

# Supporting Information

Maerkl and Quake 10.1073/pnas.0907688106

## SI Text

### Materials and Methods

**Device Fabrication.** The device was designed in AutoCAD2004 (Autodesk, Inc.) and each layer reproduced as a chrome mask at 20,000 dots per inch (Fineline-Imaging). Flow molds were fabricated on 3" silicon wafers (Silicon Quest International) coated with hexamethyldisilazane (HMDS) in a vapor bath for 2 min. The wafers were then spin-coated with SPR 220-7 (Shipley) initially at 500 rpm for 5 s followed by 4,000 rpm for 60 s, yielding a substrate height of  $\approx 6-7 \mu\text{m}$ . The molds were baked at 105° C for 90 s, followed by a 15-s I-line exposure on a MA6 contact mask aligner (Karl Suss). Next, the molds were developed with 1:5 2401 developer (Microposit) in dH<sub>2</sub>O. Finally, the molds were annealed at 120° C for 20 min. Control molds were fabricated on 3" silicon wafers by spin-coating SU-8 2025 (MicroChem) at 2,700 rpm for 80 s, followed by a 65° C bake for 2 min, 95° C for 5 min, and a final step of 65° C for 2 min. The wafers were then exposed for 10 s on the I-line, followed by a postexposure bake series of 65° C for 2 min, 95° C for 12 min, and 65° C for 2 min. The wafers were then developed in SU-8 developer for 90 s, followed by an acetone and isopropanol wash. One wafer from each control and flow wafer set was selected and used for all subsequent microfluidic device fabrication. The microfluidic devices were fabricated essentially as described previously (1, 2).

**Mutant Synthesis.** Linear-expression templates were generated by using a two-step PCR method. During the first step of the PCR, the N-terminal portion, including the basic region, is swapped out for the mutant version. The necessary 5' and 3' UTRs for efficient *in vitro* transcription/translation are added during the second step of the PCR. MAX isoform B (NCBI accession #:BC003525) obtained from Open Biosystems (Clone ID: 3607261) was used in this study, as it has a shorter 5' terminus due to a spliced-out region just upstream of the basic region. The first-step PCR contained 100 nM of the 5' BR\_mutant primer, 100 nM of the 3' gene specific primer, 200  $\mu\text{M}$  of each dNTP, 0.5 unit of DNA polymerase enzyme (Expand High Fidelity Plus, Roche), and 0.1  $\mu\text{L}$  of a plasmid preparation from the Open-Biosystems cDNA clone purified previously, in a final volume of 50  $\mu\text{L}$ . The reaction was cycled for 7 min at 94° C, followed by 30 cycles of 30 s at 94° C, 60 s at 55° C, and 90 s at 72° C, followed by an elongation step of 7 min at 72° C, and a final 4° C hold step. All PCRs were then checked for efficient synthesis on a 1% agarose gel and served directly as template for the second PCR step. The second PCR step contained 1  $\mu\text{L}$  of the first PCR product, 5 nM of 5' MAXB\_ext1, and 5 nM of 3' ext2 primers, 200  $\mu\text{M}$  of each dNTP, 1 unit of DNA polymerase enzyme (Expand High Fidelity Plus, Roche) in a final volume of 100  $\mu\text{L}$ . The reaction was then cycled for 7 min at 94° C, followed by 10 cycles of 30 s at 94° C, 60 s at 55° C, and 90 s at 72° C, followed by an elongation step of 7 min at 72° C, and a final 4° C hold step. After this round of extension, 2  $\mu\text{L}$  of 5- $\mu\text{M}$  5' finalCy3 and 5- $\mu\text{M}$  3' final in dH<sub>2</sub>O were added to each reaction, and cycling was continued immediately at 94° C for 4 min, followed by 30 cycles of 30 s at 94° C, 60 s at 50° C, and 90 s at 72° C followed by a final extension of 72° C for 7 min. The final products were then purified on PCR spin columns (Qiagen) by using the protocol supplied and eluted in 80  $\mu\text{L}$  of dH<sub>2</sub>O, pH 8.0–8.5.

**Target DNA Synthesis.** All small dsDNA oligos serving as targets for transcription factor binding were synthesized by isothermal primer extension in a reaction containing 6.7- $\mu\text{M}$  5' CompCy5, 10- $\mu\text{M}$  library primer, 1 mM of each dNTP, 5 units Klenow fragment (3'-5' *exo*<sup>-</sup>), 1-mM dithiothreitol 50-mM NaCl, 10-mM MgCl<sub>2</sub> and 10-mM Tris-HCl, pH 7.9 in a final volume of 30  $\mu\text{L}$ . All reactions were incubated at 37° C for 1 h followed by 20 min at 72° C, and a final annealing gradient down to 30° C at a rate of 0.1° C sec<sup>-1</sup>. Fifty  $\mu\text{L}$  of a 0.5% BSA dH<sub>2</sub>O solution was added to each reaction, and the entire volume was transferred to a 384-well plate.

**DNA Arraying and Device Alignment.** All target sequences were spotted with an OmniGrid Micro (GeneMachines) microarrayer by using a CMP3B pin (TeleChem International, Inc.) for delivery onto epoxy-coated glass substrates (CEL Associates). Three rounds of spotting were sequentially performed, the first round consisting only of a priming round spotting 0.5% BSA dH<sub>2</sub>O to prevent binding of the linear-expression templates and DNA targets to the epoxy surface. The linear expression templates were spotted in the second round, followed by the target DNAs in the third round. This resulted in each spot consisting of one type of linear-expression template coding for a specific TF mutant and one target DNA sequence. Device alignment was done by hand on a SMZ1500 (Nikon) stereoscope and bonded overnight in the dark on a heated plate at 40° C.

