Supplemental Material for "Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity"

Contents

EXTENDED EXPERIMENTAL PROCEDURES

Selection and cloning of DBDs for analysis. We compiled the predicted proteomes of 290 eukaryotic organisms from a variety of sources, and supplemented them with an additional 49 known TFs from organisms without fully sequenced genomes. We scanned all protein sequences for putative DNA-binding domains (DBDs) using the 81 Pfam (Finn et al., 2010) models listed in (Weirauch and Hughes, 2011) and the HMMER tool (Eddy, 2009), with the recommended detection thresholds of Per-sequence Eval < 0.01 and Per-domain conditional Eval < 0.01. Each protein was classified into a family based on its DBDs and their order in the protein sequence (e.g., bZIPx1, AP2x2, Homeodomain+Pou). We then aligned the resulting DBD sequences within each family using clustalOmega (Sievers et al., 2011), with default settings. For protein pairs with multiple DBDs, each DBD was aligned separately. From these alignments, we then calculated the sequence identity of all DBD sequence pairs (i.e. the percent of AA residues that are exactly the same across all positions in the alignment).

We selected TFs from species with available TF templates (**Table S6**) using five different selection strategies:

(1) We first selected proteins in order to achieve comparisons at different levels of protein sequence identity. We thus iteratively selected TFs from all 55 diverse species, spanning all major branches of the eukaryotic kingdom, to evenly populate bins of sequence identity among pairs of TFs (as shown in the Boxplots in **Figure 2**). We sought to obtain ten comparisons for each of nine bins (10-19.99 to 90-99.99 percent identity in the DBD AA sequences) for each DBD class.

(2) We next sought to balance numbers across kingdoms and DBD classes. For each of the 81 DBD classes, and for each of five major kingdoms (metazoans, plants, algae, fungi, and protists), we selected the single TF for which the largest number of additional proteins are >65% identical in the DBD. This level was chosen because it was the threshold for DBD AA motif identity obtained in our previous study of homeodomains (Berger et al., 2008).

(3) To increase numbers of inferred motifs in general, we also selected 250 additional TFs with DBDs that are >65% identical to those of at least 35 other TFs, regardless of family or kingdom.

(4) To improve sampling of TFs from model species, which we reasoned would be useful to the largest number of investigators, we obtained data for 222 TFs from *Arabidopsis thaliana* (plant), 124 from *Neurospora crassa* (filamentous fungus), 94 from *Mus musculus* (mouse), 41 from *Caenorhabditis elegans* (worm), 37 from *Dictyostelium discoideum* (social amoebae), and 24 from *Ostreococcus tauri* (green alga), selected to represent a diversity of TFs from these species.

(5) To improve sampling of several major DBD classes, we obtained data for TFs representing 207 homeodomains, 84 bZIPs, 84 Myb/SANTs, 76 bHLHs, 65 zinc clusters, 45 AP2s, 45 C2H2 zinc fingers, and 37 nuclear receptors.

In a first round of analysis, we cloned 2,913 DBDs by PCR-assisted cloning and successfully obtained data from 905. PBM success criteria are described below; proteins were repeated on each array type if they failed the first attempt. We then

repeated strategies (1) and (2) in an effort to fill remaining gaps, using gene synthesis (206 TFs).

For our first four selection strategies, we designed primers to clone the region encompassing all DBDs plus the 50 (or in some cases, 15) flanking endogenous AAs on either side (or until the termini of the protein) by conventional PCR methods into one of a panel of T7-GST vectors for expression in *E. coli* (referred to hereafter as "plasmid constructs"). For our fifth selection strategy, for major DBD classes we selected up to five proteins from each group with similarity above 65% (or 85% in the DNA-contacting residues, if known, following (Berger et al., 2008)). We then designed an Agilent long-oligo (OLS) pool that contained complimentary 200-base oligonucleotides encoding the DBD, which we amplified from the pool using primers complimentary to unique ends and cloned as above. These constructs did not contain any AAs flanking the DBD. All inserts were sequence verified in full. Insert sequences and other information are available in **Table S6**.

