Predicting the DNA binding specificity of mutated transcription factors using family-level biophysically interpretable machine learning

Shaoxun Liu[1], Pilar Gomez-Alcala[1], Christ Leemans[1], William J. Glassford[2], Richard S. Mann[2,3,+], Harmen J. Bussemaker[1,3,+]

[1]Department of Biological Sciences, Columbia University, New York, NY, USA.
[2]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA
[3]Department of Systems Biology, Columbia University, New York, NY, USA
[+]Corresponding authors (hjb2004@columbia.edu; rsm3@columbia.edu)

ABSTRACT

Sequence-specific interactions of transcription factors (TFs) with genomic DNA underlie many cellular processes. High-throughput *in vitro* binding assays coupled with computational analysis have made it possible to accurately define such sequence recognition in a biophysically interpretable yet mechanism-agonistic way for individual TFs. The fact that such sequence-to-affinity models are now available for hundreds of TFs provides new avenues for predicting how the DNA binding specificity of a TF changes when its protein sequence is mutated. To this end, we developed an analytical framework based on a tetrahedron embedding that can be applied at the level of a given structural TF family. Using bHLH as a test case, we demonstrate that we can systematically map dependencies between the protein sequence of a TF and base preference within the DNA binding site. We also develop a regression approach to predict the quantitative energetic impact of mutations in the DNA binding domain of a TF on its DNA binding specificity, and perform SELEX-seq assays on mutated TFs to experimentally validate our results. Our results point to the feasibility of predicting the functional impact of disease mutations and allelic variation in the cell-wide TF repertoire by leveraging high-quality functional information across sets of homologous wild-type proteins.

SIGNIFICANCE STATEMENT

Transcription factors (TFs) are DNA binding proteins that play a key role in gene expression control. Genetic mutations in the protein sequence of TFs are increasingly found to be associated with disease. Being able to predict the functional impact of such mutations in terms the quantitative changes in DNA sequence preference they cause is therefore highly useful. TFs come in families that are structurally similar but vary in terms of their sequence and function. In this study, we show that by jointly analyzing high-throughput DNA binding data for the basic helix-loop-helix (bHLH) family of transcription factors, we can successfully build a model that predicts the impact of TF protein sequence mutations.

Keywords

transcription factors; DNA binding specificity; functional impact of missense mutations; biophysically interpretable machine learning; basic helix-loop-helix (bHLH) family

INTRODUCTION

Gene expression regulation is a central aspect of cellular function that is especially important during cell differentiation and stress response, and in which transcription factors (TFs) play a critical role. The DNA binding domain of TFs enables them to perform their regulatory functions by recognizing and binding to specific DNA sequences (1). TF binding sites typically reside either in proximal promoter regions near transcription start sites or in more remote enhancer regions. They allow TFs to regulate gene expression by recruiting or inhibiting associated proteins (2). Since the range of TF binding affinity (typically represented by a dissociation constant (K_D), which equals the TF concentration at which the DNA is bound 50% of the time) covers many orders of magnitude between optimal and non-specific binding, even ultra-weak binding sites can be functionally relevant (3-6), and it is therefore of great value to be able to predict binding affinity of any DNA sequence over its full range (7).

Previous studies have provided evidence that mutations in TFs are responsible for numerous common developmental disorders. For example, anophthalmia has been linked to mutations in SOX2 (8), while autosomal dominant Rolandic epilepsy is associated with mutations in NEUROG1 (9). A single point mutation in the DNA binding domain of TFs has the potential to disrupt their DNA binding preference, leading to dysregulated developmental pathways. It is important to note that if these TF mutations cause quantitative shifts in base preference within the DNA binding site, these can be functionally relevant even when the preferred base remains the same (10). Detecting subtle changes caused by de novo mutations requires quantitative protein-DNA binding assays such as SELEX-seq (11) and high throughput SELEX (HT-SELEX) (12, 13) or chromatin immunoprecipitation coupled with sequencing (ChIP-seq) (14). Such experiments are both labor-intensive and expensive: SELEX-seq experiments involve in vitro protein synthesis and multiple rounds of affinity-based selection; ChIP-seq experiments require manual collection of animal tissues and immunoprecipitation with antibodies that can vary widely in their efficacy. Therefore, the development of a computational prediction model for mutant TF binding affinity would greatly facilitate large-scale screening of TF mutations, by reducing the number of candidates that need to be tested experimentally to only those predicted to exhibit significant changes in DNA binding specificity.

An approximation that treats the effects of base-pair mutations at different positions within the DNA binding site as independent (which means additivity of partial binding free energies or, equivalently, multiplicativity of relative binding affinities) has long been relied on to parameterize binding specificity across all possible DNA ligands (15-18). For a given TF, this amounts to summing up contributions from a position-specific energy matrix, where each of the four base types is assigned a row, the length of the DNA binding site determines the number of columns, and values are free energy differences ($\Delta\Delta G$ in units of RT) associated with base substitutions at a particular position. There have been significant recent advances in our ability to estimate such $\Delta\Delta G$ parameters with sufficient accuracy to allow functionally meaningful quantification of ultra-low-affinity binding sites (6). In the present study, we build on an interpretable machine learning framework, known as ProBound, that was recently developed by our lab (18). ProBound allows us to accurately estimate the binding free energy parameters that define a TF's base preferences using data from in vitro binding assays. While it is possible to extend this model to account for dependencies between nucleotide positions, the improvement in binding energy prediction accuracy over simple position-specific matrix representation

is often relatively modest (6, 19).

In this study, we present an analytical framework to dissect the quantitative relationships between base preference and TF protein sequence within a given TF structural family. It enables us to systematically map the protein features that are the most important determinants of differences in DNA binding specificity among TF paralogs. Our method requires as inputs for model training the protein sequences of representative transcription factors within the same family along with data from suitable in vitro DNA binding assays. Several previous studies also analyzed variation in DNA binding specificity within TF families (20-24). However, the availability of large compendia of HT-SELEX data, along with state-of-the-art biophysically interpretable machine learning methods that can reliably estimate binding energy parameters from such data, has opened up new avenues for predicting the effect of TF protein sequence variation on DNA binding specificity. As a proof of concept, we here focus on the basic helix-loop-helix (bHLH) family of TF paralogs, which is widely studied in both eukaryotes and bacteria (25).

RESULTS

A compendium of DNA binding specificity models for bHLH transcription factors

bHLH proteins (**Figure 1A**) function as homo- and/or heterodimers and typically prefer to bind a reverse-complement symmetric, or palindromic, core sequence CANNTG (with "N" denoting any of the four nucleobases) enhancer box (E-box) sequence (26). We collected 147 in vitro DNA binding datasets for the bHLH family from three different studies and used ProBound (18) to obtain biophysically interpretable and accurate binding energy models (**Figure 1B, C**). Because bHLH homodimers are expected to have palindromic binding preferences (meaning that the binding free energy differences associated with base-pair substitutions are subject to reverse complement symmetry, but not that the binding sites themselves need to be palindromic), we configured ProBound to impose reverse-complement symmetry on its binding free energy parameters (18). This yielded a binding model that matches the symmetrical CANNTG consensus E-box for 96 out of the 147 experiments analyzed, covering 54 distinct bHLH proteins (**Figure 1B, C**). Further filtering for high intrinsic model quality (see **Methods** for details) resulted in a compendium of DNA recognition models for 52 bHLH factors (**Figure 1D** and **Supplemental Data S1**). For each TF included in this training set, the cloned sequence for each protein covers the 56 amino acid bHLH HMMER profile, with a consensus Glu residue at the 9th position of the protein sequence (25). The preferred DNA binding site is centered on one of three E-box variants (CACGTG, CAGCTG, or CATATG) with only a few exceptions. Protein sequences were aligned using the HMMER (27) profile of the bHLH family (see **Methods**); alignment of the DNA binding specificity models was trivial thanks to their imposed palindromic symmetry.