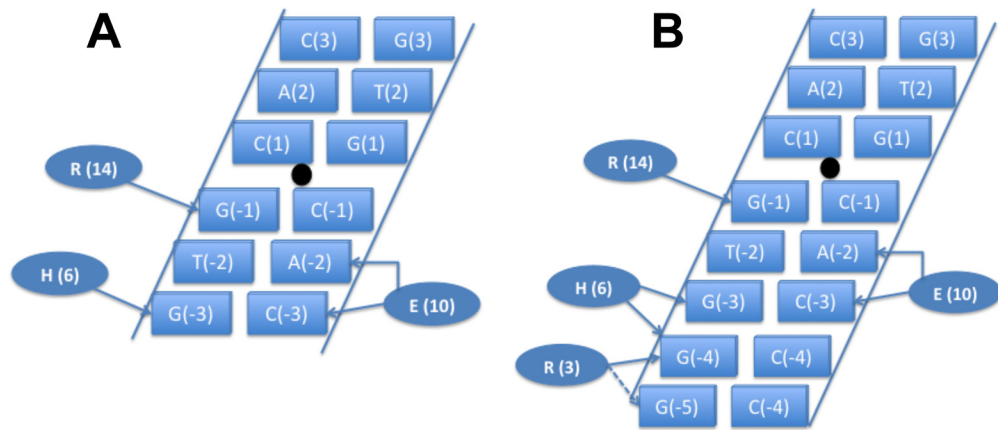
**Surface Chemistry, Protein Synthesis, and MITOMI.** All devices were driven between 12 and 18 psi in the control line and 6 psi for the flow line. For the initial surface-derivatization steps, the chamber valves remained closed to prevent liquid from entering the chambers containing the spotted DNA targets. First, all accessible surface area was derivatized by flowing a solution of biotinylated BSA (Pierce) resuspended to 1 mg/mL in dH<sub>2</sub>O for 20 min through all channels, followed by a 5-min PBS wash. Next, a 500  $\mu\text{g}/\text{mL}$  Neutravidin (Pierce) solution in PBS was flown for 20 min, followed by a 10-min PBS wash. Next, the "button" membrane was closed, and the PBS wash continued for an additional 5 min. Then the remaining accessible surface area was passivated with the same biotinylated solution as described above for 20 min, followed by a 10-min PBS wash. Finally, a 1:5 solution of biotinylated-penta-histidine antibody (Qiagen) in 2% BSA in PBS was loaded for 2–5 min, after which the "button" membrane was opened, and flow continued for 20 min, again followed by a 10-min PBS wash, completing the surface-derivatization procedure. Next, a standard 25- $\mu\text{L}$  TNT T7-coupled wheat germ extract mixture (Promega) was prepared and spiked with 1- $\mu\text{L}$  tRNA<sub>Lys</sub>-bodipy-fl (Promega). The mixture was immediately loaded onto the device and flushed for 5 min, after which the chamber valves were opened, allowing for dead-end loading of the chambers with wheat germ extract. The chamber valves were again closed, and flushing continued for an additional 5 min. Next, the segregation valves separating each unit cell were closed, followed by opening of the chamber valves to allow for equilibration of the unit cell by diffusion. The entire device was heated to 30° C on a temperature-controlled microscope stage and incubated for up to 90 min. After the incubation period, the device was imaged on a modified arrayWoRx (Applied Precision) microarray scanner.

**Data Extraction and Analysis.** For the 1mer single-base substitution data, 32, 16, 16, 12, and 12 measurements over differing DNA-

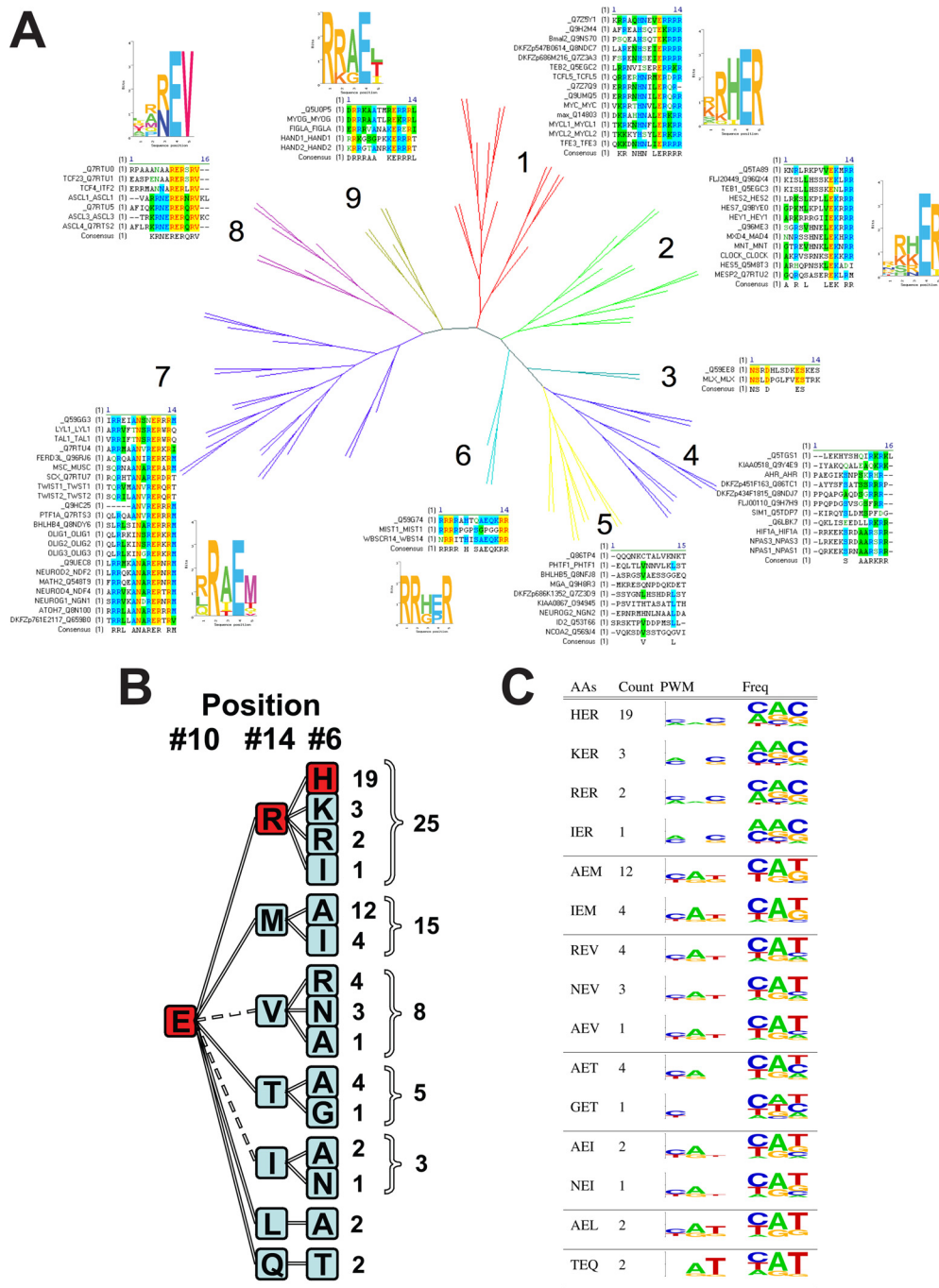
input concentrations were collected for each amino acid in positions 14, 10, 6, 3, and 2, respectively. The relative fluorescent units (RFU) values representing surface-immobilized TF, surface-bound DNA, and solution-phase DNA were extracted by using GenePix software. The surface-bound DNA values were normalized by the immobilized protein values. This ratio was then regressed against the solution-phase DNA RFU values, and a linear regression was fitted to each dataset with the origin set to zero. The slopes of these regressions are plotted in Fig. 2 with error bars indicating the standard error of the regression. The average correlation coefficient for all linear regressions was 0.84 with a standard deviation of 0.15. Care was taken to use low DNA concentrations to ascertain that the binding response was in the linear regime.

