Materials and Methods for "Diversity and Complexity in DNA Recognition by Transcription Factors"

This document contains detailed description of Materials and Methods. All data files, tables, and figures are posted separately at:

http://thebrain.bwh.harvard.edu/suppl105/

username: rver105

password: Ee4ki9Au

The PBM data itself can be accessed at:

http://the_brain.bwh.harvard.edu/pbms/webworks/index.php

username: rver105

password: kee5naiR

Table of contents:

A. Cloning transcription factors into pMAGIC1

Full-length transcription open reading frames or their DNA binding domains, consisting of the Pfam-defined DNA-binding domain (DBD) plus 15 amino acids of N-term and of C-term 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 provided in **Table S2**. 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**):

MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYI
DGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLK
VDFLSKLPEMLKMFEDRLCHKTYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKL
VCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSDLVPRPCEL**KLDVHML
VPRGSLEVLFQGPEGDATMGHMVHRPWIQ** — DBD region - **AWPQGGRTRIVSAHNSENLYFQGDLRGSITN** GSGC*

B. Protein production, Western blots, and quantification

We produced proteins by two methods that yielded essentially identical results: Expression and purification from *E. coli*, and expression by *in vitro* transcription/translation.

Expression and purification in *E. coli*. We transformed TF-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_{600} ~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 (50 mM Tris pH 8, 150 mM NaCl, 2 mM 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 4ºC for 15 minutes at 4000 rpm and soluble fraction was collected. GST resin slurry (Amersham) was added to the fraction and binding proceeded at 4ºC 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 pH7.5, 10 mM reduced glutathione, Roche protease inhibition and 2 mM DTT) at 4ºC for 1 hr. Concentration of GST-tagged DBD was calculated for each protein relative to a dilution series of GST standards on Coomassie-stained SDS-PAGE gels.

In vitro transcription/translation. For *in vitro* translation reactions, the manufacturer's protocol (Ambion ActivePro Kit) was followed. Molar concentrations of all *in vitro* translated proteins were determined by Western blot utilizing 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 precast 4-12% Bis-Tris Criterion 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 visualized using the SuperSignal West Femto Maximum Sensitivity Substrate kit (Pierce) according to the manufacturer's protocols. Primary antibody was added at a final concentration of 20 ng/ml (anti-GST produced in rabbit; Sigma), and secondary antibody was added at a final concentration of 5 ng/ml (anti-GST produced in rabbit; Sigma). Film was

scanned and concentrations were determined using Quantity One software version 4.5.0 (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.

C. Protein Binding Microarrays

Design of universal 'all 10-mer' universal protein binding microarrays (PBMs):

The design of 'all 10-mer' universal protein binding microarrays (PBMs) using a de Bruijn sequence of order 10 has been described previously (*3*) and is described in detail in conference proceedings posted at http://thebrain.bwh.harvard.edu/RECOMB2007.pdf. For this study, we created two separate designs for replicate experiments, which we optimized to achieve maximal 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 de Bruijn sequences of order 10 for our universal PBMs, we used a linear-feedback shift register corresponding to the primitive polynomial:

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

The two de Bruijn sequences for our two PBMs differ by cyclic permutations of A, C, G, and T. We empirically selected these particular de Bruijn sequences because they cover all contiguous 10-mers and all gapped 10-mers spanning 11 total positions. Furthermore, they exhibit optimal coverage of contiguous and gapped 8-mers. Any 8-mer is guaranteed to occur 32 times in a deBruijn sequence of order 10 (16 times for palindromes). Our de Bruijn sequences exhibit 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 these de Bruijn sequences *in silico*, we partitioned them into subsequences of length 36 nucleotides (nt) and overlapping by 11 nt, resulting in 41,944 36-mers for each microarray. 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'-gtctgtgttccgttgtccgtgctg-3') complementary to our primer for double-stranding (5'-cagcacggacaacggaacacagac-3') in order to create 60mer sequences that would become the probes on our custom-designed microarrays. 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 10mers 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.

Protein Binding Microarray Experiments:

Protein binding microarray (PBM) experiments were performed essentially as described previously (*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 μM HPLC-purified common primer (Integrated DNA Technologies), 40 μM dATP, dCTP, dGTP, and dTTP (GE Healthcare), 1.6 μM Cy3 dUTP (GE Healthcare), 32 Units Thermo Sequenase™ DNA Polymerase (USB), and 90 μl 10x reaction buffer (260 mM Tris-HCl, pH 9.5, 65 mM $MgCl_2$) in a total volume of 900 μl. 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, 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.

Purified TFs were diluted to 100 nM (unless otherwise specified) in a 175-μl protein binding reaction containing PBS / 2% (wt/vol) milk / 51.3 ng/μl salmon testes DNA (Sigma) / 0.2 μg/μl 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 μg/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. Every protein in this study was assayed in duplicate, once on each of our two separate microarray designs described above.

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). At least one of the two duplicate experiments for each protein was performed on a fresh slide.

