S., Stergachis,A.B., Reynolds,A., Sandstrom,R., Borenstein,E. and Stamatoyannopoulos,J.A. (2012) Circuitry and

dynamics of human transcription factor regulatory networks. *Cell*, **150**, 1274–1286.
18. Henikoff,J.G., Belsky,J.A., Krassovsky,K., MacAlpine,D.M. and Henikoff,S. (2011) Epigenome characterization at single base-pair resolution. *Proc. Natl Acad. Sci. USA*, **108**, 18318–18323.
19. Jaeger,S.A., Chan,E.T., Berger,M.F., Stottmann,R., Hughes,T.R. and Bulyk,M.L. (2010) Conservation and regulatory associations of a wide affinity range of mouse transcription factor binding sites. *Genomics*, **95**, 185–195.
20. Pique-Regi,R., Degner,J.F., Pai,A.A., Gaffney,D.J., Gilad,Y. and Pritchard,J.K. (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.*, **21**, 447–455.
21. Negre,N., Brown,C.D., Ma,L., Bristow,C.A., Miller,S.W., Wagner,U., Kheradpour,P., Eaton,M.L., Loriaux,P., Sealfon,R. *et al.* (2011) A cis-regulatory map of the *Drosophila* genome. *Nature*, **471**, 527–531.
22. Marbach,D., Roy,S., Ay,F., Meyer,P.E., Candeias,R., Kahveci,T., Bristow,C.A. and Kellis,M. (2012) Predictive regulatory models in *Drosophila melanogaster* by integrative inference of transcriptional networks. *Genome Res.*, **22**, 1334–1349.
23. Kazemian,M., Blatti,C., Richards,A., McCutchan,M., Wakabayashi-Ito,N., Hammonds,A.S., Celniker,S.E., Kumar,S., Wolfe,S.A., Brodsky,M.H. *et al.* (2010) Quantitative analysis of the *Drosophila* segmentation regulatory network using pattern generating potentials. *PLoS Biol.*, **8**, e1000456.
24. Cheng,Q., Kazemian,M., Pham,H., Blatti,C., Celniker,S.E., Wolfe,S.A., Brodsky,M.H. and Sinha,S. (2013) Computational identification of diverse mechanisms underlying transcription factor-DNA occupancy. *PLoS Genet.*, **9**, e1003571.
25. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S.A. and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.
26. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.
27. Emerson,R.O. and Thomas,J.H. (2009) Adaptive evolution in zinc finger transcription factors. *PLoS Genet.*, **5**, e1000325.
28. Laity,J.H., Dyson,H.J. and Wright,P.E. (2000) DNA-induced alpha-helix capping in conserved linker sequences is a determinant of binding affinity in Cys(2)-His(2) zinc fingers. *J. Mol. Biol.*, **295**, 719–727.
29. Elrod-Erickson,M., Rould,M.A., Nekludova,L. and Pabo,C.O. (1996) Zif268 protein-DNA complex refined at 1.6 A: a model system for understanding zinc finger-DNA interactions. *Structure*, **4**, 1171–1180.
30. Pavletich,N.P. and Pabo,C.O. (1991) Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science*, **252**, 809–817.
31. Fairall,L., Schwabe,J.W., Chapman,L., Finch,J.T. and Rhodes,D. (1993) The crystal structure of a two zinc-finger peptide reveals an extension to the rules for zinc-finger/DNA recognition. *Nature*, **366**, 483–487.
32. Houbaviy,H.B., Usheva,A., Shenk,T. and Burley,S.K. (1996) Cocrystal structure of YY1 bound to the adeno-associated virus P5 initiator. *Proc. Natl Acad. Sci. USA*, **93**, 13577–13582.
33. Kim,C.A. and Berg,J.M. (1996) A 2.2 A resolution crystal structure of a designed zinc finger protein bound to DNA. *Nat. Struct. Biol.*, **3**, 940–945.
34. Wolfe,S.A., Grant,R.A., Elrod-Erickson,M. and Pabo,C.O. (2001) Beyond the "recognition code": structures of two Cys2His2 zinc finger/TATA box complexes. *Structure*, **9**, 717–723.