Tetrahedron representation of DNA binding specificity

The DNA binding specificity model for each individual TF consists of a set of free energy differences $\Delta\Delta G$, or equivalently, a set of relative affinities, equal to $\exp(-\Delta\Delta G/RT)$, for each nucleotide position, which together constitute a position specific affinity matrix (PSAM) (28). Directly modeling the $\Delta\Delta G$ values in terms of features of the amino acid sequence of the TF would have several drawbacks: Not

only can the reference base be different for each TF, but for bases that are strongly disfavored at a given position, the inferred ΔΔG values for disfavored bases can show large fluctuations that are not biophysically meaningful. To facilitate both the visualization and the quantitative analysis of the DNA binding specificity models, we transformed each column of four relative affinity values in the PSAM to a point within a three-dimensional tetrahedron, whose vertices correspond to the four possible base identities (**Figure 2A** and **Supplemental Data S2**). Each column in a PSAM, representing the quantitative base preference at a particular position within the DNA binding site for a particular TF, can be reversibly mapped to a single 3D position within the tetrahedron (see **Methods** for details). Intuitively, the degree of proximity to each of the vertices reflects the relative preference for the corresponding base. More precisely, the binding affinity ratios among the four bases are the same as the volume ratios among the four sub-tetrahedrons formed by the internal point and each of the four sides of the tetrahedron. This elegantly maps the common redundant four-dimensional representation of base preference in terms of energies or affinities to a non-redundant three-dimensional representation. For any position within the DNA binding site, the base preferences can be easily visualized for the entire gene family, with each point inside the tetrahedron representing a different TF.

Mapping the bHLH protein sequence determinants of DNA binding specificity

As a first step towards performing family-level analysis of how the DNA binding specificity of bHLH TFs is determined by their protein sequence, we focused on DNA position –1 in the PSAM for each TF and labeled/colored each corresponding point in the tetrahedron according to the amino-acid identity at residue position 13 in the protein sequence alignment (**Figure 2A**). Visual inspection suggests that there is a trend for TFs that have an arginine at residue position 13 to prefer binding to DNA sequences with a cytosine over the other three bases at nucleotide position –1, while a methionine at the same position confers a preference for thymine. A similar visualization can be generated for other combinations of DNA binding site position and protein alignment position: When a different residue position is chosen in the alignment of TF protein sequences, the coloring of the points in the tetrahedron changes (**Figure 2B**); on the other hand, when a different position within the DNA binding site is chosen, the position of each point changes (**Figure 2C**).

To systematically dissect the relationships between amino-acid identity at residue positions in the bHLH protein sequence alignment and the base preferences at nucleotide positions within the DNA binding site in a way that addresses statistical significance, we performed a series of multidimensional analysis of variance (MANOVA) tests. Each MANOVA analyzes the positions of points within the tetrahedron that represented the binding affinities at a specific motif position. The three-dimensional tetrahedral coordinate plays the role of dependent variable in these tests, while the (categorical) independent variable is the amino-acid identity at a given position in the TF protein alignment. For each combination of TF and DNA positions, the MANOVA yields a p-value that quantifies the significance of the statistical association between the protein and DNA positions. It is important to note that since the binding model is symmetrical, the p-values for nucleotide positions –1, –2, and –3 are the same as +1, +2, and +3, respectively, for the corresponding residue position (**Figure 2D**). We found that of the 56 residue positions in the bHLH protein alignment, 12 are significantly associated with base preference at nucleotide position –1/+1 ($\alpha < 0.01$ after Bonferroni correction (29)). Visualizing the p-values from the MANOVA test along the structure of PHO4 (**Figure 2E**) shows that, as expected, the residues in

direct contact with the DNA major groove (e.g., positions 5, 13, and 14) tend to have the most significant associations. Base preference at nucleotide position –1/+1 is greatly influenced by amino acid identity at residue position 13 (**Figure 2F**), consistent with the fact that in PHO4, the sidechain of Arg13 directly interacts with the base-pairs at position –1 and +1 through hydrogen bonds that are known to be crucial for standard E-box preference (30, 31). Available structures of bHLH-DNA complexes with other amino acids at residue position 13 also provide a mechanistic rationale for the observed difference in base preference (**Figure 2G-I**).

Predicting DNA binding specificity from TF protein sequence within the bHLH family

Beyond identifying positional correlations across the protein-DNA interface, the tetrahedron representation provides a natural starting point for predicting the effect of protein mutations on DNA binding specificity in a quantitative manner. For example, to estimate the effect of a point mutation from Arg to Val at residue position 13 in the bHLH domain on base preference at nucleotide position –1/+1, we can compute the centroid of all TFs with Arg13 or Val13, respectively (**Figure 3A, B**), and then use the vector connecting the two centroids as a predictor of the change in binding specificity upon R13V mutation of any bHLH protein that contains an Arg at position 13.

To generalize this approach so that predictions can be made for any TF in the bHLH family, two issues need to be addressed. First, we would like to define a tetrahedral coordinate frame in a data-driven way, so that it optimally reflects the variation in chemical features that drives the interactions between amino-acid side chains and the DNA ligand. Second, we need to account for the fact that evolutionary selection has created dependencies between amino-acid identities at distinct residue positions, which can confound the analysis when using multiple protein features simultaneously as a predictor of base preference. For instance, it would not be accurate to simply add up the individual effects of two residue positions that are in perfect linkage disequilibrium with each other.

The solution we settled on was to first perform a principal component analysis (PCA) of the cloud of points within the tetrahedron (see **Methods** for details). Each of the three principal components defines a natural direction within the tetrahedron onto which variation in base preference can be projected (**Figure 4A**). Grouping TFs by amino-acid identity at a given residue position again reveals interpretable patterns (**Supplemental Data S3**). For instance, TFs with protein feature Arg13 and Val13, respectively, are on opposite ends of the spectrum for PC1 (**Figure 4B)**. By performing a (one-dimensional) ANOVA for each individual principal component, its informative residue positions can be mapped (**Figure 4C** and **Supplemental Data S4**). The difference between the mean across all TFs and the mean of the subset that matches a specific protein sequence feature such as Arg13 reflects a shift in base preference associated with that feature. To predict the base preference for a particular TF, we performed iterative feature selection using an F-test p-value threshold for each PC to decide if an added feature is contributing to the model. In this way, the collinearity between protein features reflecting the evolutionary history of the bHLH was appropriately dealt with (see **Methods** for details).