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 (MicroaArray 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.
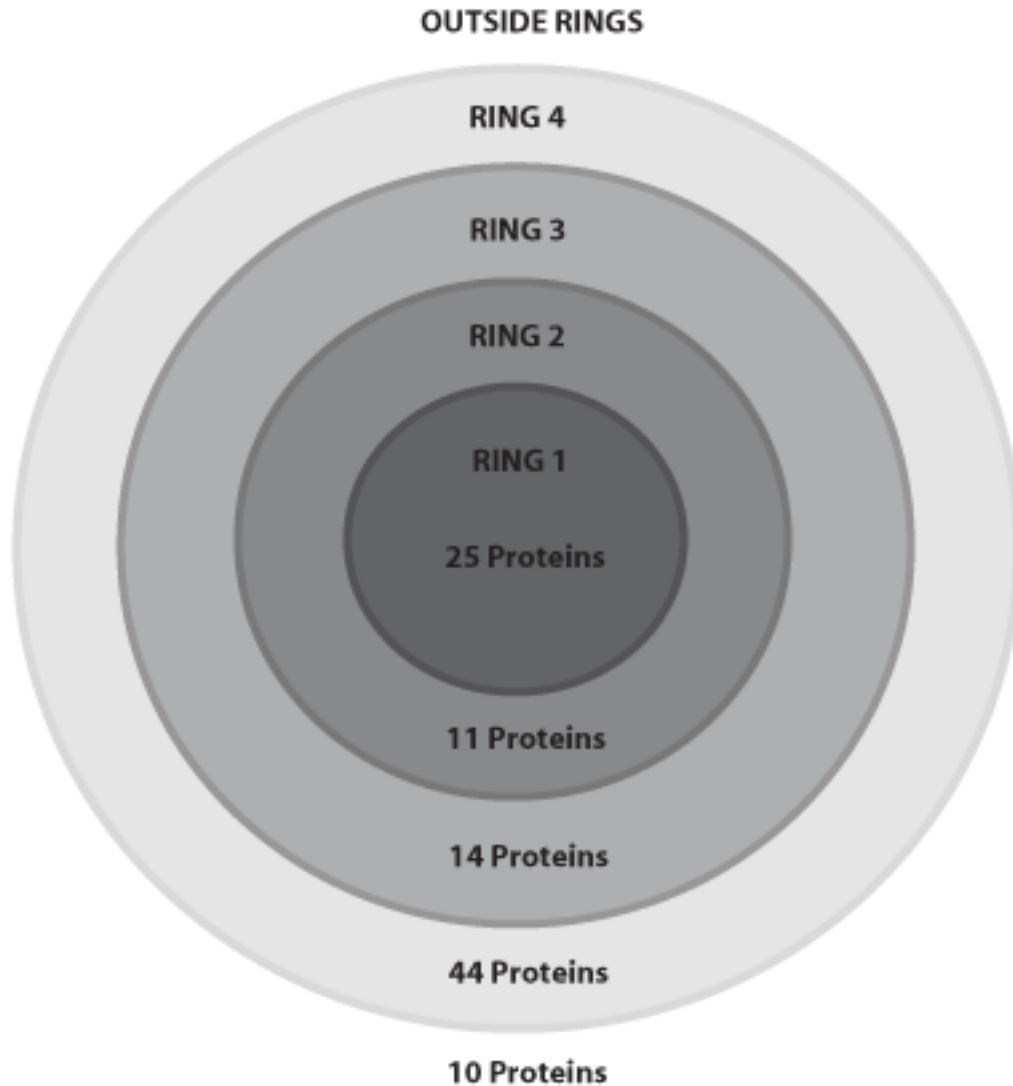
D. Comparisons to previous binding specificity data in TRANSFAC, JASPAR, and the literature.

Identification of previously annotated binding data from TRANSFAC, JASPAR, and the literature.

In order to determine the quantity and quality of previously known binding site information, we performed a comprehensive search in the TRANSFAC (*5*) and JASPAR (*6*) databases for matches to the 104 TFs used in this study. Because TRANSFAC hosts data derived from a variety of experimental procedures and from a range of model organisms, we created four categories, or "rings" to indicate (a) whether a quality matrix or single binding site was available for given TF for which we generated PBM data, and (b) whether such a matrix or sequence was derived from the specific mouse TF that we examined, a paralog of the TF, or an ortholog of the TF. The TRANSFAC database assigns a BIOBASE quality score (1-6, 1 being the most specific) to its annotated matrices and binding. In order to restrict our analysis to high quality datasets, we filtered the TRANSFAC matrices by their BIOBASE quality score. We included data of quality one and two, which were derived using purified or recombinant protein, and of quality six, which predated the quality control system. We excluded categories three, four, and five, which were attained using crude nuclear extracts. All of the matrices in the JASPAR database were included. For each protein queried, all synonyms were also searched based on the Mouse Genome Informatics (MGI) database to ensure comprehensiveness.

Binding site data from these two databases fell into one of the following 4 "rings" (see figure below): **Ring 1:** *All proteins which have a known matrix to the exact mouse protein.* We found that 25/104 (~24%) of the proteins in this study have previously annotated matrices in mouse. **Ring 2:** *All proteins which have a known matrix to a paralog in mouse.* We found that 11 additional proteins used in this study have matrices available for paralogs in mouse. If the protein being queried had no matrix for its paralogs in either database, we performed a BLASTP search against the mouse genome to identify all proteins in mouse that had $\geq$ 66% amino acid

identity, which we define as a paralog. **Ring 3:** *All proteins which have a known matrix to an orthologous protein.* We found that 14 additional proteins in this study have matrices available for orthologous factors. If the protein queried had no matrix available for known orthologs, we performed a BLASTP search against all other organisms to identify proteins with ≥66% amino acid similarity for which a matrix was available. **Ring 4:** *All proteins which have sequence data for the mouse protein, for a paralog, or for an ortholog, or has a binding sequence annotated in the literature.* We found that 44 additional proteins used in this study have known binding sequences in the TRANSFAC database. If the protein queried had no matrix for the mouse protein, paralog, or ortholog, we searched the database to see if there was a single binding sequence known for the protein or any other protein which had ≥66% amino acid similarity (paralogs and orthologs). At this point, in order to ensure that no binding site data were missed, we searched the literature for all the proteins which did not fall into rings 1-4. We found sequence data that were not annotated in either database for the mouse protein, paralog or ortholog for 9 proteins, which were then partitioned into the appropriate rings based on the data available. **Note:** If a TF is present in one ring, it is not present in any other ring. **Outside the rings**: *All proteins which did not fall into any of the 4 rings, and thus have no binding information known according to our criteria.* We found that 10/104 (~10%) of the proteins in this study had no entry for the mouse protein, paralog, or ortholog, in either database according to our criteria, and had no binding data reported in the literature.

