

# **Supplementary material for “Variation in homeodomain DNA-binding revealed by high-resolution analysis of sequence preferences”**

## **Contents:**

<b>1. Cloning</b>	<b>P. 2</b>
<b>2. Protein production and quantization</b>	<b>P. 3</b>
<b>3. Microarray methods</b>	<b>P. 5</b>
<b>4. Data representation</b>	<b>P. 8</b>
<b>5. PBM motifs and comparison to motifs in Transfac and Jaspar</b>	<b>P. 12</b>
<b>6. Evidence that binding profiles are independent</b>	<b>P. 21</b>
<b>7. Predicting 8-mer profiles and scoring the predictions</b>	<b>P. 49</b>
<b>8. Consistency between homeodomain groups derived from PBM data and homeodomain amino acid sequences</b>	<b>P. 52</b>
<b>9. Agreement between Predicted Z-score vs. measured relative affinity for the <i>Drosophila Engrailed</i> homeodomain</b>	<b>P. 55</b>
<b>10. Comparison between PBM data and ChIP-chip or ChIP-seq data</b>	<b>P. 56</b>
<b>11. References</b>	<b>P. 59</b>

## 1. Cloning

### *Cloning homeodomains into pMAGIC1*

Homeodomain open reading frames, consisting of the pfam-defined homeodomain and 15 amino acids of flanking sequence (or to the end of the full open reading frame) were cloned into pMAGIC1 (1) by either RT-PCR from pooled mouse mRNA (2) followed by ligation independent cloning, or by gene synthesis (DNA 2.0) followed by conventional cloning using BamHI and NotI restriction sites. All clones were sequence verified in pMAGIC1 and are given in the file "Protein and DNA sequence". The inserts were then transferred into a T7-GST-tagged variant of pML280 according to protocols described in (1). The resulting recipient plasmids after transfer express N-terminal GST fusion proteins fused to the DBD flanked by H3 and H4 regions used in the recombination step (**bold**):

MSPILGYWKIKGLVQPTRLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYI  
DGDVKLTQSMAIRYIADKHNLGGCPKERAISMLEGAVLDIRYGVSRIAYSKDFETLK  
VDFLSKLPFMLKMFEDRLCHKTYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKL  
VCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSDLVPRPCEL**KLDVHML**  
**VPRGSLEVLFQGPEGDATMGHMVHRPWIQ** – DBD region -  
**AWPQGGRRTRIVSAHSENLYFQGDLRGSITN** GSGC\*

## 2. Protein production and quantization

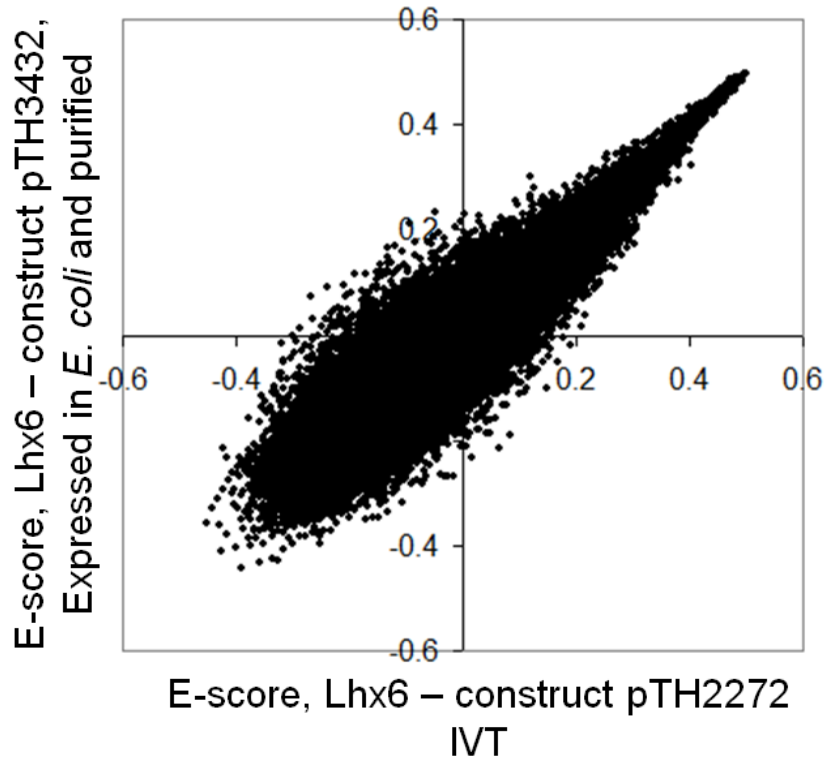
We produced proteins by two methods that yielded essentially identical results: Expression and purification from *E. coli*, and expression by *in vitro* transcription/translation (IVT). (See next page for a graph comparing results from the two systems.)

*Expression and purification in E. coli.* We transformed homeodomain-encoding constructs into *E. coli* C41 DE3 cells (Lucigen). Freshly-transformed cultures were grown overnight in LB medium containing 50 mg/ml ampicillin and diluted 1:100 in fresh LB medium. The cells were grown at 25°C to a final concentration of OD<sub>600</sub> ~0.8 and then induced with IPTG (Bioshop) to a final concentration of 1 mM. These cultures were then grown overnight at 14°C. Cell pellets were obtained by centrifugation at 4°C for 15 minutes at 4000 rpm. Each pellet was resuspended in cold lysis buffer (50mM Tris pH 8, 150 mM NaCl, 2mM DTT, and 12.8 mg of lysozyme). The resuspension was incubated in ice for 20 minutes and lysed by sonication. Cell lysates were centrifuged at 4°C for 15 minutes at 4000 rpm and soluble fraction was collected. GST resin slurry (Amersham) was added to the fraction and binding proceeded at 4°C for 45 minutes. The beads were washed 2-3 times with PBS with 2 mM DTT and then incubated with elution buffer (50 mM Tris pH 7.5, 10 mM reduced glutathione, Roche protease inhibition and 2 mM DTT) at 4°C for 1 hr. Concentration of GST-tagged DBD was estimated for each protein relative to a dilution series of GST standards on Coomassie-stained SDS-PAGE gels.

*In vitro translation.* For *in vitro* translation reactions, the manufacturer's protocol (Ambion ActivePro Kit) was followed. The molarities of all *in vitro* translated proteins were determined by Western blot using a dilution series of recombinant GST (Sigma). Equal volumes of sample and known concentrations of GST were suspended in 1x NuPAGE LDS Sample Buffer (Invitrogen), heated to 95°C for 5 minutes, and loaded on a Bio-Rad 4-12% Bis-Tris Criterion precast gel (Bio-Rad). Samples were electrophoresed at 200 V for 25 minutes and then transferred to a nitrocellulose membrane (Sigma) at 30 V for 3 hours. Membranes were labeled and developed using the SuperSignal West Femto Maximum Sensitivity Substrate kit (Pierce) according to the manufacturer's protocols. Primary antibody (anti-GST produced in rabbit; Sigma) was added at 20 ng/ml, and secondary antibody (horseradish-peroxidase-conjugated anti-rabbit IgG produced in goat; Pierce) was added at 5 ng/ml. Film was scanned and analyzed using Quantity One version 4.5.0 software (Bio-Rad) according to the GST standard curve.

Glycerol was added to a final concentration of 30% to both IVT and purified protein samples prior to storage.

**Supplementary Figure 1.** The plot below shows results from Lhx6, assayed as an IVT protein from one plasmid, and as a protein expressed in *E. coli* and purified, from a different construct, illustrating the reproducibility of the system and robustness to protein production method:



### 3. Microarray methods

#### *Microarray Design*

The construction of ‘all 10-mer’ universal protein binding microarrays (PBMs) using a de Bruijn sequence of order 10 has already been described (3) and is described in more detail in conference proceedings posted at <http://thebrain.bwh.harvard.edu/RECOMB2007.pdf>. For this study, we further optimized our design to achieve greater coverage of gapped  $k$ -mers, as described below. A de Bruijn sequence of order  $k$  is a circular string of length  $4^k$  that contains every  $k$ -mer exactly once when overlaps are considered. To generate a de Bruijn sequence of order 10 for our universal PBM, we used a linear-feedback shift register corresponding to the primitive polynomial:

$$3x^{10} + 3x^9 + 2x^8 + 1x^7 + 2x^6 + 2x^5 + 3x^4 + 3x^3 + 1x^2 + 2x$$

We empirically selected this particular de Bruijn sequence because it uniformly covers all contiguous 10-mers and all gapped 10-mers spanning 11 total positions. Further, it exhibits optimal coverage of contiguous and gapped 8-mers. Any 8-mer is guaranteed to occur 16 times in our de Bruijn sequence (32 times for non-palindromes). Our de Bruijn sequence exhibits this 16/32-fold redundancy for all gapped 8-mers spanning up to 12 total positions (except for sequence variants of the single pattern 1111-1-1--11), as well as all gapped 8-mers of the pattern 1111-gap-1111 with a gap of up to 20 positions. Thus, all  $4^8$  sequence variants for each of these 341 patterns (more than 22.3 million 8-mers) occur at least 16 times each.

After generating this de Bruijn sequence *in silico*, we partitioned it into subsequences of length 36 nucleotides (nt) and overlapping by 11 nt, resulting in 41,944 36-mers. Any 36-mer with a run of five or more consecutive guanines was replaced by its reverse complement to avoid problems in double-stranding (see below). We appended a common 24-nt sequence to each 3' end (5'-gtctgtgtccgtgtcgcgtgctg-3') complementary to our primer for double-stranding (5'-cagcacggacaacggaacacagac-3') in order to create 60-mer sequences that would become the probes on our custom-designed microarray. These microarrays were synthesized by Agilent technologies in their “4x44K” format, with probes attached to the glass slide at the 3' end. Each slide contains the entire complement of all possible 10-mers in four identical subgrids of approximately 44,000 probes each, which can be physically separated into four chambers for four separate experiments. The additional probes beyond the set of 41,944 were designated as control sequences for a variety of purposes. All microarray probe sequences used in this study are listed on our website, [http://the\\_brain.bwh.harvard.edu/pbms/webworks2/](http://the_brain.bwh.harvard.edu/pbms/webworks2/).

### ***Protein Binding Microarrays***

Protein binding microarray (PBM) experiments were performed essentially as described previously, except that four proteins were simultaneously assayed in separate sectors of a single microarray (3). First, single-stranded oligonucleotide microarrays were double-stranded by primer extension and scanned on a microarray scanner (GSI Lumonics ScanArray 5000) prior to protein incubation. Primer extension reactions consisted of 1.17  $\mu\text{M}$  HPLC-purified common primer (Integrated DNA Technologies), 40  $\mu\text{M}$  dATP, dCTP, dGTP, and dTTP (GE Healthcare), 1.6  $\mu\text{M}$  Cy3 dUTP (GE Healthcare), 32 Units Thermo Sequenase™ DNA Polymerase (USB), and 90  $\mu\text{l}$  10x reaction buffer (260 mM Tris-HCl, pH 9.5, 65 mM  $\text{MgCl}_2$ ) in a total volume of 900  $\mu\text{l}$ . The reaction mixture, microarray, stainless steel hybridization chamber, and single-chamber gasket cover slip (Agilent) were pre-warmed to 85°C in a stationary hybridization oven and assembled according to the manufacturer's protocols. After a two-hour incubation (85°C for 10 min, 75°C for 10 min, 65°C for 10 min, and 60°C for 90 min), the hybridization chamber was disassembled in a glass staining dish in 500 ml phosphate buffered saline (PBS) / 0.01% Triton X-100 at 37°C. The microarray was transferred to a fresh staining dish, washed for 10 min in PBS / 0.01% Triton X-100 at 37°C with a magnetic stir bar, washed once more for 3 min in PBS at 20°C, and spun dry by centrifugation at 40 g for 1 min.

For protein binding reactions, double-stranded microarrays were first pre-moistened in PBS / 0.01% Triton X-100 for 5 min and blocked with PBS / 2% (wt/vol) nonfat dried milk (Sigma) under LifterSlip cover slips (Erie Scientific) for 1 h. Microarrays were then washed once with PBS / 0.1% (vol/vol) Tween-20 for 5 min and once with PBS / 0.01% Triton X-100 for 2 min. Proteins were diluted to 100 nM (unless otherwise specified) in a 175- $\mu\text{l}$  protein binding reaction containing PBS / 2% (wt/vol) milk / 51.3 ng/ $\mu\text{l}$  salmon testes DNA (Sigma) / 0.2  $\mu\text{g}/\mu\text{l}$  bovine serum albumin (New England Biolabs). Preincubated protein binding mixtures were applied to individual chambers of a four-chamber gasket cover slip in a steel hybridization chamber (Agilent), and the assembled microarrays were incubated for 1 h at 20°C. Microarrays were again washed once with PBS / 0.5% (vol/vol) Tween-20 for 3 min, and then once with PBS / 0.01% Triton X-100 for 2 min. Alexa488-conjugated rabbit polyclonal antibody to GST (Invitrogen) was diluted to 50  $\mu\text{g}/\text{ml}$  in PBS / 2% milk and applied to a single-chamber gasket cover slip (Agilent), and the assembled microarrays were again incubated for 1 h at 20°C. Finally, microarrays were washed twice with PBS / 0.05% (vol/vol) Tween-20 for 3 min each, and once in PBS for 2 min. Slides were spun dry by centrifugation at 40 g for 5 min. After each hour-long incubation step, microarrays and cover slips were disassembled in a staining dish filled with 500 ml of the first wash solution. All washes were performed in Coplin jars at 20°C on an orbital shaker at 125 r.p.m. Immediately following each series of washes, microarrays were rinsed in PBS (slowly removed over approximately 10 seconds) to ensure removal of detergent and uniform drying.

### ***Microarray Stripping***

Protein and antibody were digested from double-stranded microarrays in a 70-ml stripping solution consisting of 10 mM EDTA, 10% SDS, and 290 Units of protease (from *Streptomyces griseus*; Sigma), shaking at 200 r.p.m. in a Coplin jar at 37°C for 16 hours. Microarrays were then washed 3 times for 5 minutes each in PBS / 0.5% (vol/vol) Tween-20, once for 5 minutes in PBS, and finally rinsed in PBS in a 500-ml staining dish (slowly removed over approximately 10 seconds) to ensure removal of detergent and uniform drying. All washes were performed in Coplin jars at 20°C on an orbital shaker at 125 r.p.m. Before re-use, slides were scanned once at the highest laser power for Alexa488 (488 nm excitation (ex), 522 nm emission (em)) to ensure that no protein or antibody signal remained, and once for Cy3 (543 nm ex, 570 nm em) to ensure that there was no appreciable loss in DNA quantity. For this study, all PBM experiments were performed either on a fresh slide or a slide that had been stripped exactly once, which yielded indistinguishable results (data not shown).

### ***Image Quantification and Data Normalization***

Protein-bound microarrays were scanned to detect Alexa488-conjugated antibody (488 nm ex, 522 nm em) using at least three different laser power settings to best capture a broad range of signal intensities and ensure signal intensities below saturation for all spots. Separately, slides were scanned after primer extension to quantify the amount of incorporated Cy3-conjugated dUTP (543 nm ex, 570 nm em). Microarray TIF images were analyzed using GenePix Pro version 6.0 software (Molecular Devices), bad spots were manually flagged and removed, and data from multiple Alexa488 scans of the same slide were combined using masliner (MicroArray LINEar Regression) software (4).

Our two-step method of raw data normalization was described previously (3). First, we normalize Alexa488 signal by the Cy3 signal for each spot to account for differences in the total amount of double-stranded DNA. Because Cy3-dUTP incorporation is influenced both by the total number of adenines and the sequence context of each adenine, we perform a linear regression over all 41,944 variable spots to compute the relative contributions to the total signal of all trinucleotide combinations (followed by adenine). Using these regression coefficients, we calculate the ratio of observed-to-expected Cy3 intensity and use that as a normalization factor. Second, to correct for any possible non-uniformities in protein binding, we further adjust the Cy3-normalized Alexa488 signals according to their positions on the microarray. We calculate the median normalized intensity of the 15 x 15 block centered on each spot and divide the spot's signal by the ratio of the median within the block to the median over the entire chamber. Raw and normalized forms of the data for all experiments in this study are provided on our supplementary website, [http://the\\_brain.bwh.harvard.edu/pbms/webworks2/](http://the_brain.bwh.harvard.edu/pbms/webworks2/).

## 4. Data representation

### *Sequence Analysis and Motif Construction*

Due to the 32-fold redundancy of 8-mers described above (i.e., every non-palindromic 8-mer occurs on at least 32 spots in each chamber of our universal PBM), we are able to provide a robust estimate of the relative preference of a transcription factor for every contiguous and gapped 8-mer that is covered on our array. Here, we provide several scores for each 8-mer: (1) Median Intensity, (2) Z-Score, (3) Enrichment Score (E-Score), and (4) False Discovery Rate Q-Value for the E-Score.

Median intensity refers to the median normalized signal intensity for the set of ~32 probes harboring a match to each 8-mer. We have previously shown that PBM median signal intensity is proportional to binding affinity (3). The distribution of log (median intensity) over all 8-mers is used to compute a Z-Score for each 8-mer according to the following formula:

$$Z\text{-Score} = \frac{\log \text{median intensity of kmer} - \log \text{median intensity of all kmers}}{\text{robust estimate of standard deviation}}$$

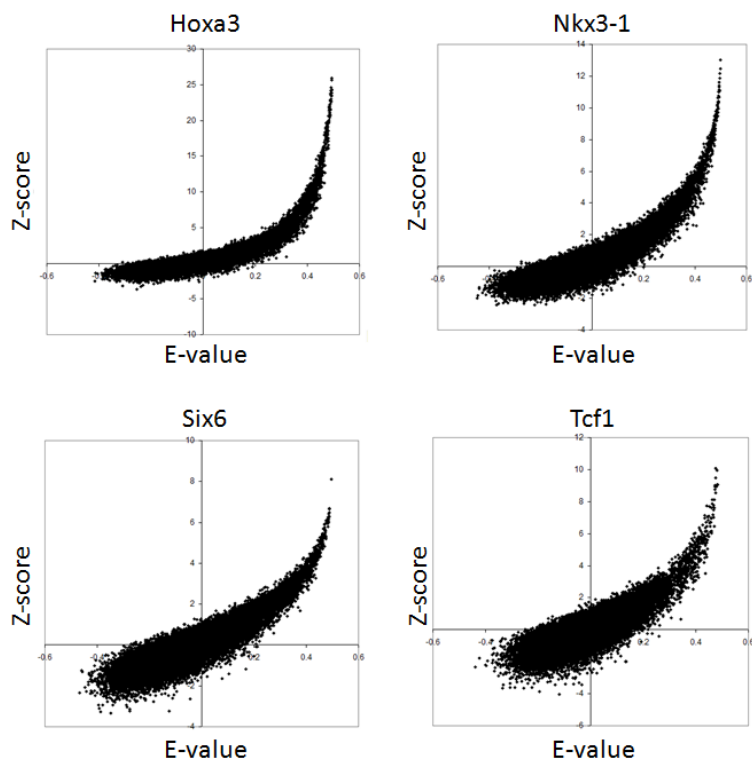
Here, our robust estimate of the standard deviation is the median absolute deviation (MAD), multiplied by 1.4826 for normally distributed data (5). (We have observed, though, that the distribution of the log median intensity is not a normal distribution. Therefore, the inferred Z-Score should not be mistaken for the usual Z-Score annotation in the literature.) Both the median intensity and Z-Score are advantageous because they retain information regarding relative differences in signal intensity, and thus probe occupancy and relative affinity as well. However, experimental variability and differences in absolute signal intensity and non-specific binding can make these measures difficult to compare for different transcription factors.