As an initial test of this scheme, we performed leave-one-out cross-validation across the bHLH family, in which we built an PCA-regression model from all TFs except one, and then used this model to predict the tetrahedral position of the held-out TF. The prediction accuracy was 0.7625 in terms of coefficient

of determination (R^2) and 0.5796 in terms of root-mean-squared deviation (RMSD) of ΔΔG/RT (**Figure 4D**). This result exceeds what we observed for a subset of 40 bHLH factors for which we had access to two HT-SELEX replicates (R^2 = 0.7167, RMSD = 0.5389, **Figure 4E**). We also compared to a previous approach (32) in which the binding specificity of the closest paralog in the same family for which data is available was taken as the prediction (**Figure 4F**). Our PCA-regression model outperforms this scheme by a significant margin (R^2 = 0.6879, RMSD = 0.6832, bootstrap t-test p-value < 10^{-32}), and in addition provides detailed information about the specific effect of individual protein features on DNA binding preference.

Experimental validation using unseen mutants of HES2 and ASCL2

Validation of our prediction method so far has focused on predicting DNA binding specificity for held-out members of a given TF family. Since high-throughput in vitro binding data is available for most human TFs (13, 33, 34), a more useful application is to predict the effect on DNA binding specificity associated with natural non-synonymous mutations in TF protein sequence or engineered sequences. The number of TF variants in the human population is too large for an experimental approach to be practical.

To validate our prediction model for TF variants, we performed SELEX-seq experiments on two wild-type bHLH proteins: HES2, which energetically prefers the standard E-box CACGTG (cytosine at position –1), and ASCL2, which prefers the E-box variant CAGCTG. Among other differences, wild-type HES2 contains Lys5 and Arg13 whereas wild-type ASCL2 has Arg5 and Val13 (see **Figure 5A**). We performed additional SELEX-seq assays for the four single mutants HES2(K5R or R13V) and ASCL2(R5K or V13R) and two double mutants HES2(K5R+R13V) and ASCL2(R5K+V13R).

As a check on data quality, we first applied ProBound (18) to the SELEX-seq data for wild-type HES2 and ASCL2. As expected, the resulting models (**Figure 5B**) were similar to the ones previously obtained from HT-SELEX data for the same TFs (**Figure 5C**). Next, we used our SVD-regression approach to predict the base preferences at nucleotide position –1/+1 for the HES2 and ASCL2 mutants. Both single mutants for HES2 are predicted to retain their preference for the standard E-box CACGTG, while the double mutant is predicted to have a more pronounced shift in preference to the alternative E-box CAGCTG (**Figure 5D**). On the other hand, for ASCL2, the individual mutations are projected to reduce the preference for the alternative E-box motif, and the double mutant to switch preference to the standard E-box.

To experimentally validate these predictions in a targeted manner, we performed electrophoretic mobility shift assays (EMSAs) on wild-type and mutant ASCL2 proteins using radioactively labeled DNA probes containing either a standard or an alternative E-box (see Methods for details). Consistent with our family-based model predictions, these assays showed that the V13R single mutant and the R5K/V13R double mutant of ASCL2 had higher affinity for the canonical E-box (CACGTG) than the alternative E-box (CAGCTG) preferred by wild-type ASCL2 and its R5K mutant (**Figure 5E**). The difference in the ratio between the bound and unbound band between the CACGTG and CAGCTG probes can be used to obtain an experimental estimate of ΔΔG (**Figure 5F**). The binding free energy difference (ΔΔG/RT) between CACGTG and CAGCTG equals –2.12 ± 0.30 for wild type ASCL2, –

2.79 ± 0.55 for the R5K mutant of ASCL2, 1.82 ± 2.02 for the V13R mutant, and 2.10 ± 1.18 for the double mutant.

Next, to more rigorously evaluate the performance of our SVD-regression approach on mutant TFs, we used a model trained on all available wild-type bHLH data to predict the DNA binding specificity of HES2 and ASCL2 mutants. Predicted ΔΔG values for the central nucleotide positions –1 and +1 were highly correlated with ΔΔG inferred from our SELEX-seq data using ProBound (R^2 = 0.5852, RMSD = 0.5798; **Figure 4G**), illustrating the promise of our approach.

DISCUSSION

In this study, we achieved an encouraging level of accuracy when predicting the binding specificity of mutant bHLH proteins. Our family-level analysis strategy can also be applied to other protein families to identify key residue positions that affect DNA binding and generate a recognition scheme for residue type-base pair interactions. Since the prediction model is structure-agnostic, future research would no longer need to rely on crystal structures to gain functional insights into DNA binding. This would enable the identification of DNA-binding-associated residues in unstructured regions of a TF, such as the linker domain consisting of random coils or a protein interaction domain far from DNA. The interpretable nature of our prediction model allows for predicting DNA binding specificities and dissecting the biophysical mechanisms underlying binding preferences. It could facilitate future research in designing mutant TFs with specific DNA binding motifs for applications in gene regulation drugs or DNA pull-down mediators.

Our approach is innovative in several regards. Previous research on protein-DNA binding has primarily focused on one of two approaches: starting from the binding structure and assessing the association between residues and bases, or constructing a mechanism-agnostic machine learning model that predicts the binding affinities of protein-DNA pairs using arbitrary features (15, 23, 35-38). These approaches often incorporate structural features and other high-level parameters, neglecting the fact that the protein sequence alone can provide sufficient information for functional prediction. The framework presented in this study takes a structure-agnostic approach yet offers mechanistic insights throughout the analysis. This is particularly valuable in identifying protein sequence determinants of DNA binding in dimerization or linker domains.

METHODS

Collecting HT-SELEX data for training bHLH binding models

The training set data used in this study consisted of HT-SELEX data for all TFs annotated as belonging to the bHLH family from three publications of the Taipale lab (13, 33, 34). For (34), we only used the unmethylated library and ignored reads corresponding to methylated DNA ligands. This yielded a total of 121 multi-round HT-SELEX datasets covering 62 distinct bHLH proteins.

Constructing binding free energy models using ProBound

For each of the 121 HT-SELEX datasets, we ran ProBound (http://github.com/RubeGroup/ProBound and Ref. (18)) using the JSON configuration file in **Supplemental Data S5**. We filtered by model quality based on the following two heuristic criteria: (i) sufficient ability of the model to predict sequence enrichment at the level of 8-mers ($R^2 > 0.15$), and (ii) a base preference pattern consistent with the E-box consensus CANNTG in the center of the binding energy model. A total of 97 models passed these criteria. An bHLH factor was assigned the binding model with the highest R^2 value if multiple models for it passed the quality filter. Our final bHLH binding model compendium comprised 52 distinct bHLH proteins.

Tetrahedron representation of DNA binding energies

In the position-specific affinity matrix (PSAM) representation for a given TF (28), each column j corresponds to a different position within the DNA binding site, each row b to a different base, and each element w_{jb} to the affinity relative to the preferred base $b_0(j)$ at position j (this implies that $w_{jb_0(j)} = 1$ for the preferred base at each position). To map the four relative affinities in a column of the PSAM to a 3D position within the tetrahedral embedding space, the relative affinities are first normalized to frequencies $f_{jb} = w_{jb}/(\sum_{b'} w_{jb'})$. In a subsequent transformation step, the transpose of the frequency matrix is multiplied by a 4x3 tetrahedral transformation matrix:

$$v_j = [v_{jx}\ v_{jy}\ v_{jz}] = [f_{jA}\ f_{jC}\ f_{jG}\ f_{jT}] \times \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix}$$

This operation maps each PSAM column containing four elements to the Cartesian coordinates (x, y, z) within a tetrahedron. Note that the tetrahedron has corners located at positions $(1, 1, 1)$, $(1, -1, -1)$, $(-1, 1, -1)$, and $(-1, -1, 1)$, corresponding to a strict preference for A, C, G, and T, respectively.

