Published online 18 February 2011 Nucleic Acids Research, 2011, Vol. 39, No. 11 4553–4563
doi:10.1093/nar/gkr070

Using a structural and logics systems approach to infer bHLH–DNA binding specificity determinants

Federico De Masi[1], Christian A. Grove[2], Anastasia Vedenko[1], Andreu Alibés[3], Stephen S. Gisselbrecht[1], Luis Serrano[3,4], Martha L. Bulyk[1,5,6,*] and Albertha J. M. Walhout[2,*]

[1]Department of Medicine, Division of Genetics, Brigham & Women's Hospital and Harvard Medical School, Boston, MA 02115, USA, [2]Program in Gene Function and Expression and Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, MA 01605, USA, [3]EMBL-CRG Systems Biology Research Unit, Center for Genomic Regulation, Universitat Pompeu Fabra, 08003 Barcelona, Spain, [4]Institució Catalana de Recerca i Estudis Avançats (ICREA) Professor, Center for Genomic Regulation, Universitat Pompeu Fabra, 08003 Barcelona, Spain, [5]Department of Pathology, Brigham & Women's Hospital and [6]Harvard-MIT Division of Health Sciences and Technology (HST), Harvard Medical School, Boston, MA 02115, USA

Received December 13, 2010; Revised January 23, 2011; Accepted January 24, 2011

ABSTRACT

Numerous efforts are underway to determine gene regulatory networks that describe physical relationships between transcription factors (TFs) and their target DNA sequences. Members of paralogous TF families typically recognize similar DNA sequences. Knowledge of the molecular determinants of protein–DNA recognition by paralogous TFs is of central importance for understanding how small differences in DNA specificities can dictate target gene selection. Previously, we determined the *in vitro* DNA binding specificities of 19 *Caenorhabditis elegans* basic helix-loop-helix (bHLH) dimers using protein binding microarrays. These TFs bind E-box (CANNTG) and E-box-like sequences. Here, we combine these data with logics, bHLH–DNA co-crystal structures and computational modeling to infer which bHLH monomer can interact with which CAN E-box half-site and we identify a critical residue in the protein that dictates this specificity. Validation experiments using mutant bHLH proteins provide support for our inferences. Our study provides insights into the mechanisms of DNA recognition by bHLH dimers as well as a blueprint for system-level studies of the DNA binding determinants of other TF families in different model organisms and humans.

INTRODUCTION

The regulation of gene expression is partially controlled by transcription factors (TFs) that bind DNA in a sequence-specific manner and that function in the context of intricate gene regulatory networks (1,2). TFs can be grouped into families according to the structural class of their DNA binding domains. Members from some but not all TF families typically bind to similar DNA sequences (3–5). For instance, most homeodomain TFs prefer AT-rich sequences (6), whereas basic helix-loop-helix (bHLH) TFs bind E-box (CANNTG), or E-box-like sequences that differ from E-boxes in 1 or 2 nt (7).

A major question in the field of transcriptional regulation is what determines the small differences in DNA binding site specificity between different members of TF families. In particular, how do differences in protein sequence and structure result in differences in DNA binding specificity and thereby target gene selection? Many studies of protein–DNA recognition have focused on C2H2 zinc fingers, for which relatively accurate prediction of DNA binding specificities can be achieved (8–10).

*To whom correspondence should be addressed. Tel: +1 508 8564364; Fax: +1 508 8565460; Email: marian.walhout@umassmed.edu
Correspondence may also be addressed to Martha L. Bulyk. Tel: +1 617 5254725; Fax: +1 617 5254705; Email: mlbulyk@receptor.med.harvard.edu
Present addresses:
Federico De Masi, Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Building 208, DK-2800 Kongens Lyngby, Denmark.
Christian A. Grove, California Institute of Technology, Pasadena, CA 91125, USA.

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

However, the existence of a simple, amino acid to nucleotide 'recognition code', in which certain amino acids specify interactions with particular bases (11,12) has been largely discounted as more intricate relationships between the biochemical and biophysical properties of TFs and their DNA binding sites have been revealed (6,8–10,13–16).

The analysis of protein–DNA binding specificity is greatly enhanced by the comprehensive determination of DNA binding specificities, measured on a single experimental platform, for a diverse set of proteins belonging to a TF family. This approach permits even very small differences in DNA binding specificity to be assessed. Such data sets were recently generated on a large scale and were used to gain insights into the DNA binding specificity determinants of *Drosophila* and mouse homeodomain TFs (6,17). However, for the vast majority of other TF families, such determinants remain largely unknown.

Many TFs-bind DNA as heterodimers that complicates the analysis of protein–DNA recognition (2). This is because each monomer contributes to different aspects of binding site specificity. In heterodimers, particularly, it is therefore important to first assess which monomer contacts which part of the bound DNA sequence. Here, we analyze the molecular mechanisms of DNA recognition by bHLH proteins which bind DNA as obligate homo- or heterodimers (18). We chose this family for several reasons. First, multiple co-crystal structures have been obtained (19–21) that provide a high-resolution view of bHLH dimerization and DNA binding. Dimerization between bHLH proteins is mediated by two α-helices from each of the bHLH monomers that face each other, and DNA interactions occur mainly with residues in the basic region, located N-terminal to the first α-helix (Supplementary Figure S1). Second, we previously determined the dimerization and DNA binding specificities of most *C. elegans* bHLH homo- and heterodimers (7) using yeast two hybrid (Y2H) assays (22) and universal protein binding microarrays (23). These data sets provide high-confidence insights into the sometimes subtle differences in DNA binding specificity and target gene selection between members of this important TF family. We first infer the half-site specificities of each of the bHLH monomers. Subsequently, we interrogate available bHLH-E-box co-crystal structures for the amino acids that contact the DNA. We identify a residue in the bHLH-basic region that partially explains the specificity for interacting with different E-box half-sites. PBM assays on three HLH-1 mutant proteins provide support for these determinants.

MATERIALS AND METHODS

bHLH-contact frequency plots

Co-crystal structures of the upstream stimulatory factor (USF) (PDB:1AN4) (24), Max (PDB:1HLO) (25), Pho4 (PDB:1A0A) (26), MyoD (PDB:1MDY) (27) homodimers and Max/Mad (PDB:1NLW), Max/Myc (PDB:1NKP) (28) and E47/NeuroD (PDB:2QL2) (21) heterodimers were analyzed for proximity of DNA-contacting residues to each nucleotide base, deoxyribose or phosphate group. Contacts within 3.5 Å were considered 'direct' contacts, and distances up to 5 Å away were considered 'partially' contacting. All observed 'direct' contacts and 'partial' contacts (counted as 1 and 0.5, respectively) were summed for each amino acid position and divided by the total number of instances of that amino acid that could be analyzed. This number depended on the total number of bHLH monomers represented in the seven crystal structures (22 total, see below) and whether or not the amino acids were represented adequately (e.g. alanine substitutions were not counted). There were either two (1HLO:Max, 1AN4:USF, 1A0A:Pho4) or four (1NLW:Max/Mad, 1NKP:Max/Myc, 1MDY:MyoD, 2QL2:E47/NeuroD) bHLH monomers in each bHLH-crystal structure depending on the crystal unit composition, resulting in 22 bHLH monomers in total. The resulting 'contact frequency score' indicates the normalized contacting frequency for each residue; a score of 100 indicates a 'direct' contact in 100% of the monomers observed. Note that a score of 50 could indicate either a 'partial' contact in 100% of the monomers observed, or a 'direct' contact in 50% of the monomers observed, or some combination of such contacts.