Because of the increased number of DNA sequences for the 3mer triple-base substitution experiment, three measurements were taken for each DNA–TF mutant at high DNA concentrations. Only the surface-bound DNA RFU values were extracted and averaged. These values and the standard deviation for each can be found in [Dataset S2](#). These data matrices for each position were then imported into Gene Cluster 3.0 (Michiel de Hoon, University of Tokyo), where the data were normalized by centering “genes” by subtracting the median from each row of data (in our case, genes refer to the 20 aa per position, and arrays are the respective 64 target sequences) and normalizing across “genes”. Complete linkage clustering was then performed on the data by using a centered Pearson correlation as the distance measure. The resulting clusters were displayed by using Java Treeview (3).

- Maerkl SJ, Quake SR (2007) A systems approach to measuring the binding energy landscapes of transcription factors. *Science* 315:233–237.
- Thorsen T, Maerkl SJ, Quake SR (2002) Microfluidic large-scale integration. *Science* 298:580–584.
- Saldanha AJ (2004) Java treeview—extensible visualization of microarray data. *Bioinformatics* 20:3246–3248.
- Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) Weblogo: A sequence logo generator. *Genome Res* 14:1188–1190.
- Hastie T, Tibshirani R, Friedman JH (2001) The elements of statistical learning: Data mining, inference, and prediction: With 200 full-color illustrations (Springer, New York).
- Ferre-D’Amare AR, Prendergast GC, Ziff EB, Burley SK (1993) Recognition by max of its cognate dna through a dimeric b/hlh/z domain. *Nature* 363:38–45.
- Shimizu T, et al. (1997) Crystal structure of pho4 bhlh domain-dna complex: Flanking base recognition. *EMBO J* 16:4689–4697.
- Choithia C (1976) The nature of the accessible and buried surfaces in proteins. *J Mol Biol* 105:1–12.
- Zamyatnin AA (1972) Protein volume in solution. *Prog Biophys Mol Biol* 24:107–123.
- Karplus PA (1997) Hydrophobicity regained. *Protein Sci* 6:1302–1307.
- Biro JC (2006) Amino acid size, charge, hydrophathy indices and matrices for protein structure analysis. *Theor Biol Med Model* 3:15.
- Blackwood EM, Eisenman RN (1991) Max: a helix–loop–helix zipper protein that forms a sequence-specific dna-binding complex with myc. *Science* 251:1211–1217.
- Blackwell TK, Kretzner L, Blackwood EM, Eisenman RN, Weintraub H (1990) Sequence-specific dna binding by the c-myc protein. *Science* 250:1149–1151.
- Takebayashi K, et al. (1994) Structure, chromosomal locus, and promoter analysis of the gene encoding the mouse helix–loop–helix factor hes-1—negative autoregulation through the multiple n-box elements. *J Biol Chem* 269:5150–5156.
- Sasai Y, Kageyama R, Tagawa Y, Shigemoto R, Nakanishi S (1992) Two mammalian helix–loop–helix factors structurally related to drosophila hairy and enhancer of split. *Genes Dev* 6:2620–2634.
- Ishibashi M, Sasai Y, Nakanishi S, Kageyama R (1993) Molecular characterization of hes-2, a mammalian helix–loop–helix factor structurally related to drosophila-hairy and enhancer of split. *Eur J Biochem* 215:645–652.
- Cairo S, Merla G, Urbinati F, Ballabio A, Reymond A (2001) Wbscr14, a gene mapping to the williams-beuren syndrome deleted region, is a new member of the mlx transcription factor network. *Hum Mol Genet* 10:617–627.
- Ma L, Sham YY, Walters KJ, Towle HC (2007) A critical role for the loop region of the basic helix–loop–helix/leucine zipper protein mix in dna binding and glucose-regulated transcription. *Nucleic Acids Res* 35:35–44.
- Stoekman AK, Ma L, Towle HC (2004) Mix is the functional heteromeric partner of the carbohydrate response element-binding protein in glucose regulation of lipogenic enzyme genes. *J Biol Chem* 279:15662–15669.
- Knofler M, et al. (2002) Human hand1 basic helix–loop–helix (bhlh) protein: Extra-embryonic expression pattern, interaction partners and identification of its transcriptional repressor domains. *Biochem J* 361:641–651.
- Hollenberg SM, Sternglanz R, Cheng PF, Weintraub H (1995) Identification of a new family of tissue-specific basic helix–loop–helix proteins with a 2-hybrid system. *Mol Cell Biol* 15:3813–3822.
- Dai YS, Cserjesi P (2002) The basic helix–loop–helix factor, hand2, functions as a transcriptional activator by binding to e-boxes as a heterodimer. *J Biol Chem* 277:12604–12612.
- Kophengnavong T, Michnowicz JE, Blackwell TK (2000) Establishment of distinct myod, e2a and twist dna binding specificities by different basic region-dna conformations. *Mol Cell Biol* 20:261–272.
- Hsu HL, et al. (1994) Preferred sequences for dna recognition by the tal1 helix–loop–helix proteins. *Mol Cell Biol* 14:1256–1265.
- Miyamoto A, Cui XM, Naumovski L, Cleary ML (1996) Helix–loop–helix proteins lyl1 and e2a form heterodimeric complexes with distinctive dna-binding properties in hematolymphoid cells. *Mol Cell Biol* 16:2394–2401.
- Shimizu C, Akazawa C, Nakanishi S, Kageyama R (1995) Math-2, a mammalian helix–loop–helix factor structurally related to the product of drosophila proneural gene atonal, is specifically expressed in the nervous-system. *Eur J Biochem* 229:239–248.
- MacIsaac KD, Fraenkel E (2006) Practical strategies for discovering regulatory dna sequence motifs. *PLoS Comput Biol* 2:e36.
- Marsich E, Vetere A, Piazza MD, Tell G, Paoletti S (2003) The pax6 gene is activated by the basic helix–loop–helix transcription factor neurod/beta2. *Biochem J* 376:707–715.
- Meierhans D, et al. (1995) Binding-specificity of the basic helix–loop–helix protein mash-1. *Biochemistry* 34:11026–11036.
- Kunne AGE, Sieber M, Meierhans D, Allemann RK (1998) Thermodynamics of the dna binding reaction of transcription factor mash-1. *Biochemistry* 37:4217–4223.
- Ellenberger T, Fass D, Arnard M, Harrison SC (1994) Crystal-structure of transcription factor e47 - e-box recognition by a basic region helix–loop–helix dimer. *Genes Dev* 8:970–980.
- Blackwell TK, 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.
- Edmondson DG, Cheng TC, Cserjesi P, Chakraborty T, Olson EN (1992) Analysis of the myogenin promoter reveals an indirect pathway for positive autoregulation mediated by the muscle-specific enhancer factor mef-2. *Mol Cell Biol* 12:3665–3677.
- Ma P, Rould M, Weintraub H, Pabo C (1994) Crystal structure of myod bhlh domain-dna complex: Perspectives on dna recognition and implications for transcriptional activation. *Cell* 77:451–459.