Alignment of bHLH protein sequences

The protein sequences of bHLH transcription factors were obtained from the cloned sequence tables provided by (13, 33, 34) and aligned using the multisequence alignment function provided by Clustal Omega. The resulting sequence alignment is consistent with PFAM entry PF00010 (https://www.ebi.ac.uk/interpro/entry/pfam/PF00010/), with amino-acid positions in the range 1-56.

MANOVA analysis

A series of MANOVA tests were performed to examine the association across all bHLH factors between amino acid identity (independent variable) at a given residue position in the protein alignment and tetrahedral position (dependent variable) reflecting the base preferences for a given position in the DNA binding model. For each combination of residue position and DNA position, we used the summary.manova() function in R, with the F-statistic computed using the Pillai-Barlett trace test. The null hypothesis is that the mean tetrahedral position is the same for all amino acids.

bHLH replicate comparisons used as reference for predictions

Of the 52 unique bHLH proteins, 36 had two replicates available (when more than two replicates were available, we chose the two with best correlation between predicted and observed 8-mer enrichment). The ΔΔG/RT estimates from the binding model for these replicates were compared to obtain a reference for predictive performance.

Definition of closest paralog for predicting binding specificity

For each of the 52 bHLH proteins that was held out, the protein with the smallest Levenshtein distance in terms of protein sequence was selected from the remaining 51 sequences. These ΔΔG/RT estimates from the binding model for this closest paralog were then used as predicted values.

Construction of SVD-regression model

To perform SVD regression for a given DNA position, we first constructed a tetrahedral coordinate matrix M whose three columns correspond to the coordinates of the 3D space in which the tetrahedron is embedded, and whose rows correspond to the set of bHLH proteins analyzed. The 52x3 matrix M was centered by subtracting the column mean from each value. Next, we performed singular value decomposition (SVD): $M = UDV^T$ using the base function svd() in R. The columns of 3x3 matrix V define a natural data-driven basis for the vector space inside the tetrahedron, while the columns of U correspond to the projection of the points inside the tetrahedron along each of these principal component (PC) directions.

For each PC we separately constructed a linear regression model that predicts the position of an unseen bHLH factor from its protein sequence. Each residue position in the bHLH multiple alignment corresponds to a feature that could be used as a predictor. The importance of each of the 168 features (56 residue positions × 3 PCs) was assessed using an ANOVA test implemented using the aov() function in R, with the amino acid identity as the independent variable and the values in each column of U as the dependent variable. Each residue position is associated with three p-values, one for each PC. For each PC, we ranked the protein features by their ANOVA p-value, and iteratively selected for the features to include. For each PC, we started by fitting a linear regression model based on a single feature. Next, we compared with a linear model using an additional feature. Along as the p-value of the F-test was below 0.05, indicating a significant improvement in variance explained, we included the

additional feature in our model, and proceed to the next feature.

To compare the significance level between the prediction to label R^2 and RMSD values between the SVD-regression model and the closest sequence method. A set of 52 trials each with one left-out sample was performed using each model, and a t-test was used to obtain the p-value between the set of 52 paired R^2 and RMSD values.

Wildtype and mutant HES2 and ASCL2 protein expression

ASCL2 and HES DNA binding domain (DBD) sequences were expressed based on the cloned sequences described in (34). To create mutant proteins, site-directed PCR was performed using primers containing mutations at the 5th and 13th residue positions. Both ASCL2 wildtype and mutants were tagged with mScarlet to enhance solubility. Affinity purification was carried out for both HES2 and ASCL2 wildtype and mutants using a poly-histidine tag.

For HES2, the wildtype and mutant sequences were cloned into the pQE30 vector and expressed in BL21 cells. Standard procedures were followed for protein purification, with the exception of using a low Imidazole wash buffer instead of the regular wash buffer. The lysis buffer remained unchanged throughout the process. The final concentration of HES2 obtained was 17±3μM for wildtype and R5K mutant and 70±20μM for R13V and double mutant.

For ASCL2, the wildtype and mutant sequences were cloned into the pET11 vector and expressed in BL21 cells. Similar to HES2, the proteins were purified using standard procedures, with the use of a low Imidazole wash buffer. The lysis buffer for ASCL2 also remained unchanged. The final concentrations obtained were 14±1 μM for ASCL2 wildtype and mutants.

SELEX assays

SELEX experiment were performed following the protocol outlined in Riley*, et al.* (11) and a SELEX library with a 16-mer randomized region whose full sequence is as follows:

`GTTCAGAGTTCTACAGTC-CGACCTAA-16N-TTAGG-ACTCGGACCTGGACTAGG`

The libraries were double strand DNA annealed and extended with Klenow polymerase. 66nM of HES2 and ASCL2 wildtype and mutant proteins was added to each DNA binding reaction.

EMSA assays were performed alongside the SELEX reactions to define the position of the bound band in the gel. The high-affinity probes for HES2 and ASCL2 binding, respectively, were as follows:

HES2 `CTCTCCTCCGTCAACACGTGTTGAGCAGCGCAGTCGTATGCCGTCTTCTGCTTG`

ASCL2 `CTCTCCTCCGTCAACAGCTGTTGAGCAGCGCAGTCGTATGCCGTCTTCTGCTTG`

50nM of probe was added to 150 nM of ASCL2 wildtype protein and 150 nM of HES2 wildtype protein to locate the bands. The bound band was cut out, purified, PCR amplified, and sequenced with Illumina sequencing according to protocol of a previous study (39).

SELEX data analysis

The raw FASTQ files for each of the eight protein samples were analyzed using the R/Bioconductor package SELEX (11). An 8mer enrichment analysis was performed on all protein samples, and each unique 8mer sequence was assigned an affinity score relative to the most enriched 8mer.

To compare the relative enrichment of the CACGTG 6mer and CAGCTG 6mer for each protein sample, 8mer pairs were selected based on the following criteria: both contain a CANNTG E-box and are identical except for a CG/GC difference in the NN position. For instance, an example of an 8mer pair would be ATCACGTGAA and ATCAGCTGAA. All 8mer pairs that fulfilled these requirements were collected for each protein sample, and their affinity scores were plotted against each other.

ProBound (18) was applied for all 8 protein samples to construct binding free energy models using the JSON configuration file in **Supplemental Data S5**, resulting in symmetrical motif matrices with a CANNTG E-box core.

Wildtype and mutant ASCL2 EMSA experiment

EMSA assays were performed using ASCL2 wildtype and mutant protein and 32P-labeled DNA probes. Protein samples were prepared using the same procedure as above. The DNA probes used, however, were different to avoid affinity towards the flanking regions:
CACGTG probe: `TAGCCAATAACTTCGTCCCTCACGTGCATATAAGGAAGATCTAACCACCAATTTGG`
CAGCTG probe: `TAGCCAATAACTTCGTCCCTCAGCTGCATATAAGGAAGATCTAACCACCAATTTGG`
Each probe was synthesized with P32 labeled SRI sequence through annealing and Klenow reactions.