Our E-Score is a rank-based, non-parametric statistical measure that is invariant to protein concentration and readily allows different experiments to be compared on the same scale. This score has already been described in detail (3). Briefly, for each 8-mer (contiguous or gapped) we consider the collection of all probes harboring a match as the “foreground” feature set and the remaining probes as a “background” feature set. We compare the ranks of the top half of the foreground with the ranks of the top half of the background by computing a modified form of the Wilcoxon-Mann-Whitney (WMW) statistic scaled to be invariant of foreground and background sample sizes. The E-Score ranges from +0.5 (most favored) to -0.5 (most disfavored). Finally, we compute a False Discovery Rate Q-Value for the E-Score by comparing it to the null distribution of E-Scores (over 32,896 8-mers) calculated by randomly shuffling the mapping among the 41,944 probe sequences and intensities (repeated 20 times) (6) We note that in



computing all of the above scores, we do not consider probes for which the 8-mer occupies the most distal position on the probe (5' with respect to the template strand) or for which the 8-mer overlaps the 24-nt primer region.

**Supplementary Figure 2.** Shown below is a scatter plot comparing E and Z for four homeodomains (Hoxa3, Tcf1, Six6, Nkx3-1) illustrating how the two measures relate to one another. In essence, the E-score and Z-score capture essentially the same information, but the E-score representation is flattened as values approach 0.5.



In addition to reporting scores for each individual 8-mer, we compactly represent these data in a position weight matrix (PWM) for each TF using our “Seed-and-Wobble” algorithm, which has been described previously (3). The algorithm works in two stages. In the first “Seed” stage, we identify the single 8-mer (contiguous or gapped) with the greatest enrichment score. For this study, we considered all 8-mers spanning up to 10 total positions as candidate seeds. In the second “Wobble” stage, we systematically test the relative preference of each nucleotide variant at each position, both within and outside the seed. This is accomplished by examining each of the four nucleotides at each position within the 8-mer seed (keeping the other 7 positions fixed) and computing the modified WMW statistic using the entire set of probes containing one of the four variants. For positions outside the 8-mer seed, we first identify the single position within the seed with the lowest information content, treat it as a gapped position, and query every other

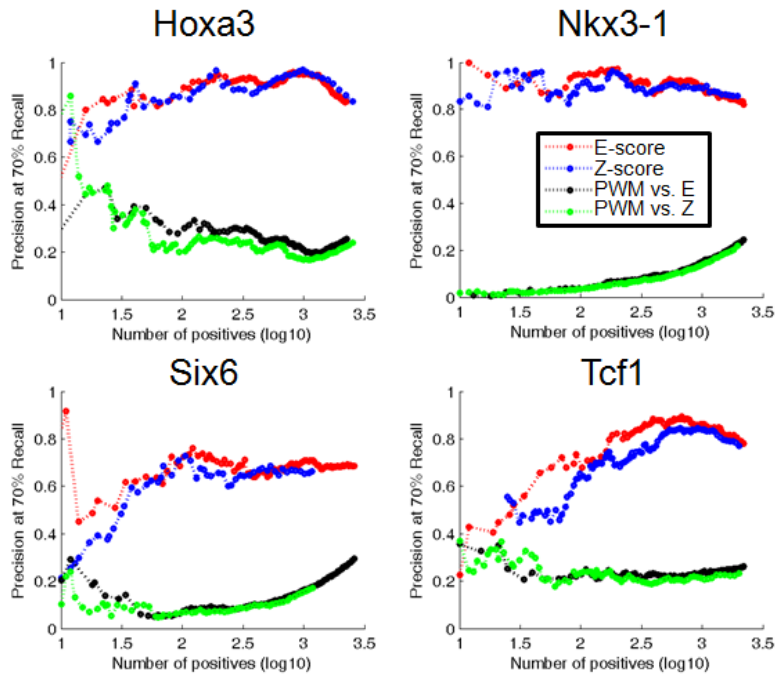
position outside the seed for which the resulting 8-mer is covered in our de Bruijn sequence (i.e., all  $4^8$  sequence variants of that pattern exhibit 32-fold redundancy). Finally, we transform the motif derived from this method into a PWM using Boltzmann distribution. Importantly, this method takes advantage of the fact that all sequence variants occur an equal number of times on the microarray, and it considers all features without applying any arbitrary cutoffs.

In the analysis below, we applied the Seed-and-Wobble algorithm to the construction of “TF-specific” motifs that capture the binding sites that are preferred by one TF relative to another TF or group of TFs. To derive a TF-specific PWM for one factor ( $TF_A$ ) relative to a single other factor ( $TF_B$ ), we first rank all microarray probes according to the ratio of their ranks for the two experiments (i.e.,  $\text{rank } TF_A / \text{rank } TF_B$ ). We then identify the 8-mer with the highest enrichment score in the new probe ranking and use Seed-and-Wobble to generate a TF-specific PWM in the new ranking. To derive a TF-specific PWM for one factor ( $TF_A$ ) relative to a group of other factors ( $TF_{\text{group}}$ ), our approach is similar except that we rank probes based on the average rank for the entire group (i.e.,  $\text{rank } TF_A / \text{rank } (\text{avg rank } TF_{\text{group}})$ ). In both cases, we limit to only those probes for which we have data in all experiments under consideration.

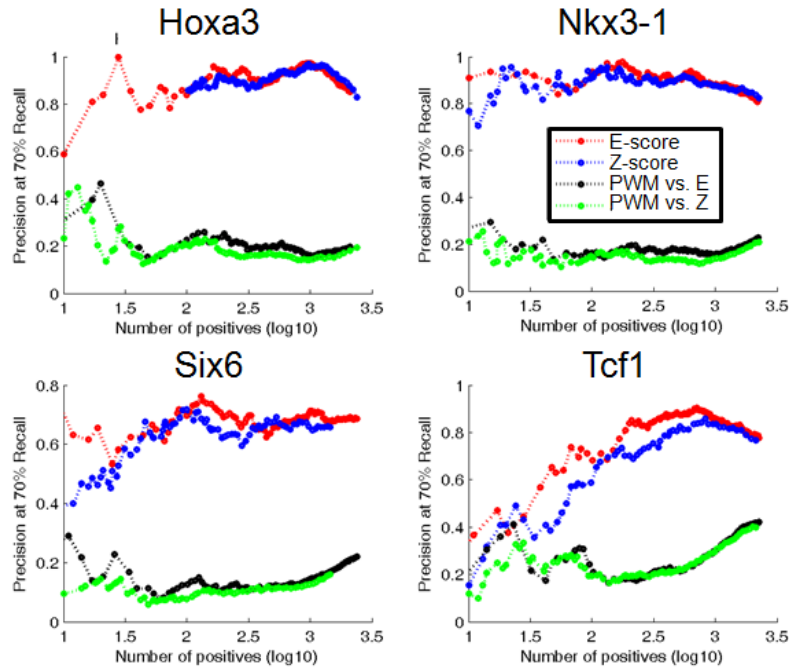
In order to establish which representation should be used as a standard, we replicated four experiments on a second array with completely different probe sequences (but still containing all possible 10-mers). We found that 8-mer E- and Z-scores determined from one array were substantially better predictors of the 8-mer E- and Z-scores on the other array, in comparison to the relative preference of each 8-mer predicted from the dominant motif obtained from the data on the other array.

Moreover, E-scores were typically a slightly better predictor than Z-scores. The following graphs show Precision at 70% Recall. Positives were defined by a moving threshold on E-scores (or Z-scores, if Z-scores are used as training/test) for the second independent array. At each threshold (i.e. each point on the graph) the precision statistic ( $\text{True Positives} / (\text{True Positives} + \text{False Positives})$ ) was determined for the value at which the recall statistic ( $\text{True Positives} / (\text{True Positives} + \text{False Negatives})$ ) is 70%. E-score (red), Z-score (blue), and SW (Seed-and-Wobble) (3) motifs (black and green) were defined on the original array. For the precision-recall analysis, motif scores were calculated for all 8-mers using the Gomer scoring method (7), which is an estimate of the probability of transcription factor binding. These graphs demonstrate the reproducibility in our measurements for separate 8-mers and emphasize the importance of retaining individual scores for all 8-mers rather than compressing these data into a motif which cannot fully recapitulate these preferences.

**Supplementary Figure 3.** Results from training on the version of the array used for all homeodomains, and testing on an alternative array design.







































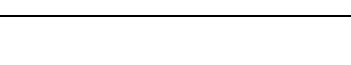







**Supplementary Figure 4.** Results from training on the alternative array design, and testing on the version used for all homeodomains:
































## 5. PBM motifs and comparison to motifs in Transfac and Jaspas

**Supplementary Table 1.** Logos constructed using our Seed-and-Wobble algorithm are shown for all proteins presented in this study. When available, the corresponding position weight matrix (PWM) assembled from the literature is also shown for the mouse protein or its closest ortholog in any other metazoan species. PWMs were retrieved from the JASPAR (8) and TRANSFAC (9) databases, as indicated. These PWMs were mainly derived either from *in vitro* selection (Selex) experiments or from a compilation of validated binding sites from a variety of sources and experimental methods. Of the 168 mouse proteins examined here, 6 had a PWM in the JASPAR database (10 to 59 sequences), 28 more had a PWM in the TRANSFAC database (5 to 56 sequences), and 23 had between 1 and 5 individual binding sequences listed. For the remaining 111 proteins, we searched for additional binding data corresponding to the closest ortholog of each protein in human, zebrafish, *D. melanogaster*, *C. elegans*, and sea urchin. The ortholog was considered to be the best ungapped BLAST match in the other species over at least 35 AAs. We identified 40 proteins possessing an ortholog with a reported binding matrix or sequence, bringing our total to 97 proteins with binding data. In cases where binding data existed for both the mouse protein and an ortholog, we considered the more comprehensive data set. (h = human; m = mouse; r = rat).

TF Name	PBM Seed-and-Wobble	Best Hit	Database	Database PWM
Alx3				
Alx4		Alx4 (h,m)	TRANSFAC 5 compiled seq.	
Arx				
Bapx1		Bapx1 (m)	JASPAR 24 Selex seq.	
Barhl1				
Barhl2				
Barx1				
Barx2				
Bsx				
Cart1		Cart1 (h)	TRANSFAC 25 Selex seq.	

















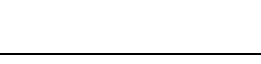













Cdx1		Cdx1 (h,m)	TRANSFAC 13 compiled seq.	
Cdx2		Cdx2 (m)	TRANSFAC 9 compiled seq.	
Cphx		Eve (fly)	TRANSFAC 9 compiled seq.	
Crx		Crx (h,m,r)	TRANSFAC 26 compiled seq.	
Cutl1		Cutl1 (h)	TRANSFAC 86 Selex seq.	
Dbx1				
Dbx2				
Dlx1				
Dlx2				
Dlx3				
Dlx4				
Dlx5				
Dmbx1				
Dobox4				
Dobox5				
Duxl <small>(motif corrected 12-27-09)</small>				
Emx2				
En1		En1 (m)	JASPAR 10 Selex seq.	
En2		En (fly)	TRANSFAC 11 compiled seq.	
Esx1				
Evx1		Eve (fly)	TRANSFAC 9 compiled seq.	
Evx2		Eve (fly)	TRANSFAC 9 compiled seq.	

Gbx1				
Gbx2				
Gsc				
Gsh2				
Hdx				
Hlx1				
Hlxb9				
Hmbox1				
Hmx1				
Hmx2				
Hmx3		Hmx3 (m)	TRANSFAC 11 Selex seq.	
Homez				
Hoxa1				
Hoxa2				
Hoxa3		Hoxa3 (m)	TRANSFAC 14 Selex seq.	
Hoxa4		Hoxa4 (h,m)	TRANSFAC 6 compiled seq.	
Hoxa5		Hoxa5 (m)	TRANSFAC 10 compiled seq.	
Hoxa6		Antp (fly)	TRANSFAC 13 compiled seq.	
Hoxa7		Hoxa7 (h,m)	TRANSFAC 18 Selex seq.	
Hoxa9		Abd-B (fly)	TRANSFAC 45 Selex seq.	
Hoxa10				
Hoxa11				

Hoxa13				
Hoxb3		Dfd (fly)	TRANSFAC 57 selected seq.	
Hoxb4		Dfd (fly)	TRANSFAC 57 selected seq.	
Hoxb5				
Hoxb6		Antp (fly)	TRANSFAC 13 compiled seq.	
Hoxb7		Antp (fly)	TRANSFAC 13 compiled seq.	
Hoxb8		Antp (fly)	TRANSFAC 13 compiled seq.	
Hoxb9		Antp (fly)	TRANSFAC 13 compiled seq.	
Hoxb13		Abd-B (fly)	TRANSFAC 45 Selex seq.	
Hoxc4		Dfd (fly)	TRANSFAC 57 selected seq.	
Hoxc5				
Hoxc6		Antp (fly)	TRANSFAC 13 compiled seq.	
Hoxc8		Antp (fly)	TRANSFAC 13 compiled seq.	
Hoxc9		Antp (fly)	TRANSFAC 13 compiled seq.	
Hoxc10		Abd-A (fly)	TRANSFAC 40 compiled seq.	
Hoxc11		Abd-B (fly)	TRANSFAC 45 Selex seq.	
Hoxc12		Abd-B (fly)	TRANSFAC 45 Selex seq.	
Hoxc13		Cad (fly)	TRANSFAC 13 compiled seq.	
Hoxd1				
Hoxd3				
Hoxd8		Antp (fly)	TRANSFAC 13 compiled seq.	
Hoxd10		Abd-A (fly)	TRANSFAC 40 compiled seq.	





Hoxd11		Abd-B (fly)	TRANSFAC 45 Selex seq.	
Hoxd12		Abd-B (fly)	TRANSFAC 45 Selex seq.	
Hoxd13		Abd-B (fly)	TRANSFAC 45 Selex seq.	
lpx1		Pdx1 (h)	JASPAR 31 Selex seq.	
Irx2				
Irx3				
Irx4				
Irx5				
Irx6				
Isl2				
Isx				
Lbx2				
Lhx1				
Lhx2				
Lhx3		Lhx3 (h)	JASPAR 21 Selex seq.	
Lhx4				
Lhx5				
Lhx6				
Lhx8				
Lhx9				
Lmx1a				
Lmx1b				



Meis1		Meis1 (m)	TRANSFAC 32 Selex seq.	
Meox1				
Mrg1				
Mrg2				
Msx1		Msx1 (m)	TRANSFAC 13 Selex seq.	
Msx2				
Msx3				
Nkx1-1				
Nkx1-2				
Nkx2-2		Nkx2-2 (m)	TRANSFAC 23 Selex seq.	
Nkx2-3				
Nkx2-4				
Nkx2-5		Nkx2-5 (m)	JASPAR 17 Selex seq.	
Nkx2-6				
Nkx2-9				
Nkx3-1		Nkx3-1 (h)	JASPAR 20 Selex seq.	
Nkx6-1		Nkx6-1 (r)	TRANSFAC 17 Selex seq.	
Nkx6-3				
Obox1		Abd-B (fly)	TRANSFAC 45 Selex seq.	
Obox2		Abd-B (fly)	TRANSFAC 45 Selex seq.	
Obox3				
Obox5				

Obox6				
Og2x		Nobox (m)	JASPAR 38 Selex seq.	
Otp				
Otx1		Otx1 (m)	TRANSFAC 10 compiled seq.	
Otx2		Otx1 (m)	TRANSFAC 10 compiled seq.	
Pax4		Pax4 (m)	TRANSFAC 20 Selex seq.	
Pax6		Pax6 (m)	TRANSFAC 47 Selex seq.	
Pax7		Prd (fly)	TRANSFAC 9 compiled seq.	
Pbx1		Pbx1 (h)	JASPAR 18 Selex seq.	
Phox2a				
Phox2b				
Pitx1				
Pitx2		Pitx2 (h,m)	TRANSFAC 9 compiled seq.	
Pitx3				
Pknox1				
Pknox2				
Pou1f1		Pou1f1 (h,m)	TRANSFAC 17 compiled seq.	
Pou2f1		Pou2f1 (h)	TRANSFAC 56 Selex seq.	
Pou2f2				
Pou2f3				
Pou3f1				
Pou3f2		Pou3f2 (h)	TRANSFAC 7 Selex seq.	

Pou3f3				
Pou3f4				
Pou4f3		Unc-86 (worm)	TRANSFAC 5 compiled seq.	
Pou6f1		Pou6f1 (m)	TRANSFAC 16 Selex seq.	
Prop1				
Prrx1				
Prrx2		Prrx2 (m)	TRANSFAC 59 Selex seq.	
Rax				
Rhox11				
Rhox6				
Shox2				
Six1				
Six2				
Six3				
Six4				
Six6				
Tcf1		Tcf1 (h,m)	TRANSFAC 26 compiled seq.	
Tcf2				
Tgif1		Tgif1 (h)	TRANSFAC 15 Selex seq.	
Tgif2				
Titf1		Titf1 (h,m)	TRANSFAC 7 compiled seq.	
Tlx2		Tlx2 (h)	TRANSFAC 40 Selex seq.	

Uncx4.1				
Vax1				
Vax2				
Vsx1				

## 6. Evidence that binding profiles are independent

We evaluated whether any two 8-mer profiles were the same or different using two basic criteria. First, we considered the degree of overlap among the top-scoring 100 8-mers (determined using the E-value), as shown in Figure 2B. Two homeodomains that had an overlap within the distribution of that for duplicated experiments (see below), as well as a similar overlap with other homeodomains, were considered to be inseparable by this criterion. Second, for groups of homeodomains that were not distinct by this criterion, we considered whether differences in binding profiles across the composite set of 8-mers with  $E > 0.45$  for any member of the group could be attributed to (a) differences in the primary motif obtained; (b) differences in the “TF-specific” motif obtained as described above, or (c) an otherwise clear theme among the 8-mers preferred or not preferred by one or more homeodomains. If at least one of these criteria supported the dendrogram obtained by clustering analysis of the 8-mer profiles, we considered it to be evidence for separable binding activities. The decisions made are outlined in the following series of figures, for each group of 8-mers. The individual class assignments are given in the Supplementary document “Homeodomain subclass assignments” in the Supplementary data. Although the assignment process here is *ad hoc*, the process is validated by the fact that the resulting groups of indistinguishable binding activities closely follow both the overall sequence similarity among the homeodomains (shown in the figures on the following pages) and the 15AA-defined groups as described in the main text, which are also listed in the document “Homeodomain subclass assignments”, together with the 65-class merged PBM/15AA classifications described in the main text.