35. Segal,D.J., Crotty,J.W., Bhakta,M.S., Barbas,C.F. 3rd and Horton,N.C. (2006) Structure of Aart, a designed six-finger zinc finger peptide, bound to DNA. *J. Mol. Biol.*, **363**, 405–421.
36. Desjarlais,J.R. and Berg,J.M. (1993) Use of a zinc-finger consensus sequence framework and specificity rules to design specific DNA binding proteins. *Proc. Natl Acad. Sci. USA*, **90**, 2256–2260.
37. Wolfe,S.A., Greisman,H.A., Ramm,E.I. and Pabo,C.O. (1999) Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code. *J. Mol. Biol.*, **285**, 1917–1934.
38. Dreier,B., Beerli,R.R., Segal,D.J., Flippin,J.D. and Barbas,C.F. 3rd. (2001) Development of zinc finger domains for recognition of the 5′-ANN-3′ family of DNA sequences and their use in the construction of artificial transcription factors. *J. Biol. Chem.*, **276**, 29466–29478.
39. Sander,J.D., Zaback,P., Joung,J.K., Voytas,D.F. and Dobbs,D. (2009) An affinity-based scoring scheme for predicting DNA-binding activities of modularly assembled zinc-finger proteins. *Nucleic Acids Res.*, **37**, 506–515.
40. Choo,Y. (1998) End effects in DNA recognition by zinc finger arrays. *Nucleic Acids Res.*, **26**, 554–557.
41. Zhu,C., Smith,T., McNulty,J., Rayla,A.L., Lakshmanan,A., Siekmann,A.F., Buffardi,M., Meng,X., Shin,J., Padmanabhan,A. *et al.* (2011) Evaluation and application of modularly assembled zinc-finger nucleases in zebrafish. *Development*, **138**, 4555–4564.
42. Wolfe,S.A., Nekludova,L. and Pabo,C.O. (2000) DNA recognition by Cys2His2 zinc finger proteins. *Ann. Rev. Biophys. Biomol. Struct.*, **29**, 183–212.
43. Klug,A. (2010) The discovery of zinc fingers and their applications in gene regulation and genome manipulation. *Ann. Rev. Biochem.*, **79**, 213–231.
44. Maeder,M.L., Thibodeau-Beganny,S., Osiak,A., Wright,D.A., Anthony,R.M., Eichtinger,M., Jiang,T., Foley,J.E., Winfrey,R.J., Townsend,J.A. *et al.* (2008) Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. *Mol. Cell*, **31**, 294–301.
45. Maeder,M.L., Thibodeau-Beganny,S., Sander,J.D., Voytas,D.F. and Joung,J.K. (2009) Oligomerized pool engineering (OPEN): an 'open-source' protocol for making customized zinc-finger arrays. *Nat. Protoc.*, **4**, 1471–1501.
46. Christensen,R.G., Gupta,A., Zuo,Z., Schriefer,L.A., Wolfe,S.A. and Stormo,G.D. (2011) A modified bacterial one-hybrid system yields improved quantitative models of transcription factor specificity. *Nucleic Acids Res.*, **39**, e83.
47. Badis,G., Berger,M.F., Philippakis,A.A., Talukder,S., Gehrke,A.R., Jaeger,S.A., Chan,E.T., Metzler,G., Vedenko,A., Chen,X. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.
48. Jolma,A., Kivioja,T., Toivonen,J., Cheng,L., Wei,G., Enge,M., Taipale,M., Vaquerizas,J.M., Yan,J., Sillanpaa,M.J. *et al.* (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.*, **20**, 861–873.
49. Noyes,M.B., Meng,X., Wakabayashi,A., Sinha,S., Brodsky,M.H. and Wolfe,S.A. (2008) A systematic characterization of factors that regulate *Drosophila* segmentation via a bacterial one-hybrid system. *Nucleic Acids Res.*, **36**, 2547–2560.
50. Enuameh,M.S., Asriyan,Y., Richards,A., Christensen,R.G., Hall,V.L., Kazemian,M., Zhu,C., Pham,H., Cheng,Q., Blatti,C. *et al.* (2013) Global analysis of *Drosophila* Cys2-His2 zinc finger proteins reveals a multitude of novel recognition motifs and binding determinants. *Genome Res.*, **23**, 928–940.