PBMs and data processing. PBM laboratory methods were identical to those described previously (Lam et al., 2011; Weirauch et al., 2013). Each plasmid was analyzed in duplicate on two different arrays with differing probe sequences. Microarray data were processed by removing spots flagged as 'bad' or 'suspect', and employing spatial de-trending (using a 7x7 window centered on each spot) as in (Weirauch et al., 2013). Calculation of 8-mer Z- and E-scores was performed as previously described (Berger et al., 2006). Z-scores are derived by taking the average spot intensity for each probe containing the 8-mer, then subtracting the median value for each 8-mer, and dividing by the standard deviation, thus yielding a distribution with a median of zero and

a standard deviation of one. E-scores are a modified version of the AUROC statistic, which considers the relative ranking of probes containing a given 8-mer, and range from -0.5 to +0.5, with E>0.45 taken as highly statistically significant (Berger et al., 2008). Experiments were deemed successful if at least one 8-mer had an E-score > 0.45 on both arrays, the complimentary arrays produced highly correlated E- and Z-scores, and the complimentary arrays yielded similar PWMs based on the PWM_align algorithm, which aligns the top 10 8-mer E-scores and tallies the frequency at each position to generate a PWM (Weirauch et al., 2013).

Motif derivation and scanning. To obtain a single representative motif for each protein, we generated motifs for each array using four different algorithms that performed well in a recent study from our group (Weirauch et al., 2013). (i) BEEML-PBM (Zhao and Stormo, 2011) is a biophysical-based method that obtains maximum likelihood estimates of parameters to a PWM using a model that incorporates the TF's chemical potential, non-specific binding affinity, and probe position-specific effects. (ii) FeatureREDUCE (manuscript in prep, source code available at http://rileylab.bio.umb.edu/content/software) combines a biophysical free energy PWM model with a contiguous k-mer background model (length 4 to 8) in a robust regression framework. (iii) PWM_align is described above. (iv) PWM_align_Z aligns the top 10 scoring 8-mers based on their Z-scores, weighting them by their Z-scores (Ray et al., 2013).

For each plasmid construct, we ran each algorithm on the normalized PBM intensity data from both array designs (ME and HK). We then scored each motif on the complimentary array for the same plasmid construct by scoring each probe sequence

using BEEML's energy-based scoring function (with the mu parameter set to 0) (Zhao and Stormo, 2011), which sums the PWM score at each position in the probe. (This same approach was used throughout the paper for scanning DNA sequences with a PWM – when comparing sequences of different lengths, we use the average score instead of the sum). We then compared these PWM-based probe score predictions with the actual probe intensities using (1) the Pearson correlation coefficient (PCC) and (2) the AUROC of "bright probes" (defined by transforming all probe intensities to Z-scores, and selecting probes with Z-scores >= 4), following (Weirauch et al., 2013).

We then divided all plasmid constructs into two groups: those for which at least one algorithm's PWM achieved cross-platform probe PCC > 0.70 (i.e. instances where the PWM scores can predict at least ~half of the variance of the intensities of all probes), and those for which all algorithms had PCC < 0.70. For plasmids in the PCC > 0.70 group, we chose a single PWM (from the eight candidates) based on a relative mix of the PCC and AUROC scores – we first divided each of the two scores by the highest score obtained by any algorithm for the plasmid, and then took the mean of these scores, as in (Weirauch et al., 2013). If no PWM achieved a cross-array PCC of 0.70 or higher, we first removed all uninformative PWMs (PWMs with less than one bit of total information, which occasionally perform well, presumably by capturing low-magnitude sequence biases in the PBM assay). We then chose the single PWM that performed best based only on its AUROC.

To estimate the number of novel motifs determined in this study, we first scored 50,000 random 100 base DNA sequences of uniform base composition using each motif from this and other studies (utilizing BEEML's scoring function, and summing across all

positions). We then calculated the Pearson correlation coefficient *r* across all 50,000 sequences for each pair of motifs. We considered a motif from this study 'novel' if no previously determined motif has a value of r > 0.70 (i.e., no previously determined motif can explain at least half the variance across these sequences).

Inference scheme. We established a separate motif inference threshold for each DBD class. For each DBD class, we only compared constructs with the same count and order of DBDs. We aligned the DBD sequences of all constructs using clustalOmega (Sievers et al., 2011), as described above. Alignments are available in **Supplementary Data 1.** We then calculated the AA %ID for all construct pairs (i.e. the number of identical AA sequences in the alignments).