**Supplementary Table 2.** The Top 100 overlaps among duplicated experiments were as follows:

**Different gene, same homeodomain (identical at all 57 positions within the pfam-defined homeodomain but with different flanking residues):**

Hoxa5-Hoxb5	86
Lhx2-Lhx9	90
Lmx1a-Lmx1b	85
Vax1-Vax2	90
Irx2-Irx5	76
Lhx1-Lhx5	96
Phox2a-Phox2b	91
Pitx2-Pitx3	83
Gbx1-Gbx2	88
Evx1-Evx2	83
Nkx2-4-Titf1	84

**Same gene, different clones (identical at all 57 positions within the pfam-defined homeodomain but with different amounts of flanking sequence included):**

Cart1-Cart1	90
Hoxa7-Hoxa7	76
Irx3-Irx3	81
Lhx6-Lhx6	86
Obox5-Obox5	95
Pou6f1-Pou6f1	94
Rhox11-Rhox11	86

**Same clone, different batch of protein:**

Nkx6-1-Nkx6-1	91
Six6-Six6	66
Cutl1-Cutl1	73

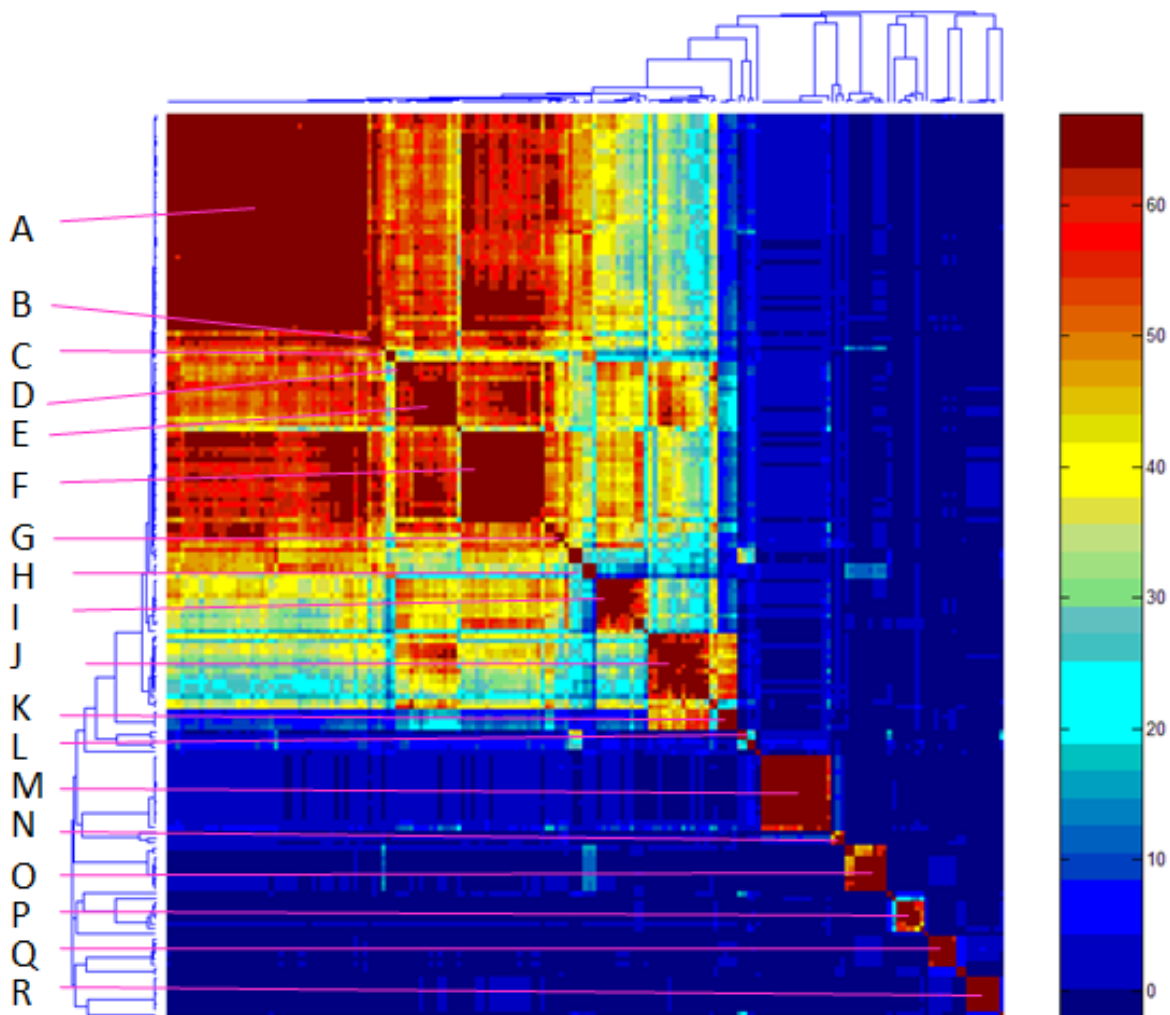
Mean	85.2
Standard Deviation	7.6
Left 0.5% of normal distribution	65.7

Note that proteins in these duplicate experiments were occasionally expressed by separate protocols (*in vitro* transcription/translation vs. *E. coli*). Clone sequences can be found in our supplementary table “Protein and DNA sequence”, where other experimental information and the groupings derived below are also found. Expression methods for all samples are listed on [http://the\\_brain.bwh.harvard.edu/pbms/webworks2/](http://the_brain.bwh.harvard.edu/pbms/webworks2/)

and at

<https://hugheslab.cabr.utoronto.ca/twiki/bin/view/MouseHomeodomain/WebHome>.

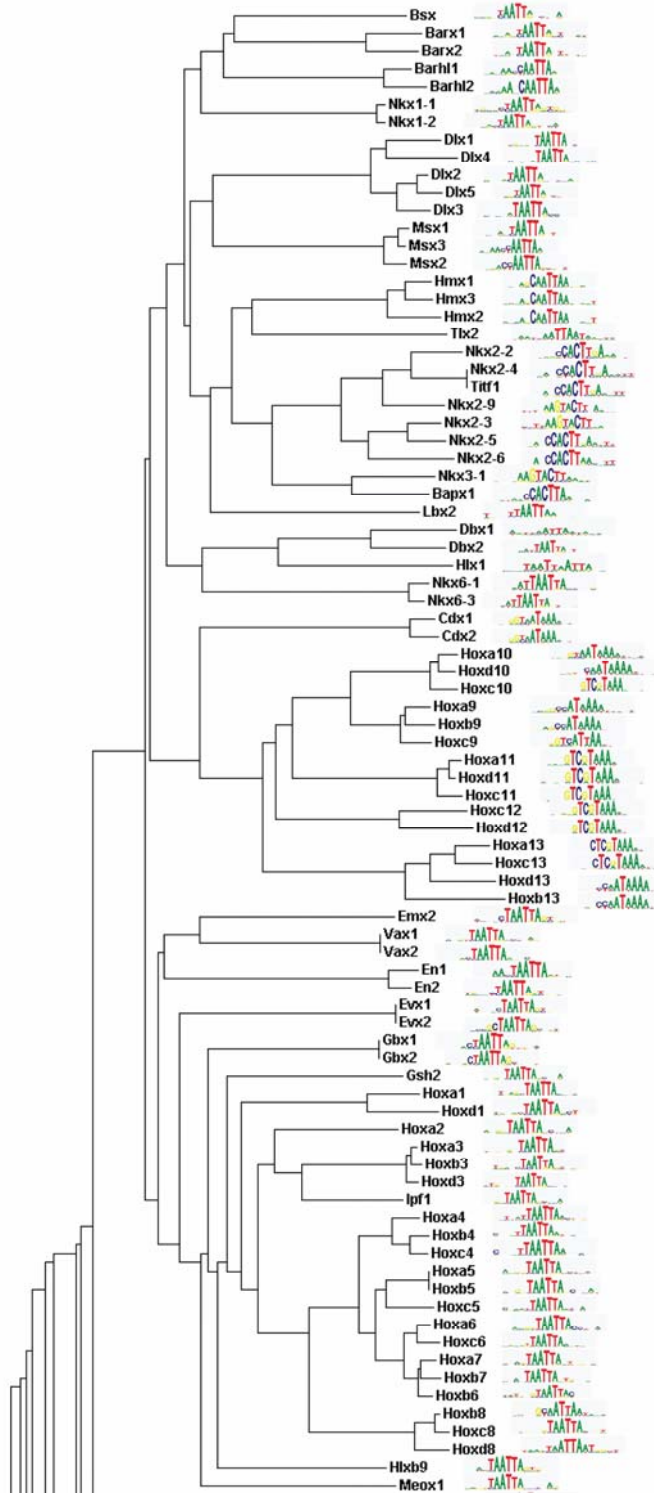
**Supplementary Figure 5.** From the diagram in Figure 2B, we identified 18 groups requiring further investigation were identified (numbered A through R in the graph below, which is identical to that in Figure 2B; the color scale is the same as shown in Figure 2B, i.e. the units refer to Top 100 overlap):



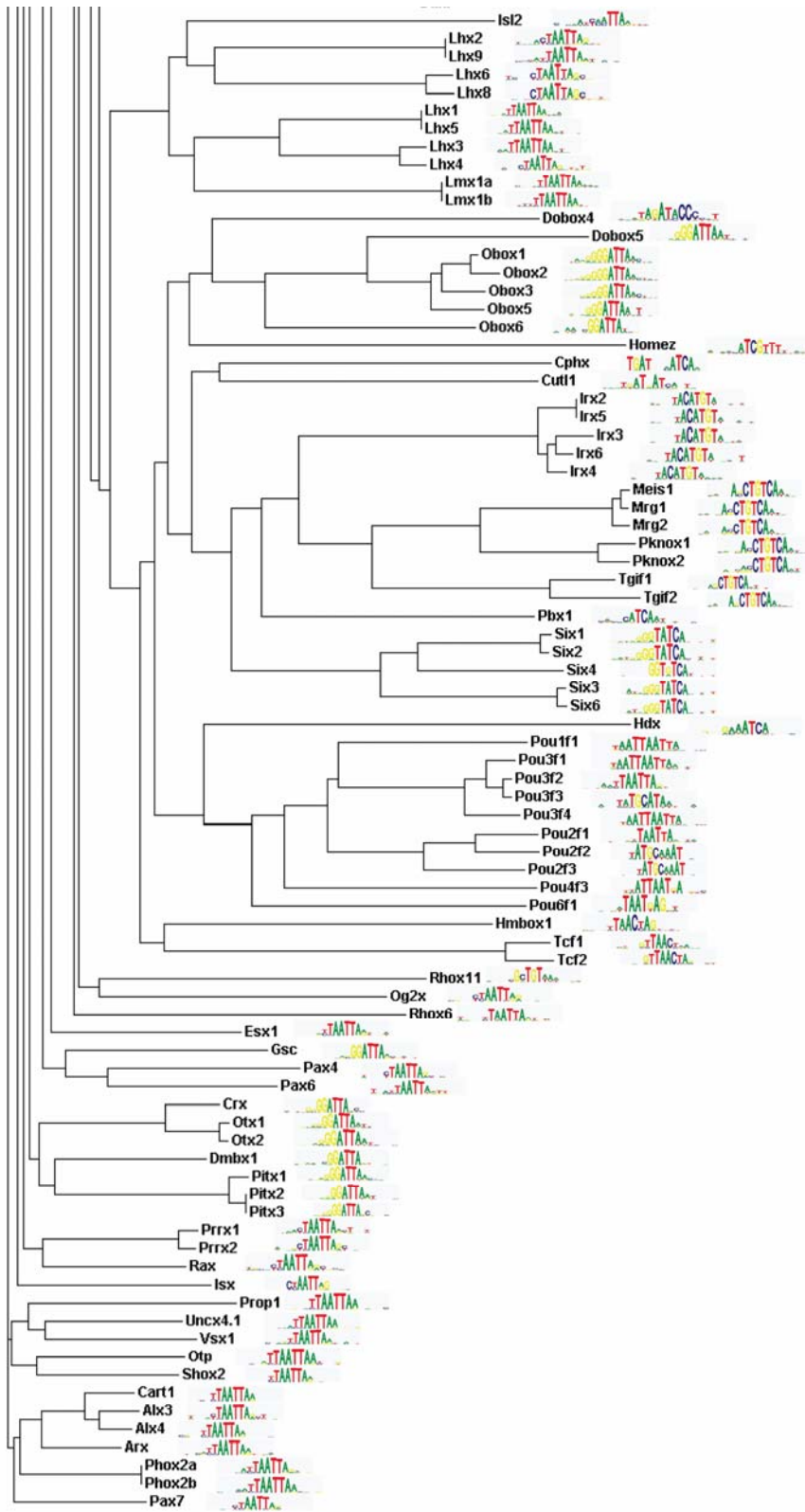
Evidence for subclassification of Group A is shown in Figure 2C.

In the figures on the following pages, the dendrogram for the full homeodomain amino acid sequence is shown, but it was not relied upon in the classification procedure. A “legend” for each of the diagrams is found on the first page, followed by two pages showing the full ClustalW diagram and “dominant motif” logos.

**Supplementary Figure 6.** This page and the next show the ClustalW phylogram tree for the homeodomain sequences for the 168 proteins for which binding specificity was determined. The dominant motif for each is also shown.







Supplementary Figure 7.

**The figures on the following pages contain the following for each of the groups with similar top 100 8-mer profiles (larger groups are spread over more than one page):**

Group name and number of 8mers in the analysis – selected to have at least one instance of  $E > 0.45$  among the experiments in the group

Heat-map (i.e. clustergram) with dendrograms showing hierarchical agglomerative clustering results, using E-scores for all 8-mers that have, and with Pearson correlation as the distance metric

Portion of the ClustalW dendrogram, derived from the full homeodomain amino acid sequence (in order to illustrate the relationships among the protein sequences), together with the dominant motif (these are taken from the full ClustalW tree shown below)

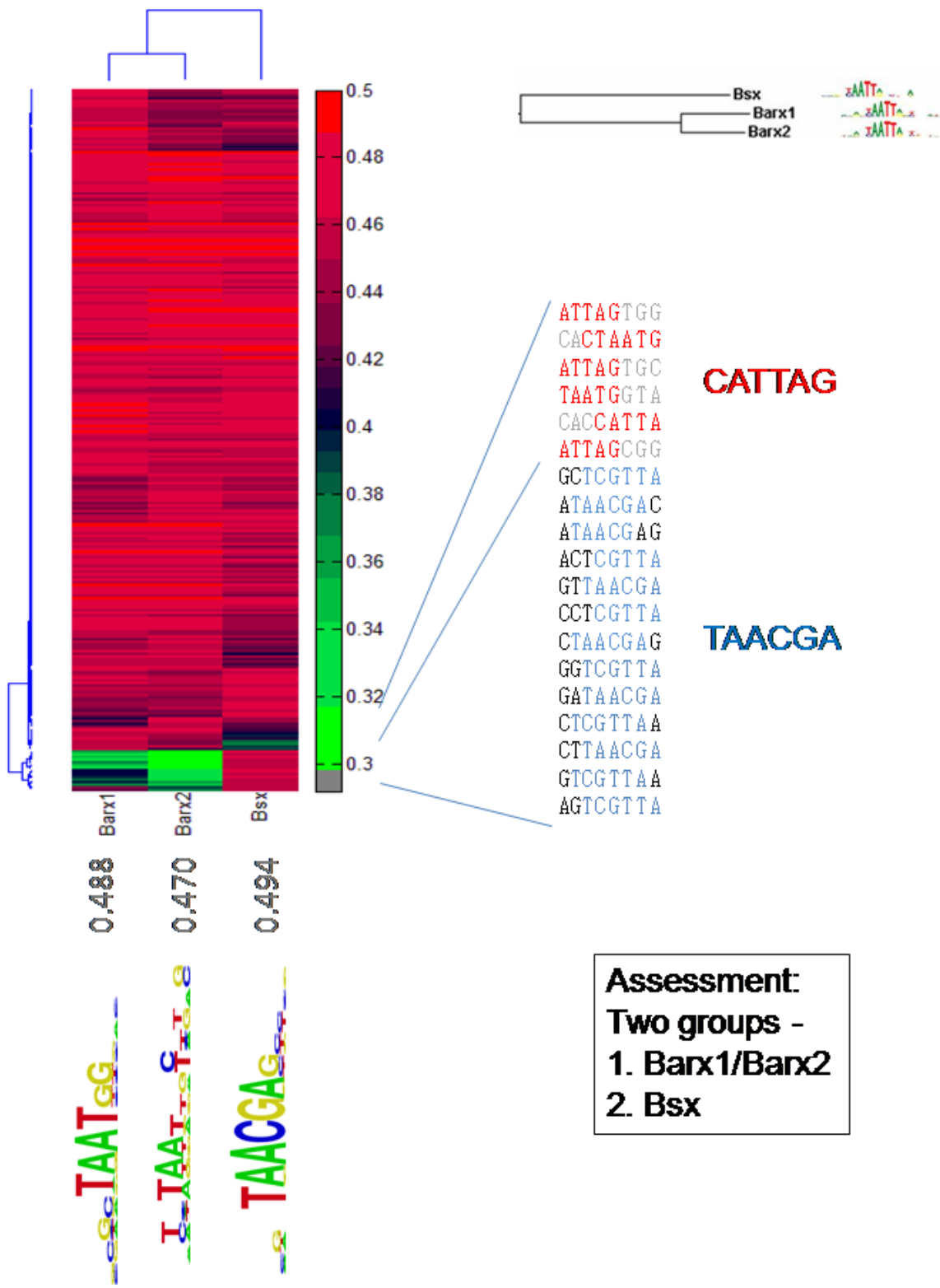
Pull-outs showing individual 8-mers in order to illustrate sequence elements in the clusters that distinguish one or more sub-categories of proteins within the clustergram. Sequence themes here were identified manually.