OUTSIDE RINGS

RING 4

RING 3

RING 2

RING 1

25 Proteins

11 Proteins

14 Proteins

44 Proteins

10 Proteins

Survey of known binding information for 104 TFs. In order to determine the extent of binding information known for the 104 TFs, a systematic search was performed in the TRANSFAC and JASPAR databases as well as the literature, and broken down into 4 "rings" (see above for criteria). **Ring 1** includes each protein for which there was a matrix available to the exact mouse protein. **Ring 2** includes each protein for which there was a matrix available to a paralogous protein in mouse. **Ring 3** includes each protein for which there was a matrix available for an orthologous protein. **Ring 4** includes each protein for which binding *sequence* data was available for the mouse protein, paralog, or ortholog. **Outside** these rings represents proteins for which there is no binding data according to our criteria (see above). A protein was assigned to only the most specific ring possible, such that no protein is included in multiple rings.

Comparison of 104 TFs' PBM *k*-mers versus prior binding data

We calculated area under ROC curve (AUC) values to assess the similarity between each of the 104 TFs and each of the previously annotated matrices in TRANSFAC, JASPAR, and the literature. For these comparisons, we obtained all 834 matrices in TRANSFAC Professional 11.3, all 138 matrices in JASPAR_CORE version 3.0, and 3 matrices from literature, where a single matrix was taken from each of the following papers: Osaki *et al.*, 1999 (*7*), Pengue *et al.*, 1993 (*8*), and Wolfe *et al.*, 2005 (*9*). We derived PWMs from Pengue *et al.* and Wolfe *et al*., which did not publish a PWM or position frequency matrix, by running the available binding sequence data through AlignACE 3.0 (*10, 11*) (http://atlas.med.harvard.edu/cgi-bin/alignace.pl) and then passing the AlignACE output to enoLOGOS (*12*) (http://biodev.hgen.pitt.edu/cgi-bin/enologos/enologos.cgi) for nucleotide position frequency determination.

In brief, the following procedure was used to compare matrices. For each of our 104 TFs, we generated an indicator matrix across all contiguous 8-mers using a PBM E-score threshold ≥ 0.37 (corresponding to a PBM Q-value threshold of ~0.001). For a given TF, every 8-mer above the threshold was labeled as a positive hit, and all other 8-mers were labeled as negative hits. We likewise generated GOMER (*13*) scores across all 8-mers for each previously annotated matrix, which allowed us to rank the 8-mer preferences of these matrices. In order to assess the relationship between a given TF indicator matrix and a set of matrix GOMER scores, used the Lever software package (*14*) to obtain AUC statistics and Q-values via multiple hypothesis testing for every paired combination of PBM TF with previously annotated matrix.

E. EMSAs

EMSA probe sequences:

The sequences of the oligonucleotides that we used for EMSA probes were as follows:

Novel Motif Validation

Zfp740 positive probe

5' - NNNNNNNNNNNNNNNNNNNNCCCCCCCCNNNNNNNNNNNNNGAAAGGATGGGTGCGACGCG - 3'

Zfp161 positive probe

5' - NNNNNNNNNNNNNNNNNNNGCGCGCGCNNNNNNNNNNNNNGAAAGGATGGGTGCGACGCG - 3'

GGCC repeat probe

5' - NNNNNNNNNNNNNNNNNNNGGCCGGCCNNNNNNNNNNNNNGAAAGGATGGGTGCGACGCG - 3'

Osr2 positive probe

5' - NNNNNNNNNNNNNNNNNNTACAGTAGCNNNNNNNNNNNNNGAAAGGATGGGTGCGACGCG - 3'

Sp100 positive probe

5' - NNNNNNNNNNNNNNNNNNTTCTCGCGAAAANNNNNNNNNNNNGAAAGGATGGGTGCGACGCG - 3'

Secondary Motif Validation

Hnf4a primary probe

 5' - NNNNNNNNNNNNNNNAGGGGTCACCNNNNNNNNNNNNNNNGAAAGGATGGGTGCGACGCG - 3'

Hnf4a secondary probe

5' - NNNNNNNNNNNNNNNAGGGGTCCACCNNNNNNNNNNNNNNNGAAAGGATGGGTGCGACGCG - 3'

Hnf4a hybrid probe

5' - NNNNNNNNNNNNNNNAGGGGTCCCACCNNNNNNNNNNNNNNNGAAAGGATGGGTGCGACGCG - 3'

Nkx1.3 primary probe

5' - NNNNNNNNNNNNNNNNNNNCCACTTAANNNNNNNNNNNNNNNGAAAGGATGGGTGCGACGCG - 3'

Nkx1.3 secondary probe

5' - NNNNNNNNNNNNNNNNNNAAGTACTTNNNNNNNNNNNNNNNGAAAGGATGGGTGCGACGCG - 3'

Mybl1 and Myb primary probe

5' - NNNNNNNNNNNNNNNNNNAACCGTTANNNNNNNNNNNNNNNGAAAGGATGGGTGCGACGCG - 3'

Mybl1 and Myb secondary probe

5' - NNNNNNNNNNNNNNNNNCCAACTGCCNNNNNNNNNNNNNNNGAAAGGATGGGTGCGACGCG - 3'

Foxj3 primary probe

5' - NNNNNNNNNNNNNNNNNNGTAAACAANNNNNNNNNNNNNNNGAAAGGATGGGTGCGACGCG - 3'

Foxj3 secondary probe

5' - NNNNNNNNNNNNNNNNNNNCAAAACAANNNNNNNNNNNNNNNGAAAGGATGGGTGCGACGCG - 3'