50nM of probe was combined with 500nM, 1μM, and 2μM of protein, respectively, in each binding reaction, and the same reagents as used in Ref. (40). The mixture was loaded into a polyacrylamide gel after 30 minutes of reaction time. An EMSA gel was run for 90 minutes at 4 degrees with 150mA current, and dried and imaged using a phosphor-imaging plate and Typhoon gel imager.

ACKNOWLEDGEMENTS

The research reported in this publication was supported by NIMH award R01MH106842 and NHGRI award R01HG003008 to H.J.B. and NIGMS award R35GM118336 to R.S.M. We thank members of the Bussemaker and Mann labs for useful discussions, and Xiang-Jun Lu for advice on 3DNA/DSSR.

AUTHOR CONTRIBUTIONS

MPG and HJB conceived of the project; XL, MPG, and HJB developed computational methods; XL, MPG, and CL wrote computer code; XL performed all computational analyses under the supervision of HJB; XL performed all wet lab experiments with input from WJG and under the supervision of RSM; XL and HJB wrote the paper with significant input from RSM; all authors edited and approved the paper.

SOFTWARE

All computer code developed for and used in this study, include scripts to generate all figure panels from scratch, is available at http://github.com/BussemakerLab/FamilyCode.

COMPETING INTERESTS

H.J.B. is a co-founder and shareholder of Metric Biotechnologies, Inc.

SUPPLEMENTAL DATA

Supplemental Data S1: DNA recognition models for the 52 bHLH factors analyzed in this study.

Supplemental Data S2: Interactive 3D representations of tetrahedrons. Open HTML files in browser to view and manipulate. Related to Figure 2A-C.

Supplemental Data S3: Empirical cumulative distribution of tetrahedral position along each principal component direction for various amino-acid positions. Related to Figure 4B.

Supplemental Data S4: Statistical significance of ANOVA test of PCs at DNA position –2/+2 and –3/+3. Related to Figure 4C.

Supplemental Data S5: JSON configuration file used for all ProBound analyses.

REFERENCES

1. D. S. Latchman, Transcription factors: an overview. *Int J Biochem Cell Biol* **29**, 1305-1312 (1997).
2. J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, N. M. Luscombe, A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* **10**, 252-263 (2009).
3. A. Tsai *et al.*, Nuclear microenvironments modulate transcription from low-affinity enhancers. *Elife* **6** (2017).
4. J. Crocker *et al.*, Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell* **160**, 191-203 (2015).
5. E. K. Farley *et al.*, Suboptimization of developmental enhancers. *Science* **350**, 325-328 (2015).
6. C. Rastogi *et al.*, Accurate and sensitive quantification of protein-DNA binding affinity. *Proc Natl Acad Sci U S A* **115**, E3692-E3701 (2018).
7. J. F. Kribelbauer, C. Rastogi, H. J. Bussemaker, R. S. Mann, Low-Affinity Binding Sites and the Transcription Factor Specificity Paradox in Eukaryotes. *Annu Rev Cell Dev Biol* **35**, 357-379 (2019).
8. N. K. Ragge *et al.*, SOX2 anophthalmia syndrome. *Am J Med Genet A* **135**, 1-7; discussion 8 (2005).
9. J. Dupont *et al.*, Adding evidence to the role of NEUROG1 in congenital cranial dysinnervation disorders. *Clin Genet* **99**, 588-593 (2021).
10. L. Jen-Jacobson, Protein-DNA recognition complexes: conservation of structure and binding energy in the transition state. *Biopolymers* **44**, 153-180 (1997).
11. T. R. Riley *et al.*, SELEX-seq: a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes. *Methods Mol Biol* **1196**, 255-278 (2014).
12. S. Ruan, S. J. Swamidass, G. D. Stormo, BEESEM: estimation of binding energy models using HT-SELEX data. *Bioinformatics* **33**, 2288-2295 (2017).
13. A. Jolma *et al.*, DNA-binding specificities of human transcription factors. *Cell* **152**, 327-339 (2013).
14. P. J. Park, ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* **10**, 669-680 (2009).
15. P. V. Benos, A. S. Lapedes, G. D. Stormo, Probabilistic code for DNA recognition by proteins of the EGR family. *J Mol Biol* **323**, 701-727 (2002).
16. G. D. Stormo, D. S. Fields, Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem Sci* **23**, 109-113 (1998).
17. B. C. Foat, A. V. Morozov, H. J. Bussemaker, Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* **22**, e141-149 (2006).
18. H. T. Rube *et al.*, Prediction of protein-ligand binding affinity from sequencing data with interpretable machine learning. *Nat Biotechnol* **40**, 1520-1527 (2022).
19. G. D. Stormo, Z. Zuo, Y. K. Chang, Spec-seq: determining protein-DNA-binding specificity by sequencing. *Brief Funct Genomics* **14**, 30-38 (2015).
20. M. F. Berger *et al.*, Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133**, 1266-1276 (2008).

21. J. Liu, G. D. Stormo, Context-dependent DNA recognition code for C2H2 zinc-finger transcription factors. *Bioinformatics* **24**, 1850-1857 (2008).
22. M. B. Noyes *et al.*, Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell* **133**, 1277-1289 (2008).
23. R. G. Christensen *et al.*, Recognition models to predict DNA-binding specificities of homeodomain proteins. *Bioinformatics* **28**, i84-89 (2012).
24. R. Pelossof *et al.*, Affinity regression predicts the recognition code of nucleic acid-binding proteins. *Nat Biotechnol* **33**, 1242-1249 (2015).
25. W. R. Atchley, W. Terhalle, A. Dress, Positional dependence, cliques, and predictive motifs in the bHLH protein domain. *J Mol Evol* **48**, 501-516 (1999).
26. R. Gordan *et al.*, Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep* **3**, 1093-1104 (2013).
27. S. R. Eddy, A new generation of homology search tools based on probabilistic inference. *Genome Inform* **23**, 205-211 (2009).
28. H. J. Bussemaker, B. C. Foat, L. D. Ward, Predictive modeling of genome-wide mRNA expression: from modules to molecules. *Annu Rev Biophys Biomol Struct* **36**, 329-347 (2007).
29. R. A. Armstrong, When to use the Bonferroni correction. *Ophthalmic Physiol Opt* **34**, 502-508 (2014).
30. T. Shimizu *et al.*, Crystal structure of PHO4 bHLH domain-DNA complex: flanking base recognition. *EMBO J* **16**, 4689-4697 (1997).
31. A. K. Aditham, C. J. Markin, D. A. Mokhtari, N. DelRosso, P. M. Fordyce, High-Throughput Affinity Measurements of Transcription Factor and DNA Mutations Reveal Affinity and Specificity Determinants. *Cell Syst* **12**, 112-127 e111 (2021).
32. M. T. Weirauch *et al.*, Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431-1443 (2014).
33. L. Yang *et al.*, Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Mol Syst Biol* **13**, 910 (2017).
34. Y. Yin *et al.*, Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356** (2017).
35. W. Yang, L. Deng, PreDBA: A heterogeneous ensemble approach for predicting protein-DNA binding affinity. *Sci Rep* **10**, 1278 (2020).
36. J. Si, R. Zhao, R. Wu, An overview of the prediction of protein DNA-binding sites. *Int J Mol Sci* **16**, 5194-5215 (2015).
37. K. Harini, D. Kihara, M. Michael Gromiha, PDA-Pred: Predicting the binding affinity of protein-DNA complexes using machine learning techniques and structural features. *Methods* **213**, 10-17 (2023).
38. H. Yuan, M. Kshirsagar, L. Zamparo, Y. Lu, C. S. Leslie, BindSpace decodes transcription factor binding signals by large-scale sequence embedding. *Nat Methods* **16**, 858-861 (2019).
39. J. F. Kribelbauer, X. J. Lu, R. Rohs, R. S. Mann, H. J. Bussemaker, Toward a Mechanistic Understanding of DNA Methylation Readout by Transcription Factors. *J Mol Biol* **432**, 1801-1815 (2020).
40. S. Feng *et al.*, Transcription factor paralogs orchestrate alternative gene regulatory networks by context-dependent cooperation with multiple cofactors. *Nat Commun* **13**, 3808 (2022).