**Fig. S1.** Overview of the known base-specific contacts for MAX (*A*) and Pho4 (*B*), adapted from refs. 6 and 7, respectively. Only base-specific contacts of one half-site are shown. Amino acid and DNA-base numbering have been adjusted to the scheme used in this report.



**Fig. S2.** The diversity and sequence specificity of the naturally occurring nonredundant bHLH basic regions. (A) All basic regions fall into a total of nine clusters, two of which (4, 5) constitute basic regions known to be unable to bind DNA. For the remaining functional basic-region branches, an amino acid logo was generated showing the amino acid diversity and preference in the five positions investigated experimentally. (B) A schematic showing the natural diversity of the three most important positions in the basic region (positions 14, 10, and 6). As only 1 aa is viable in position 10, it served as the origin. Seven different amino acids are possible in position 14, which further diversify into a number of combinations in position 6. The two most dominant combinations are ERH and EMA, with 19 and 12 counts each. The diversity seen in positions 14 and 10 correspond to the experimentally determined values, with R, M, T, L, and Q being functional in vitro. Valine and isoleucine only show baseline affinities. In position 6, H, K, and R were the most functional amino acids. (C) A summary of the predicted DNA sequence specificities for all naturally observed basic regions. DNA sequence specificities are shown both as PWMs and frequency plots. It can be seen that the highest diversity exists in the third base, which can be CA[CT/G]. The second base is predominantly adenine or guanine