Rfxdc2 primary probe

5' - NNNNNNNNNNNNNNNNNNCCTAGCAACGNNNNNNNNNNNNNGAAAGGATGGGTGCGACGCG - 3'

Rfxdc2 secondary probe

5' - NNNNNNNNNNNNNNNNNNCCTGGATACGNNNNNNNNNNNNNGAAAGGATGGGTGCGACGCG - 3'

Universal Biotinylated Primer

5' – Biotin-CGCGTCGCACCCATCCTTTC - 3'

Probe Design:

EMSA probes were ordered as single-stranded 60 bp oligonucleotides from Integrated DNA Technologies (IDT) containing a common 20 bp sequence at the 3' end with which to anneal a universal biotinylated primer. Probes were designed so that the binding sequence of interest was imbedded in random flanking sequence (corresponding to "N" in sequence, or roughly 25% of each of dATP, dCTP, dGTP, dTTP incorporated at that position for the pool of oligos).

Primer Extension

Lyophilized oligonucleotides from IDT were resuspended in TE pH 8.0 to a working stock of 100 µM. Extensions reactions were performed in 1x Thermopol Buffer (NEB; 20 mM Tris-HCl, 10 mM $(NH_4)_2SO_4$, 10 mM KCl, 2 mM $MgSO_4$, 0.1 % Triton X-100) using final concentrations of 0.8 mM dNTPs (Amersham), 4 µM primer, and 4 µM oligo template in a 25 µL reaction. Primer extension was performed in a thermocycler according the following protocol:

1) 95° C for 3 minutes
2) Ramp down to 60° C (0.1° C per second)
3) Hold at 60° C

At step 3, 8 units of BST polymerase (NEB) in 1x Thermopol buffer was added to each reaction as a "hot start". After polymerase was added the reaction was allowed to continue to step 4:

4) Incubate at 60° C for 90 minutes
5) Hold at 3° C

Oligos were filtered using MinElute PCR Purification Kit (Qiagen) according to the manufacturer's instructions, and diluted to a working concentration of 10 nM after concentrations were determined using a spectrophotometer.

Electrophoretic Mobility Shift Assay (EMSA) experiments

EMSAs were performed using the LightShift© Chemiluminescent EMSA Kit (Pierce), essentially according to the manufacturer's protocols. Each 20 µL binding reaction contained 1x Binding Buffer (10 mM Tris, 50 mM KCl, 1 mM DTT), 2.5% glycerol, 0.5 µg Salmon Testes DNA (Sigma), 10 mM KCl, 4 µg BSA (NEB), 0.05 % NP-40, and 50 µM zinc acetate (Sigma). 0.5 nM DNA probe was used, and 0.2 µM protein. The binding reactions were allowed to incubate at room temperature for 1 hour. A precast 6% polyacrylamide DNA retardation gel (Invitrogen) was pre-run for 30 minutes at 100 V, and then 5 µL of 5x loading buffer was added to the binding reaction, and subsequently 20 µL of the reaction was run on the gel at 100 V for 45 minutes.   The gel was then transferred to a charged modified 0.45 µm nylon membrane (Sigma) for 1.5 hours at 100 V, and subsequently UV-crosslinked to the membrane at 120 µJ/cm2. The membrane was then treated with developing buffers (Lightshift Blocking Buffer with stabilized Streptavidin-Horse Radish Peroxidase conjugate, Wash Buffer, Substrate Equilibration Buffer, Luminol/Enhancer Solution and Peroxide Solution) according to manufacturer's protocol, then exposed to film and developed.

F. PBM *k*-mer scoring, combining replicate array data, and motif construction:

PBM *k*-mer scores

Every non-palindromic 8-mer occurs on at least 32 spots in each chamber of our universal PBM. Because of this redundancy, 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 normalized signal intensity, (2) Z-score, (3) enrichment score (E-score), and (4) False Discovery Rate Q-value.

Median normalized signal intensity refers to the median normalized signal intensity for the set of probes containing a match to each 8-mer (usually ~32 probes, but some might be flagged occasionally because of dust flecks, etc., and therefore removed from further consideration). We have shown previously that higher PBM median signal intensity corresponds to stronger protein-DNA 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 (*15*). Both the median signal 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 intensities and nonspecific 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 enrichment score has been described previously in detail (*3*). Briefly, for each 8-mer (ungapped or gapped) we consider the collection of all probes containing a match as the "foreground" feature set and the remaining probes as the "background" feature set. We compare the ranks of

17

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 each k-mer E-score by comparing it to the null distribution of E-scores calculated from the distribution of E-scores from negative control PBMs performed using GST in binding buffer (duplicate negative control PBMs on array version 1 plus duplicate negative control PBMs on array version 2). Q-values for the *k*-mer data from array version 1 were calculated using the GST negative control PBM data from array v1, Q-values for the *k*-mer data from array version 2 were calculated using the GST negative control PBM data from array v2, and Q-values for the *k*-mer data from the combined array data (from version 1 and version 2) were calculated using the GST negative control PBM data from the combined array data. Negative control PBM data using GST in binding buffer versus a mock IVT reaction gave indistinguishable distributions by a Kolmogorov-Smirnov (KS) test (data not shown). 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 (relative to the slide surface) on the probe or for which the 8-mer overlaps the 24-nt constant primer region.

In essence, the E-score and Z-score capture essentially the same information, but the E-score representation is compressed as E-scores approach 0.5.

Combining data from replicate arrays
As described above, we performed duplicate experiments for every transcription factor using microarrays created with independent sequence designs. In order to combine the data from multiple experiments, we first computed E-scores and Z-scores for all 8-mers for each separate experiment. We then calculated the mean E-score for each 8-mer directly and calculated the mean Z-score for each 8-mer after first performing variance stabilizing normalization (*16*) on the Z-score measurements for the set of arrays.