Highest TF-specific 8-mer E-score for each protein

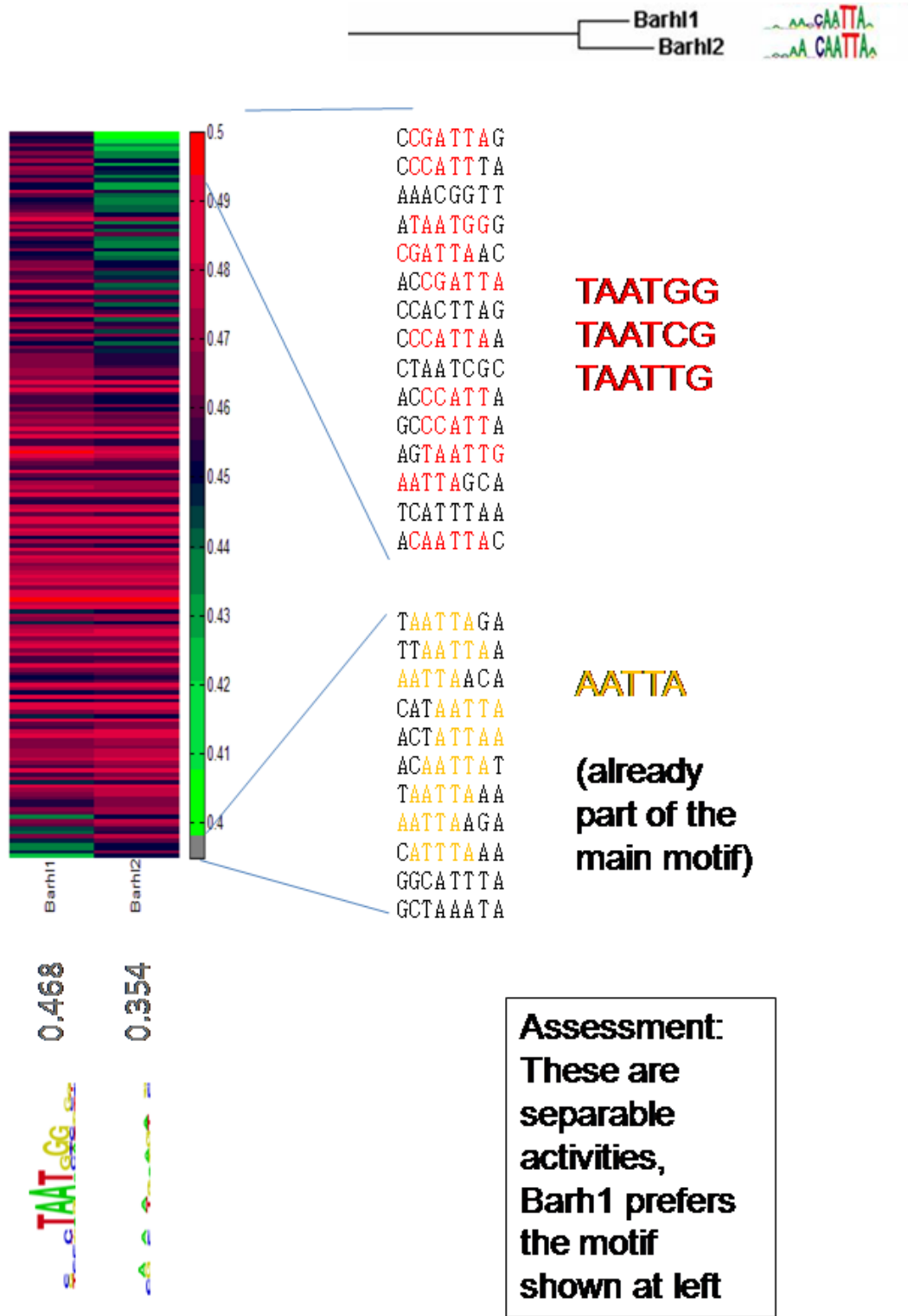
TF-specific motifs

Assessment of the results shown

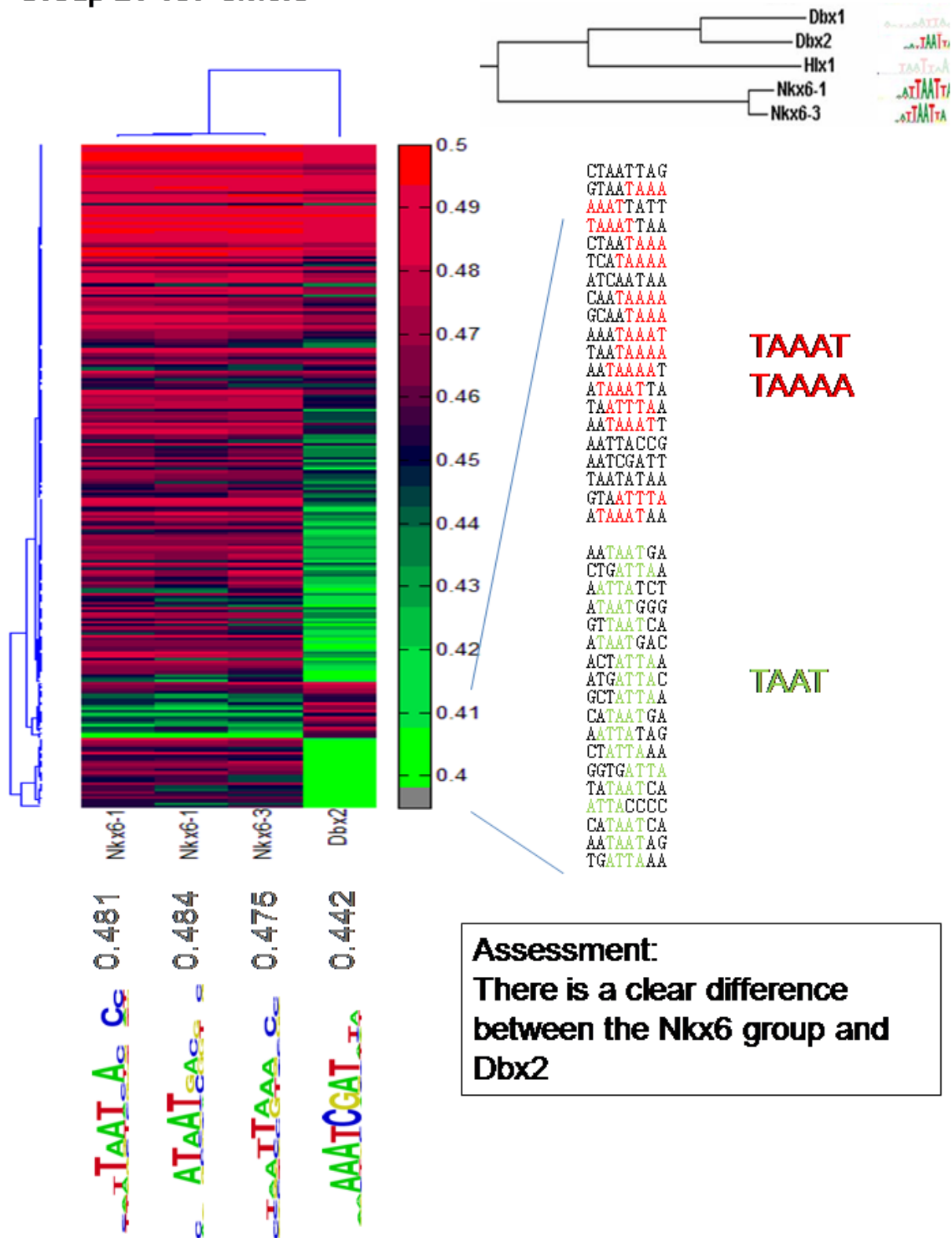
Group B: 257 8mers



# Group C: 187 8mers

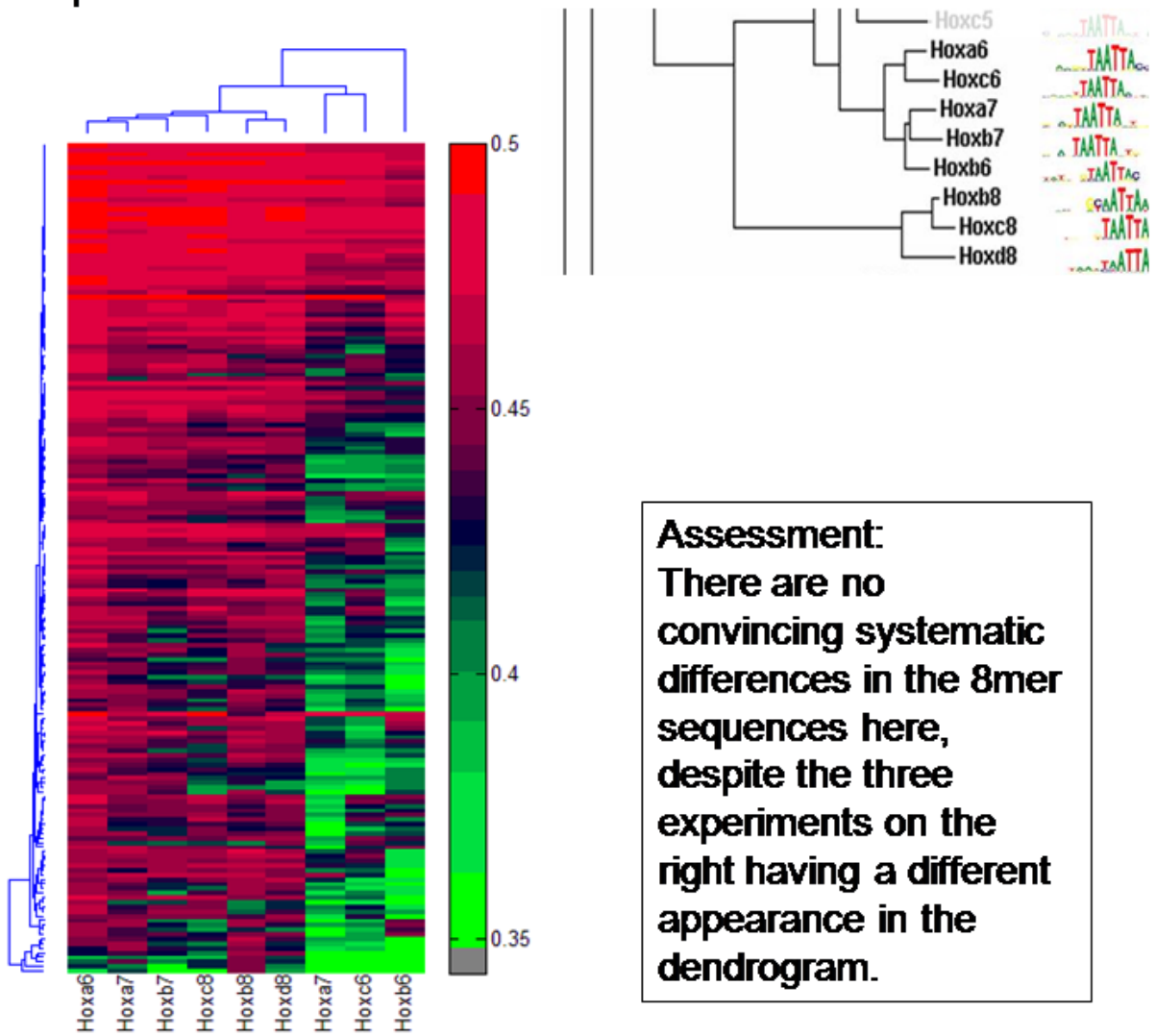


# Group D: 187 8mers



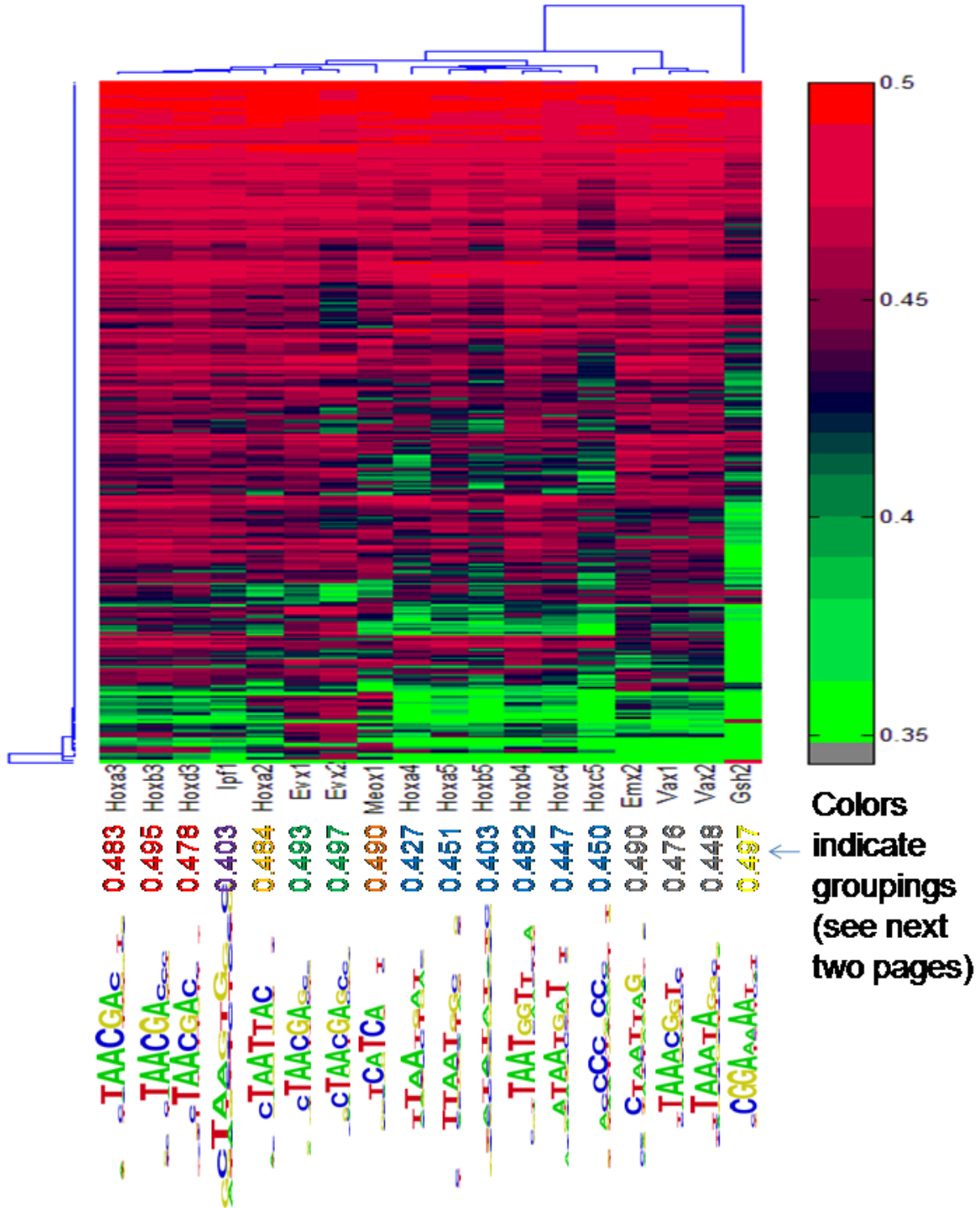
**Assessment:**  
 There is a clear difference between the Nkx6 group and Dbx2

## Group E: 181 8mers

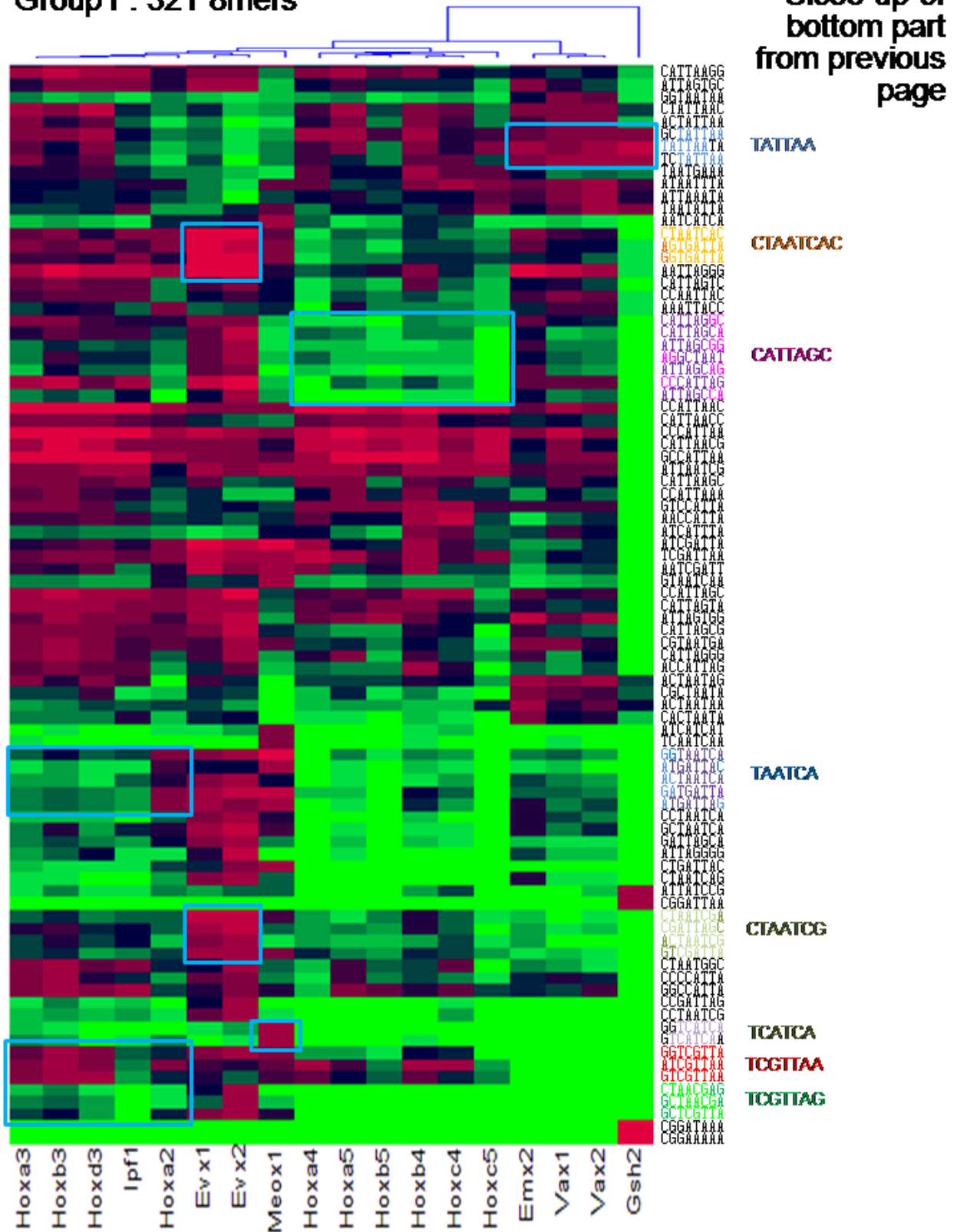


Hoxa6		0.495
Hoxa7_rep1		0.475
Hoxa7_rep2		0.381
Hoxb6		0.361
Hoxb7		0.417
Hoxb8		0.434
Hoxc6		0.424
Hoxc8		0.467
Hoxd8		0.471