# Position 14, Arginine

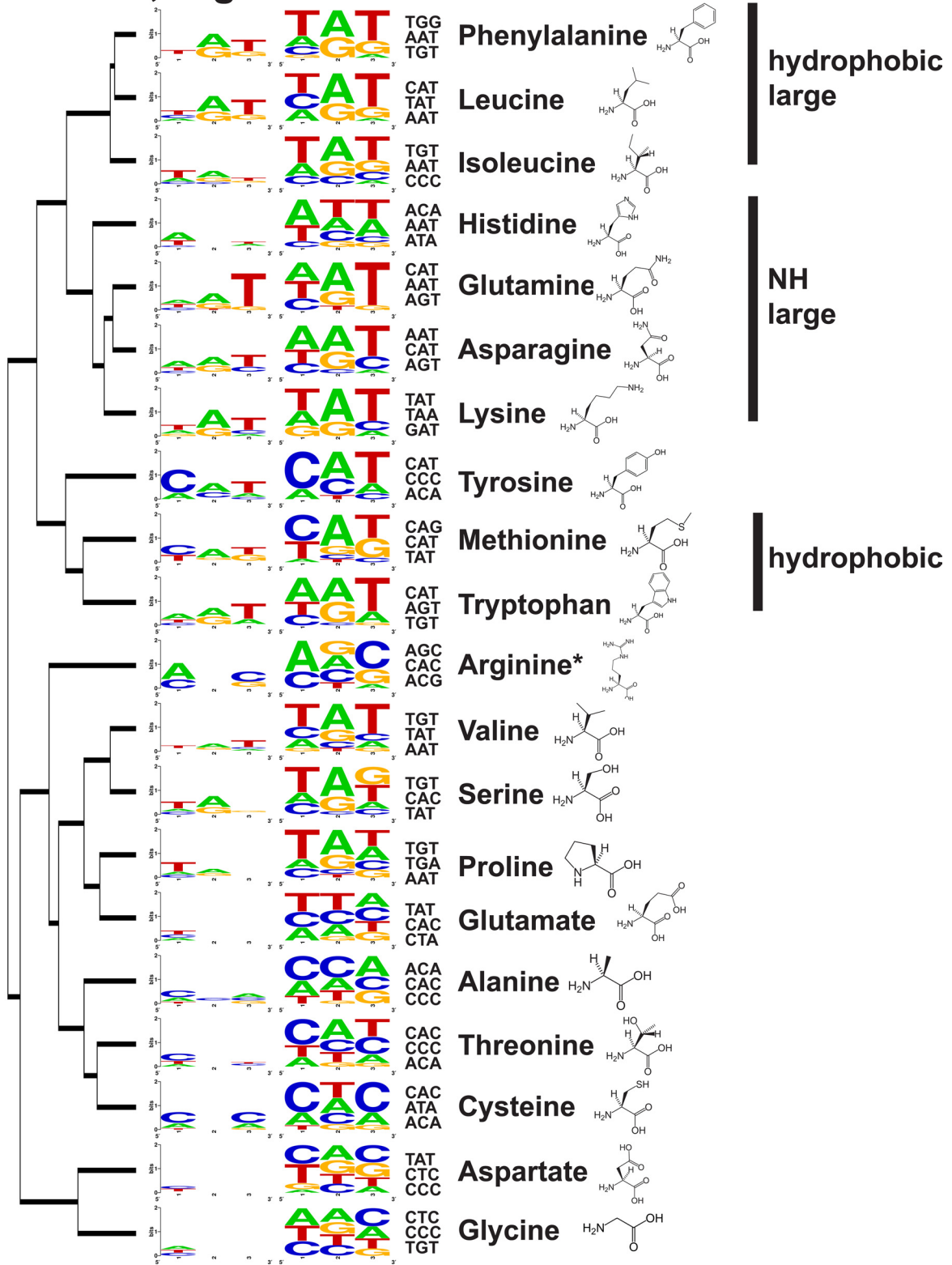


Fig. S3. PWM representation of the data of position 14 shown in Fig. 3. The DNA sequence specificities were replaced with PWMs, but the vertical ordering remained the same as in Fig. 3. The structure for each amino acid is shown, and clusters of similar amino acids are indicated by black columns.