Motif construction using Seed-and-Wobble

In addition to reporting scores for each individual 8-mer, we compactly represent these data as position weight matrices (PWMs) for each TF.

Our "Seed-and-Wobble" algorithm has been described previously (*3*). The algorithm works in two stages. In the first stage (the "Seed" stage), we identify the single 8-mer (ungapped 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 stage (the "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 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 a Boltzmann distribution (*17, 18*). 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 order to derive a single motif combined from separate experiments, we choose the 8-mer with the greatest average E-score as a seed, build a PWM on each separate array, and average the matrix elements, as described previously (*3*).

To derive "secondary" motifs (and in the case of Oct-1, a "tertiary" motif) using Seed-and-Wobble, we first score all microarray probes according to how well they match the primary motif. For this we use the GOMER scoring framework, which calculates binding probabilities over the entire length of the probe according to position weight matrices (*13*). We then re-rank all probe sequences by their ratios of observed-to-expected ranks, based on the scores assigned

by the primary motif. Consequently, the top of the re-ranked list is populated by probes with high signal intensity but without a strong match to the primary motif. We identify the 8-mer with the highest enrichment score in the new list and use Seed-and-Wobble to construct a secondary PWM in the new ranking. This process can be iterated multiple times to generate additional motifs.

To derive secondary motifs combined from separate experiments, we use the combined primary PWM to separately re-rank each set of probes, identify the 8-mer with the greatest average E-score, build a PWM on each separate array, and average the matrix elements as before.

"Trimming" Seed-and-Wobble motifs

Due to the extensive set of gapped patterns covered by our universal PBM designs, the PWMs generated by our Seed-and-Wobble algorithm can contain up to 17 columns. Since many of these positions often show very little preference for any nucleotide, we devised two methods for trimming these distal positions in order to more compactly represent the binding specificity. The first method is based on optimizing AUC statistics to minimize misclassification of k-mers over all possible trimmed motifs, and the second method utilizes an optimized information content (IC) threshold of the distal nucleotide positions of the motif.

In the first method, in order to calculate AUC statistics, we had to first define a foreground (class 1) and background (class 0) set of 8-mers. We defined class 1 and class 0 as all 8-mers with False Discovery Rate Q-values less than 0.005 and greater than 0.5, respectively. Separately, each PWM under consideration (see below) was used to score all 8-mers according to the GOMER framework (*13*). We tested every possible trimmed version of the original PWM, where trimming proceeded in single-nucleotide increments from the 5' and 3' ends (up to any position with an information content of 1). 8-mers were ranked according to their PWM scores, and we calculated AUC statistics according to their class 1 and class 0 assignments.

In the second method we sought to determine an optimal IC threshold for trimming that could be uniformly applied to all PWMs. For any single IC threshold, PWMs were trimmed from both the 5' and 3' ends until a position exceeding the threshold was reached. We inspected all 111 TFs in our dataset using a wide range of information content thresholds (0 to 1, increments of 0.05). We calculated AUC statistics for each trimmed motif as described above, and for each IC threshold, we determined the difference between the mean AUC for the 111 trimmed motifs and the mean AUC for the 111 untrimmed motifs. By this measure, we found that an IC threshold of 0.3 was optimal.

Primary PWMs trimmed by each method were used to re-rank the microarray probes in order to derive a secondary PWM according to the approach described above. We then trimmed these secondary PWMs using an IC threshold of 0.3.

Motif construction using RankMotif++

RankMotif++ is a motif finding algorithm that learns PWMs from PBM intensity data. These binding preferences are represented using probe preference pairs: pairs of probes where the first probe has significantly higher intensity than the second probe and, as such, is much more likely to be bound by the DBD than the second probe. RankMotif++ fits PWMs that are, to the greatest extent possible, consistent with these pairs by maximizing the average log likelihood of these pairs given a PWM assuming a particular model of probe binding.

Using RankMotif++ on single arrays

We used the RankMotif++ version 3.0 software (*19*) to fit PWMs to the PBM intensity data for individual arrays.

Extracting probe preference pairs

We used robust Z-scores derived from the PBM intensity data to identify the probe preference pairs. The calculation of the robust Z-scores and the subsequent identification of the probe preference pairs are based on the settings of four input parameters to the RankMotif++ software: the log-transform flag "-u", the positive probe threshold "-p", the confidence interval scale "-c", and the number of negative probes "-n".

To run RankMotif++ on the PBM intensity data from single arrays, we set the log-transform flag to "-u 1", indicating to the RankMotif++ software that it should log-transform the normalized intensities before converting them into robust Z-scores. After the log-transformation, the software sets the Z-score of the probe with the median log intensity to 0 and it sets the Z-scores of the other probes by first calculating the difference of the log intensity of the probe and the median log intensity and then dividing this difference by a robust estimate of the standard deviation of the log intensities (see (*19*) for details). We set the positive probe threshold to "-p 3", indicating that only probes with Z-score > 3 can appear as the first probe in a binding preference pair. We set the confidence interval scale to 3, using "-c 1.5", to indicate that only pairs of probes whose Z-scores difference is at least 2 x 1.5 = 3 can appear at probe pairs. For efficiency, we also restrict the number of probes with Z-score < 0 that can appear in preference pairs by setting the number of negative probes to 400, using "-n 400", indicating to the software that it should select 400 random probes with Z-scores < 0 to appear in preference pairs. The software then identifies all pairs of probes that satisfy these constraints, i.e. the first probe must have Z-score > 3, the Z-score difference must be at least 3, and if the second probe has a Z-score < 0 then it must be one of the 400 negative probes, and fits the PWM model to these probe pairs. Note that it is possible for both probes in the pair to have Z-scores > 0, or even >3.

Note that for one of the proteins, Sp100, the above parameter settings yielded no probe preference pairs for either of the two different De Bruijn sequence universal array designs. In order to run RankMotif++ for Sp100, we set "-p 2.0" and "-c 0.5".