Group F: 321 8mers

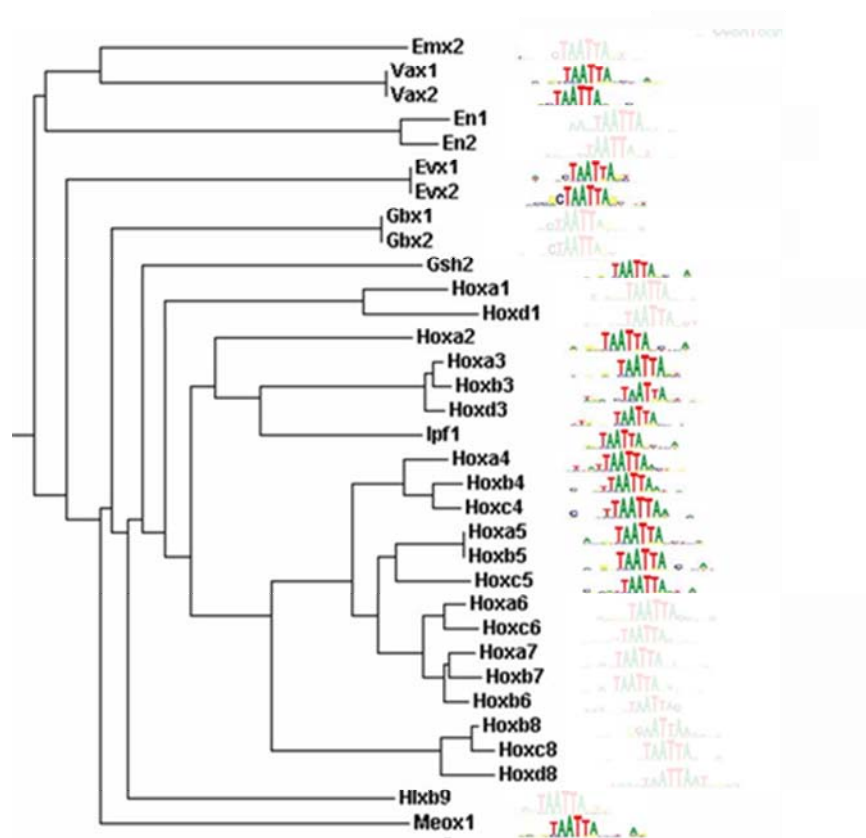


Group F: 321 8mers





Group F: 321 8mers with E > 0.45

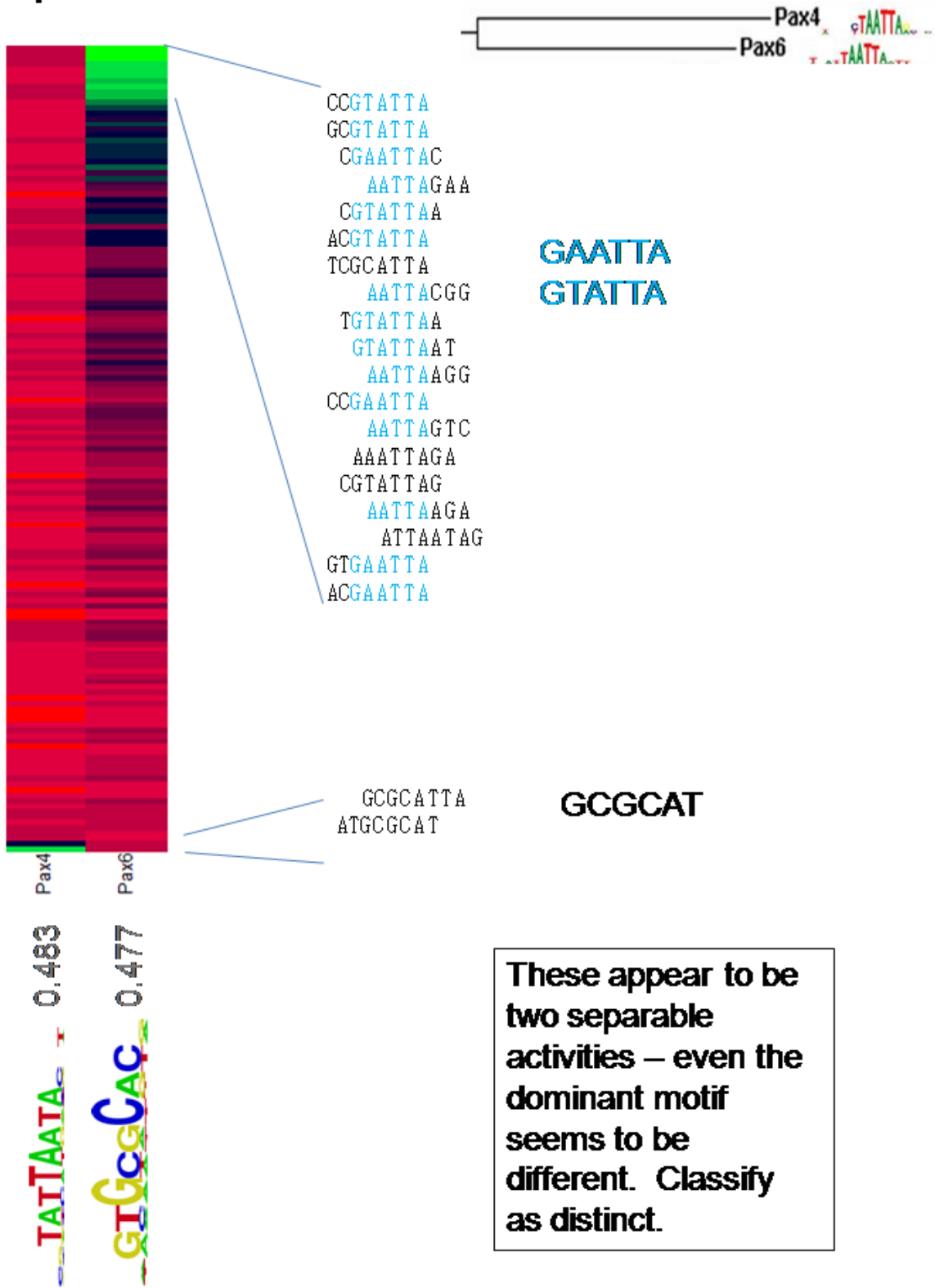


Data on previous pages support the following groups:

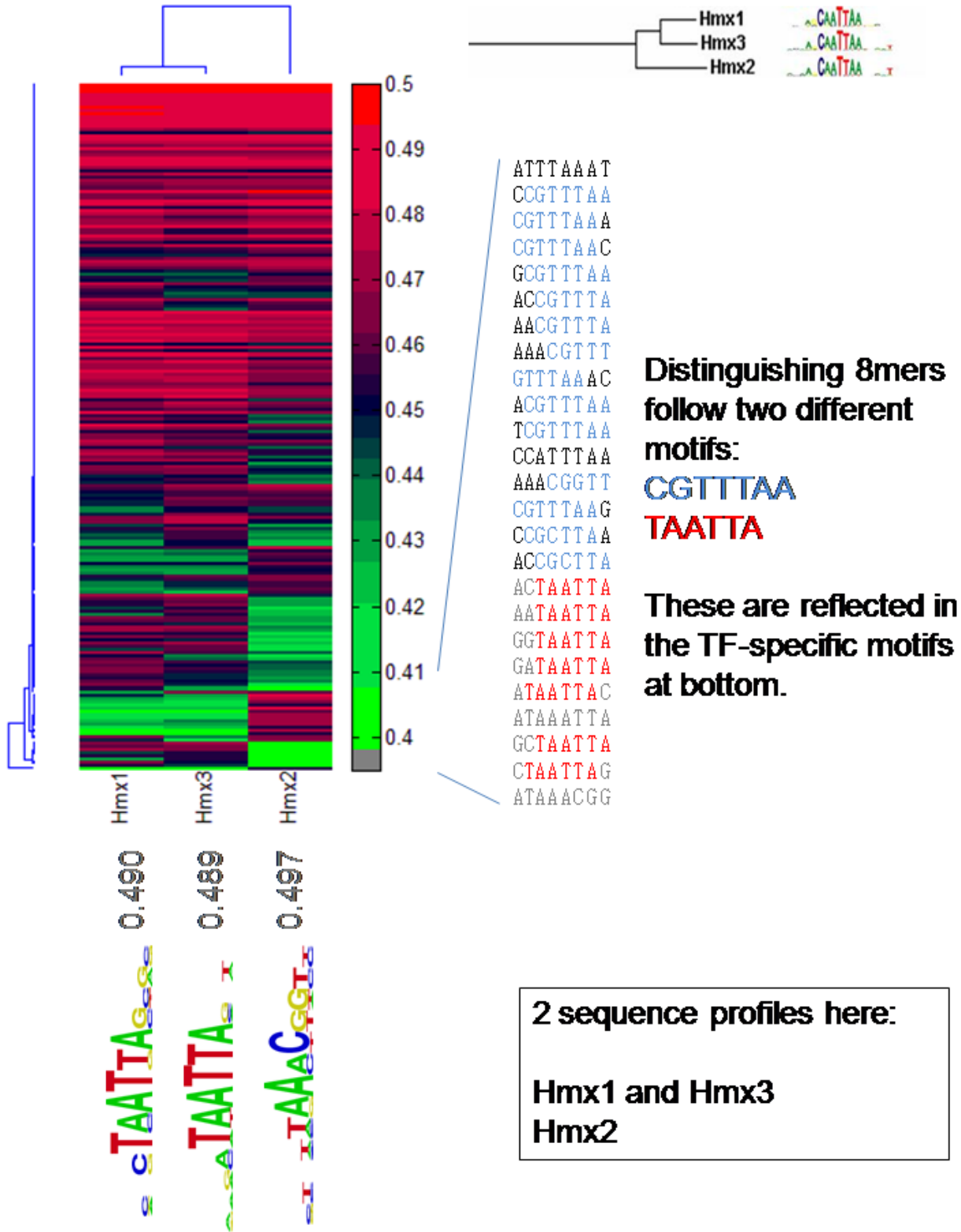
- Hoxa2
- Hox3
- Ipf1
- Evx1/2
- Meox1
- Hox4/5
- Emx2/Vax1/2
- Gsh2

Most of the TF-specific motifs agree with the 8mer clustering. Some of the distinguishing features in the clustergrams are *unpreferred* sequences.

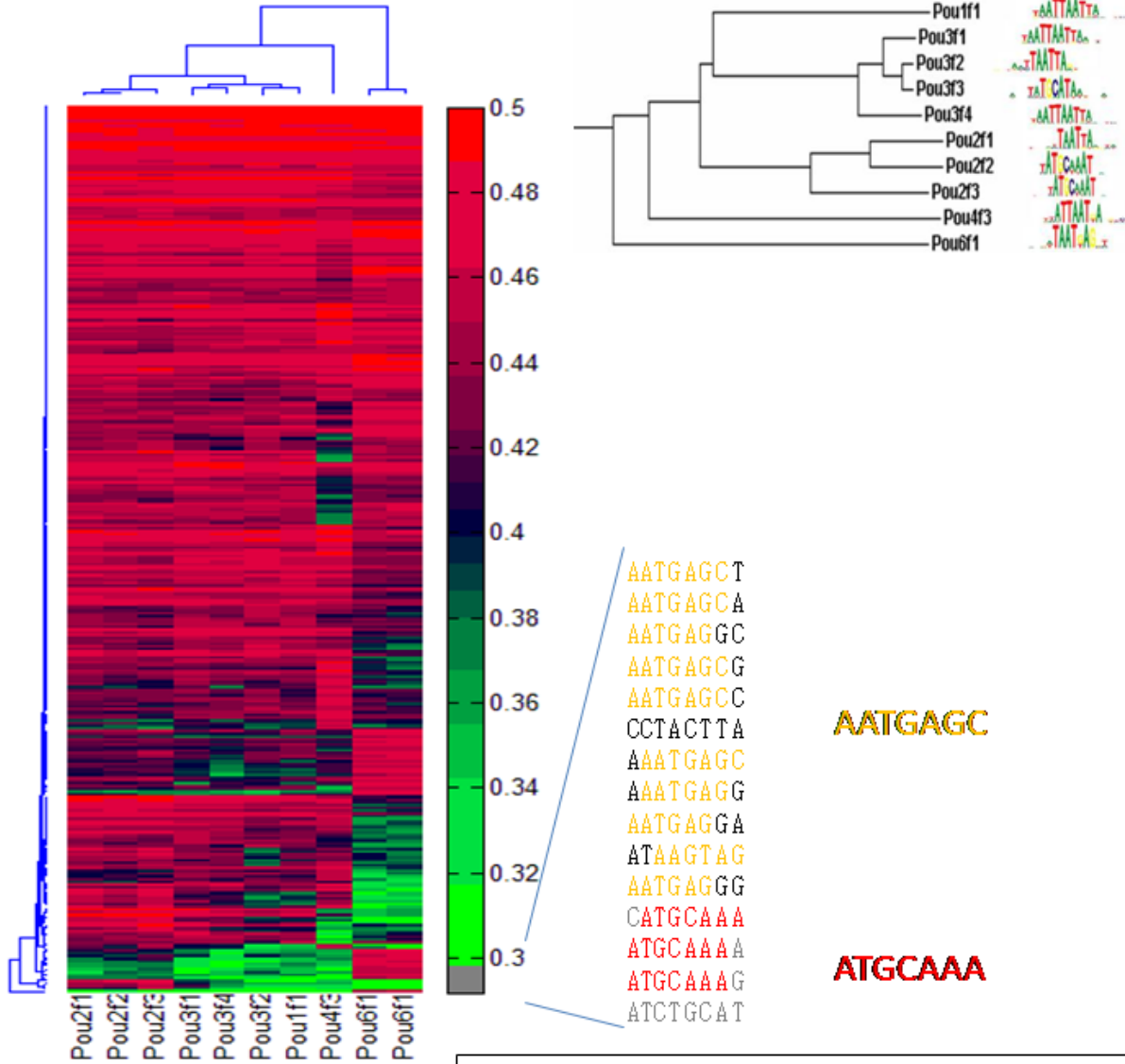
**Group G: 149 8mers**



Group H: 149 8mers



# Group I: 324 8mers

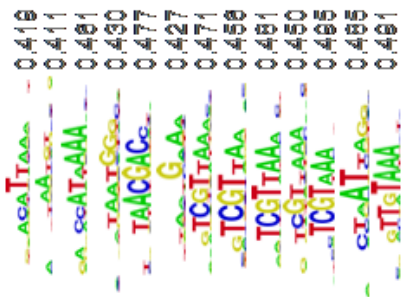
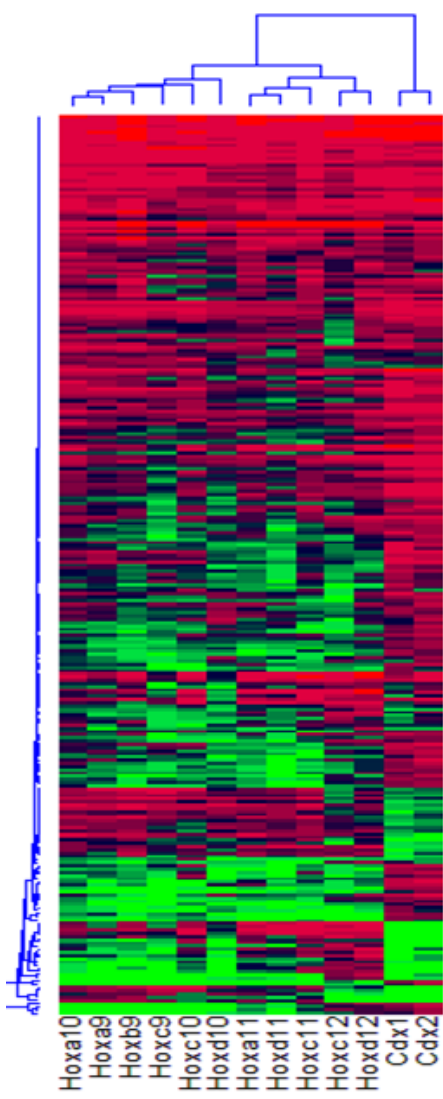


0.498 ATGCAAAT  
 0.499 ATCAAAT  
 0.499 ATCAAAT  
 0.497 ATCAAAT  
 0.483 ATCAAAT  
 0.447 ATCAAAT  
 0.447 ATCAAAT  
 0.497 ATCAAAT  
 0.493 ATCAAAT  
 0.494 ATCAAAT

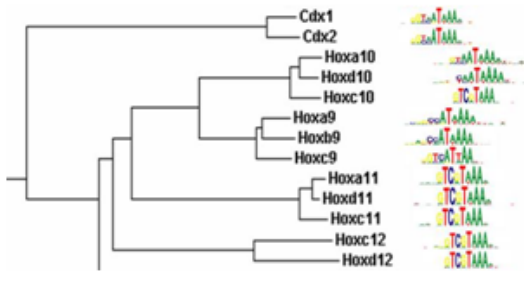
Groups derived from 8mer scores and TF-specific motifs *\*almost\** exactly agree. Pou3f1 is the only ambiguity – go with the dendrogram.

Four groups:  
 Pou2  
 Pou3/1  
 Pou4  
 Pou6

# Group J: 281 8mers



AATTACCC  
 CAATTACC  
 CCATAAAAG  
 GGCCATAA  
 CCATAAAC  
 GTTATAAAA  
 ATTAATAAA  
 CCATATAAA  
 TCATATAAA  
 AATAAACT  
 AATAAAATC  
 GTAAAATTA  
 AATAAAACG  
 CTAAAATTA  
 CTAAATTA  
 GCAAATAC  
 CTAAATGC  
 CCCAATTA  
 TCGTTAAA  
 ATCGTTAA  
 GTCGTTAA  
 GTCGTTAA  
 CGATAAAA  
 ATAACGAC  
 AGTCGTAA  
 GCGATAAA  
 CTTACGA  
 GTCGTAAA  
 GATCGTAA  
 ATCGTAAAT  
 ATTACGAG  
 CATCGTAA  
 GTAACGA  
 CTCGTTAA  
 GTGATAAAA  
 TGATAAAA  
 GCAATATA  
 ATATAATG  
 CCCATTAAC  
 CCATTAAC  
 TACATTAAC  
 GACATTAAC  
 ACATTAAC  
 GTTGTAAA  
 ATTGTAAA  
 TTGTAAA  
 ATTTACAA