# Position 10, Glutamate

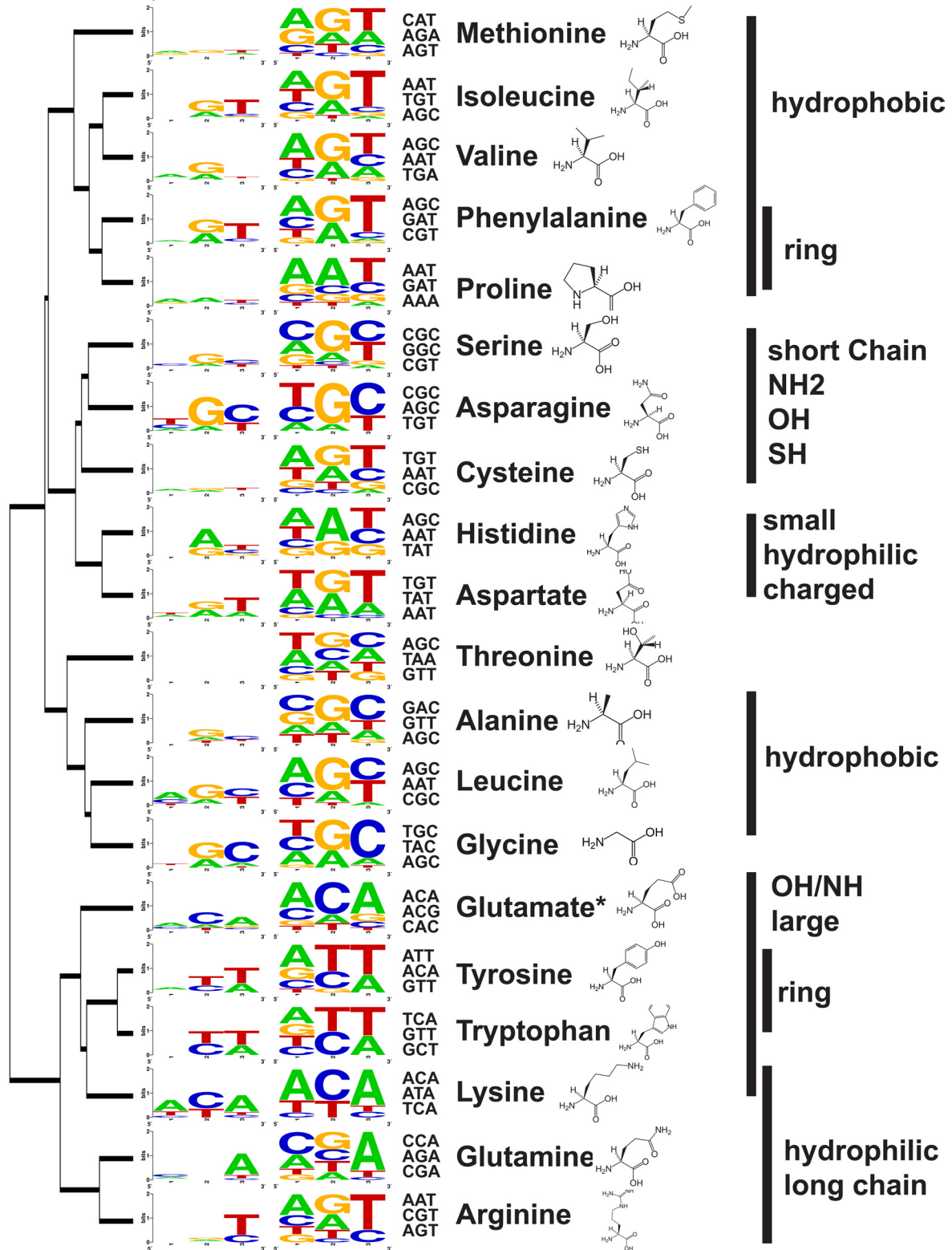


Fig. S4. PWM representation of the data of position 10 shown in Fig. 3. The DNA sequence specificities were replaced with PWMs, but the vertical ordering remained the same as in Fig. 3. The structure for each amino acid is shown, and clusters of similar amino acids are indicated by black columns.

# Position 6, Histidine

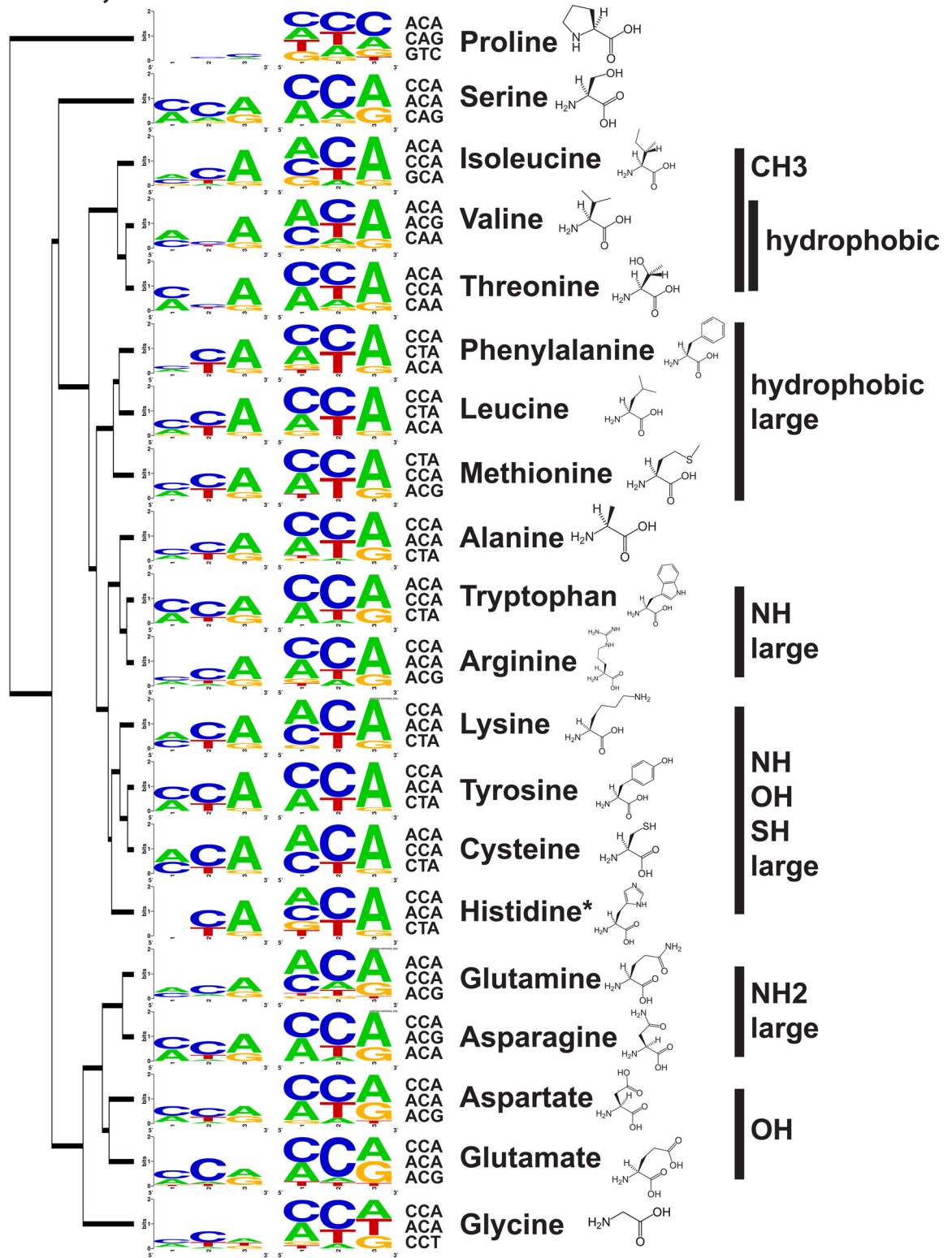


Fig. S5. PWM representation of the data of position 6 shown in Fig. 3. The DNA sequence specificities were replaced with PWMs, but the vertical ordering remained the same as in Fig. 3. The structure for each amino acid is shown, and clusters of similar amino acids are indicated by black columns.