FIGURES

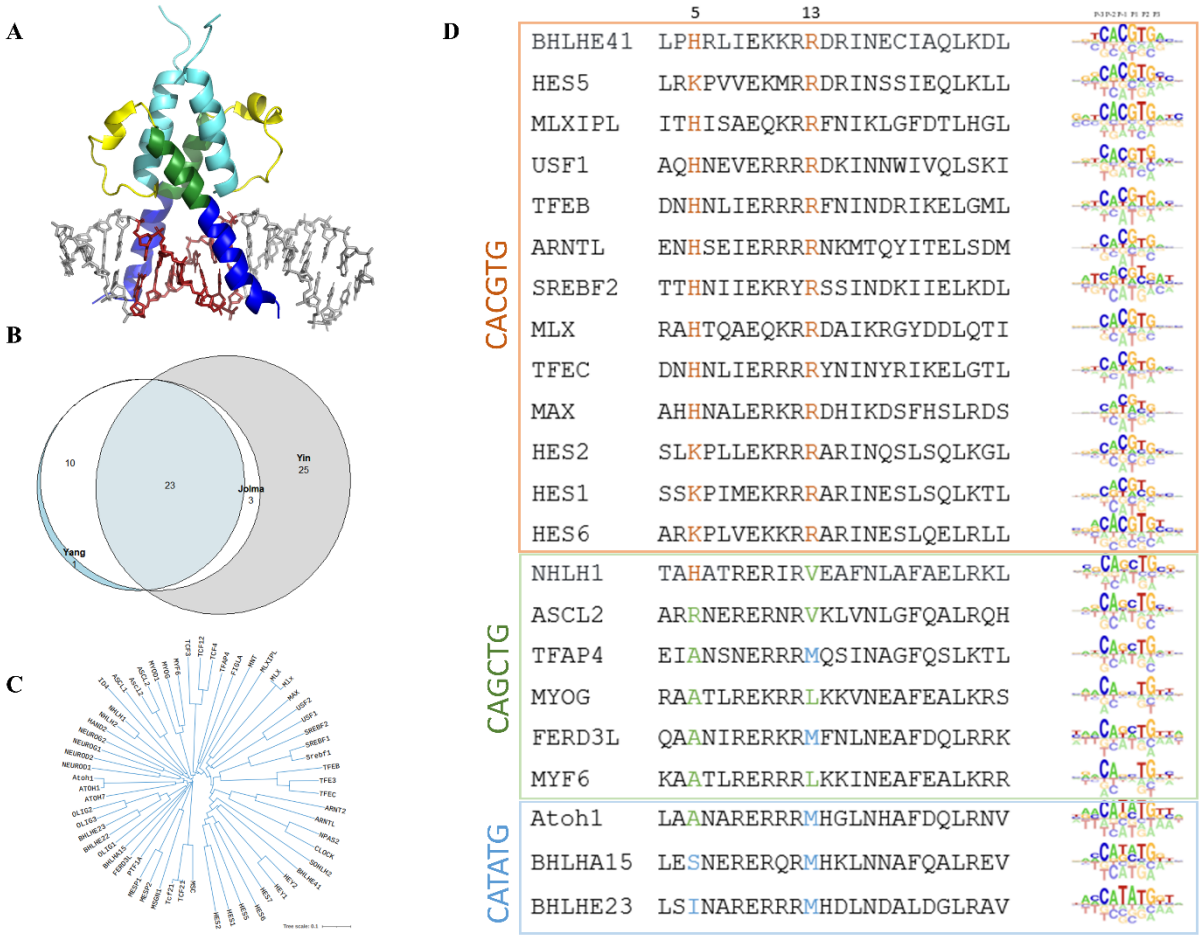

Figure 1: Variation in DNA binding preference within the helix-loop helix (bHLH) family of transcription factors. (**A**) Structure of the representative homodimeric bHLH transcription factor Pho4 bound to DNA. (**B**) We analyzed high-throughput (HT-SELEX) *in vitro* DNA binding data for 62 distinct human bHLH proteins. The Venn diagram shows the overlap among the TFs covered by each of the data sources (13, 33, 34). (**C**) Phylogenetic relationship among the bHLH proteins in our training set. (**D**) Aligned protein sequences and DNA binding energy logos for a representative set of bHLH factors, grouped by preferred E-box core.