In silico modeling

The change in interaction energy upon mutation of each of the 10 DNA-contacting residues of the bHLH-PDB structures listed above were calculated with the protein design tool FoldX (29,30). The side chain positions were first optimized with the RepairPDB command, to correct residues with positive internal energy. Next, each of the 10 DNA-contacting residues was mutated to each of 20 amino acids and the resulting change in interaction energy upon mutation ($\Delta\Delta G_{int}$) was calculated. We defined $\Delta\Delta G_{int}$ as the difference in interaction energy between the structures of the wild-type and mutated proteins plus the difference (if >0.6 kcal/mol) in intramolecular clashes (both for DNA and for the protein) in order to penalize potentially destabilizing mutations. FoldX simulations were performed for each mutation 5×, to increase the conformational space explored, and the results were averaged. We have increased the default number of rotamers used in FoldX 20-fold, in order to improve the accuracy of the results. In the case of homodimers, the mutation was introduced in both monomers and its $\Delta\Delta G_{int}$ effect was divided in half, while for heterodimers each of the individual mutations was evaluated separately. Values of $\Delta\Delta G_{int} > 1.0$ kcal/mol might significantly affect the DNA binding properties of the mutant, while values >2.0 kcal/mol may completely disrupt the DNA binding capabilities of the mutant. All FoldX calculations were performed assuming temperature equal to 298 K, pH 7.0 and 150 mM ionic strength. The relative binding strength of each pair of monomers was calculated as the ratio of the monomers' K_ds, which were calculated from their $\Delta\Delta G_{int}$ values (K_d is proportional to $e^{-\Delta G/RT}$).

Cloning, protein expression and quantification

A wild-type *C. elegans hlh-1* open reading frame was synthesized by gene synthesis and cloned into the pGS21a vector (GenScript USA, Inc). This clone was used to generate the L13R, L13T and L13V mutants (GenScript USA, Inc). All clones used to produce protein for PBM experiments were full-length DNA sequence verified. All protein constructs were expressed using the PURExpress coupled *in vitro* transcription and translation system (New England BioLabs, Inc.) following the manufacturer's instructions. Protein concentrations were estimated by western blotting by comparison to a dilution series of recombinant glutathione-S-transferase (Sigma).

PBM experiments and data analysis

Microarray design, double stranding of the oligonucleotide arrays by primer extension, and PBM experiments were performed essentially as described previously using custom-designed 'all 10-mer' arrays synthesized in the Agilent '4 × 44K' array format (Agilent Technologies, Inc.; AMADID #015681) (31). All PBM experiments were performed in duplicate at a final concentration of 400 nM wild-type or mutant HLH-1 protein, except for the HLH-1 L13T mutant, which was assayed at 460 nM in one of the PBM experiments. Microarray scanning, spot quantification, data filtering, normalization and analysis were performed as described previously (31).

Statistical analysis of HLH-1-mutant binding preferences for E-box sequences

For each of the four HLH-1 constructs (wild-type and three mutants), ungapped 8-mers were ranked by PBM E-score. Statistical significance of preferential or disfavored binding of 8-mers containing either all E-boxes or particular E-box sequence variants was assessed by a Wilcoxon–Mann–Whitney U-test. Preferential or disfavored binding of each E-box variant by each of the three mutant proteins was compared with wild-type HLH-1 by computing a U-statistic within the context of all 8-mers and, separately, of just E-box-containing 8-mers. The boxplots shown in Figure 4d were created using the 'boxplot' function in the 'graphics' library of the R-project statistical software (http://www.r-project.org). Significantly bound E-boxes were identified as those E-boxes for which the 75th percentile of their E-score distribution is >0.4 and, as previously described (7), their q-value is <0.001 and their area under the curve (AUC) is >0.85.

RESULTS

DNA binding specificity of individual bHLH proteins

All *C. elegans* bHLH dimers, with the exception of HLH-27 and HLH-29, bind E-boxes, defined as CANNTG. Here, we investigated the mechanisms of DNA binding by 19 *C. elegans* bHLH proteins, using PBM data that we recently generated (7). We focused our analysis exclusively on E-boxes, since only E-boxes were represented in available bHLH–DNA co-crystal structures (19,21,24–28).

Within a bHLH dimer, each monomer contacts a CAN half-site (Supplementary Figure S1). We reasoned that the DNA binding specificities of each bHLH dimer would enable the inference of the half-site preferences of the participating monomers by the following logic (Figure 1). When a bHLH homodimer binds a palindromic DNA sequence, each of the monomers binds to an identical half-site. We hypothesized that if only a single palindrome is bound by a homodimer, then that half-site should be optimal for its corresponding monomer. For instance, HLH-30 homodimers bind only the CACGTG palindrome, indicating that each HLH-30 monomer prefers a CAC half-site over the other three possible half-sites (Table 1). We first performed this type of analysis for each of the nine *C. elegans* bHLH homodimers' PBM data (7). We subsequently extended this analysis to the bHLH proteins that form the 10 heterodimers that we analyzed (7).

There are 10 non-redundant E-boxes when both the forward and reverse complementary orientations are considered (e.g. CACCTG and CAGGTG are the same E-box). These include four palindromes, one of which is not bound by any bHLH dimer (CAATTG, Table 1). This strongly suggests that bHLH proteins disfavor the CAA

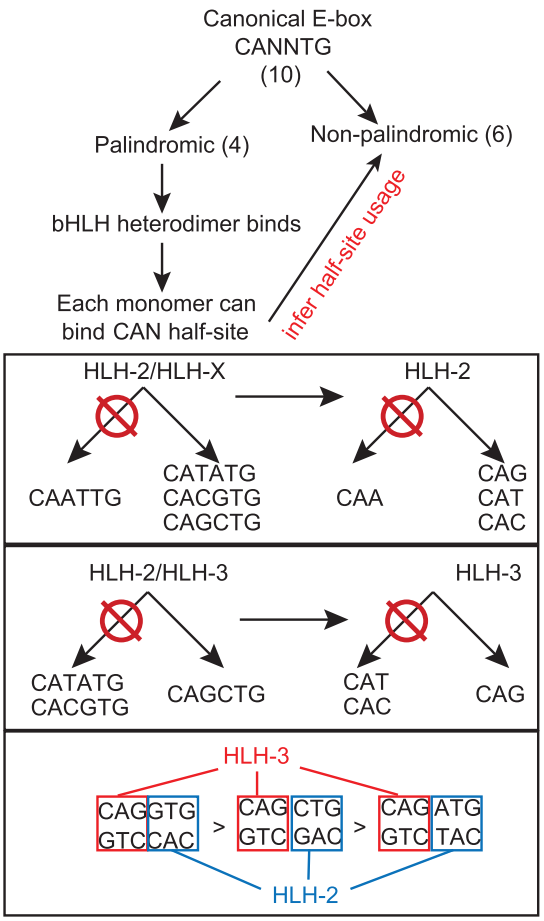

Figure 1. Deduction of interactions between bHLH monomers and CAN E-box half-sites. HLH-2/HLH-3 heterodimers are shown as an example.