Fitting PWMs to probe preference pairs

To guide RankMotif++ in fitting the PWM model, we set three further parameters: the range of PWM widths to try, the number of PWMs fit at each width, and the reverse complement flag "-r". The width range was set to 6-13, "-w 6-13", indicating that RankMotif++ should fit PWMs of 6 to 13 columns; the number of PWMs fit at each width was set to 5 using "-s 5"; and the reverse complement flag was set to "true" indicating that RankMotif++ should scan both the probe sequence and its reverse complement for binding site. The flag was set using "-r TGCA" which indicates the complements for nucleotides "ACGT", respectively.

PWMs for trimmed and untrimmed RankMotif++ primary motifs

RankMotif++ returns the PWM at each width that assigns the highest average log likelihood to the probe preference pairs. For the untrimmed RankMotif++ motifs learned from 59-mers (removing the most distal nucleotide, relative to the slide surface, from the 60-mer probe sequences on the arrays), we chose the PWM with the highest likelihood. For the trimmed RankMotif++ motifs learned from 59-mers, we subtracted a complexity penalty equal to 0.007 times the width from the average log likelihood of each PWM and chose the PWM with the highest penalized likelihood.

PWMs for trimmed and untrimmed RankMotif++ secondary motifs

In order to calculate the likelihood for a PWM, RankMotif++ uses the PWM to assign each probe preference pair a probability that reflects how likely a DBD with the given PWM is to show a binding preference for the first probe over the second probe; as such this probability is a measure of how well the observed preference is explained by the PWM. In order to compute the trimmed and untrimmed RankMotif++ secondary motifs, we assigned the probe preference pairs weights equal to one minus the probability for the preference pair under the corresponding primary RankMotif++ PWM (trimmed or untrimmed) and used the RankMotif++ software to fit the best PWM to the re-weighted preference pairs. To derive the trimmed secondary

RankMotif++ motif, we selected the PWM with the highest penalized likelihood as described above.

Using RankMotif++ on the combined array data

In order to fit the RankMotif++ PWMs to both arrays at once, we took the union of the sets of probe preference pairs identified on each of the single arrays and followed the same PWM procedures described above except that we used a complexity penalty of 0.005 times the PWM width instead of 0.007.

Availability of RankMotif++

The RankMotif++ code and AMD x64 Linux binaries used above are available from http://morrislab.med.utoronto.ca as Bash shell scripts that implement all the steps described above.

Motif construction using Kafal

Kafal (K-mer affinity align) (Pena-Castillo *et al.*, manuscript in preparation) finds motifs models by clustering DNA sequences using affinity propagation (*20*) and then aligning all sequences in each cluster found using ClustalW (*21*). The similarity metric used during clustering is a modified edit distance where insertions are highly penalized. During clustering, we set affinity propagation's maximum number of iterations to 3000, the number of iterations required to converge to 20, and the damping factor to 0.99. Default values where used for ClustalW. For the PBM data, Kafal considers all 8-mers with either an enrichment score of at least 0.45 or, if less than 200 8-mers have an enrichment score of at least 0.45, the top two hundred 8-mers.

Assessing the predictive power of different motif representations

The predictive power of primary to n-ary PWMs trained by Seed-and-Wobble (*3*) (and trimmed as described above), RankMotif++ (*19*), and Kafal (Pena-Castillo *et al.*, manuscript in

preparation) on one array (array design #1) were evaluated by testing for their ability to predict the 8-mer intensities on a replicate array (array design #2) and vice versa using precision-recall (PR) statistics and using area under receiver operator characteristic curve (AUC) statistics. Briefly, given a PWM model for a TF binding, a value for each 8-mer was calculated, using the GOMER scoring function (*13*), that measures the probability of transcription factor binding to any site in the genome. For the PR analysis, the precision at a fixed 70% recall was calculated for 100 bins representing class 1 from a minimum E-score value of 0.2 (all of the 8-mers with E-score < 0.2 were classified as class 0) or a minimum Z-score of 2 (all of the 8-mers with Z-score < 2 were classified as class 0), to the maximum E- or Z-score for each PBM experiment (thus, the bins were of varying sizes across the different TFs but of equal size within each TF). At each E-score or Z-score threshold (i.e., each point on the graph) the Precision (= True Positives / (True Positives + False Positives)) was determined for the value at which the Recall (= True Positives / (True Positives + False Negatives)) is 70%. Similarly, for the AUC analysis, the AUC statistics measuring the probability that an example from class 1 scores higher than an example from class 0 were calculated for 40 bins representing class 1 from a minimum of 10 top to a maximum of 10,000 top 8-mers (0.1 increment on a $\log_{10}$ scale), and the bottom 22,896 8-mers represent class 0. At each threshold (i.e., each point on the graph), the Sensitivity (= True Positives / (True Positives + False Negatives)) and 1-Specificity (= false positive rate = False Positives / (False Positives + True Negatives)) were calculated.

We asked how well the motif models derived from each method (SW, RM, K, 8-mer scores and regression-derived combination of PWM motif models) learned from one array, were able to recapitulate array 8-mer binding data of the other array by precision/recall analysis, by measuring the percent of variance explained, and by area under receiver operating characteristic plots (AUC) analysis. In both PR and AUC analyses, 8-mer E-scores from one array replicate were evaluated against the E-scores of the other array replicate.

Our overall conclusion is that, in virtually every case, none of the PWMs learned from either the original 35-mer data or the 8-mer transformed data is capable of reproducing the 8-mer ranks on the opposite array (for array A vs. B, or B vs. A comparisons) nearly as well as the 8-mer scores derived from the first array, although virtually all of the dominant motifs vastly outperform

random guessing (i.e., the dominant motifs account for a good proportion of the variation in the other replicate and almost all motifs weighted by Lasso explain 10% or greater the variation in the replicate 8-mer data; if there was no relationship between the motif and the 8-mer data, the correlation and percent of variance explained would be at or around zero).