51. Berger,M.F., Badis,G., Gehrke,A.R., Talukder,S., Philippakis,A.A., Pena-Castillo,L., Alleyne,T.M., Mnaimneh,S., Botvinnik,O.B., Chan,E.T. *et al.* (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*, **133**, 1266–1276.
52. Noyes,M.B., Christensen,R.G., Wakabayashi,A., Stormo,G.D., Brodsky,M.H. and Wolfe,S.A. (2008) Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell*, **133**, 1277–1289.
53. Grove,C.A., De Masi,F., Barrasa,M.I., Newburger,D.E., Alkema,M.J., Bulyk,M.L. and Walhout,A.J. (2009) A multiparameter network reveals extensive divergence between C. elegans bHLH transcription factors. *Cell*, **138**, 314–327.
54. Wei,G.H., Badis,G., Berger,M.F., Kivioja,T., Palin,K., Enge,M., Bonke,M., Jolma,A., Varjosalo,M., Gehrke,A.R. *et al.* (2010) Genome-wide analysis of ETS-family DNA-binding *in vitro* and *in vivo*. *EMBO J.*, **29**, 2147–2160.
55. Tadepally,H.D., Burger,G. and Aubry,M. (2008) Evolution of C2H2-zinc finger genes and subfamilies in mammals: species-specific duplication and loss of clusters, genes and effector domains. *BMC Evol. Biol.*, **8**, 176.

56. Thomas,J.H. and Emerson,R.O. (2009) Evolution of C2H2-zinc finger genes revisited. *BMC Evol. Biol.*, **9**, 51.
57. Baudat,F., Buard,J., Grey,C., Fledel-Alon,A., Ober,C., Przeworski,M., Coop,G. and de Massy,B. (2010) PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science*, **327**, 836–840.
58. Myers,S., Bowden,R., Tumian,A., Bontrop,R.E., Freeman,C., MacFie,T.S., McVean,G. and Donnelly,P. (2010) Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science*, **327**, 876–879.
59. Zhu,C., Byers,K.J., McCord,R.P., Shi,Z., Berger,M.F., Newburger,D.E., Saulrieta,K., Smith,Z., Shah,M.V., Radhakrishnan,M. *et al.* (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res.*, **19**, 556–566.
60. Badis,G., Chan,E.T., van Bakel,H., Pena-Castillo,L., Tillo,D., Tsui,K., Carlson,C.D., Gossett,A.J., Hasinoff,M.J., Warren,C.L. *et al.* (2008) A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol. Cell*, **32**, 878–887.
61. Bae,K.H., Kwon,Y.D., Shin,H.C., Hwang,M.S., Ryu,E.H., Park,K.S., Yang,H.Y., Lee,D.K., Lee,Y., Park,J. *et al.* (2003) Human zinc fingers as building blocks in the construction of artificial transcription factors. *Nat. Biotechnol.*, **21**, 275–280.
62. Zhu,C., Gupta,A., Hall,V.L., Rayla,A.L., Christensen,R.G., Dake,B., Lakshmanan,A., Kuperwasser,C., Stormo,G.D. and Wolfe,S.A. (2013) Using defined finger-finger interfaces as units of assembly for constructing zinc-finger nucleases. *Nucleic Acids Res.*, **41**, 2455–2465.
63. Dreier,B., Fuller,R.P., Segal,D.J., Lund,C.V., Blancafort,P., Huber,A., Koksch,B. and Barbas,C.F. 3rd (2005) Development of zinc finger domains for recognition of the 5′-CNN-3′ family DNA sequences and their use in the construction of artificial transcription factors. *J. Biol. Chem.*, **280**, 35588–35597.
64. Dreier,B., Segal,D.J. and Barbas,C.F. 3rd (2000) Insights into the molecular recognition of the 5′-GNN-3′ family of DNA sequences by zinc finger domains. *J. Mol. Biol.*, **303**, 489–502.
65. Segal,D.J., Dreier,B., Beerli,R.R. and Barbas,C.F. 3rd (1999) Toward controlling gene expression at will: selection and design of zinc finger domains recognizing each of the 5′-GNN-3′ DNA target sequences. *Proc. Natl Acad. Sci. USA*, **96**, 2758–2763.