Within each DBD class, we grouped all PBM construct pairs into bins, based on AA %ID. We used bins of size 10, ranging from 0 to 100, increasing by 5 (i.e., 0-9.99%, 5-14.99%....90-99.99%, 95-99.99%, 100%). We calculated the precision of each bin by comparing the DNA sequence preferences obtained from all characterized protein pairs contained in the bin, as described below. We quantified the similarity of the DNA sequence preferences of two proteins P_1 and P_2 as the fraction of shared high-scoring 8-mers $F(P_1,P_2)$, using the following procedure: (1) identify the number of 8-mers (N_1 and N_2) that exceed the given threshold in each experiment; (2) calculate N as max (N_1,N_2); (3) calculate the similarity between the two experiments as the number of identical top N 8-mers, divided by N. We identified high-scoring 8-mers using eight different methods/thresholds: E-scores exceeding 0.45, 0.46, 0.47, or 0.48, and the top 10, 20, 30, or 40 Z-scores. Values of $F(P_1,P_2)$ for the E-score>0.45 scheme are

depicted on the Y-axis of the boxplots in **Figure 2**. Results from other 8-mer schemes are provided in **Supplementary Data 1.**

We considered a prediction to be correct only if its value for $F(P_1,P_2)$ exceeded the value obtained at the 25th percentile of experimental replicates (i.e., $F(P_1,P_2)$ calculated between the ME and HK arrays for the same protein). In order to minimize the impact of noise on setting the thresholds, we used a single threshold calculated across all DBD classes. This stringent threshold setting means that the 8-mers of the given protein pair are more similar than the 8-mers of the lower 25% of all experimental replicates. The proportion of predictions for non-replicates scored as correct (i.e. precision) for each bin of each DBD class is shown as magenta stars in **Figure 2** and **Supplementary Data 1.** We considered eight different E-score or Z-score thresholds, and obtained very similar results for each of them (**Figure S6** and **Supplementary Data 1**).

We chose inference thresholds for each DBD class based on the precision scores of each AA %ID bin. Since we used the 25th percentile threshold to define precision, we would expect a precision of 0.75 or higher in each AA %ID bin. We therefore chose an inference threshold for each DBD class by identifying the final AA %ID bin before precision drops below 0.75 (vertical bars in **Figure 2**). Similar thresholds were obtained regardless of the E- and Z-score thresholds used, and also regardless of the replicate overlap percentile considered (i.e., 25th percentile, requiring 0.75 precision or 20th percentile, requiring 0.80 precision) (**Figure S6**).

The final threshold for a DBD class was then chosen as the median threshold across the eight 8-mer similarity measures (see **Figure S6** and **Supplementary Data 1**). We

found this scheme to be appropriate for most DBD classes (all of which are depicted in **Figure 2**). For three DBD classes (IRF, CXXC zinc fingers, and Dof zinc fingers), we could not establish a threshold – these therefore received a threshold of 100%. For the AT-hook class, whose members recognize AT-rich sequences, we chose a 40% threshold based on manual inspection of the pairwise 8-mer E-score scatterplots, which illustrate that all tested constructs in this class have similar overall 8-mer preferences (scatterplots provided in **Supplementary Data 1**). For the remaining classes, with suggestive but insufficient data, we chose a threshold of 70%, which is the mean, median, and mode threshold across all DBD classes. We used the AA %ID of all pairs of proteins to infer motifs, 8-mer scores, and consensus sequences within each DBD class by simple transfer (i.e., aligning the DBD sequences of all proteins and all constructs in a given DBD class, as described above, and calculating the AA %ID of each protein with each construct).

We evaluated the effectiveness of our inference scheme in a leave-one-out cross validation framework. For each DBD class, we calculated the cross validation success rate of each AA %ID bin using the following procedure:

(1) For each characterized protein, choose the single closest protein that has been characterized using PBMs (based on AA %ID).

(2) Consider the prediction a "success" if this protein has a $F(P_1,P_2)$ value (i.e., the 8-mer overlap score defined above) exceeding the value obtained at the 25th percentile of experimental replicates (i.e., $F(P_1,P_2)$ calculated between the ME and HK arrays for the same protein).

(3) Report the cross validation success rate as the fraction of successes for all characterized proteins for the given AA %ID bin of the given DBD class.