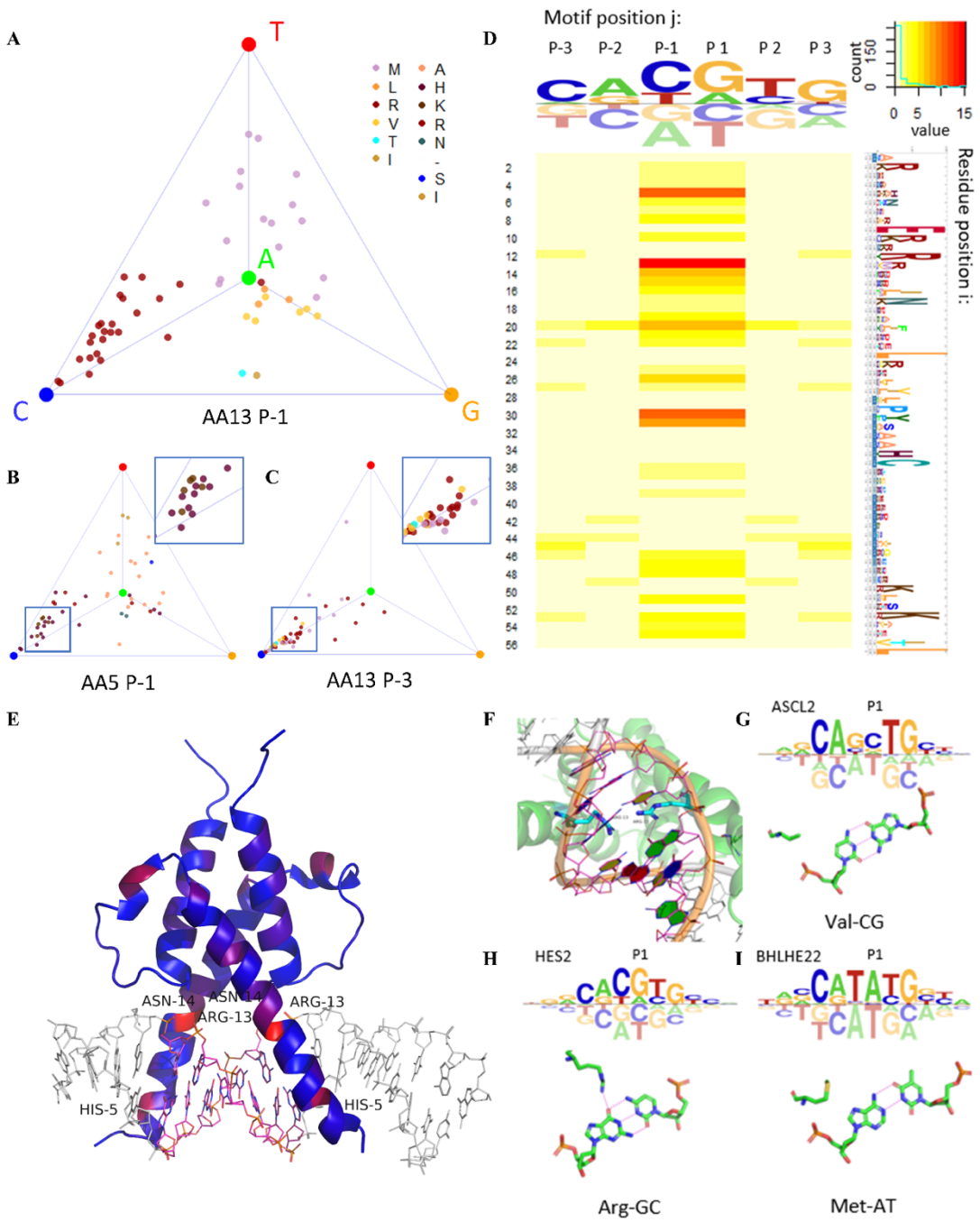

Figure 2: Family level analysis uncovers protein sequence determinants of DNA binding specificity. (**A-C**) Tetrahedron representation of base preference at DNA position –1 (panel A, B) or –3 (panel C). Each point represents a different bHLH factor, and is colored according to the amino acid identity at protein position 13 (panel A, C) or 5 (panel B). (**D**) Heatmap showing the p-value from a MANOVA test between amino-acid identity and position within the tetrahedron. (**E**) Statistical significance at position –1 shown in the context of the co-crystal structure for Pho4 (red denotes significant p-values). (**F**) Detailed view of the interaction between Arg13 and a GC base pair at motif position +1. The Arg13 sidechain is colored in cyan. The E-box is shown in magenta using DNA blocks. (**G-I**) Additional examples of interactions between amino-acids and base pairs at residue position 13 and DNA position +1. Structural images generated using 3DNA/DSSR (x3dna.org). Hydrogen bonds are shown as pink dashed lines.

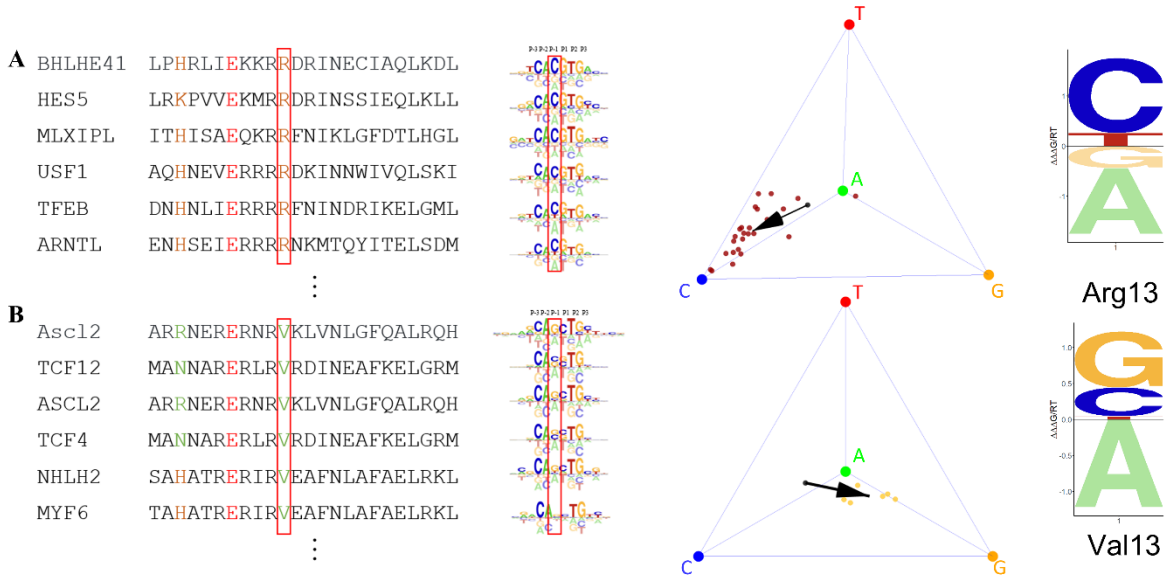

Figure 3: Predicting shifts in DNA binding preference associated with specific protein sequence features. (**A-B**) Protein alignment and binding motif of select bHLH proteins with either Arg (panel A) or Val (panel B) at residue position 13 (highlighted by red box). The energy logos and tetrahedron plots and the right show the base preference at DNA position –1 (red box). Arrows in the tetrahedrons indicate the shift relative to the overall centroid associated with the Arg13 and Val13 subset, which can again be represented by an energy logo (far right).