66. Greisman,H.A. and Pabo,C.O. (1997) A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites. *Science*, **275**, 657–661.
67. Isalan,M., Klug,A. and Choo,Y. (1998) Comprehensive DNA recognition through concerted interactions from adjacent zinc fingers. *Biochemistry*, **37**, 12026–12033.
68. Isalan,M., Klug,A. and Choo,Y. (2001) A rapid, generally applicable method to engineer zinc fingers illustrated by targeting the HIV-1 promoter. *Nat. Biotechnol.*, **19**, 656–660.
69. Liu,Q., Xia,Z., Zhong,X. and Case,C.C. (2002) Validated zinc finger protein designs for all 16 GNN DNA triplet targets. *J. Biol. Chem.*, **277**, 3850–3856.
70. Sander,J.D., Dahlborg,E.J., Goodwin,M.J., Cade,L., Zhang,F., Cifuentes,D., Curtin,S.J., Blackburn,J.S., Thibodeau-Beganny,S., Qi,Y. *et al.* (2011) Selection-free zinc-finger-nuclease engineering by context-dependent assembly (CoDA). *Nat. Methods*, **8**, 67–69.
71. Gupta,A., Christensen,R.G., Rayla,A.L., Lakshmanan,A., Stormo,G.D. and Wolfe,S.A. (2012) An optimized two-finger archive for ZFN-mediated gene targeting. *Nat. Methods*, **9**, 588–590.
72. Lam,K.N., van Bakel,H., Cote,A.G., van der Ven,A. and Hughes,T.R. (2011) Sequence specificity is obtained from the majority of modular C2H2 zinc-finger arrays. *Nucleic Acids Res.*, **39**, 4680–4690.
73. Bulyk,M.L., Huang,X., Choo,Y. and Church,G.M. (2001) Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc. Natl Acad. Sci. USA*, **98**, 7158–7163.
74. Persikov,A.V., Rowland,E.F., Oakes,B.L., Singh,M. and Noyes,M.B. (2013) Deep sequencing of large library selections allows computational discovery of diverse sets of zinc fingers that bind common targets. *Nucleic Acids Res.*, **42**, 1497–1508.
75. Workman,C.T., Yin,Y., Corcoran,D.L., Ideker,T., Stormo,G.D. and Benos,P.V. (2005) enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic Acids Res.*, **33**, W389–W392.
76. Benos,P.V., Lapedes,A.S. and Stormo,G.D. (2002) Probabilistic code for DNA recognition by proteins of the EGR family. *J. Mol. Biol.*, **323**, 701–727.
77. Kaplan,T., Friedman,N. and Margalit,H. (2005) Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Comput. Biol.*, **1**, e1.
78. Liu,J. and Stormo,G.D. (2008) Context-dependent DNA recognition code for C2H2 zinc-finger transcription factors. *Bioinformatics*, **24**, 1850–1857.
79. Cho,S.Y., Chung,M., Park,M., Park,S. and Lee,Y.S. (2008) ZIFIBI: Prediction of DNA binding sites for zinc finger proteins. *Biochem. Biophys. Res. Commun.*, **369**, 845–848.
80. Persikov,A.V., Osada,R. and Singh,M. (2009) Predicting DNA recognition by Cys2His2 zinc finger proteins. *Bioinformatics*, **25**, 22–29.
81. Persikov,A.V. and Singh,M. (2011) An expanded binding model for Cys2His2 zinc finger protein-DNA interfaces. *Phys. Biol.*, **8**, 035010.
82. Persikov,A.V. and Singh,M. (2014) *De novo* prediction of DNA-binding specificities for Cys2His2 zinc finger proteins. *Nucleic Acids Res.*, **42**, 97–108.
83. Vapnik,V.N. (1999) An overview of statistical learning theory. *IEEE Trans. Neural Netw.*, **10**, 988–999.
84. Breiman,L. (2001) Random Forests. *Mach. Learn.*, **45**, 5–32.
85. Christensen,R.G., Enuameh,M.S., Noyes,M.B., Brodsky,M.H., Wolfe,S.A. and Stormo,G.D. (2012) Recognition models to predict DNA-binding specificities of homeodomain proteins. *Bioinformatics*, **28**, i84–i89.