The overall cross validation success rate for each DBD class was calculated across all bins exceeding the inference threshold for the given DBD class. The final cross validation score for a DBD class was then calculated as the mean score across all eight 8-mer similarity measures. We also report a single cross validation rate across all DBD classes in an analogous manner, by grouping the results from all DBD classes together. Results are provided in **Supplementary Data 1.**

Cis-BP database. To house the motifs, and other related information about the TFs, we created the Cis-BP database (**http://cisbp.ccbr.utoronto.ca**). The Cis-BP data is stored in a mySQL relational database accessible through a browsable web interface written in PHP. Cis-BP currently incorporates 10 predefined or built-in tools for specific tasks, including scanning DNA or protein sequences against PWMs, comparing a submitted PWM to our dataset, and predicting a DNA motif based on a given AA sequence. Cis-BP data can be downloaded for the complete set of TFs with known or inferred motifs, or for specific subsets of TFs. All analyses in this paper used version 0.90 of Cis-BP.

Comparison to ChIP-seq data. We calculated AUROC scores on real and permuted ChIP-seq peak sequences (maintaining dinucleotide frequencies), following (Weirauch et al., 2013). We obtained ENCODE consortium human ChIP-seq data from the UCSC Genome Browser (Rosenbloom et al., 2012). For each ChIP experiment, we extracted the top 500 scoring peak region sequences, and scored them using all direct and

inferred PWM models for the given TF, using the average BEEML-based score across all sequence positions (described above). We also scored a corresponding negative sequence set for each experiment (created using an algorithm that randomly permutes each peak sequence, while maintaining all dinucleotide frequencies). For each PWM/experiment pair, we then calculated the AUROC using these sets of 500 positives and 500 negatives. Since experiments are available for multiple cell types and antibodies for many TFs, we report the final score for a PWM/TF pair as its average AUROC score across all cell types and antibodies. To extract PWMs from ChIP-seq peaks for comparison to our motifs in **Figure 1B**, we ran the ChIPMunk algorithm (Kulakovskiy et al., 2010) with default settings, using the top 500 scoring peak sequences.

Positional bias of motifs in eukaryotic promoters. We obtained promoter sequences (1000 upstream bases) from the following sources: Human and *D. rerio*, Ensembl Biomart (Kinsella et al., 2011); *D. discoideum,* DictyMart (Fey et al., 2009); *N. crassa*, The Broad Institute (build NC12) (Galagan et al., 2003); *S. cerevisiae*, de Boer et al. (de Boer et al., 2014); *T. vaginalis*, TrichDB (Aurrecoechea et al., 2009), *O. sativa*, Phytozome Biomart (Kinsella et al., 2011). We used the transcription start site (TSS), if known; if not, we used the start codon. For *D. discoideum* and *T. vaginalis*, the majority of genes have no known TSS. We used BEEML's scoring method (Zhao and Stormo, 2011) (setting the 'mu' parameter to 0) to score each PWM (incorporating all PBM-derived direct and inferred PWMs for the given organism) at each base position of each promoter. We then placed the resulting scores into 20 bp bins, summed the scores for each bin, and took the average across all promoters for the given species for each bin.

To correct for mono- and dinucleotide biases, we also scored shuffled promoter sequences, which were created by shuffling the sequences within each 20 bp bin (while maintaining dinucleotide frequencies). For each PWM, we then calculated the ratio of each bin's real score relative to the score of the shuffled sequence. The resulting ratios were then normalized across all bins for the given PWM using a standard Z-score transformation. We also created a negative control set of TF PWMs for each organism using a collection of random motifs from species in other clades that were unrelated to any PWM from the given species (i.e., no PWM had a Pearson correlation across 10,000 random sequences exceeding 0.30.). We scored each promoter sequence for each PWM, and calculated Z-scores for this negative set using the same procedure described above.

We also performed these analyses on a reduced set of non-redundant PWMs for each organism. To obtain a set of non-redundant motifs for each organism, we used the matrix of pairwise motif similarity scores (Pearson correlation across 50,000 random sequences) described above. To identify groups of related motifs within each DBD class, we clustered this matrix using Affinity Propagation (Frey and Dueck, 2007), setting the "q" parameter to 0.80. We chose this setting from several we tested (0.05, 0.10, 0.20, 0.50, 0.80, 0.90) because it consistently resulted in visually similar E-score profiles upon manual examination (data not shown). The final set of non-redundant motifs for an organism was then constructed from the exemplars of each cluster (i.e., the single member of each cluster that best explains the values of its members was chosen as the cluster representative). This procedure resulted in a reduction of motif counts ranging from 2.4-fold (*T. vaginalis*) to 8.2-fold (human).