Table 1. Deduced half-site preferences for 19 *C. elegans* bHLH proteins

bHLH monomer	Deduced half-site preference	Dimerization partner	Predicted binding sites of bHLH dimer	Observed binding sites of bHLH dimer
HLH-2	CAG, CAC, CAT	Several	See below	See below
HLH-3	CAG	HLH-2	CAGCTG, CACCTG, CATCTG	CAGCTG (0.96), CACCTG (0.98), CATCTG (0.91)
HLH-4	CAG	HLH-2	CAGCTG, CACCTG, **CATCTG**	CAGCTG (0.98), CACCTG (0.99)
HLH-10	CAG, CAC, CAT	HLH-2	CAGCTG, CACCTG, CATCTG, CACGTG, **CATGTG**, CATATG	CAGCTG (1.00), CACCTG (1.00), CATCTG (0.95), CACGTG (0.93), CATATG (0.94)
HLH-15	CAG	HLH-2	CAGCTG, CACCTG, **CATCTG**	CAGCTG (0.96), CACCTG (0.95)
HLH-8	CAT	HLH-2	CAGATG, CACATG, CATATG	CATCTG (0.86), CACATG (0.88), CATATG (0.90)
LIN-32	CAG, CAT	HLH-2	CAGCTG, CACCTG, CATCTG, **CACATG**, CATATG	CAGCTG (0.98), CACCTG (0.94), CATCTG (0.93), CATATG (0.88)
HLH-14	CAG, CAT	HLH-2	CAGCTG, CACCTG, CATCTG, **CACATG**, CATATG	CAGCTG (0.93), CACCTG (0.94), CATCTG (0.87), CATATG (0.88)
HLH-19	CAG, CAT	HLH-2	CAGCTG, CACCTG, CATCTG, **CACATG**, CATATG	CAGCTG (0.99), CACCTG (0.94), CATCTG (0.89), CATATG (0.95)
CND-1	CAG, CAT	HLH-2	CAGCTG, **CACCTG**, CATCTG, **CACATG**, CATATG	CAGCTG (0.86), CATCTG (0.91), CATATG (0.98)
HLH-11	CAG, CAT	Self	CAGCTG, **CATCTG**, CATATG	CAGCTG (0.93), CATATG (0.96)
HLH-1	CAG	Self	CAGCTG	CAGCTG (0.95), **CACCTG (0.92)**, **CAACTG (0.86)**
REF-1	CAC	Self	CACGTG	CACGTG (0.99)
HLH-25	CAC	Self	CACGTG	CACGTG (0.98), **CACATG (0.88)**
HLH-26	CAC	Self	CACGTG	CACGTG (0.86)
MXL-3	CAC	Self	CACGTG	CACGTG (1.00), **CACATG (0.93)**
MDL-1	CAC	MXL-1	CACGTG	CACGTG (1.00), **CACATG (1.00)**
MXL-1	CAC	MDL-1	Same as MDL-1	Same as MDL-1
HLH-30	CAC	Self	CACGTG	CACGTG (0.98)

The table indicates the bHLH TF, its (deduced) preferred E-box half-sites and dimerization partner(s). Additionally shown is a comparison of E-boxes predicted to be bound based on these deduced half-site preferences (E-boxes predicted but not observed are indicated in bold text) versus E-boxes observed to be bound based on actual experimentation (E-boxes observed but not predicted are indicated in bold text). Previously published AUC values (7) for each observed E-box are indicated in parentheses to illustrate the differences in the relative binding affinities of each bHLH to the different E-boxes bound. Note that the predictions are based only on observed palindromic E-box binding.

half-site. Only a single CAA-containing, non-palindromic E-box was bound, and only by the HLH-1 homodimer (7). This indicates that only HLH-1 monomers, and no other bHLH proteins, tolerate a CAA half-site but only when another, more optimal half-site is present. HLH-30, MXL-3, HLH-25, HLH-26 and REF-1 homodimers can bind to only one palindromic E-box, CACGTG, and so each of these proteins must be able to recognize the CAC half-site. Similarly, HLH-1 binds only the CAGCTG palindrome, and so each HLH-1 monomer likely prefers CAG. The fact that HLH-1 binds CAA only in the context of the CAACTG E-box (7) (Figure 4d below) provides further support that CAG is its preferred half-site (i.e. CAACTG is composed of a CAA and a CAG half-site). The observation that most bHLH homodimers bind only a single palindrome further suggests that these proteins indeed prefer the corresponding half-site. In fact, in most cases these palindromes exhibited the highest PBM-enrichment scores (E-scores) (7), a relative measure of DNA binding preference (23). Moreover, most of the non-palindromic bound E-boxes contain one half-site from the preferred palindromic E-boxes and palindromes of less preferred half-sites are typically not strongly preferred. For example, the MXL-3 homodimer significantly prefers the CACGTG E-box, but also binds CACATG, composed of CAC and a CAT half-sites, but not the CATATG palindrome (Table 1) (7). This strongly suggests that MXL-3 monomers prefer CAC over CAT.

We used available co-crystallographic data for bHLH dimers bound to an E-box to estimate, *in silico,* the relative binding affinities (K_ds) of each bHLH monomer to DNA using the protein design tool FoldX (29,30,32). The available co-crystallographic data are for the mammalian homodimers Max (19,25), USF (24) and MyoD (27); the heterodimers Myc/Max, Mad/Max (28) and E47/NeuroD (21); and the yeast homodimer Pho4 (26). This FoldX analysis predicted that the two monomers within USF and Max homodimers exhibit ~100-fold and ~10-fold differences, respectively, in half-site binding affinities. In other words, one USF monomer binds to its half-site 100× more strongly than the other USF monomer to the other half-site. Individual Pho4 and MyoD monomers, on the other hand, appear to bind to each half-site with similar affinities (Supplementary Table S1). Similarly, within the three available heterodimer structures bound to the CACGTG palindromic E-box (Myc/Max, Mad/Max (28) and E47/NeuroD (21)), the K_ds of each monomer/half-site interaction are predicted by FoldX to exhibit differences of two to six orders of magnitude within the same structure (Supplementary Table S1). These observations suggest that the two monomers that comprise a bHLH dimer can bind each E-box half-site with different affinities. In non-palindromic E-boxes, the presence of a high-affinity half-site could provide a platform for one bHLH monomer and enable the other bHLH monomer to tolerate a lower affinity half-site.

We next investigated the half-site preferences of *C. elegans* bHLH proteins that bind DNA exclusively as heterodimers. HLH-2-containing heterodimers collectively

bind all palindromes except CAATTG. Thus, the HLH-2 monomer must be able to bind CAT, CAC and CAG half-sites (Figure 1 and Table 1). This is in agreement with crystal structures of dimers containing E47, the human HLH-2 ortholog, which collectively detected E47 binding to each of these half-sites (20,21). HLH-2 dimerizes with 14 other bHLH proteins, and the resulting heterodimers have partially overlapping DNA-binding specificities (Table 1) (7). The fact that the DNA binding specificities of HLH-2 heterodimers are non-identical indicates that each HLH-2 partner must, at least in part, be responsible for the specification of particular E-boxes, presumably by preferring particular half-sites. For example, the HLH-2/HLH-3 heterodimer binds only the CAGCTG palindrome, indicating that HLH-3 prefers CAG (Figure 1 and Table 1). The non-palindromic sequences bound by the HLH-2/HLH-3 heterodimer (CAGGTG and CAGATG) both contain a CAG half-site. We, therefore, propose that HLH-3 specifies the CAG half-site, and that HLH-2 specifies the other half-site, although we cannot exclude the possibility that HLH-3 half-site preference may be E-box dependent. By similar reasoning, we propose that HLH-8, which together with its partner HLH-2 binds the CATATG palindrome and the CATCTG and CATGTG non-palindromes, specifies a CAT half-site. All deduced half-site preferences for HLH-2 and its partners are provided in Table 1.

HLH-2 half-site specification can depend on its dimerization partner

Our half-site model suggests that an HLH-2 monomer can bind to CAC, CAG or CAT half-sites. Surprisingly, however, we found that the half-site preference of HLH-2 can depend on its dimerization partner. For instance, when HLH-2 is dimerized with HLH-3, the bound E-boxes are in order of higher to lower relative affinity, CACCTG, CAGCTG and CATCTG (Table 1) (7). Each of these E-boxes contains the CAG half-site likely specified by HLH-3, and thus HLH-2 likely prefers CAC, followed by CAG and then CAT in the context of this heterodimer. When HLH-2 is dimerized with HLH-8, however, CATATG, CATCTG and CACATG are bound, with the latter two E-boxes being bound with similar PBM-enrichment scores (Table 1) (7). Since we propose that HLH-8 specifies CAT, this suggests that HLH-2 prefers CAT in the context of this heterodimer, followed by CAG and CAC. Thus, HLH-2 may prefer either a CAC or CAG half-site in the context of the HLH-2/HLH-3 heterodimer, but prefers CAT when dimerized with HLH-8. In sum, bHLH proteins that can dimerize with multiple partners appear to prefer different half-sites in the context of different heterodimers.