86. Gupta,A., Meng,X., Zhu,L.J., Lawson,N.D. and Wolfe,S.A. (2011) Zinc finger protein-dependent and -independent contributions to the *in vivo* off-target activity of zinc finger nucleases. *Nucleic Acids Res.*, **39**, 381–392.
87. Ihaka,R. and Gentleman,R. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Stat.*, **5**, 299–314.
88. Benson,G. (2002) A new distance measure for comparing sequence profiles based on path lengths along an entropy surface. *Bioinformatics*, **18(Suppl. 2)**, S44–S53.
89. Tanaka,E., Bailey,T., Grant,C.E., Noble,W.S. and Keich,U. (2011) Improved similarity scores for comparing motifs. *Bioinformatics*, **27**, 1603–1609.
90. Wang,T. and Stormo,G.D. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, **19**, 2369–2380.
91. Mahony,S., Auron,P.E. and Benos,P.V. (2007) DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies. *PLoS Comput. Biol.*, **3**, e61.
92. Narlikar,L. and Hartemink,A.J. (2006) Sequence features of DNA binding sites reveal structural class of associated transcription factor. *Bioinformatics*, **22**, 157–163.
93. Sandelin,A. and Wasserman,W.W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J. Mol. Biol.*, **338**, 207–215.
94. Schones,D.E., Sumazin,P. and Zhang,M.Q. (2005) Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics*, **21**, 307–313.
95. Finn,R.D., Clements,J. and Eddy,S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, **39**, W29–W37.
96. Barish,G.D., Yu,R.T., Karunasiri,M., Ocampo,C.B., Dixon,J., Benner,C., Dent,A.L., Tangirala,R.K. and Evans,R.M. (2010) Bcl-6 and NF-kappaB cistromes mediate opposing regulation of the innate immune response. *Genes Dev.*, **24**, 2760–2765.
97. Goecks,J., Nekrutenko,A. and Taylor,J. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.
98. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME

SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
99. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.
100. Isalan,M., Choo,Y. and Klug,A. (1997) Synergy between adjacent zinc fingers in sequence-specific DNA recognition. *Proc. Natl Acad. Sci. USA*, **94**, 5617–5621.
101. Alleyne,T.M., Pena-Castillo,L., Badis,G., Talukder,S., Berger,M.F., Gehrke,A.R., Philippakis,A.A., Bulyk,M.L., Morris,Q.D. and Hughes,T.R. (2009) Predicting the binding preference of transcription factors to individual DNA k-mers. *Bioinformatics*, **25**, 1012–1018.
102. Abdi,H. (2010) Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Interdiscip. Rev. Comput. Stat.*, **2**, 97–106.
103. Wood,A.J., Lo,T.W., Zeitler,B., Pickle,C.S., Ralston,E.J., Lee,A.H., Amora,R., Miller,J.C., Leung,E., Meng,X. *et al.* (2011) Targeted genome editing across species using ZFNs and TALENs. *Science*, **333**, 307.
104. Hockemeyer,D., Soldner,F., Beard,C., Gao,Q., Mitalipova,M., DeKelver,R.C., Katibah,G.E., Amora,R., Boydston,E.A., Zeitler,B. *et al.* (2009) Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases. *Nat. Biotechnol.*, **27**, 851–857.
105. Soldner,F., Laganiere,J., Cheng,A.W., Hockemeyer,D., Gao,Q., Alagappan,R., Khurana,V., Golbe,L.I., Myers,R.H., Lindquist,S. *et al.* (2011) Generation of isogenic pluripotent stem cells differing exclusively at two early onset Parkinson point mutations. *Cell*, **146**, 318–331.
106. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
107. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
108. Stormo,G.D. and Zhao,Y. (2010) Determining the specificity of protein-DNA interactions. *Nat. Rev. Genet.*, **11**, 751–760.
109. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of *in vivo* protein-DNA interactions. *Science*, **316**, 1497–1502.
110. Otto,S.J., McCorkle,S.R., Hover,J., Conaco,C., Han,J.J., Impey,S., Yochum,G.S., Dunn,J.J., Goodman,R.H. and Mandel,G. (2007) A new binding motif for the transcriptional repressor REST uncovers large gene networks devoted to neuronal functions. *J. Neurosci.*, **27**, 6729–6739.