For the majority of the proteins in this study, RankMotif++ PWMs yielded better performance measures over the full range of preference scores (i.e., highest through lowest affinities) as compared to Seed-and-Wobble PWMs or Kafal PWMs, although for some proteins there was no clear distinction which algorithm captured the whole range of data best – sometimes Seed-and-Wobble PWMs would capture the highest affinity sites best while the PWMs from other algorithms captured the lower affinities better. We used TomTom (*22*) to evaluate the pairwise similarity between PWMs, using an E-value threshold of 0.01 to classify motifs as related or un-related. Surprisingly, motifs that are clearly related can yield very different performance measures. 37% of pairwise PWM similarity comparisons (for the same TF by the three methods) with a TomTom E-value ≤ 0.01 (i.e. the PWMs are considered to be the same) have differences $\geq$ 10% in variance explained on the 8-mer ranks of the opposite array, suggesting that small differences between motif models have an effect on their explanatory power. Conversely, motifs that did not meet this similarity threshold may score highly on the same data set. 42% of pairwise PWM similarity comparisons with a TomTom E-value > 0.01 (i.e., the PWMs are considered to be distinct) differed no more than 10% in variance explained on the 8-mer ranks when considering PWMs derived from one array replicate compared to the data from the other array. Kafal most often yielded diverse PWMs that captured the binding data well as compared to other motif finders, as 35% of Kafal PWMs are given positive weights by the linear regression model, while 30% and 19% respectively of RM and SW PWMs have non-zero weights.

Lasso analysis to build multiple-motif model representations of the PBM data

Weighted combinations of PWMs were built by using the least absolute shrinkage and selection operator (Lasso) algorithm (*23*), which learns the weighting of each PWM by linear regression, where the independent variables were the 8-mer GOMER scores (*13*), which is an estimate of the

probability of transcription factor binding, derived from the PWMs learned from each array experiment and the dependent variables were the 8-mer Z-scores and E-scores. Lasso only weights the motifs that contribute to explaining the variance. A bootstrap procedure was employed to assess the stability of the learned weights.

G. Analysis of simulated 14-bp motif PBM data

We performed Seed-and-Wobble analysis on two different sets of 50 simulated, 14-bp motifs, to search for primary and secondary motifs in simulated data for simulated 'long' (14-bp) motifs. The 1st set of 50 simulated motifs corresponded to motifs that we assembled by stitching together various existing motifs, in one of 3 different ways:

1. Combined the primary and secondary motifs from the 6 TFs (Hnf4a, Nkx3.1, Mybl1, Foxj3, Foxk1 and Rfxdc2) in our manuscript for which we verified the secondary motifs by EMSA, trimming the flanking positions so that the resulting motifs were 14 bp long.

2. Took 24 long motifs from JASPAR (widths 11 to 17) and either removed the flanking positions or added new columns to the end of the motif from columns derived from the middle of the motif to get a final width of 14. Some motifs had exactly 14 positions, in which case nothing was done to the columns.

3. Combined 40 pairs of shorter motifs (widths 5 to 11) and either removed the flanking positions or added new columns to the end of the motif from columns derived from the middle of the motif to get a final width of 14. Some concatenated motif pairs totaled exactly 14 positions, in which case nothing was done to the columns. The ordering of the motif pair concatenation was random.

We used real motifs in these ways to generate simulated motifs, since we have found in the past (Bulyk Lab, unpublished results) that purely synthetically simulated motifs are not 'round' in terms of 'Hamming ball' motif space, and thus are not a highly accurate simulation of true motifs. The 2nd set of 50 simulated motifs corresponded to shuffled versions (positions, or 'columns', of the PWM were shuffled) of the 1st set of 50 simulated motifs.

Since PBM data from different de Bruijn sequences are highly reproducible, and since the requested analysis does not depend on the particular de Bruijn sequence used, we used the GOMER scoring scheme to score each of these 100 simulated motifs against the probe sequences of array #1 (de Bruijn sequence #1, arbitrarily). For both sets of motifs, the data were 'noised' to simulate PBM data. Specifically, the simulated motifs were scored against all the 60-mer array probes using GOMER. The log(GOMER) scores were then converted into z-scores and split into 100 bins. The standard deviation within each bin was calculated and noise was added by generating random Gaussian z-scores using the mean and standard deviation in the bin. The standard deviation was scaled with a multiplicative factor such that the scores on the low end were made noisier than the scores on the high end, in order to simulate the noise distribution in real PBM data. This procedure outputted a set of scores representing the binding probabilities of a TF to each of the array probes as specified by its simulated motif. These outputted scores correspond to the simulated PBM data. This resulted in a ranking of the probes according to GOMER scores (probabilities), which we then analyzed for motif content. Since our identification of "secondary" motifs was based upon analysis with Seed-and-Wobble and not by analysis with RankMotif++, we performed all subsequent analysis of these simulated motif data using just Seed-and-Wobble.