Analysis of bHLH monomer–DNA contacts

We next analyzed all available co-crystallographic data for bHLH–DNA complexes, i.e. eight different bHLH monomers bound to three different E-boxes. Structural alignment of these co-crystal structures indicated that all eight bHLH dimers exhibit extremely similar DNA binding geometry (mean RMSD = 1.49 ± 1Å) (Supplementary Figure S2). To inspect the protein–DNA interfaces of these bHLHs in more detail, we devised a strategy to determine the relative contribution of contacts between each amino acid in the entire bHLH domain and the bound E-box. By inspecting each bHLH–DNA co-crystal structure, we generated a frequency map of molecular contacts made by bHLH residues and each nucleotide position in and around the E-box. From this frequency map, we inferred which residues in the bHLH domain are most likely to participate in directing DNA binding specificity.

bHLH proteins contact DNA with residues in their basic regions (indicated in green in Figure 2a). Although each E-box half-site is contacted by a bHLH monomer, there are also extensive contacts across the E-box, i.e. with the other half-site, and most frequently these occur on the

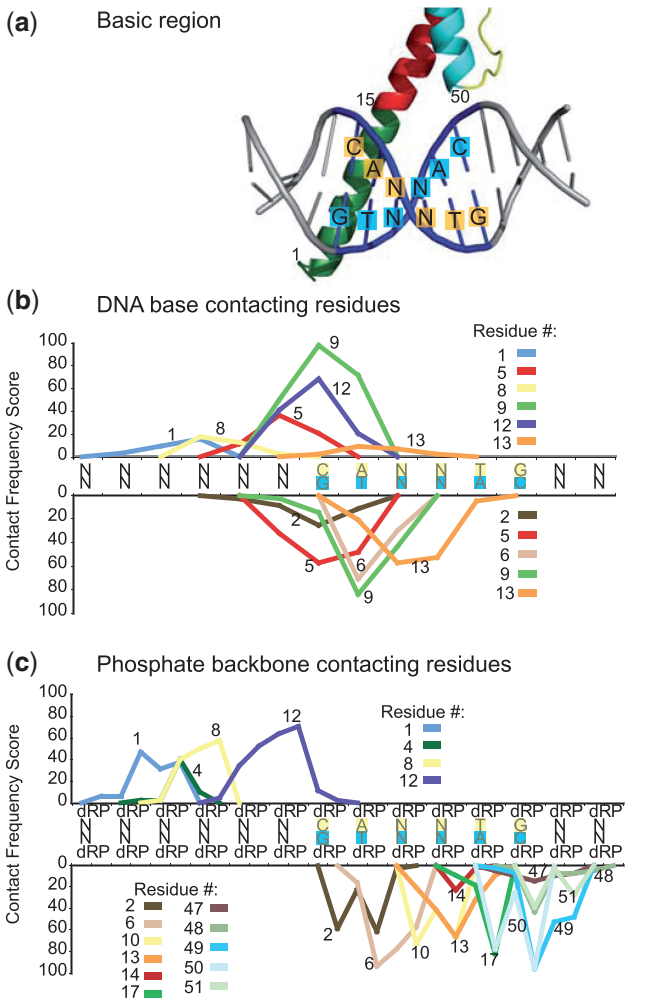

Figure 2. Analysis of bHLH residue/DNA-contact frequency. (**a**) Canonical structure of a bHLH monomer contacting the CANNTG E-box. The E-box nucleotides on each DNA strand are indicated in blue, the bHLH basic region is shown in green and bHLH helices 1 and 2 are indicated in red and cyan, respectively. Residues 1, 15 and 50 indicate the start of the basic region, helices 1 and 2, respectively. (**b**) Plot indicating the frequency of contacts between bHLH residues and the bases in or flanking the E-box based on seven co-crystal structures. (**c**) As in (b) but indicating contacts between bHLH residues and the DNA backbone. P, phosphate; dR, deoxyribose. bHLH residue consensus numbering scheme based on Atchley *et al.* (33).

opposite strand of DNA (Figure 2b and c) (20,21). For instance, residue 13 in the basic region often contacts the base at position 4 of the CANNTG E-box (Figure 2b). This suggests that residue 13 may contribute to half-site specificity on the opposite side of the E-box.

The most critical residue for E-box binding is a nearly invariant glutamate at position 9 that contacts both the C and the A in the CAN half-site (33). Our analysis revealed that a large number of base contacts are also made by residues 5, 6 and 12, which contact either the C or the A in the CAN half-site, the 5′ flanking nucleotide, or a combination of these (Figure 2b). In addition to base contacts, multiple amino acids make contacts with the DNA backbone; each bHLH monomer contacts the backbone of the entire E-box and a number of 5′ and 3′ flanking nucleotides (Figure 2c). Our observations support the role of extensive backbone contacts in stabilizing bHLH–DNA interactions (21).

bHLH half-site recognition determinants

To identify correlations between particular amino acid residues and bound DNA sequences for our *C. elegans* data set, we focused on the positions identified in the above frequency map and coupled bHLH half-site specificity to the amino acids at each of those 17 DNA-contacting positions (Figure 3a). In order to obtain a comprehensive understanding of the half-site recognition mechanism of each bHLH protein, we included all half-sites present at least once in each significantly bound E-box (e.g. HLH-1 is associated with CAA, CAC and CAG) (Table 1). As reported previously, bHLH proteins containing an arginine at position 13 (Arg13) prefer CAC (33–35). However, while HLH-30, HLH-26 and REF-1 exclusively specify CAC, MDL-1, MXL-1, MXL-3 and HLH-25 also specify CAT (Figure 3a). Interestingly, these proteins exhibit a difference at position 2 in the basic region: proteins with an arginine (or a glutamine in case of the first bHLH domain of HLH-25) at position 2 bind both CAC and CAT, whereas the remaining proteins bind exclusively to CAC. This suggests that Arg2/Gln2 may enable CAT specification in the context of Arg13-containing proteins. PBM data obtained for bHLH proteins from other species (available from the UniPROBE database) (36) further support our model; of the five high-quality PBM data sets for Arg13-containing bHLH proteins (*Saccharomyces cerevisiae* bHLH TFs Cbf1p, Tye7p and Pho4p; and mouse Max and bHLHb2), three exactly matched our residue/half-site correlations. Tye7p has Arg13 but not Arg2 and show specificity for CAC only, while Max and Pho4p, having both Arg2 and Arg13, bind to both CAC and CAT (Supplementary Figure S3).

We found that bHLH proteins that do not have an arginine at position 13, with the exception of HLH-8, which has a threonine at that position (see below), can specify the CAG half-site either exclusively or in combination with any of the other possible half-sites (Figure 3a). In addition, we inferred that proteins with Val13 specify CAG as their default half-site. These can also specify CAC and CAT if they contain Arg1 and Arg2 (e.g. HLH-2 and HLH-10). Publicly available PBM data (5,36) for another Arg1 + Arg2 + Val13 protein, mouse Tcfe2a, shows specificity for CAG, CAC and CAT (Supplementary Figure S3). Of all 560 bHLH proteins listed in UniProt, only 17 contain the Arg1 + Arg2 + Val13 combination. Alignment of these bHLH proteins shows extremely strong conservation of their basic regions (Supplementary Figure S4), supporting the idea that these bHLHs share similar DNA binding mechanisms.

Two *C. elegans* bHLH proteins have unique DNA binding specificities as well as unique amino acid residues at position 13. The first is HLH-1, the *C. elegans* MyoD ortholog, which can bind a CAACTG E-box. HLH-1 is the only protein in our data set to have both Leu13 and tolerate a CAA half-site. Support for the importance of Leu13 in specifying CAA comes from homologs of HLH-1, such as myogenin and MyoD, that also bind C AACTG (37,38) and that also have a leucine at position 13 (Supplementary Figure S5). The mouse Myf6 bHLH protein also possesses a leucine at position 13 and also binds to CAA, CAC and CAG (Supplementary Figure S3). The second *C. elegans* bHLH protein that has unique DNA binding specificity as well as a unique amino acid residue at position 13 is HLH-8, the *C. elegans* twist ortholog, which has a threonine at position 13 and specifies CAT (Table 1). Homologs of twist in other species also bind CATATG (39,40) and also have Thr13 (Supplementary Figure S5). Thus, comparisons between the amino acid identities at DNA-contacting positions and the DNA binding specificities of *C. elegans* bHLH proteins and their homologs in other species provide support for our inference that Leu13 and Thr13 may specify CAA and CAT half-sites, respectively.