**TAAA**

**TCGTTAA**

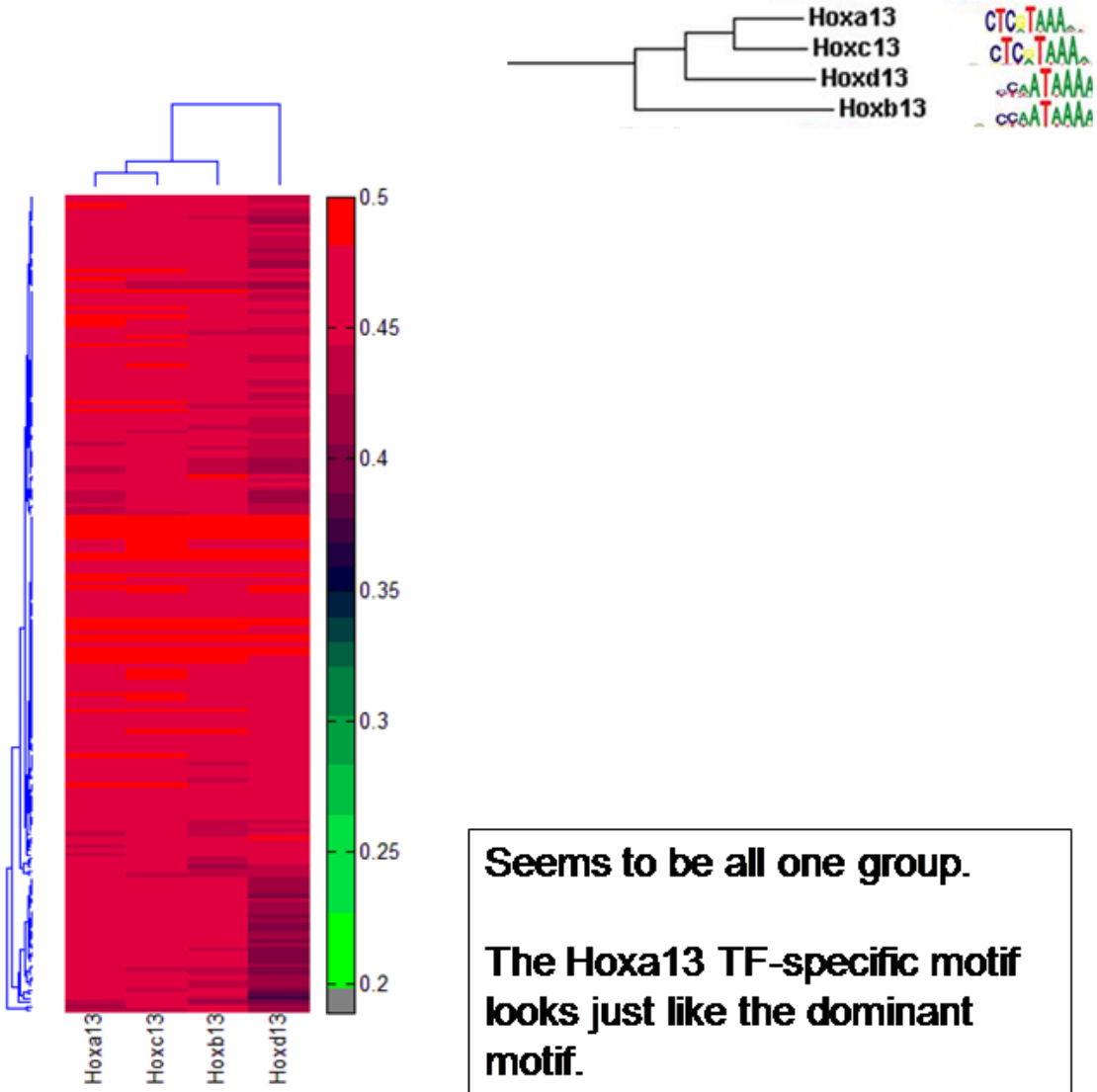
**(A)TCGT**

**CATTAA**

**TGTAAA**

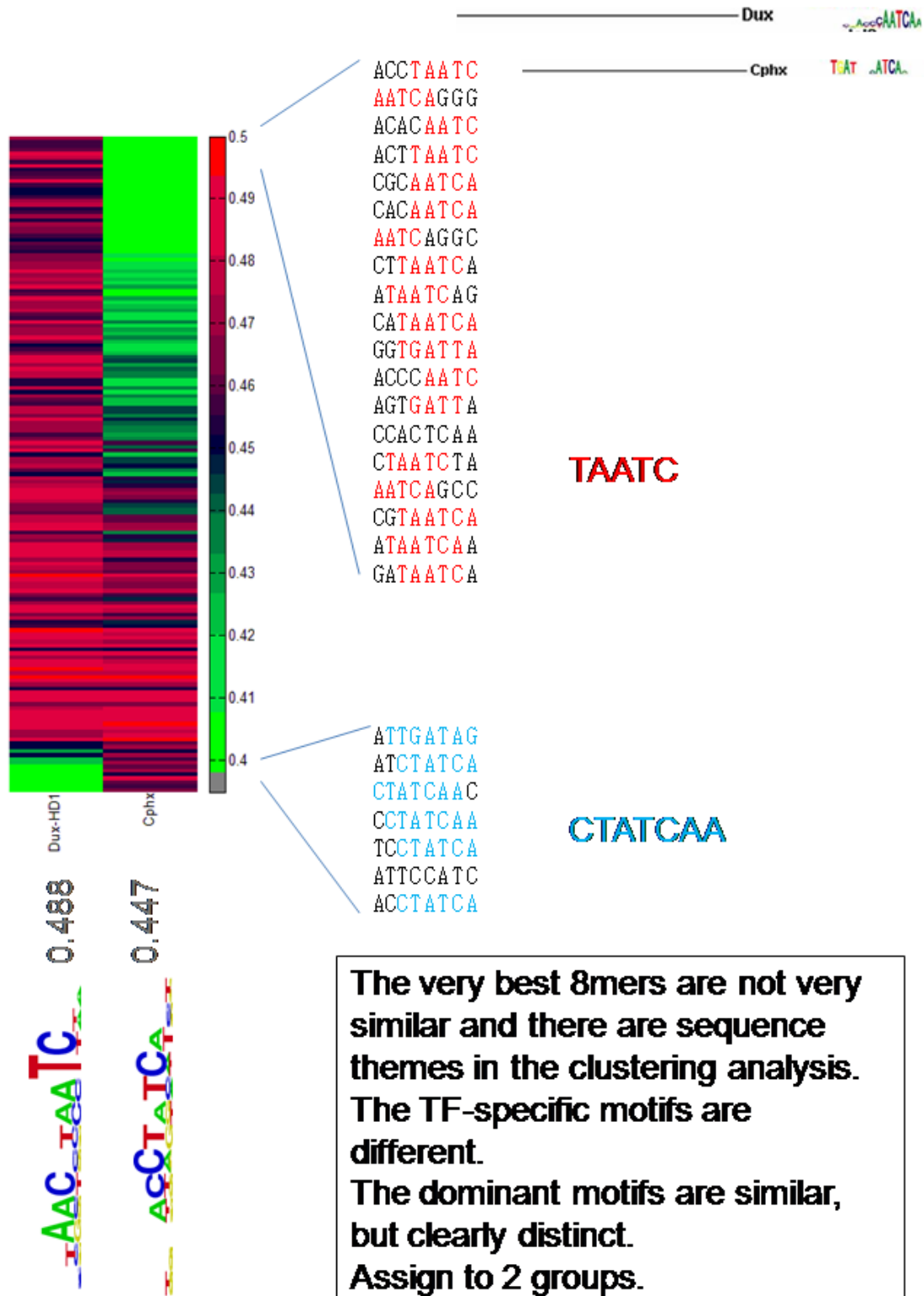
**Dominant motifs, dendrograms, and 8mers derived from clusters all agree on four groups:**  
**Cdx1/2**  
**Hox9/10**  
**Hox11**  
**Hox12**  
**TF-specific motifs largely agree.**

# Group K: 199 8mers

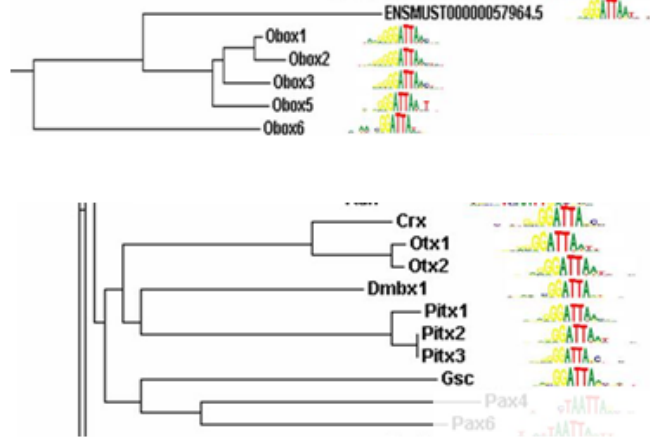
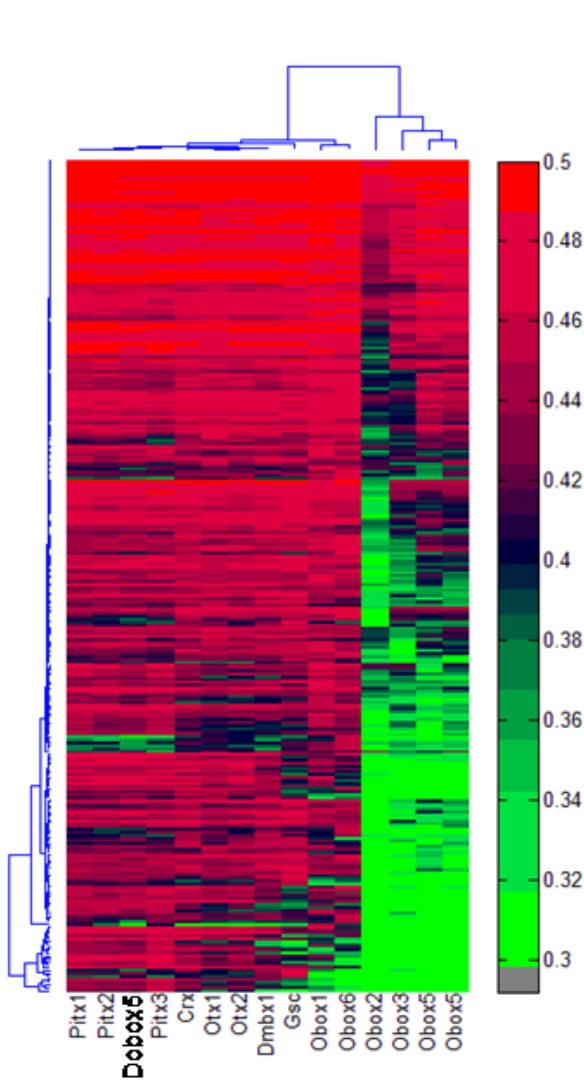


<b>Hoxa13</b>		0.497
<b>Hoxb13</b>		0.424
<b>Hoxc13</b>		0.471
<b>Hoxd13</b>		0.455

# Group L: 199 8mers



# Group M: 288 8mers

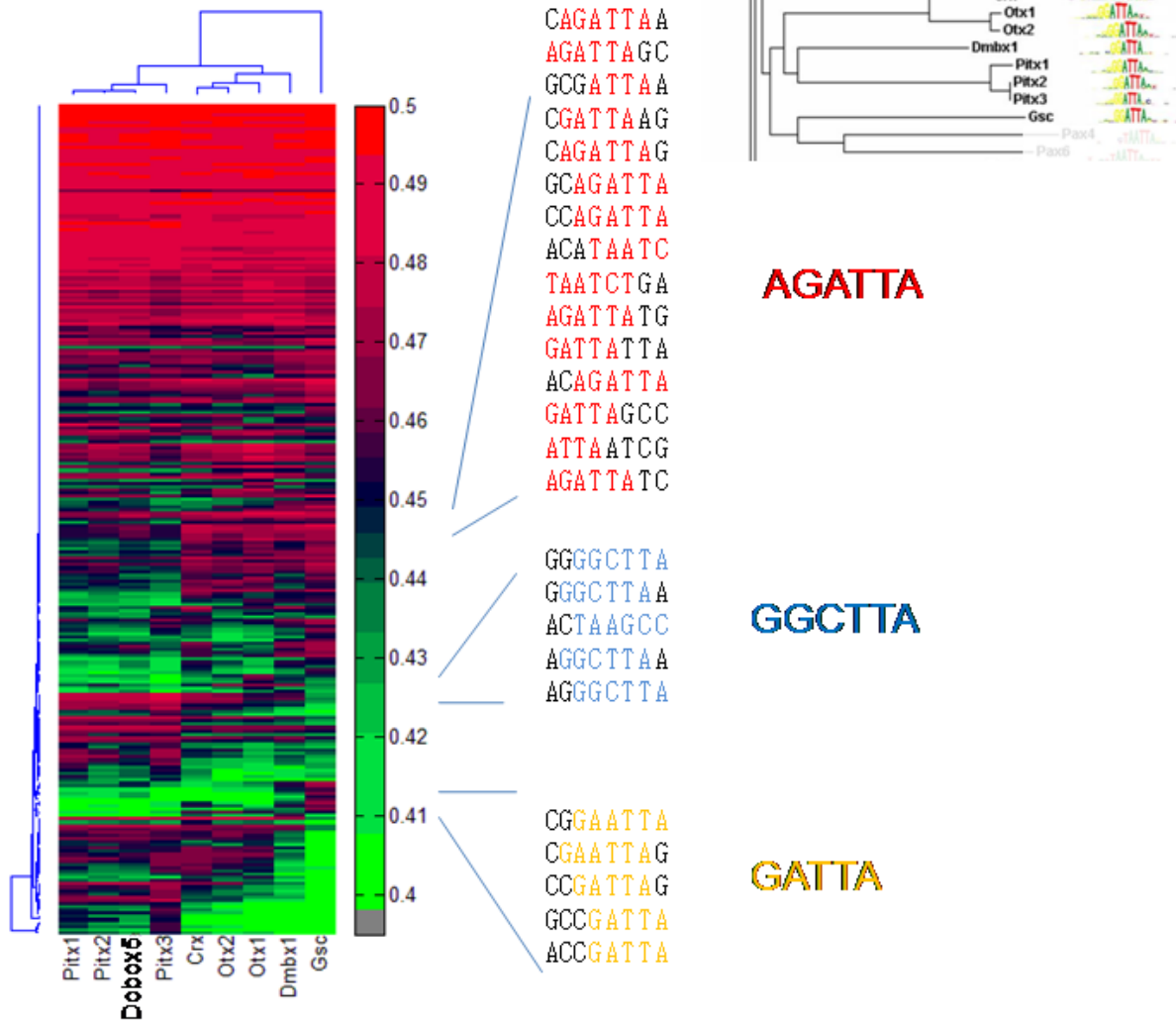


The Obox proteins seem to require more G's.

Try repeating the analysis without the Obox proteins, which are obscuring subtler differences between the other proteins. Over ->



## Group M2: 255 8mers



TF-specific motifs are shown in the next page ->

<u>Crx</u>		0.492
Dmbx1		0.490
Otx1		0.490
Otx2		0.480
<u>Gsc</u>		0.497
Dobox5		0.494
Pitx1		0.497
Pitx2		0.490
Pitx3		0.497
Obox1		0.484
Obox2		0.431
Obox3		0.365
Obox5_rep1		0.384
Obox5_rep2		0.397
Obox6		0.491

Evidence here indicates four groups:

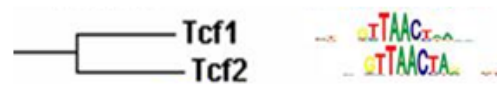
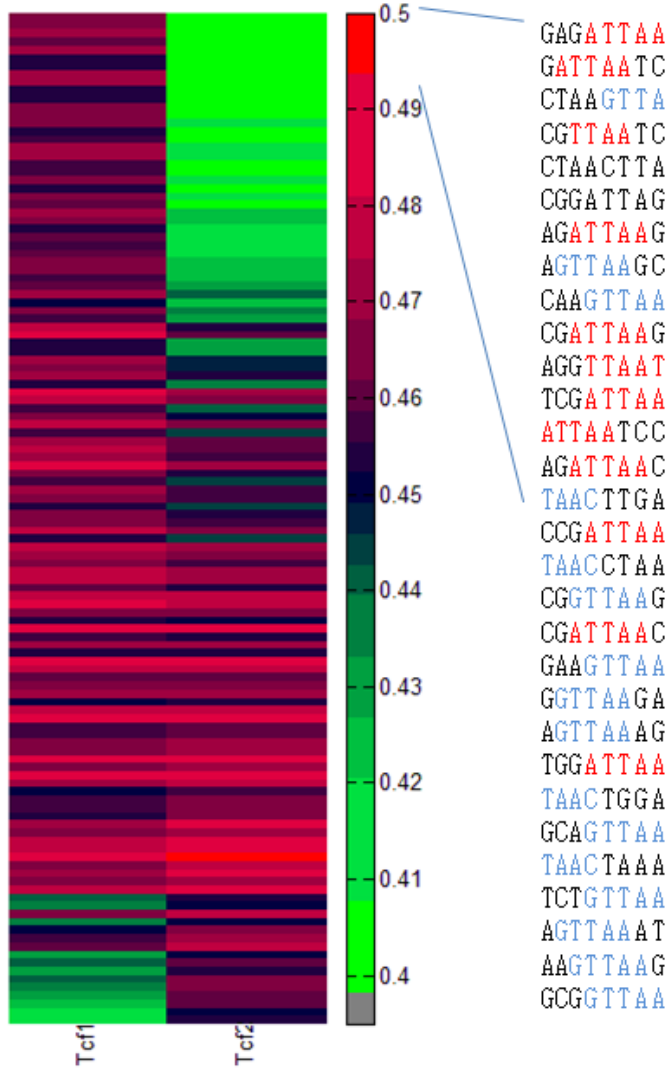
Gsc

Obox1/2/3/5/6

Pitx1/2/3/Dobox5

Otx1/2/Crx/Dmbx1 (Dmbx1 is a candidate for its own group)

Group N: 124 8mers with E > 0.45



Dominant motifs look similar.

The 8mers that are different mostly contain **ATTAA** or **GTTAA**.

These are preferred by Tcf1 but apparently not Tcf2. There is no such theme found at the bottom of the graph.

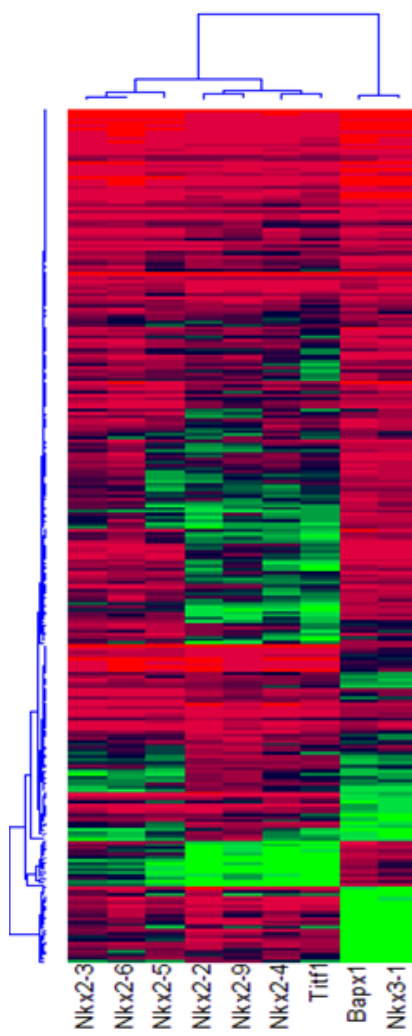
The TF-specific motifs show roughly the same thing: TTAA preferred by Tcf1 but not Tcf2.

Supports two different activities.