# Position 3, Arginine

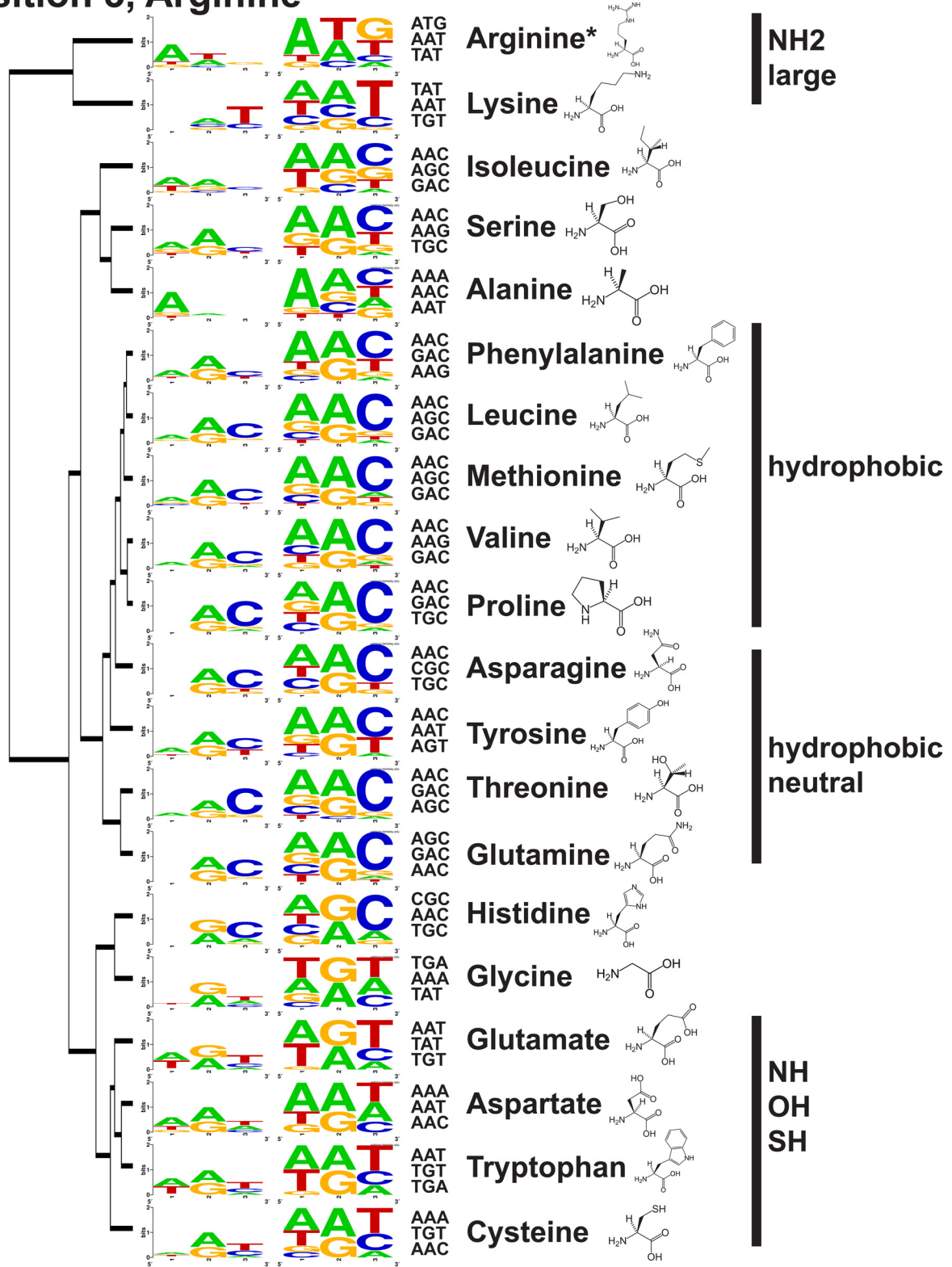


Fig. S6. PWM representation of the data of position 3 shown in Fig. 3. The DNA sequence specificities were replaced with PWMs, but the vertical ordering remained the same as in Fig. 3. The structure for each amino acid is shown, and clusters of similar amino acids are indicated by black columns.