In silico analysis of the structural determinants of bHLH–DNA interactions

In order to comprehensively evaluate which protein residues are energetically compatible with E-box binding, we performed structural modeling of DNA-bound bHLH proteins using FoldX. We mutated, *in silico,* all eight base-contacting residues identified by our frequency map (positions 1, 2, 5, 6, 8, 9, 12 and 13) to each of the 20 amino acids in the context of the seven available bHLH–DNA co-crystal structures. To evaluate the impact of mutations in the base-contacting residues, we used position 3, which makes no DNA contact (data not shown), as a negative control. We also compared mutation of the base-contacting residues against mutation analysis of position 10, which makes only phosphate backbone contacts (Figure 3d and Supplementary Table S2) and thus may not be a major specificity determinant.

As expected, a proline is not tolerated in any of the eight base-contacting positions, likely because of the destabilizing effect that proline has on α-helices (41). Most of the *in silico* mutations predicted to result in a loss of DNA binding introduce amino acids rarely observed at that position in naturally occurring bHLH proteins (Supplementary Table S3). However, two detrimental mutations—Pro6 and Ala9—are observed in >5% of

(a)

bHLH	bHLH domain position																	Half-site specificities
	1	2	4	5	6	8	9	10	12	13	14	17	47	48	49	50	51	
HLH-1	R	R	A	A	T	R	E	R	R	L	R	N	R	L	P	K	V	CAG,CAC,CAA
HLH-2	R	R	Q	N	N	R	E	R	R	V	R	N	N	Q	T	K	L	CAC,CAG,CAT
HLH-10	R	R	E	A	N	R	E	R	R	V	Q	S	K	L	S	K	L	CAC,CAG,CAT
HLH-14	N	Q	A	R	N	R	E	R	R	V	H	N	K	W	S	K	A	CAG,CAT
HLH-19	S	R	R	A	N	R	E	R	R	Q	K	G	K	P	S	K	A	CAG,CAT
LIN-32	R	R	A	A	N	R	E	R	R	M	N	N	K	L	S	K	F	CAG,CAT
CND-1	R	R	K	A	N	R	E	R	R	M	H	N	K	L	S	K	I	CAG,CAT
HLH-11	R	R	I	A	N	N	E	R	R	M	Q	N	K	L	S	K	A	CAG,CAT
HLH-3	T	K	K	R	N	R	E	R	R	V	D	N	K	L	S	K	V	CAG
HLH-4	V	V	K	R	N	R	E	R	R	V	H	N	R	V	S	K	L	CAG
HLH-15	Y	R	L	H	A	R	E	R	R	V	E	N	K	L	S	K	I	CAG
HLH-8	Q	R	C	A	N	R	E	R	R	T	K	N	K	M	S	K	I	CAT
HLH-30	K	K	I	H	N	I	E	R	R	R	Y	N	K	L	N	K	G	CAC
HLH-26A	S	H	H	R	S	T	E	K	R	R	D	N	S	–	E	K	L	CAC
HLH-26B	I	K	I	K	S	R	E	Q	R	R	N	V	E	I	D	K	M	CAC
REF-1A	A	H	R	K	T	Q	E	K	R	R	D	N	N	–	E	K	M	CAC
REF-1B	R	K	V	K	K	R	E	Q	R	R	D	G	Q	V	E	K	M	CAC
MXL-3	R	R	H	H	N	L	E	R	R	R	D	K	D	G	E	K	S	CAC,CAT
MDL-1	S	R	A	H	N	L	E	K	R	R	A	R	D	A	T	R	N	CAC,CAT
MXL-1	A	R	Q	H	N	L	E	R	R	R	D	K	E	R	V	Q	A	CAC,CAT
HLH-25A	N	Q	P	K	S	S	E	R	R	R	N	N	E	–	E	H	I	CAC,CAT
HLH-25B	R	R	V	K	T	R	E	K	R	R	K	D	Q	R	L	K	L	CAC,CAT

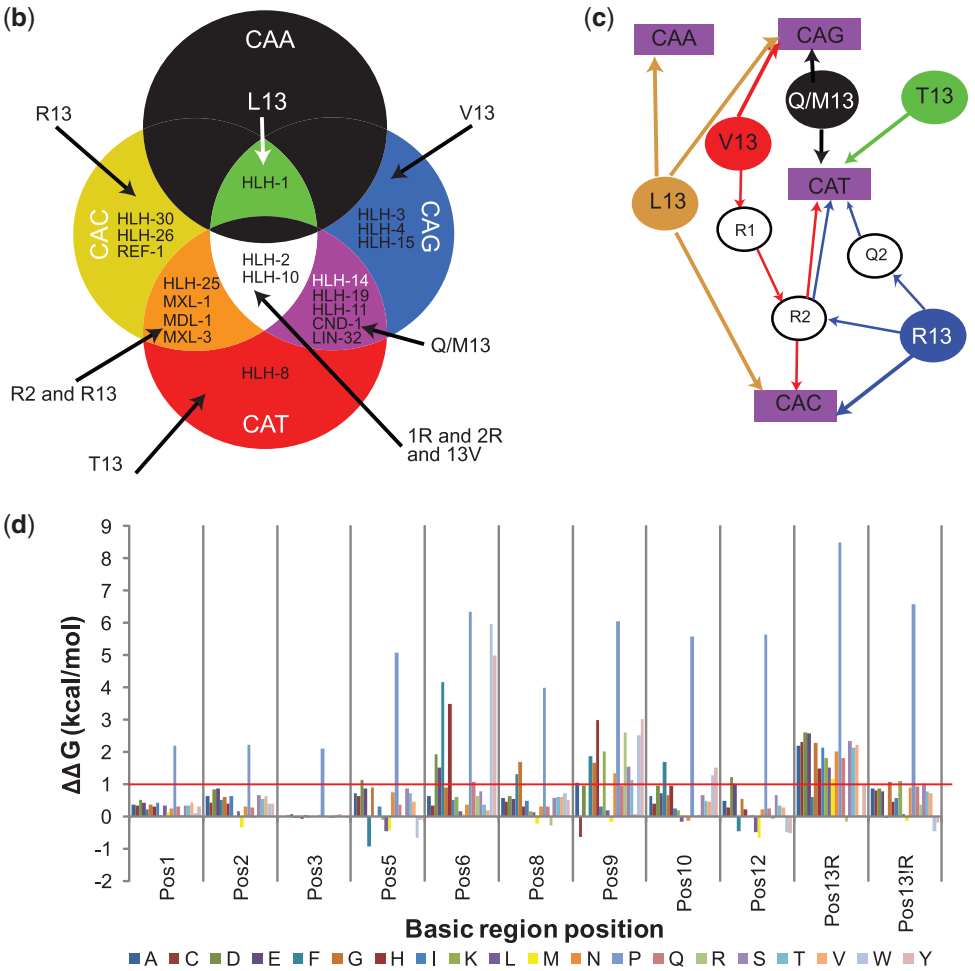

Figure 3. bHLH domain amino acids correlate with E-box half-site recognition (**a**) Amino acid residues at each position in *C. elegans* bHLH proteins that are involved in DNA contacts. The final column shows the list of recognized half-sites for each bHLH (7). 'A' and 'B' (e.g. HLH-26A and HLH-26B) refer to the first and second bHLH domain of two within the same protein. (**b**) Diagram illustrating the distribution of bHLH proteins according to their half-site specification. (**c**) Schematic representation of the rules inferred from (a) and illustrated in (b). Filled circles represent residues at position 13; purple rectangles represent half-sites and open circles represent specific sequence determinants for linking residue 13 to a particular half-site. (**d**) Graphical representation of the average $\Delta\Delta G_{int}$ per mutation for the seven bHLH structures analyzed. Mutations for which $\Delta\Delta G_{int}$ is larger than 1.0 kcal/mol (indicated by the red line) may have considerably diminished, or even abolished, the binding capabilities of the mutant. Position 13 is split between the average $\Delta\Delta G_{int}$ for the scaffold of bHLH proteins that normally do or do not have an arginine in that position.