Arabidopsis eQTL analysis. We used a publicly available dataset (Gan et al., 2011) containing genome-wide RNA-seq variance-stabilized expression levels (Huber et al., 2002) taken from 19 strains of seedling *Arabidopsis thaliana*, and matching genome sequences. We obtained 1000 base upstream sequences for each *A. thaliana* gene using the TAIR-10 genome build annotation, and scored every position within these regions with each *A. thaliana* PWM using BEEML's scoring method (setting mu to 0). We considered a sequence in a promoter as a putative binding site for a PWM if its score was at least as high as that of the 25th highest scoring unique sequence found in a set of 1000 randomly chosen promoters, where each PWM-sized window is taken as a sequence.

We restricted the set of genetic variants considered in our analysis using the following filters: (1) We only included variants with genotypes available in all 19 strains; (2) To include only the strongest associations for each gene, we only included variants whose cis-eQTL p-value was not greater than $p_{min}^{0.9}$, where p_{min} is the minimum p-value in a 30 kb window around the corresponding gene; (3) We only considered promoters with no more than five variants fulfilling the above two criteria.

We used the resulting set of variants to calculate the percentage of variants that affect putative binding sites, as a function of cis-eQTL p-value (red line, **Figure 7**). We also created a null distribution (blue line and blue shaded region, **Figure 7**) to exclude the possibility that the observed percentages might solely be due to the higher density of TF binding sites in promoter regions. This distribution was created by multiplying the

density of PWM hits at each position by the number of variants at the corresponding position, summing these products across all promoter regions, and dividing this sum by the total number of significant variants. The variance of these percentages was estimated by first computing 100 times the fraction of variants that overlap the PWM hits of randomly chosen genes, and then calculating the variance across these values.

Human disease SNP/TF analysis. We devised a system for utilizing our collection of PBM data to identify candidate human TFs whose binding might be affected by the allelic sequences of genetic variants. In this system, we score each variant using 8-mer E-scores taken from the 3,132 PBM experiments contained in our database. For a given variant, we first determine all 8-mers that each allele overlaps in the human reference genome sequence (for example, a SNP will overlap eight 8-mers, plus their reverse complements, for each of its alleles). For each PBM experiment, we then identify the highest scoring 8-mer E-score attained by any of the risk allele sequences (E_{risk}), and the highest attained by any non-risk allele ($E_{non\text{-}risk}$). We then identify all PBM experiments where only one of E_{risk} and $E_{non\text{-}risk}$ has an E-score value exceeding 0.45 (values above this threshold will likely be strongly bound by the given TF (Berger et al., 2008)). All experiments meeting this criterion are then assigned a final score E_{final}, which is the maximum value of (E_{risk} and $E_{non\text{-}risk}$). For each human TF, a single value for E_{final} is chosen as the maximum E_{final} value attained for all PBM experiments assaying a TF whose AA %ID exceeds the inference threshold for the given DBD class (i.e., all PBM-based inferences are considered for the given TF). This procedure thus produces a ranked list of human TFs whose binding is likely to be affected by the alleles of a given SNP (e.g., strongly binding to one allele, but not binding to the other). We

applied this system to a set of 15 SNPs taken from previous studies, where a TF has been experimentally confirmed (e.g., via supershift EMSA or ChIP) to differentially bind the SNP alleles (one SNP affects two TFs, bringing the total number of TF/SNP pairs to 16). We only considered cases for inclusion in which the known TF has available PBM data (either directly determined, or for a related TF within the same DBD class). The results of this analysis, and references for the disease SNP studies, are contained in **Supplementary Data 1**.

REFERENCES

Aurrecoechea, C., Brestelli, J., Brunk, B.P., Carlton, J.M., Dommer, J., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G.*, et al.* (2009). GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens Giardia lamblia and Trichomonas vaginalis. Nucleic acids research *37*, D526-530.

Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A., Pena-Castillo, L., Alleyne, T.M., Mnaimneh, S., Botvinnik, O.B., Chan, E.T.*, et al.* (2008). Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell *133*, 1266-1276.