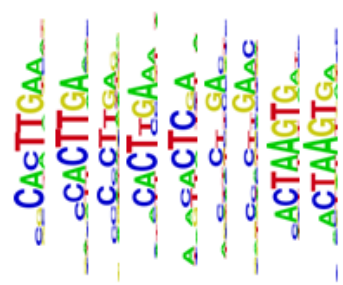
Tcf1 0.455

Tcf2 0.411

Group O: 279 8mers



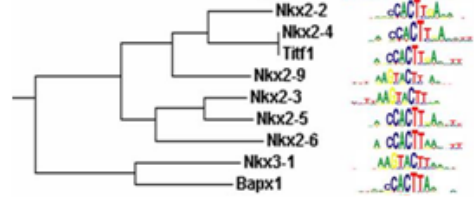
0.483  
0.494  
0.445  
0.493  
0.491  
0.451  
0.460  
0.497  
0.479



TC**ACTTAA**  
CA**CTTAGG**  
CA**CTTAGC**  
CT**AAGTAC**  
CT**AAGTGC**  
CA**CTTAGA**

CC**ACTTGA**  
CA**CTTCAA**  
CA**CTTCAG**  
CA**CTCAAG**  
CA**CTTGAA**  
CA**CTTGAG**  
CA**CTTGAC**  
GC**ACTTGA**  
GT**ACTTGA**

ATCA**CTTA**  
GTCA**CTTA**  
CTCA**CTTA**  
ATA**CTTAG**  
CAAT**TAAG**  
ACT**TAGTG**  
ACT**TAGTA**  
CTTA**ATTA**  
ACTA**ATTA**  
CCA**ATTAA**  
GTC**ATTAA**  
TTA**ATTAA**  
GCA**ATTAA**  
ATTA**ATTG**  
CTA**AGTGA**



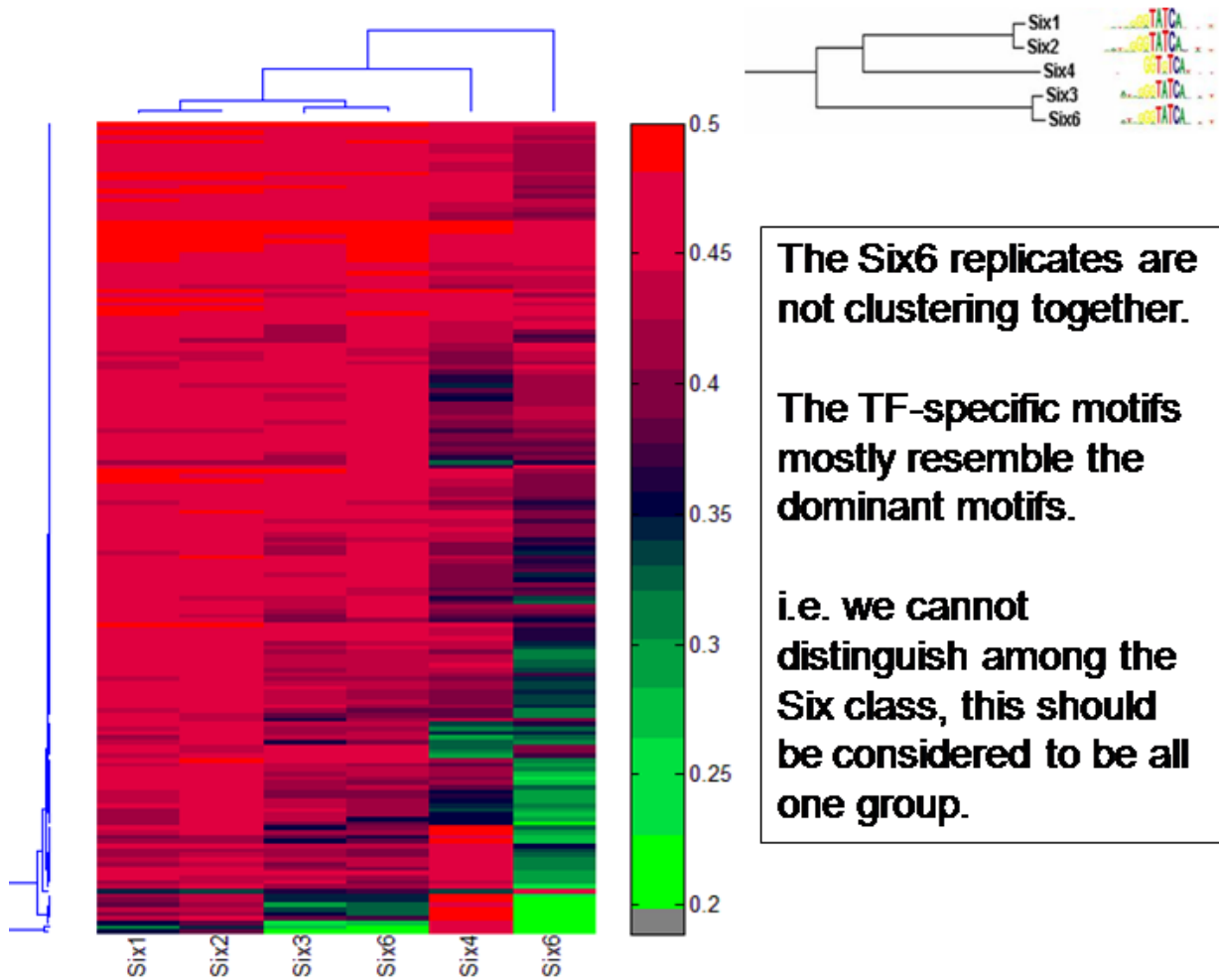
**ACTTAG**

**CACTT(GA)**

**CTAA**  
**Or**  
**AATTAA**

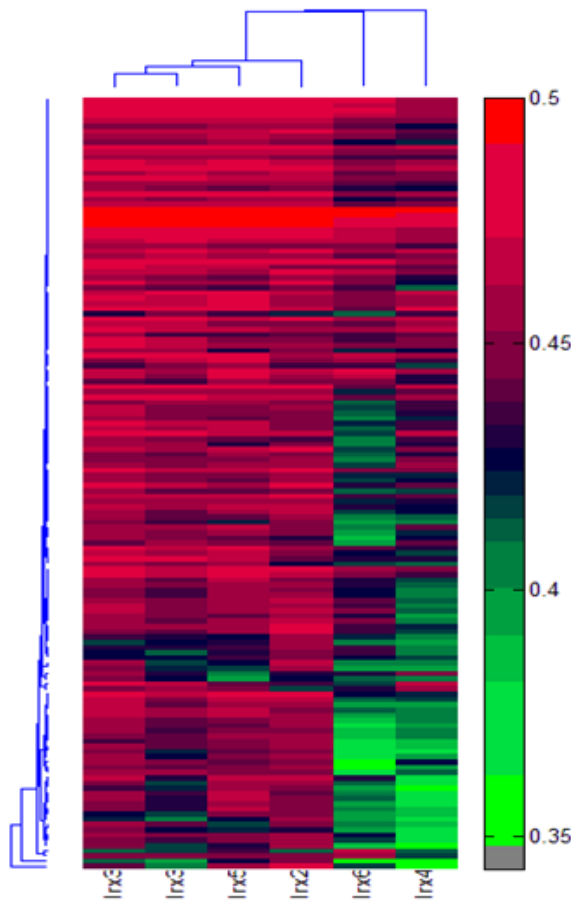
The dendrogram and the 8mer clusters support three groups:  
 Nkx3-1 and Bapx1  
 Nkx2-3,5,6  
 Nkx2-2,4,9, and Tif1  
 The TF-specific motifs are consistent although less informative. The dominant motifs for this class do not reflect the original data very well as shown previously.

## Group P: 180 8mers



Six1		0.466
Six2		0.492
Six3		0.451
Six4		0.493
Six6_rep1		0.450
Six6_rep2		0.428

# Group Q: 148 8mers

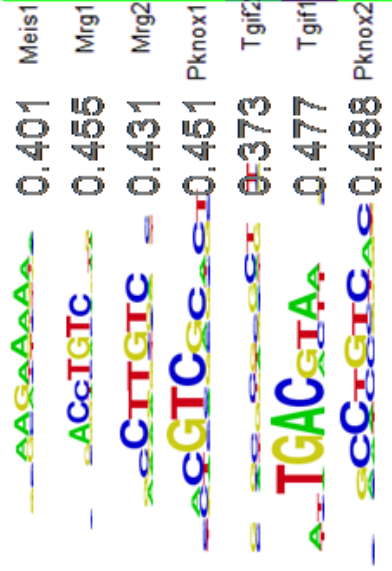
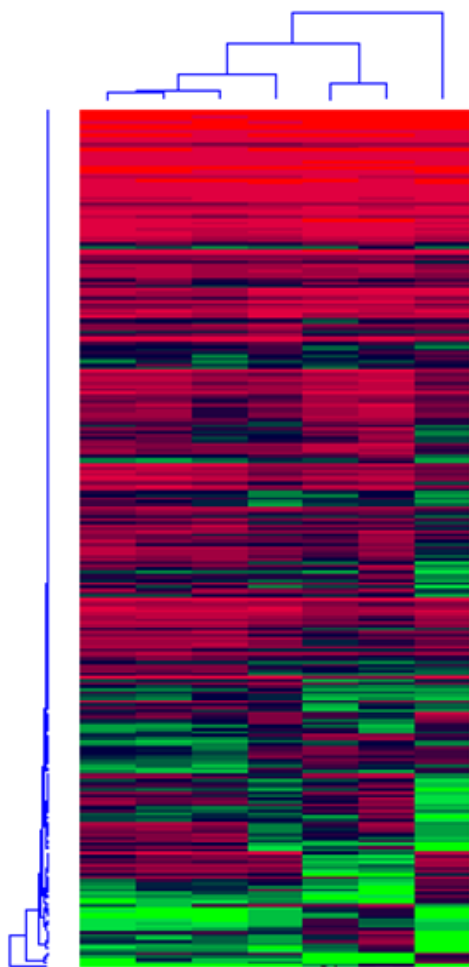


There are no obvious sequence motifs among the clusters here.

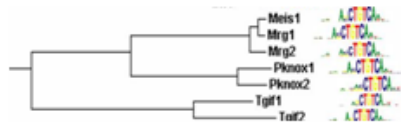
Classify as being all in the same group.

Ir2		0.477
Ir3_rep1		0.442
Ir3b_rep2		0.412
Ir4		0.406
Ir5		0.454
Ir6		0.424

# Group R: 283 8mers



TCTGACAA  
 ATTGTCAG  
 ATGTCATA  
 AGTGTCA  
 GTTGACAA  
 GATGACAC  
 ATGACATA  
 AATGACAA  
 AATGTCAT  
 ACAATGAC  
 AATGTCAA  
 AATGTCAG  
 GATTGTCA  
 GTGACATA  
 ATGACAA  
 ATGACAA  
 TATGACAA  
 GATGACAA  
 CTGACAA  
 CTGACAA  
 CCTGACAA  
 CCTTGTCA  
 GCTGACAA  
 TGACACA  
 CATGACAA  
 GACAGGTC  
 AGCTGTGC  
 CGCCTGTC  
 GACAGGCC  
 GACAGGTA  
 GACACGTC  
 AACCTGTC  
 ACCTGTCT  
 CACCTGTC  
 GACAGGCC  
 ACCCGTCA  
 ACGACAGC  
 GACAGGCA  
 CGACAGCC  
 ACAGTGT  
 AGCCTGTC  
 CGTGTGTC  
 AAACGTCA  
 TACGTCAA  
 AACGTCAA  
 ATACGTCA  
 ATGACATA  
 ATTTTGAC  
 CGTCAAAG  
 CCGTCAA  
 ACGTCAA  
 AATTGTCA  
 CAATGTCA  
 TAATGTCA  
 ATTGTCAA  
 TGACATA  
 ATTTGTCA  
 ATGACAA  
 ACCTGTGC  
 ACGACAGG  
 CCTGTGCA  
 CGACAGGC  
 TTGACAAA



**ATGACA**

**GACAGG**

**ACGTCA**

**TGTCA**

**CCTGTGC**

It appears as if there are three discrete groups in the clustergram:

**Meis1/Mrg1/2/Pknox1**  
**Tgif1/2**  
**Pknox2**  
 (Pknox1 and 2 may differ primarily by magnitude)

These are the TF-specific motifs for Group A, shown in Figure 2D

Alx3	T TAAITAA	0.469
Alx4	TAATIAA	0.463
Phox2a	TAATTCG	0.462
Phox2b	TAAITAA	0.474
Prrx2	TAATTA	0.484
Arx	TAATC	0.483
Uncx4.1	TAATTA	0.494
Shox2	CTAATA	0.466
Esx1	CAATAA	0.492
Prrx1	TAATTCG	0.477
Cart1	TAATTA	0.451
Otp	TAATTA	0.499
Rax	CGC	0.485
Vsx1	TAATTA	0.485
Isx	TAATTCG	0.483
Lhx1	TAATTA	0.492
Lhx5	TAATIAA	0.497
Lhx3	TAATTA	0.497
Lhx4	TAATCA	0.492
Lhx2	TAACTG	0.496
Lhx9	TAACTG	0.492
Nlhx1-1	CTAATCG	0.495
Nlhx1-2	CTAATAG	0.484
Msx1	CAATAAAA	0.464
Msx2	TAATTCG	0.464
Msx3	CAATAAAA	0.498
Gbx1	CTAATTCG	0.494
Gbx2	TAATCG	0.415
Hlxb9	TAATCG	0.493
Hoxd1	ATCGAT	0.496
Pax7	TAATCG	0.494
En1	ATGATCG	0.434
En2	CTCAATCA	0.496
Dlx1	TAATTCG	0.481
Dlx2	TAATTCG	0.485
Dlx3	TAATTCG	0.487
Dlx4	TAATTCG	0.478
Dlx5	TAATTCG	0.496
Prop1	TAATTA	0.463
Lbx2	TAACAG	0.497
Og2x	CAATGA	0.491
Hoxa1	CCCCIT	0.439



## 7. Predicting 8-mer profiles and scoring the predictions

We considered two general methods for predicting 8-mer binding profiles on the basis of the primary amino acid sequence: nearest-neighbor and linear regression. These both have the advantages of being able to make quantitative predictions from categorical features, and fast run times.

### *Nearest Neighbor*

In the nearest neighbor (NN) approach, we predict the 8-mer profile of any given homeodomain protein by taking the 8-mer profile(s) of its nearest neighbor(s) (averaging E-scores in the case of a tie). We tried several variations, which differed from each other in the distance metric used and/or the residues considered to determine the nearest neighbor relationship. The two distance metrics used were: (a) the number of different AAs between two aligned homeodomains, and (b) the sum of the negative of the similarities (as given by the PAM250 matrix) between the AAs of two aligned homeodomains. With respect to the residues considered, we tried a “full-homeodomain” version where all 57 residues in the pfam-defined homeodomain model were used to determine the nearest neighbor(s) of a protein and several “specificity residues” versions where four different sets of residues known in the literature (10, 11) to make direct contact with the DNA were used to determine the nearest neighbor(s) of a protein (see below).

### *Linear Regression*

Linear regression techniques require a vector of numerical values as input, and the number of features should be at least five-fold smaller than the number of examples. In our case, the features must represent the protein sequence, and should preferably number less than ~30, since we analyzed 168 homeodomain proteins. The full homeodomain family protein sequence alignment was downloaded from Pfam (Pfam Accession Number: PF00046), from which we extracted only the alignments of the 168 mouse homeodomains. This alignment was then converted to a binary representation by replacing all 20 standard amino acids in any of the canonical residue positions with unique 20-bit binary flags. Regions of the Pfam alignment that correspond to insertions were treated as separate binary variables, with a value of 1 indicating the presence of an insertion in that particular region. We then applied Principal Components Analysis (PCA) (12) to the alignment encoding in order to reduce the number of variables per protein in this encoding strategy, and to eliminate correlation between these variables. The number of components retained (23) was selected using Parallel Analysis (PA), which is essentially a random permutations test that asks whether the  $N^{\text{th}}$  component explains more of the variance than the  $N^{\text{th}}$  component would in a permuted version of the same data (reviewed in reference (13)). These 23 components together account for 70% of the variance in the binary vectors.

For a given 8-mer, we used Partial Least Squares (PLS) regression (14) to model the relationship between the retained principal components and the 8-mer Z-score, which are the independent variables and the response variable, respectively (we used Z-scores for this analysis, under the assumption that the model would learn a linear relationship between sequence features and affinity, which, as described above, is more likely to be reflected by Z-scores than E-scores). An internal round of cross-validation (15) was used to determine both the optimal model and the optimal number of latent variables (see reference (16) for details of PLS). Finally, for each of the 168 proteins, we predicted its Z-score profile across all 8-mers using a leave-one-out cross-validation strategy, wherein a distinct model learned for each 8-mer and for each homeodomain (i.e. a separate model for all 157 x 32,896 entries in the data table) was used to predict its Z-score entry.

We also tested other regression approaches (Support Vector Regression, Lasso, and Ridge Regression) and other encodings of the amino acid sequences; the PLS results are shown because they were the highest-performing. A full description of our efforts to use regression techniques to model PBM data will be published elsewhere.

### ***Comparison of prediction methods***

Each NN version and the regression outputs were evaluated by leave-one-out cross-validation using the median and the mean of the following performance measures: the Spearman correlation coefficient over all 32,896 E-scores (Z-scores in the case of linear regression), the number of top-100 8-mers in common, and the Root Mean Square Error (RMSE) between the predicted and the measured profiles (E-scores for all methods except linear regression which used Z-scores). The 1<sup>st</sup> and 2<sup>nd</sup> place in each scoring category (rows) are indicated by dark and light green, respectively. In the table below, “A vs. A” indicates that the evaluations were done between the actual real and predicted homeodomain (for example, comparing the predicted Hoxa13 profile to the real Hoxa13 profile). “All vs. all” indicates all possible pairings among all homeodomains in the data set, i.e. the aggregate of all predicted A vs. real B homeodomains (for example, comparing the predicted Pdx1 8-mer profile to the actual Hoxa13 8-mer profile). This statistic is used as a negative control to ensure that the models are not simply learning an average 8-mer profile over all homeodomains, which would result in (low) positive values for all measures because on average there is a positive correlation between randomly-selected experiments. We also considered the difference between the “A vs. A” statistics and the “All vs. all” statistics as a measure of specificity (“Difference between A vs. all medians” in the table). Wins were tallied as 1 for first-place (including ties) and 0.5 for second place (allowing ties). We note that the high 15AA-selected NN win tally is robust to most variants of our scoring scheme that include the Top 100 criterion.