naturally occurring bHLH TFs (Supplementary Table S3). Our results predict that Pro6 results in a combination of loss of DNA binding and destabilization of the protein. The fact that Pro6 is found in naturally occurring bHLH proteins suggests that the conformation of the α-helix may be different in those proteins than in the ones whose co-crystal structures were available (i.e. structures for Pro6-containing bHLH dimers were not available) for our analysis. Some bHLH proteins that have Pro6 are members of the Hairy class of bHLH proteins (34) and these proteins can bind to the N-Box, which is an E-box-like sequence (CACNAG) (42,43). Other Pro6-containing bHLH proteins are from the HAND1 group of bHLH TFs, have Thr13 and bind NRTCTG (44). Proteins with an Ala9 mainly belong to the bHLH–Per-Arnt-Sim class (bHLH-PAS) and include HIF-1α and SIM. The bHLH–PAS protein ARNT has a glutamate at position 9 and binds the CAC half-site. HIF-1α and SIM proteins dimerize with ARNT and bind to non-canonical half-sites-like (A/G)C and GT (A/G)C, respectively [JASPAR database (45)]. Thus, bHLH–PAS heterodimers bind DNA sequences comprising a conventional E-box half-site and a non-canonical half-site. We postulate that proteins with Ala9 specify the 'non-canonical' half-site in the context of bHLH–PAS dimers. Altogether, these observations suggest that the DNA recognition rules we present here pertain only to E-box binding bHLH proteins.

Our *in silico* mutation analysis indicates that positions 6, 9 and 13 are the least tolerant to mutations (Supplementary Table S2). These residues all directly contact the core NN sequence of the E-box; residues six and nine also have a direct physical interaction with the CA sequence (Figure 2b). Our analysis supports previous experimental observations for mutant Max homodimers (46) and predictions provided by computational studies (33,34).

The energy calculations using the scaffold of bHLH proteins that have Arg13 show that the majority of the mutations in position 13 result in significant decreases in binding energy. Only mutations to the aromatic amino acids phenylalanine, tryptophan and tyrosine do not significantly affect binding energy. The arginine at position 13 makes an H-bond with the guanine opposite from the cytosine in the CAN half-site. Mutations at position 13 in the bHLH protein's basic region disrupt this H-bond, leading to a decrease in binding affinity and specificity. This may explain why phenylalanine, tryptophan and tyrosine do not occur naturally at position 13 in any bHLH protein (Supplementary Table S3), as these residues are not able to form an H-bond to stabilize the TF–DNA interaction and therefore will not discriminate among the four bases since the binding energy will come mainly from desolvation upon burial of these bulky side chains. Interestingly, the scaffold of bHLH proteins that do not have Arg13 tolerate mutations at positions 13–15 other amino acids in the FoldX analysis. This could be because the amino acid at position 13 that is normally present in these bHLH proteins is smaller than an arginine and does not make specific contacts with DNA, allowing for mutations to occur without negatively impacting the overall bHLH–DNA binding properties. However, although these mutations do not affect the overall DNA binding propensity, they may affect DNA binding specificity (see below).

Experimental confirmation of bHLH-E-box recognition

In order to experimentally test our inferred bHLH–E-box recognition determinants, focusing position 13 because it appears to play a crucial role in bHLH–E-box interactions. We created three mutant versions of HLH-1 and determined each of their DNA binding specificities. We selected this bHLH dimer because it is a homodimer (which simplifies interpretations of DNA binding) and it has broad DNA-binding specificity, as it is able to bind to the CAA, CAC and CAG half-sites Figure 3b). We created three mutants at position 13—in particular, we mutated Leu13 to arginine (L13R), threonine (L13T) and valine (L13V)—and measured their DNA binding specificities by universal PBM assays (23). For each of these three mutants, we predicted changes in their half-site preferences based on the determinants we derived above (Figure 4a and b). Briefly, we expected all three mutants to lose specificity for CAA. Furthermore, we predicted that the L13R mutant would lose its specificity for CAG, gain specificity for CAT and strongly favor CAC. We also predicted that the L13T mutant would lose its specificity for both CAG, gain binding to CAT, and also lose specificity for CAC, and that the L13V mutant would gain specificity for CAT while maintaining specificity for both CAC and CAG (Figure 4a and b).

The PBM data obtained for each mutant HLH-1 protein are consistent with our predictions for their altered half-site binding specificities (Figure 4c). As expected, none of the HLH-1 mutants binds a CAA half-site-containing E-box (Figure 4d). The L13T mutant gained specificity for the CAT half-site (as indicated by binding to CATATG); however, unexpectedly, the L13T mutant also retained the ability to bind CAG, but with a lower specificity compared to the wild-type protein, and gained binding capacity for CAC (Figure 4d). These results suggest that HLH-8 (which also has a threonine at position 13) specifies CAT, CAG and/or CAC half-sites when bound to HLH-2 as a heterodimer (while HLH-2 is, in fact, restricted to the CAT half-site). The L13V mutant retained specificity for both CAG and CAC half-sites (as indicated by binding to CAGCTG and CACCTG E-boxes, Figure 4d), and gained weak specificity for CAT (as indicated by weak binding to CATATG and C ATCTG, Figure 4d); all these observations agree with our predictions. The L13R mutant gained specificity for the CAC half-site (as indicated by strong binding to the CAC GTG E-box, Figure 4d), and gained some specificity for CAT (as indicated by binding to CACATG, Figure 4d), while losing specificity for the CAG half-site (as indicated by complete loss of binding to CAGCTG, Figure 4d). Finally, this L13R mutant strongly preferred CACATG and CACGTG over any of the other E-boxes (Figure 4d), which is in agreement with our observation that Arg13-containing bHLH proteins strongly prefer CACGTG and CACATG (7) (see above).