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., 3rd, and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nature biotechnology *24*, 1429-1435.

de Boer, C.G., van Bakel, H., Tsui, K., Li, J., Morris, Q.D., Nislow, C., Greenblatt, J.F., and Hughes, T.R. (2014). A unified model for yeast transcript definition. Genome Res *24*, 154-166.

Eddy, S.R. (2009). A new generation of homology search tools based on probabilistic inference. Genome Inform *23*, 205-211.

Fey, P., Gaudet, P., Curk, T., Zupan, B., Just, E.M., Basu, S., Merchant, S.N., Bushmanova, Y.A., Shaulsky, G., Kibbe, W.A.*, et al.* (2009). dictyBase--a Dictyostelium bioinformatics resource update. Nucleic acids research *37*, D515-519.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K.*, et al.* (2010). The Pfam protein families database. Nucleic Acids Res *38*, D211-222.

Frey, B.J., and Dueck, D. (2007). Clustering by passing messages between data points. Science *315*, 972-976.

Galagan, J.E., Calvo, S.E., Borkovich, K.A., Selker, E.U., Read, N.D., Jaffe, D., FitzHugh, W., Ma, L.J., Smirnov, S., Purcell, S.*, et al.* (2003). The genome sequence of the filamentous fungus Neurospora crassa. Nature *422*, 859-868.

Gan, X., Stegle, O., Behr, J., Steffen, J.G., Drewe, P., Hildebrand, K.L., Lyngsoe, R., Schultheiss, S.J., Osborne, E.J., Sreedharan, V.T.*, et al.* (2011). Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nature *477*, 419-423.

Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics *18 Suppl 1*, S96-104.

Kinsella, R.J., Kahari, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A.*, et al.* (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database : the journal of biological databases and curation *2011*, bar030.

Kulakovskiy, I.V., Boeva, V.A., Favorov, A.V., and Makeev, V.J. (2010). Deep and wide digging for binding motifs in ChIP-Seq data. Bioinformatics *26*, 2622-2623.

Lam, K.N., van Bakel, H., Cote, A.G., van der Ven, A., and Hughes, T.R. (2011). Sequence specificity is obtained from the majority of modular C2H2 zinc-finger arrays. Nucleic Acids Res *39*, 4680-4690.

Ray, D., Kazan, H., Cook, K.B., Weirauch, M.T., Najafabadi, H.S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A.*, et al.* (2013). A compendium of RNA-binding motifs for decoding gene regulation. Nature *499*, 172-177.

Rosenbloom, K.R., Dreszer, T.R., Long, J.C., Malladi, V.S., Sloan, C.A., Raney, B.J., Cline, M.S., Karolchik, D., Barber, G.P., Clawson, H.*, et al.* (2012). ENCODE whole-genome data in the UCSC Genome Browser: update 2012. Nucleic acids research *40*, D912-917.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J.*, et al.* (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol *7*, 539.
Weirauch, M.T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T.R., Saez-Rodriguez, J., Cokelaer, T., Vedenko, A., Talukder, S.*, et al.* (2013). Evaluation of methods for modeling transcription factor sequence specificity. Nat Biotechnol *31*, 126-134.
Weirauch, M.T., and Hughes, T.R. (2011). A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. Sub-cellular biochemistry *52*, 25-73.
Zhao, Y., and Stormo, G.D. (2011). Quantitative analysis demonstrates most transcription factors require only simple models of specificity. Nat Biotechnol *29*, 480-483.

Supplemental Figure Legends

Figure S1 (Related to Figure 1A). Pictorial overview of TF choosing strategy and motif inferences. A. Network schematic depicting TFs (nodes), their related TFs (edges between nodes), and their motif status (node color) – see key. This figure depicts all 3,715 TFs (across 246 species) that contain a single bZIP domain. Motif inferences represent cases in the network where an orange node is linked to a red one – in such instances, a motif can be inferred for the orange TF, since its DNA binding domain (DBD) is "similar enough" to a TF with a known motif (based on the inference threshold for the given DBD class). Most TFs were chosen for characterization in this study (blue nodes) if they were members of large uncharacterized TF groups, or if they could be used to determine thresholds for the given DBD class (see **Experimental Procedures**). **B.** Zoom-in of boxed region in (A). Here, motifs are shown for characterized TFs. Note that motifs from the left group strongly resemble one another, as do motifs within the right group (as predicted by their DBD AA %ID). However, the motifs from the left group are not related to those of the right group, as predicted by the fact that the DBD %ID of their TF members fall below the inference threshold for bZIPs (i.e., there are no links between the two groups). Motifs with blue outlines were determined using PBMs; red outlined motifs are from the Transfac database.