**Supplementary Table 3.** Scoring metrics for predicting 8-mer profiles.

Approach		NN	Residue-selected NN	Residue-selected NN	Residue-selected NN	Residue-selected NN	Residue-selected NN	Residue-selected NN	Residue-selected NN	Residue-selected NN	Residue-selected NN	Partial Least Squares
Residues		All	Set 1	Set 2	Sets 2,3	Sets 2,3,4	Set 1	Set 2	Sets 2,3	Sets 2,3,4	NA	
Similarity metric		match = 1, mismatch = 0	match = 1, mismatch = 0	match = 1, mismatch = 0	match = 1, mismatch = 0	match = 1, mismatch = 0	PAM250	PAM250	PAM250	PAM250	NA	
Spearman predicted A vs real A	median	0.83	0.88	0.86	0.86	0.84	0.88	0.86	0.87	0.85	0.85	
	mean	0.8	0.85	0.83	0.84	0.82	0.85	0.83	0.84	0.82	0.83	
Spearman all predicted vs all real (control)	median	0.65	0.72	0.74	0.69	0.68	0.72	0.74	0.7	0.68	0.74	
	mean	0.63	0.69	0.7	0.66	0.65	0.69	0.7	0.66	0.65	0.7	
Spearman difference A vs. all medians		69	65	36	72	72	67	40	72	69	52	
Top100-overlap predicted A vs real A	median	81	77	69	82	82	76	70	82	82	74	
	mean	72.9	69.1	62.4	75.2	75	68	62.3	74.5	74.7	67.2	
Top100-overlap all predicted vs all real (control)	median	12	12	33	10	10	9	30	10	13	22	
	mean	26.7	28.4	31.6	26.5	26.3	27.6	31.3	26.7	27.1	28.5	
Top100-overlap difference A vs. all median		69	65	36	72	72	67	40	72	69	52	
RMSE predicted A vs real A	median	0.65	0.57	0.63	0.56	0.62	0.57	0.63	0.57	0.61	0.62	
	mean	0.76	0.63	0.68	0.63	0.66	0.63	0.68	0.64	0.67	0.66	
	range	[0.34-1.69]	[0.33-1.55]	[0.37-1.54]	[0.30-1.35]	[0.31-1.38]	[0.33-1.55]	[0.38-1.47]	[0.30-1.38]	[0.31-1.42]	[0.35-1.50]	
RMSE all predicted vs all real (control)	median	1.17	1.02	0.97	1.07	1.09	1.01	0.96	1.06	1.08	0.96	
	mean	1.22	1.06	1.03	1.12	1.14	1.05	1.01	1.11	1.13	1.03	
	range	[0.0-3.19]	[0.0-2.82]	[0.0-2.88]	[0.0-3.19]	[0.0-3.19]	[0.0-2.82]	[0.0-2.87]	[0.0-3.19]	[0.0-3.19]	[0.31-2.78]	
RMSE difference A vs. all median		-0.52	-0.45	-0.34	-0.51	-0.47	-0.44	-0.33	-0.49	-0.47	-0.34	
No. of proteins with an overlap (predicted A vs real A) < 50		17	31	41	14	15	31	41	15	16	28	
Wins		5	4	0	9	7	5	0	7	3.5	0	

Residues considered were as follows:

- 1) 47, 50, 54 (traditional specificity residues)
- 2) 3, 5, 47, 50, 51 (major or minor groove contacts from Engrailed structure (10))
- 3) 6, 25, 31, 44, 46, 48, 53, 54, 55, 57 (phosphate backbone contacts from Engrailed structure (10))
- 4) 7, 8, 28, 43, 52 (positions that contact DNA in other homeodomain structures)

## **8. Consistency between homeodomain groups derived from PBM data and homeodomain amino acid sequences.**

Our initial grouping of homeodomains on the basis of 8-mer profiles, in which we first compared the top 100 8-mers and then investigated 19 groups for systematic differences (i.e. TF-specific motifs, or clear differences in the primary motif) among all binding 8-mers within the group, resulted in definition of 71 different groups (all group IDs are found in the supplementary document “Homeodomain subclass assignments”) among which 31 have only one member. Our initial grouping on the basis of identity among the 15 selected amino acid positions (allowing no mismatches) resulted in 74 different groups, among which 40 have only one member.

The correspondence between these initial groups was as follows:

- Of the 71 different groups defined on the basis of 8-mer preferences, 59 of them consisted entirely of proteins that are identical among all 15 amino acids, i.e. they are in the same initial amino acid-based group.
- Of the 74 different initial amino acid groups, i.e. sets of proteins with the same residues at all 15 DNA-contacting residues, 50 of them consisted entirely of proteins that are in the same category as defined by the 8-mer profile.
- There are 42 groups that are completely consistent between the two categorizations. 27 of these groups consist of a single protein.

We reasoned that disagreement between the two categorizations might be due to the divisions between categories being more stringent in one classification scheme relative to the other. This would be unavoidable (a) if more than one configuration of the 15 amino acids could result in the same 8-mer binding profile, or (b) if additional amino acids besides the 15 selected influence the 8-mer binding profile.

We therefore asked whether simply merging groups in one categorization or the other could result in a more uniform set of categories, by grouping homeodomains that have either 8-mer profiles that are more highly correlated with the group they are merging into than they are to any other homeodomain in the data set, or by grouping those with the most closely related (but not identical) set of 15 amino acids (the groupings above allowed no mismatches among the 15 amino acids, and therefore reflect only identity, and not similarity). Indeed, by allowing the following set of merges, and one reassignment (Hoxc12, which is an unusual case) we obtained 65 groups that are entirely consistent between the fifteen amino acid residues and the 8-mer binding profiles.

1. Merge Pou1f1 with Pou3f1, Pou3f2, and Pou3f4. Pou1f1 is one amino acid different from the others among the 15AA but has an indistinguishable DNA-binding activity.

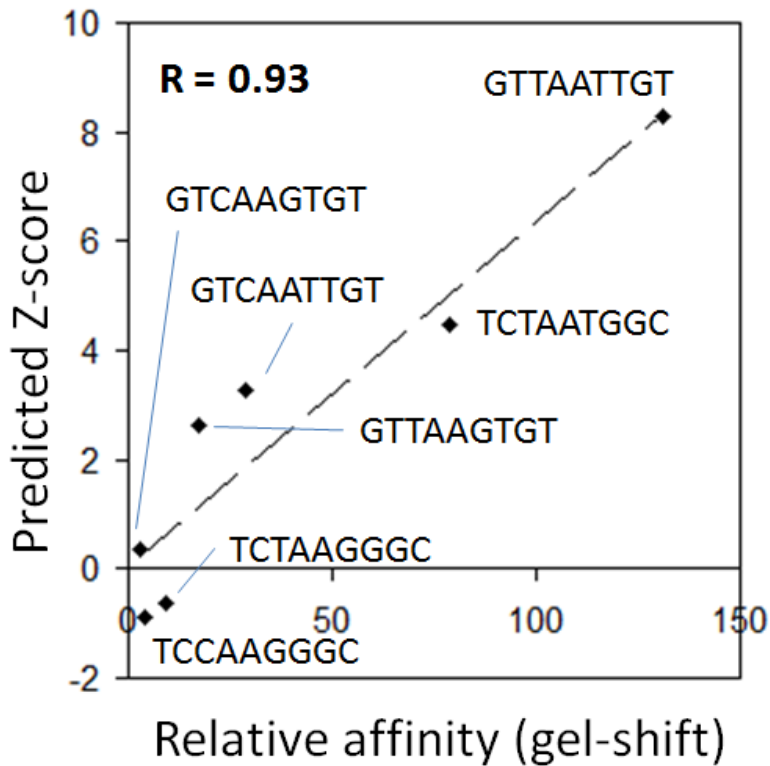
2. Group Hoxc12 with the Hox9, 10, and 11 group, instead of with Hoxd12. Hoxc12 is more similar to Hoxd12 over the entire homeodomain, but it is identical to the Hox9,10,11 group on the basis of the 15 amino acids, whereas it has two amino acid differences among the 15 DNA-contacting residues. The data figure in Section 5 suggests that its 8-mer binding profile has characteristics of the Hox9,10,11 group, as does its TF-specific motif.
3. Group Hoxb13 with the rest of the Hox13 group. It contains an amino acid mismatch among the 15 selected residues, but the 8-mer profile is indistinguishable by our criteria.
4. Group Esx1, Isx, and Otp with Alx4, Arx, Cart1, Phox2a, Phox2b, Prrx1, Prrx2, Rax, Shox2, and Uncx4.1. The 8-mer profile is indistinguishable by our criteria and these three homeodomains have only one amino acid mismatch from the rest of this group among the 15 selected residues.
5. Group Prop1 with Alx4, Arx, Cart1, Phox2a, Phox2b, Prrx1, Prrx2, Rax, Shox2, and Uncx4.1 group. Its 8-mer profile appears to be distinct, indicating that additional amino acids influence binding specificity, but it is indistinguishable by the Top 100 overlap and has the same 15 DNA-contacting residues as the remainder of this group.
6. Group Obox6 with Obox1, Obox2, Obox3, and Obox5. It has three amino acid differences among the 15 DNA-contacting residues, but the 8-mer profile is indistinguishable by our criteria.
7. Merge all members of the Six class (Six1, Six2, Six3, Six4, and Six6). The 8-mer profiles are indistinguishable by our criteria, and they all share 12 common amino acids among the 15 DNA-contacting residues.
8. Merge all members of the Irx class (Irx2, Irx3, Irx4, Irx5, Irx6). The 8-mer profiles are indistinguishable, and they all share 12 common amino acids among the 15 DNA-contacting residues.
9. Merge Pknox1 and Pknox2. The DNA-binding profiles were called as distinct because they correlate more highly with other experiments than they do with each other, but do they appear in the figure in Section 5 to have similar activities, and they have no amino acid differences among the 15 DNA-contacting residues.
10. Merge Hlxb9, and all members of the Hox3, 4, and 5 families. These clearly form three distinct groups on the basis of the 8-mer profile, but there is substantial overlap among the top 100 8-mers, and they are all identical at the 15 selected amino acid residues. This is a group in which the 15 amino acids clearly do not completely govern the entire DNA binding specificity.
11. Merge Hoxa2 and Ipfl. These have only a single amino acid difference among the 15 DNA-contacting residues (Valine vs. Isoleucine at P47) and their 8-mer binding profiles are indistinguishable.
12. Group Pou3f3 with Pou2f1, Pou2f2, and Pou2f3. There are no differences among these four at the 15 DNA-contacting residues.
13. Group all members of the Hox9, 10, and 11 families. These clearly form three distinct groups on the basis of the 8-mer profile, but there is substantial overlap among the top 100, and they are all identical at the 15 selected amino acid residues. This is a group in which the 15 amino acids clearly do not completely govern the entire DNA binding specificity.

14. Group Dbx1 and Dbx2. These appear distinct on the basis of their 8-mer binding profiles, but there is substantial overlap among the Top-100 8-mers (62) and they are identical at the 15 DNA-contacting residues.
15. Group En1 and En2 together with Gbx1 and Gbx2. These two pairs appear distinct on the basis of their 8-mer binding profiles, but there is substantial overlap among the Top100 (they are identical by the top 100 criterion) and they are identical at the 15 DNA-contacting residues.
16. Group Lmx1a and Lmx1b together with Lhx3 and Lhx4. These two pairs appear distinct on the basis of their 8-mer binding profiles, but they are identical at the 15 DNA-contacting residues. This is a group in which the 15 amino acids clearly do not completely govern the entire DNA binding specificity.
17. Group all members of the Nkx2 class (including Ttf1). There are clearly two sub-classes on the basis of the 8-mer binding profiles, but they are all identical at the 15 DNA-contacting residues. This is a group in which the 15 amino acids clearly do not completely govern the entire DNA binding specificity.

The results of these operations, including tentative names for each of these subclasses, are included in the document entitled “Homeodomain subclass assignments”.

## 9. Agreement between Predicted Z-score vs. measured relative affinity for the *Drosophila* Engrailed homeodomain

**Supplementary Figure 8.** The scatter plot shows the predicted Z-score on the X-axis, and the relative affinity from Damante *et al.* (17) on the Y-axis. Z-scores (which we take to be our most accurate inference of binding affinity) were inferred following the same protocol as for E-scores (nearest-neighbor over the 15 DNA-contacting residues, with ties averaged) with the Z-score for each 9mer taken as the average of the Z-score for the two overlapping 8-mer components.



## 10. Comparison between PBM data and ChIP-chip or ChIP-seq data

We analyzed six ChIP-chip or ChIP-seq datasets for homeodomain proteins available in the literature (see Figure 7 in the main paper and the Table and Figure on the next two pages). Bound sequences were scanned to determine enrichment of highly-bound PBM 8-mers. To standardize the length of the bound sequences across all datasets, we took either the ChIP peak (if known) or the center of the identified bound sequence and add 1 kb on each side. Enrichment ratio was determined with respect to 2kb-length random genomic regions taken from the genome (same version as the one used for the chip experiments) of the corresponding organism. For each dataset, the number of random genomic regions sampled was 10 times the number of bound sequences.

In the plots on the following page, random and bound sequences were scanned with the predicted PBM E-score 8-mer profile (since none of these data are from mouse) using a 500 bp length moving window with a 50 bp tiling distance. For each dataset, the E-score with the highest enrichment ratio in the central portion were selected by sampling 20 intervals starting from  $E = 0.43$  to the maximum value. Enrichment ratio was calculated as the ratio of number of 8-mers above the cutoff value found in the bound sequences to one tenth of the number of 8mers above the cutoff value found in the random genomic sequences.

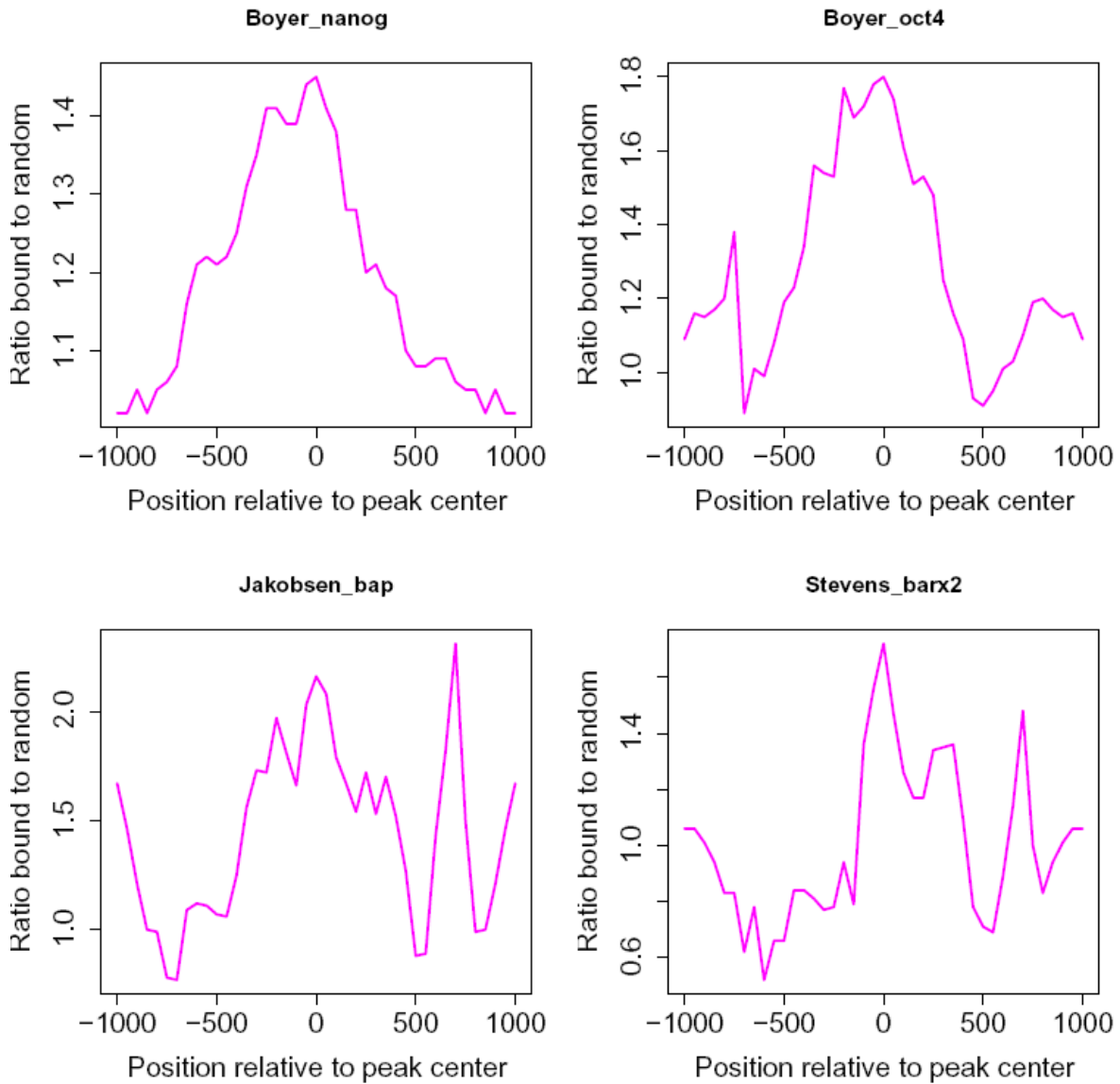


**Supplementary Table 4.** Chip-chip data sets analysed.

<i>Homeodomain</i>	<i>ChIP data reference</i>	<i>Organism ChIP experiments</i>	<i># of bound sequences</i>	<i>Is ChIP peak identified in original reference?</i>	<i>8-mer profile</i>	<i># mismatches with NNs</i>	<i>E-score threshold</i>
Tcf1/Hnf1a	Odom et al, MSB 2006 (PMID: 16738562)	Human	427 *	No	Predicted	1	0.456
Caudal/Cdx2	Li et al, PloS Biol. 2008 (PMID: 18271625)	Fly	1331	Yes	Predicted	0	0.493
Pou5f1/Oct4	Boyer et al, Cell 2005 (PMID: 16153702)	Human	603	No	Predicted	1	0.487
Nanog	Boyer et al, Cell 2005 (PMID: 16153702)	Human	1554	No	Predicted	1	0.477
Bagpipe/Bapx1	Jakobsen et al, Genes Dev. 2007 (PMID: 17908931)	Fly	78	No	Predicted	0	0.491
Barx2	Stevens et al, J Biol. Chem 2004 (PMID: 14744868)	Human	42	No	Measured	0	0.489

\* Bound sequences not provided by Odom *et al.* Bound sequences were determined by analyzing their raw microarray data with an in-house software similar to the one described in Boyer *et al.*

**Supplementary Figure 9.** For the data sets in Table 4, random and bound sequences were scanned with the predicted PBM E-score 8-mer profile using a 500 bp length moving window with a 50 bp tiling distance.



## 11. References

1. M. Z. Li, S. J. Elledge, *Nat Genet* **37**, 311 (Mar, 2005).
2. W. Zhang *et al.*, *J Biol* **3**, 21 (2004).
3. M. F. Berger *et al.*, *Nat Biotechnol* **24**, 1429 (Nov, 2006).
4. A. M. Dudley, J. Aach, M. A. Steffen, G. M. Church, *Proc Natl Acad Sci U S A* **99**, 7554 (May 28, 2002).
5. P. J. Huber, *Robust Statistics* (John Wiley, New York, 1981), pp.
6. A. Subramanian *et al.*, *Proc Natl Acad Sci U S A* **102**, 15545 (Oct 25, 2005).
7. J. A. Granek, N. D. Clarke, *Genome Biol* **6**, R87 (2005).
8. J. C. Byrne *et al.*, *Nucleic Acids Res* **36**, D102 (Jan, 2008).
9. V. Matys *et al.*, *Nucleic Acids Res* **31**, 374 (Jan 1, 2003).
10. C. R. Kissinger, B. S. Liu, E. Martin-Blanco, T. B. Kornberg, C. O. Pabo, *Cell* **63**, 579 (Nov 2, 1990).
11. E. Fraenkel, M. A. Rould, K. A. Chambers, C. O. Pabo, *J Mol Biol* **284**, 351 (Nov 27, 1998).
12. I. T. Jolliffe, *Principal Component Analysis 2nd Edition*. P. Bickel *et al.*, Eds., Springer Series in Statistics (Springer-Verlag, New York NY, 2002), pp. 1-487.
13. S. B. Franklin, D. J. Gibson, P. A. Robertson, J. T. Pohlmann, J. S. Fralish, *Journal of Vegetation Science* **6**, 99 (1995).
14. P. Geladi, B. R. Kowalski, *Analytica Chimica Acta* **185**, 1 (1986).
15. M. Stone, *Journal of the Royal Statistical Society. Series B (Methodological)* **36**, 111 (1974).
16. S. Wold, *Technometrics* **20**, 397 (1978).
17. G. Damante *et al.*, *Embo J* **15**, 4992 (Sep 16, 1996).