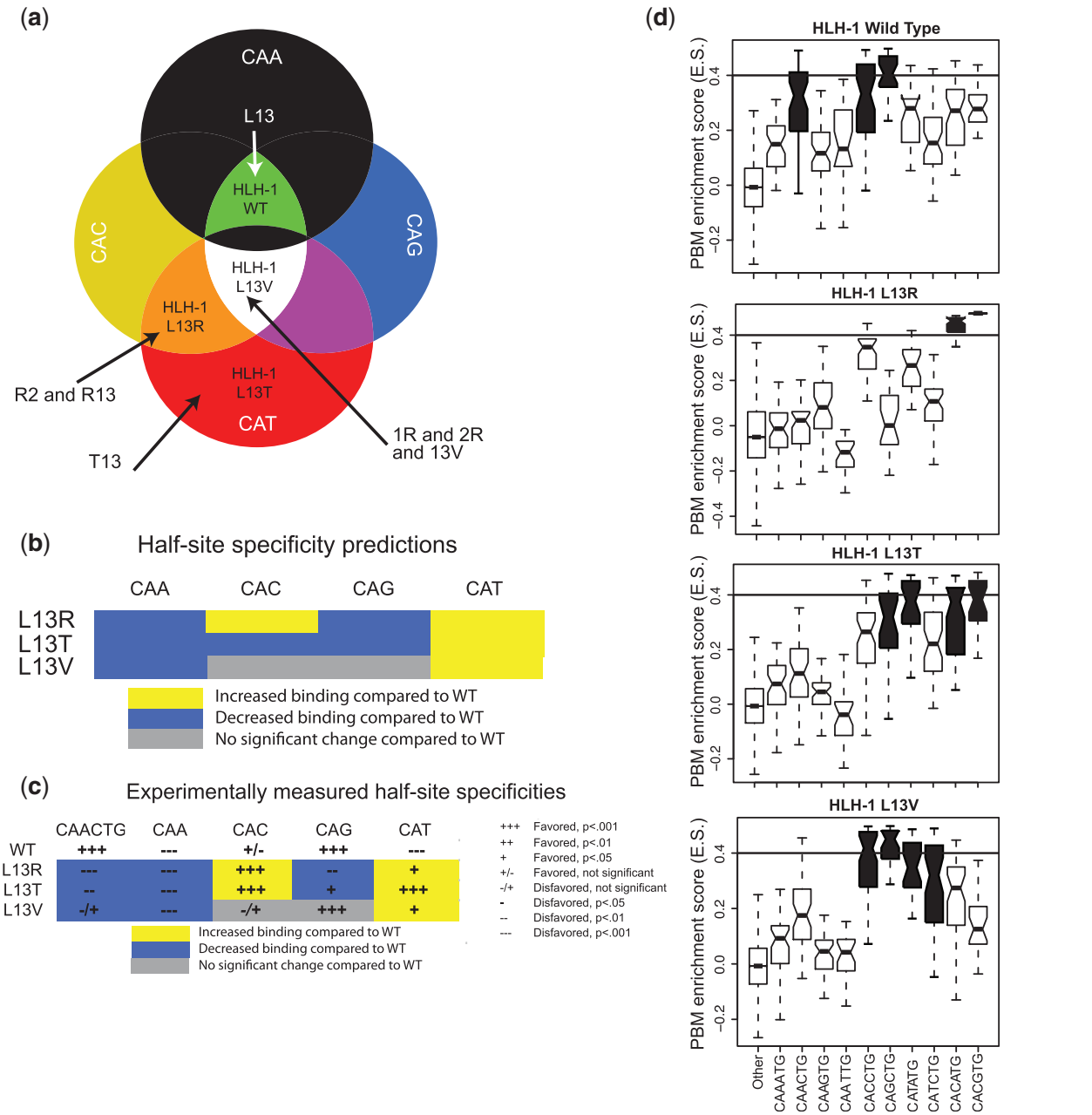

Figure 4. Prediction and experimental validation of HLH-1 Leu13 mutant E-box half-site specificities. (**a**) Classification of the three HLH-1 mutant proteins based on the E-box half-site specificity diagram in Figure 3b. (**b**) Predictions of the half-site specificities for each of the HLH-1 Leu13 mutants. (**c**) Summary of the statistical analysis of the observed half-site specificities determined from PBM data for each of the HLH-1 mutants. (**d**) Boxplot representation of the PBM-derived E-box specificities of HLH-1 wild-type and mutant proteins. Significantly bound E-boxes are colored in black (see 'Materials and Methods' section). Eight-mers that do not contain an E-box are marked as 'other'. For each box, the central horizontal bar shows the median of the distribution, the box's edges mark the 25th and 75th percentile and the whiskers represent the most extreme points of the distribution, which were not determined as being outliers. The horizontal bar of ES value of 0.4 shows our significance ES threshold, as previously determined (7).

Altogether, these experimental results confirm the role of position 13 in the determination of bHLH E-box specificity. Moreover, these results confirm that bHLH proteins containing Arg13 exhibit clear preferences for the CACGTG E-box and CAC half-sites in general. Importantly, we demonstrate that mutating the residue at position 13 in the bHLH basic region does not significantly disrupt overall DNA binding but can alter the DNA-binding specificities of bHLH proteins in a predictable manner.

DISCUSSION

In this study, we used a systems-level protein–DNA interaction network, together with logics, computational modeling and co-crystal structures to investigate the

mechanisms of E-box recognition by bHLH TFs. We first determined the CAN half-site preferences of individual bHLH monomers that bind E-boxes either as homodimers or as heterodimers. We then identified residue 13 in the bHLH basic region as a main selector of half-site specificity. Interestingly, when an arginine is present at position 13 in the bHLH protein, different specificities can be attained when a glutamine is present at position 1 and an arginine at position 2, compared with other residues at these positions. These amino acids, however, do not directly contact the bases within the CANNTG E-box, but rather makes contacts with the DNA backbone (Figure 2) and *in silico* modeling indicates that positions 1 and 2 are not of critical importance for E-box binding (Figure 3). Together, this suggests that residue 13 functionally interacts with other amino acids in the basic region. It is also possible that the residues that make primarily phosphate backbone contacts with the DNA could contribute to DNA binding affinity and/or specificity via indirect readout (47,48).

Our data overall agree that there is no straightforward protein–DNA recognition code in which one amino acid always specifies a particular base. Indeed, the 'recognition rules', we have described in this manuscript are not absolute, but rather have exceptions; for example, HLH-1 has Leu13 yet can tolerate a CAA half-site. Various features of protein–DNA interfaces contribute to complexity in recognition rules. Our analyses treated the DNA-contacting residues essentially independently. Position interdependence among amino acid residues involved in DNA recognition has been observed for EGR zinc-finger proteins (49) and likely also plays a role in other structural classes of TFs. In addition, structural studies of C2H2 zinc-finger proteins have shown that even modest rearrangements of protein side chains' docking geometries upon binding DNA can make it difficult to predict DNA binding specificities with high accuracy consistently (50,51). Future studies involving determination of DNA binding specificities for proteins with greater coverage of combinations of amino acid variants at DNA-contacting positions may permit higher order statistical models of DNA binding specificity determinants to be learned that capture the context dependence of specificity determinants, both for bHLH proteins and for TFs of other structural classes.

The determinants we identified here are not the sole factors that dictate target gene selection by bHLH dimers. Other factors that are important include the specific spatiotemporal co-expression of bHLH dimerization partners, interactions with TFs or cofactors and, importantly, accessibility to different E-boxes that can depend on the chromatin state. Further, DNA binding by TFs occurs along a spectrum of affinities and, therefore, the concentration of dimerization partners, as well as their affinity and specificity for different recognition sequences will ultimately dictate which genes they regulate under different developmental or physiological conditions.

Our study demonstrates that comprehensive DNA binding specificities that were determined on a single experimental platform can be integrated with available co-crystal structures to gain insight into the molecular mechanisms of protein–DNA interactions. This approach will likely be powerful for similar analyses of other important TF families such as nuclear hormone receptors and bZip proteins and in a variety of model organisms and humans. In the future, the derivation of predictive recognition rules may further facilitate the design of synthetic TFs with precisely engineered DNA binding specificities.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank members of the A.J.M.W. and M.L.B. laboratories and Job Dekker for advice and critical reading of the manuscript.

FUNDING

The National Institutes of Health (DK068429 to A.J.M.W. and HG003985 to M.L.B.); European Union's PROSPECTS (HEALTH-F4-2008-201648 to L.S.). Funding for open access charge: The National Institutes of Health (DK068429).

Conflict of interest statement. None declared.

REFERENCES

1. Walhout,A.J.M. (2006) Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping. *Genome Res.*, **16**, 1445–1454.
2. Grove,C.A. and Walhout,A.J.M. (2008) Transcription factor functionality and transcription regulatory networks. *Mol. Biosyst.*, **4**, 309–314.
3. Reece-Hoyes,J.S., Deplancke,B., Shingles,J., Grove,C.A., Hope,I.A. and Walhout,A.J.M. (2005) A compendium of *C. elegans* regulatory transcription factors: a resource for mapping transcription regulatory networks. *Genome Biol.*, **6**, R110.
4. Kummerfeld,S.K. and Teichmann,S.A. (2006) DBD: a transcription factor prediction database. *Nucleic Acids Res.*, **34**, D74–D81.
5. 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.
6. 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.
7. Grove,C.A., deMasi,F., Barrasa,M.I., Newburger,D., Alkema,M.J., Bulyk,M.L. and Walhout,A.J.M. (2009) A multiparameter network reveals extensive divergence between *C. elegans* bHLH transcription factors. *Cell*, **138**, 314–327.
8. Siggers,T.W. and Honig,B. (2007) Structure-based prediction of C2H2 zinc-finger binding specificity: sensitivity to docking geometry. *Nucleic Acids Res.*, **35**, 1085–1097.
9. Paillard,G., Deremble,C. and Lavery,R. (2004) Looking into DNA recognition: zinc finger binding specificity. *Nucleic Acids Res.*, **32**, 6673–6682.
10. 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.