# Position 2, Lysine

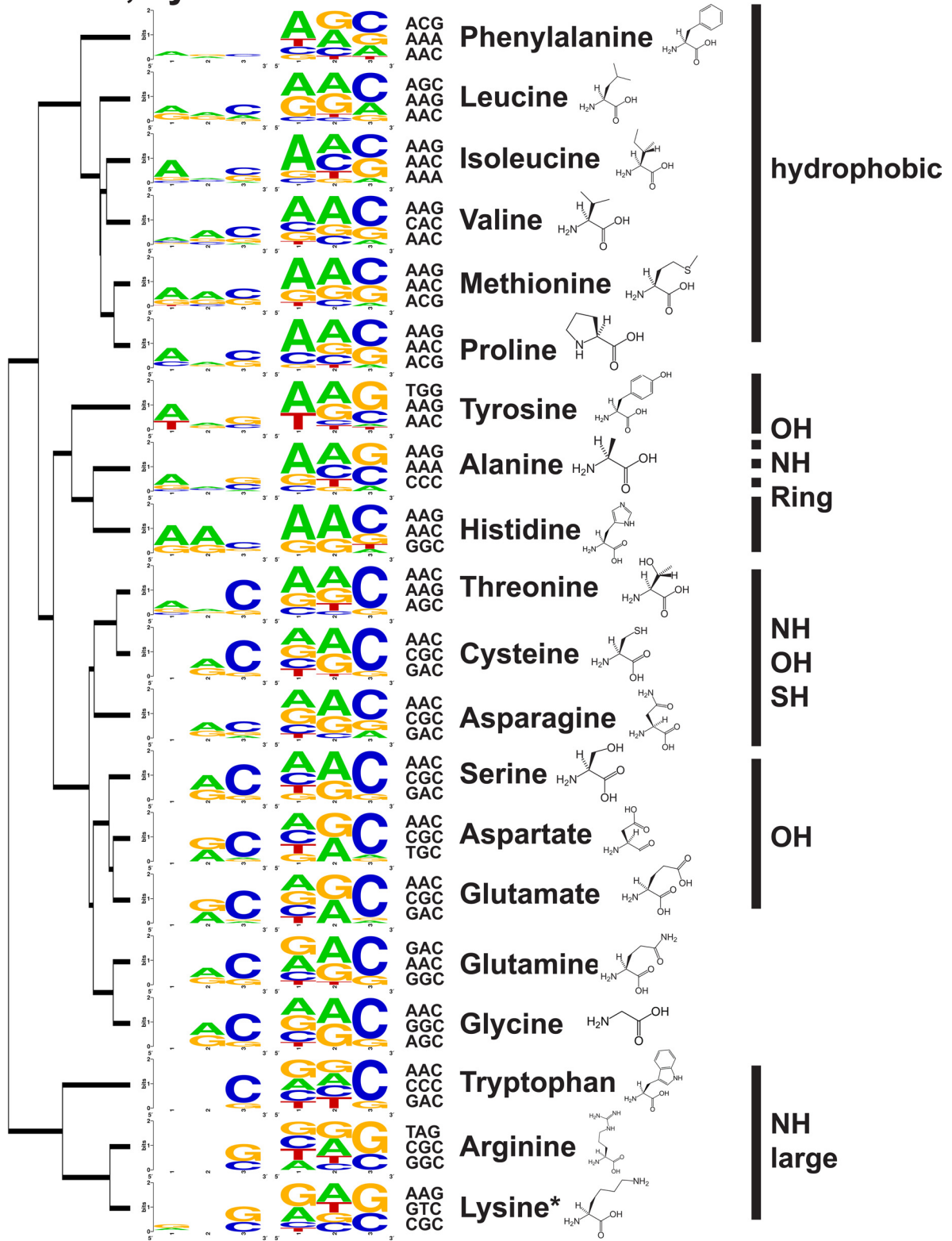


Fig. S7. PWM representation of the data of position 2 shown in Fig. 3. The DNA sequence specificities were replaced with PWMs, but the vertical ordering remained the same as in Fig. 3. The structure for each amino acid is shown, and clusters of similar amino acids are indicated by black columns.



**Table S1. Results of a principal component analysis to determine which amino acid properties best explain the observed differences in sequence specificity across the five positions assayed**

Amino Acid Property	Pos 14	Pos 10	Pos 6	Pos 3	Pos 2
Mass	1.14E-01	<b>3.27E-03</b>	7.93E-02	7.25E-01	1.69E-01
Surface*	1.96E-01	<b>1.08E-03</b>	<b>4.31E-02</b>	6.65E-01	6.77E-02
Volume†	3.95E-01	<b>1.10E-03</b>	<b>2.86E-02</b>	8.52E-01	1.06E-01
Residue nonpolar surface area‡	7.25E-01	<b>4.39E-03</b>	4.11E-01	7.32E-01	<b>3.62E-02</b>
Estimated hydrophobic effect for residual burial (kcal/mol)‡	3.58E-01	<b>3.11E-03</b>	5.10E-01	4.41E-01	5.45E-02
Estimated hydrophobic side-chain burial‡	3.34E-01	<b>2.98E-03</b>	4.82E-01	4.49E-01	5.24E-02
Hydrophobe moment§	<b>1.00E-02</b>	9.10E-02	9.51E-01	<b>3.67E-02</b>	<b>8.43E-03</b>
pI§	<b>1.28E-02</b>	4.95E-02	<b>1.75E-02</b>	<b>3.48E-02</b>	<b>1.04E-03</b>

Shown are *p*-values resulting from a F-test of a multiple linear regression of the first three principal components against the indicated amino acid properties. Significant *p*-values are shown in boldface (*p*-value < 0.05). The isoelectric point is the dominant parameter affecting specificity in all positions of the basic region, except for position 14 where size is dominant. Surface and volume also affect specificity in position 6.














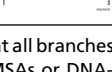
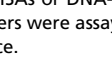


\*Ref. 8.

†Ref. 9.

‡Ref. 10.

§Ref. 11.

**Table S2. Summary of known sequence specificities of bHLH TFs covering all branches of the basic region alignment tree shown in Fig. S2, which are known to bind to DNA**

Clade	Name	AAs	Specificity	Predicted Specificity	Assay	Refs.
1	MAX	RHER	CACGTG		Crystal structure, EMSA	6, 12
1	c-Myc	RHER	CACGTG		SAAB, EMSA	13
2	HES-1	KKER	CACR[A/T]G		DNA footprinting, EMSA	14, 15
2	HES-2	KKER	CACRTG		DNA footprinting, EMSA	16
3	Mlx (ChREBP)	LGEK	CACGTG		EMSA, expression analysis	17–19
6	WBSCR14/Mlx	RHER	CACGYG	See MAX	EMSA	17
7	HAND1 (E47)	RGET	<b>NRTCTG</b>		SAAB, EMSA, expression analysis	20, 21
7	HAND2 (E12)	RAET	<b>CATCTG</b>		EMSA, SAAB	22
7	TWIST (E12)	RAET	CATATG		EMSA	23
7	TAL1 (E47)	RTEQ	<b>CAGATG</b>		SAAB, EMSA	24
7	Ly11 (E47)	RTEQ	<b>CAGATG</b>		SAAB, EMSA	25
7	MATH2 (E47)	RAEM	<b>CAGVTG</b>		EMSA	26
7	NeuroD1	RAEM	<b>CAGCTG</b>		EMSA, Bioinformatics	27, 28
8	MASH-1	RREV	<b>CANNTG</b>		EMSA, expression analysis	29, 30
8	E47	RREV	CACGTG		SAAB, EMSA, crystal structure	31, 32
8	E12 (MyoD)	RNEV	<b>CACCTG</b>		EMSA	23, 32
9	MyoG (E12)	RAEL	<b>CAGTTG</b>		EMSA	33
9	MyoD (E12)	RAEL	<b>CAGCTG</b>		SAAB, EMSA, crystal structure	23, 32, 34

Our specificity predictions are shown as base frequency plots. One can see that all branches except 3 and 9 are correctly predicted. For many of the basic regions shown here, only marginal experimental data exists, such as limited-scale EMSAs or DNA-footprinting studies. Nonetheless, we observe a good agreement between our predictions and observed specificities. In cases where heterodimers were assayed or the experimentally obtained specificity was asymmetric, we indicated the half-site corresponding with our predictions with a bold typeface.

## Other Supporting Information Files

[Dataset S1](#)

[Dataset S2](#)