Figure S2 (Related to Figure 2). Precisions obtained when combining together all DBD classes with insufficient available data to choose a cutoff. Boxplots are identical to those shown in **Figure 2**, and include data for the 32 DBD classes lacking sufficient data to establish thresholds. Numbers underneath each bin indicate the number of data points (i.e., TF pairs); values less than five (which indicate decreased confidence in the associated bin) are indicated in red. The eight plots correspond to the eight methods we used to compare 8-mer binding preferences (see **Experimental Procedures**). Bins between 50 and 70 are omitted because no data are available in these ranges for any of these DBD classes.

Figure S3 (Related to Figure 6). Motifs enriched at specific positions in promoter regions, by species and DBD class. Number of PWMs enriched (Z-score greater than 3) in at least one position in promoter regions, by species and DBD class. Trees indicate results of hierarchical clustering, using Euclidean distance with average linkage. Key: Blue, 0 motifs enriched; Grey, 1; Yellow, 2; Orange, 3; Red, 5 or more.

Figure S4 (Related to Figure 6). Results of motif TSS location enrichment, using sets of non-redundant motifs. Enrichment of motifs in promoter regions in a variety of eukaryotes. For this figure, a reduced set of non-redundant motifs were used (as opposed to the full set of all motifs for all TFs in an organism) – see **Experimental Procedures**. See Figure 6 legend for key and further information.

Figure S5 (Related to Figure 7A). Arabidopsis eQTL analysis, using only inferred motifs. Enrichment of the overlap between Arabidopsis eQTLs and motifs. This figure is identical to **Figure 7A**, except in this case only TFs with motifs inferred from species other than Arabidopsis thaliana were included (this resulted in a largely reduced set of 65 total motifs). See **Figure 7** legend for more details.

Figure S6 (Related to Figure 2). DBD class thresholds obtained for different measures of 8-mer similarity and different replicate percentiles. For each DBD class, DBD AA %ID inference threshold (based on precisions across DBD AA %ID bins) is plotted for each of the eight measures of 8-mer similarity (see key at top – "Escore45" considers 8-mers with E-scores >= 0.45; "Zscore10" considers the top 10 8-mers based on Z-scores, etc.). DBD classes are sorted (left to right) in decreasing order of number of PBM experiments performed in this study. Two plots are shown for each DBD class – the top plot ("25/75") shows the threshold obtained when using the 8-mer similarity score obtained for the 25th percentile of experimental replicates, and requiring a precision of at least 75% in each DBD AA %ID bin (see text for details). "20/80" refers to the 20th percentile, requiring 80% precision. See **Additional Data File 1** for the boxplots from which these thresholds were derived (same as those depicted in **Figure 2**), for all DBD classes.

Figure S7 (Related to Figure 5A). ChIP-seq comparison results obtained using an alternative null model. We repeated the procedure used to produce **Figure 5A** using an alternative set of background sequences. Instead of scrambled peak sequences (maintaining dinucleotide frequencies), we used real ChIP-seq peak sequences from an unrelated TF with the closest matching overall GC content in its peaks (see **Extended Experimental Procedures**).

Supplemental Tables

Source[1]	Data Type[2]	PMID[3]	Num Motifs[4]	Num TFs[5]	Description[6]
Matys 2006	Transfac	16381825	1366	847	Literature-compiled
Jolma 2013	SELEX	23332764	830	453	Human
Zhu 2011	B1H	21097781	564	298	Drosophila
Portales-Casamar 2010	JASPAR	19906716	301	286	Literature-compiled
Gerstein 2012	Chip-Seq	22955619	242	70	Human ChIP-seq
DeBoer 2011	YetFasco	22102575	232	198	Yeast Literature-compiled
Berger 2008	PBM	18585359	175	170	Mouse Homeodomains family
Badis 2008	PBM	19111667	110	110	Yeast
Badis 2009	PBM	19443739	106	105	Mouse
Zhu 2009	PBM	19158363	89	89	Yeast
Weirauch 2013	PBM	23354101	86	83	Mouse
Chen 2011	Chip-Seq	21450710	25	18	Mouse/Human ChIP-seq, compiled
Campbell 2010	PBM	21060817	23	18	Plasmodium
Lam 2011	PBM	21321018	23	23	Synthetic C2H2 ZFs
Wei 2010	PBM	20517297	22	22	Ets family
Chang 2013	PBM	23795294	16	16	Arabidopsis
Grove 2009	PBM	19632181	10	10	Worm bHLH family
Berger 2006	PBM	16998473	5	5	Five TFs (original PBM study)