Seed-and-Wobble was highly successful in identifying the long, 14-bp simulated motifs. Indeed, the motif was successfully recovered as the primary motif with a top seed 8-mer E-score ≥ 0.45 for 97 out of the 100 14-bp simulated motifs. Approximately 41% of the primary motifs technically had a 'secondary motif' that could be identified by Seed-and-Wobble with a top seed 8-mer E-score ≥ 0.45. However, for the vast majority of those cases, given their low quality (low information content and/or similarity to the primary motif), we likely would not have reported those motifs as significant secondary motifs; we note that in analyzing the real data for the 104 nonredundant mouse TFs, the primary and secondary motifs never did look very similar, but rather had important differences between them. For 10 cases, the secondary motif could be misinterpreted as suggesting position interdependence, if they were to be considered as significant secondary motifs; however, given the lengths of these motifs and our knowing that we are underpowered to fully determine the DNA binding preferences of motifs >10 bp, for at

least 8 of these cases we would not have had confidence in stating that they were indicative of position interdependence. In only 1 case might the secondary motif suggest 'variable spacer length'; however, in that particular case both the primary and secondary Seed-and-Wobble motifs were of low information content motif (in both motifs all positions had information content < 1), and they likely would not have passed our quality control criteria for considering them as significant motifs. None of the secondary motifs were indicative of 'alternate recognition interfaces'. Importantly, in 10 cases the primary motif and the secondary motif could be assembled to accurately re-create the full, 14-bp simulated motif, and in another 3 cases assembly of the primary and secondary motifs would re-create the full-length motif plus inaccurate extraneous flanking sequence. In only 1 case might the secondary motif suggest 'multiple effects'; in that case, the primary and secondary motifs are capturing half-sites (one of which is somewhat lengthy), and sufficient information is not captured to instead argue for assembling the full-motif from the primary and secondary motifs.

The results of this motif analysis of the 100 simulated motifs support our conclusion that essentially all (if not all) of the secondary motifs we found in analyzing the real PBM data for 104 nonredundant mouse TFs are indeed real and are highly unlikely to be attributable to a motif finding artefact due to long motifs. Moreover, nearly all of the 104 TFs we analyzed in this present manuscript belong to TF structural classes known to bind relatively compact motifs (in general 10 bp or shorter) based upon prior experimental studies, including traditional protein-DNA biochemical analyses, *in vitro* selections ('SELEX'), TF-DNA co-crystal structures, and ChIP-chip.

H.   Analysis of ChIP-chip data

Relative enrichment of *k*-mers corresponding to the primary versus secondary Seed-and-Wobble motifs within bound genomic regions in ChIP-chip data as compared to randomly selected sequences was calculated for Bcl6 (*24*) (GEO accession #GSE7673) and for Hnf4a (Neilsen *et al.*, submitted; GEO accession #GSE7745). ChIP-chip 'bound' peaks were identified according to the criteria of the respective studies (*24*)(Neilsen *et al.*, submitted). Briefly, the sequences of the peaks were extracted such that if the peak location was not specified, then flanking sequence from the midpoints of positive probes (usually defined as the middle probe of a string of five consecutive positive probes) was extracted. A window size of 500 bp with a step size of 100 bp was used. Regions 'bound' *in vivo* by ChIP-chip were split according to whether or not they had a primary or secondary 8-mer above the score threshold within the -250 to +250 windows considered in our analysis. This was done essentially as in a prior study (*25*), except that here we employed thresholds for 8-mers based on GOMER scores rather than on E-scores. This resulted in 4 types of regions: 1) those with only a primary motif 8-mer; 2) those with only a secondary motif 8-mer only; 3) those with both primary and secondary motif 8-mers; and 4) those with neither primary nor secondary motif 8-mers. 8-mer enrichment was calculated relative to a background sequence set containing ten times the number of randomly selected genomic regions. *P*-values were calculated for the interval (-250 to $+250$) by the Wilcoxon-Mann-Whitney rank sum test, comparing the number of occurrences per sequence in the bound set versus the background set.

Supplementary 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.	V. Matys *et al.*, *Nucleic Acids Res* **34**, D108 (Jan 1, 2006).
6.	D. Vlieghe *et al.*, *Nucleic Acids Res* **34**, D95 (Jan 1, 2006).
7.	E. Osaki *et al.*, *Nucleic Acids Res* **27**, 2503 (Jun 15, 1999).
8.	G. Pengue, P. Cannada-Bartoli, L. Lania, *FEBS Lett* **321**, 233 (Apr 26, 1993).
9.	X. Meng, M. H. Brodsky, S. A. Wolfe, *Nat Biotechnol* **23**, 988 (Aug, 2005).
10.	J. D. Hughes, P. W. Estep, S. Tavazoie, G. M. Church, *J Mol Biol* **296**, 1205 (Mar 10, 2000).
11.	F. P. Roth, J. D. Hughes, P. W. Estep, G. M. Church, *Nat Biotechnol* **16**, 939 (Oct, 1998).
12.	C. T. Workman *et al.*, *Nucleic Acids Res* **33**, W389 (Jul 1, 2005).
13.	J. A. Granek, N. D. Clarke, *Genome Biol* **6**, R87 (2005).
14.	J. Warner *et al.*, *Nature Methods* **5**, 347 (Epub March 2, 2008., 2008).
15.	P. J. Huber, *Robust Statistics* (John Wiley, New York, 1981), pp.
16.	W. Huber, A. von Heydebreck, H. Sultmann, A. Poustka, M. Vingron, *Bioinformatics* **18 Suppl 1**, S96 (2002).
17.	O. G. Berg, P. H. von Hippel, *J Mol Biol* **193**, 723 (Feb 20, 1987).
18.	G. D. Stormo, *Bioinformatics* **16**, 16 (Jan, 2000).
19.	X. Chen, T. R. Hughes, Q. Morris, *Bioinformatics* **23**, i72 (Jul 1, 2007).
20.	B. J. Frey, D. Dueck, *Science* **315**, 972 (Feb 16, 2007).
21.	R. Chenna *et al.*, *Nucleic Acids Res* **31**, 3497 (Jul 1, 2003).
22.	S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey, W. S. Noble, *Genome Biol* **8**, R24 (2007).
23.	R. Tibshirani, *Journal of the Royal Statistical Society Series B-Methodological* **58**, 267 (1996).
24.	S. M. Ranuncolo *et al.*, *Nat Immunol* **8**, 705 (Jul, 2007).
25.	M. F. Berger *et al.*, *Cell* **133**, 1266 (Jun 27, 2008).