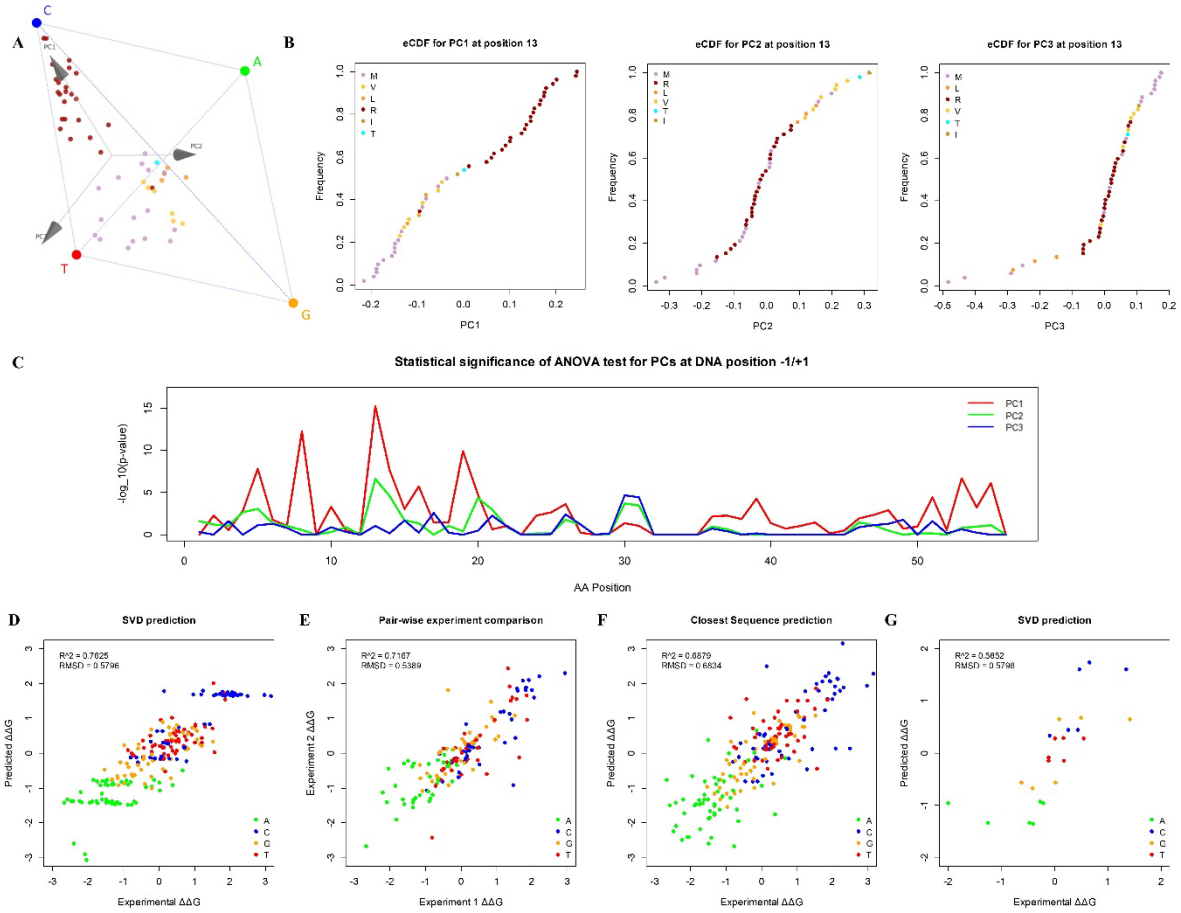

Figure 4: Using PCA-regression to predict the binding specificity of mutated bHLH proteins. (**A**) Principal component analysis of the set of tetrahedral coordinates defines a natural coordinate frame for each DNA position. (**B**) Empirical cumulative distribution of tetrahedral position along each principal component direction for DNA position –1, with points colored according to the amino-acid identity at protein position 13. (**C**) Statistical significance of ANOVA test covering all PCs over each protein position at DNA position –1/+1. (**D**) Direct comparison between predicted and observed binding free energy parameters . (**E**) Comparison between observed values for two replicates of the same TF. (**F**) Comparison in which the predicted value of the binding energy parameters is that of the closest paralog in the dataset. (**G**) Direct comparison between predicted and observed binding free energy parameters for HES2 and ASCL2 mutants.

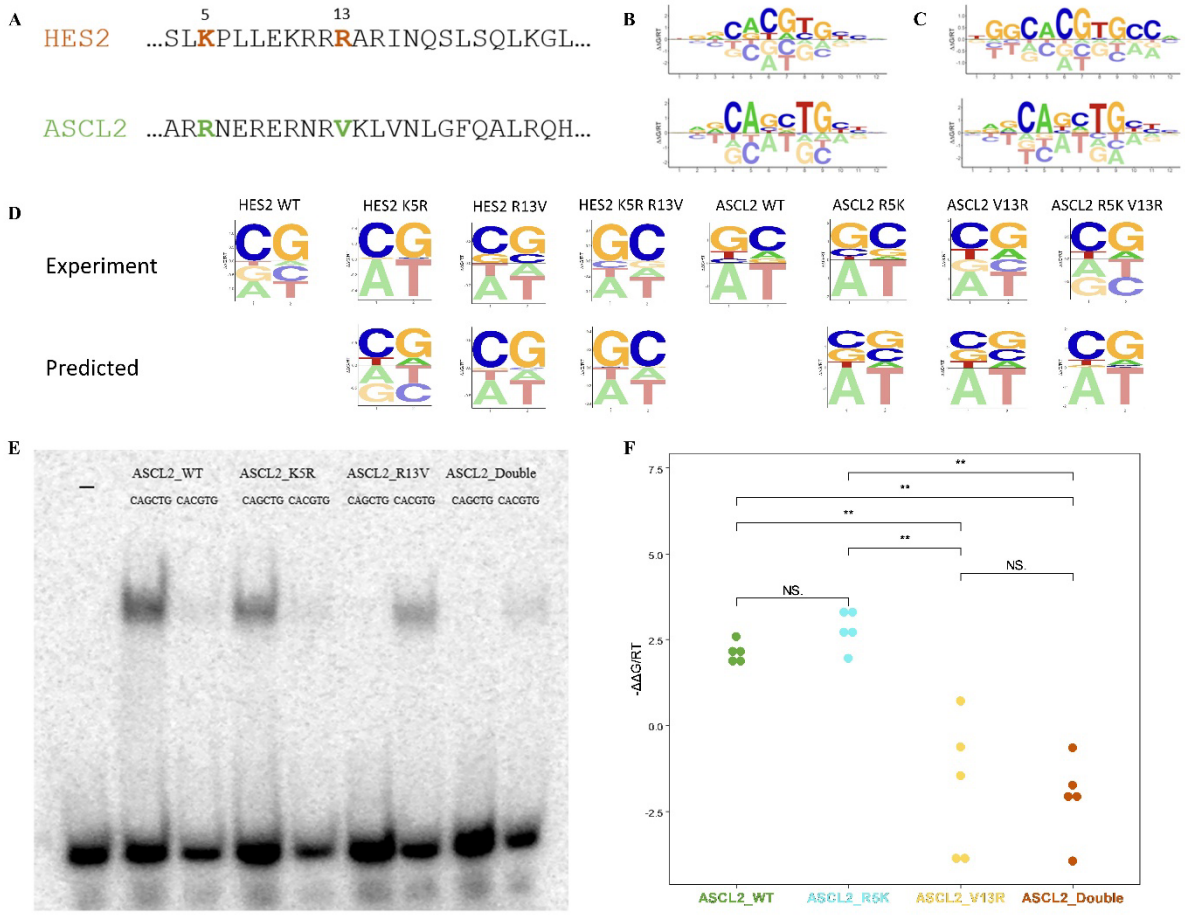

Figure 5: in vitro binding assays of single/double mutation in HES2 and ASCL2. (**A**) Illustration of HES2 and ASCL2 sequences and positions of mutations. (**B**) Binding energy calculation by ProBound using 1st round SELEX of High-throughput SELEX performed by Yin et al. (34). (**C**) Binding energy calculation by ProBound on WT HES2 and ASCL2. (**D**) Experimental and predicted binding energy of HES2 and ASCL2 single and double mutants at motif position –1/+1. (**E**) EMSA experiment on WT and mutant ASCL2 binding with CACGTG and CAGCTG probes. (**F**) Scatter plot of ΔΔG value by mutating from CACGTG to CAGCTG calculated with ProBound models from SELEX-seq experiment against ΔΔG values calculated from EMSA gel band intensity.