Table S1 (Related to Figure 1B). Sources of other TF motifs.
[1] First author and year of publication
[2] Category of experimental data
[3] Pubmed ID of publication
[4] Number of motifs characterized in the study
[5] Number of unique TFs characterized in the study
[6] Brief description of TFs characterized in the study

Table S2 (Related to Figure 1B). Full results of motif comparisons to other studies.
Each row compares one pair of motifs for a single TF. Motif IDs correspond to the CisBP database (build 0.90). Final column indicates the negative log of the p-value, corresponding to the significance of the similarity of the motif pair, as calculated by TomTom.
Available as "TabS2_Full_motif_comparisons_to_other_studies.xlsx".

Table S3, part 1 (Related to Figure 4). Motif coverage by species and DBD class.
Species (and DBD classes) are sorted by the total number of motifs characterized in this study (then by the total number of TFs). Some species only have a small number of TFs because their proteome was not available for this study, even though they have some TFs with motifs characterized in this or another study. TFs characterized that have < 97% DBD identity to any TF in any species are categorized as "PBM CONSTRUCTS".
Available as "TabS3_Motif_coverage_by_species_and_DBD_class.xlsx".

Table S4 (Related to Figure 5). Full ChIP-seq AUROC results.
AUROC scores measuring PWM performance on human ChIP-seq data from ENCODE (results summarized in Figure 5). For Figure 5A, we compared the performance of PBM-derived PWMs, as a function of %DBD identity. These results correspond to rows with "Data Type" PBM, at the various ranges of "DBD ID" for each "Family". For Figures 5B and 5C, we compared the performance of directly determined motifs obtained from PBMs, the Transfac database, and a recent human HT-SELEX study. These results correspond to rows with each associated "Data Types", only considering directly determined motifs (i.e., ones with "DBD ID" equal to 1).
Available as "TabS4_Full_ChIP-seq_AUROC_results.xlsx".

Organism	**r^2**	**N**
Trichomonas vaginalis	0.30	17
Dictyostelium discoideum	0.16	46
Oryza sativa	0.13	277
Arabidopsis thaliana	0.13	337
Neurospora crassa	0.11	217
Homo sapiens	0.04	633
Danio rerio	0.03	587
Saccharomyces cerevisiae	0.00	255

Table S5 (Related to Figure 6). Motif GC content vs mean enrichment Z-score.
For each organism depicted in **Figure 6**, the r^2 value was calculated (across all of its motifs) between the motif's total GC content (i.e., the average across all of its positions) and its mean enrichment across all promoter positions relative to the TSS. 'N' indicates the total number of motifs.

Table S6, DBD clone source material.
This spreadsheet provides information on the clone source material, the experimental construct sequences, and the clone source contributors.
Available as "TabS6_DBD_clone_information.xlsx",

List of files available in “Supplementary Data 1”

AT-hook E-score Scatterplots (.png image)
This file depicts pairwise scatterplots of all 32,896 E-scores, for all pairs of TF plasmids assaying AT-hook TFs. The Pearson correlation, and its rank within all AT-hook pairs is indicated below each plot.

Boxplots for all DBD classes (zipped directory of .png images)
These files provide boxplots similar to those depicted in **Figure 2**, for all DBD classes and all eight measures of 8-mer similarity.

DBD amino acid alignments for all DBD classes (gzipped tarball of directories)
These files provide the DBD amino acid alignments for all DBD classes included in this study. See README.txt in root directory for more information.

Full results from leave-one-out cross validation analysis (.xlsx).
This file contains the results of cross validation analysis of motif inferences.

Full results from human disease-associated genetic variants analysis (.xlsx).
This file contains all human TFs predicted to differentially bind each disease-associated SNP.