11. 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.
12. Suzuki,M. (1994) A framework for the DNA-protein recognition code of the probe helix in transcription factors: chemical and stereochemical rules. *Structure*, **2**, 317–326.
13. Kauffman,C. and Karypic,G. (2008) An analysis of information content present in protein-DNA interactions. *Pac. Symp. Biocomput.*, 477–488.
14. Rohs,R., West,S.M., Liu,P. and Honig,B. (2009) Nuance in the double-helix and its role in protein-DNA recognition. *Curr. Opin. Struct. Biol.*, **19**, 171–177.
15. Contreras-Moreira,B., Sancho,J. and Angarica,V.E. (2010) Comparison of DNA binding across protein superfamilies. *Proteins*, **78**, 52–62.
16. Gao,M. and Skolnick,J. (2009) From nonspecific DNA-protein encounter complexes to the prediction of DNA-protein interactions. *PLoS Comput. Biol.*, **5**, e1000341.
17. 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.
18. Massari,M.E. and Murre,C. (2000) Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. *Mol. Cell. Biol.*, **20**, 429–440.
19. Ferre-D'Amare,A.R., Prendergast,G.C., Ziff,E.B. and Burley,S.K. (1993) Recognition by Max of its cognate DNA through a dimeric b/HLH/Z domain. *Nature*, **363**, 38–45.
20. Ellenberger,T., Fass,D., Arnaud,M. and Harrison,S.C. (1994) Crystal structure of transcription factor E47: E-box recognition by a basic region helix-loop-helix dimer. *Gene. Dev.*, **8**, 970–980.
21. Longo,A., Guanga,G.P. and Rose,R.B. (2008) Crystal structure of E47-NeuroD1/beta2 bHLH domain-DNA complex: heterodimer selectivity and DNA recognition. *Biochemistry*, **47**, 218–229.
22. Walhout,A.J.M., Sordella,R., Lu,X., Hartley,J.L., Temple,G.F., Brasch,M.A., Thierry-Mieg,N. and Vidal,M. (2000) Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science*, **287**, 116–122.
23. Berger,M.F., Philippakis,A.A., Qureshi,A.M., He,F.S., Estep,P.W. III and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.
24. Ferre-D'Amare,A.R., Pognonec,P., Roeder,R.G. and Burley,S.K. (1994) Structure and function of the b/HLH/Z domain of USF. *EMBO J.*, **13**, 180–189.
25. Brownlie,P., Ceska,T., Lamers,M., Romier,C., Stier,G., Teo,H. and Suck,D. (1997) The crystal structure of an intact human Max-DNA complex: new insights into mechanisms of transcriptional control. *Structure*, **5**, 509–520.
26. Shimizu,T., Toumoto,A., Ihara,K., Shimizu,M., Kyogoku,Y., Ogawa,N., Oshima,Y. and Hakoshima,T. (1997) Crystal structure of PHO4 bHLH domain-DNA complex: flanking base recognition. *EMBO J.*, **16**, 4689–4697.
27. Ma,P.C., Rould,M.A., Weintraub,H. and Pabo,C.O. (1994) Crystal structure of MyoD bHLH domain-DNA complex: perspectives on DNA recognition and implications for transcriptional activation. *Cell*, **77**, 451–459.
28. Nair,S.K. and Burley,S.K. (2003) X-ray structures of Myc-Max and Mad-Max recognizing DNA. Molecular bases of regulation by proto-oncogenic transcription factors. *Cell*, **112**, 193–205.
29. Guerois,R., Nielsen,J.E. and Serrano,L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, **320**, 369–387.
30. Schymkowitz,J., Borg,J., Stricher,F., Nys,R., Rousseau,F. and Serrano,L. (2005) The FoldX web server: an online force field. *Nucleic Acids Res.*, **33**, W382–W388.
31. Berger,M.F. and Bulyk,M.L. (2009) Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat. Protoc.*, **4**, 393–411.
32. Alibes,A., Nadra,A.D., De Masi,F., Bulyk,M.L., Serrano,L. and Stricher,F. (2010) Using protein design algorithms to understand the molecular basis of disease caused by protein-DNA interactions: the Pax6 example. *Nucleic Acids Res.*, **38**, 7422–7431.
33. Atchley,W.R. and Fitch,W.M. (1997) A natural classification of the basic helix-loop-helix class of transcription factors. *Proc. Natl Acad. Sci. USA*, **94**, 5172–5176.
34. Atchley,W.R. and Zhao,J. (2006) Molecular architecture of the DNA-binding region and its relationship to classification of basic helix-loop-helix proteins. *Mol. Biol. Evol.*, **24**, 192–202.
35. Maerkl,S.J. and Quake,S.R. (2007) A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, **315**, 233–237.
36. Newburger,D.E. and Bulyk,M.L. (2009) UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **37**, D77–D82.
37. Wright,W.E., Binder,M. and Funk,W. (1991) Cyclic amplification and selection of targets (CASTing) for the Myogenin consensus binding site. *Mol. Cell. Biol.*, **11**, 4104–4110.
38. Blackwell,T.K. and Weintraub,H. (1990) Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. *Science*, **250**, 1104–1110.
39. Yin,Z., Xu,X.L. and Frasch,M. (1997) Regulation of the twist target gene tinman by modular cis-regulatory elements during early mesoderm development. *Development*, **124**, 4971–4982.
40. Kophengnavong,T., Michnowicz,J.E. and Blackwell,T.K. (2000) Establishment of distinct MyoD, E2A, and Twist DNA-binding specificities by different basic region-DNA conformations. *Mol. Cell. Biol.*, **20**, 261–272.
41. Pace,C.N. and Scholtz,J.M. (1998) A helix propensity scale based on experimental studies of peptides and proteins. *Biophys. J.*, **75**, 422–427.
42. Davis,R.L. and Turner,D.L. (2001) Vertebrate hairy and enhancer of split related proteins: transcriptional repressors regulating cellular differentiation and embryonic patterning. *Oncogene*, **20**, 8342–8357.
43. Fischer,A. and Gessler,M. (2007) Delta–Notch—and then? Protein interactions and proposed modes of repression by Hes and Hey bHLH factors. *Nucleic Acids Res.*, **35**, 4583–4596.
44. Firulli,A.B. (2003) A HANDful of questions: the molecular biology of the heart and neural crest derivatives (HAND)-subclass of basic helix-loop-helix transcription factors. *Gene*, **312**, 27–40.
45. Bryne,J.C., Valen,E., Tang,M.H., Marstrand,T., Winther,O., da Piedade,I., Krogh,A., Lenhard,B. and Sandelin,A. (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, **36**, D102–D106.
46. Maerkl,S.J. and Quake,S.R. (2009) Experimental determination of the evolvability of a transcription factor. *Proc. Natl Acad. Sci. USA*, **106**, 18650–18655.
47. Lesser,D.R., Kurpieski,M.R. and Jen-Jacobson,L. (1990) The energetic basis of specificity in the Eco RI endonuclease—DNA interaction. *Science*, **250**, 776–786.
48. Gillette,W.K., Martin,R.G. and Rosner,J.L. (2000) Probing the *Escherichia coli* transcriptional activator MarA using alanine-scanning mutagenesis: residues important for DNA binding and activation. *J. Mol. Biol.*, **299**, 1245–1255.
49. Liu,J. and Stormo,G.D. (2005) Quantitative analysis of EGR proteins binding to DNA: assessing additivity in both the binding site and the protein. *BMC Bioinformatics*, **6**, 176.
50. Elrod-Erickson,M., Benson,T.E. and Pabo,C.O. (1998) High-resolution structures of variant Zif268-DNA complexes: implications for understanding zinc finger-DNA recognition. *Structure*, **6**, 451–464.
51. Miller,J.C. and Pabo,C.O. (2001) Rearrangement of side-chains in a Zif268 mutant highlights the complexities of zinc finger-DNA recognition. *J. Mol. Biol.*, **313**, 309–315.