Supplementary Methods

Correction of GC bias

In the process of ChIP sequencing, GC bias may occur during the PCR amplification (33). DNA fragments with specific GC contents may be over amplified. Since the control library and the ChIP-seq library are generated under the same conditions, we assume that the ChIP-seq library has the same GC bias as in the control library. To correct the GC bias in the ChIP-seq library, firstly the GC content distribution of DNA fragments in a window size of 300 bp (50 bp upstream and 250 bp downstream of the starting position of the tag) is calculated in the control library and in random genomic regions. Secondly the tag enrichment at each GC content (Figure 1A), i.e., the ratio of the GC content distribution in the control library to the GC content distribution in random genomic regions, is calculated. Finally for each tag in the ChIP-seq library, the number of occurrences can be corrected by dividing the tag enrichment according to the tag's GC content in the 300 bp window size. For extreme low and high GC contents where only few tags can be found, the tag enrichment is controlled to be above 1/3 to avoid over correction.

Construction of smoothing weights (Tag distribution at binding sites)

Since the tags are only the ends of the ChIPed DNA fragments, to generate the ChIP-seq profile (Y_{obs}) from the GC-corrected tags, we need a tag distribution relative to the real TF-DNA binding sites. Thus the tag distribution can be used as a smoothing weights to smooth tags to get the actual ChIPed DNA fragments profile. The ChIP-seq profile is then generated by piling up the ChIPed DNA fragments in the 1-kb regions. Starting from a uniform distribution, the smoothing weights can be iteratively improved by minimizing the SPE of the constructed ChIP-seq profile based on the smoothing weights until the SPE is smaller than a specified threshold. The smoothing weights is derived only in those regions which have a single ChIP-seq peak to ensure that the derived smoothing weights represents the real tag distribution at binding sites.

Evaluation of other motif discovery algorithms

In general, parameters with default values were used for the algorithms we evaluated. To improve the performance of MatrixREDUCE, an equal number of background sequences which are 500 bp upstream of the input sequences in ChIP-seq binding regions were also provided as the input. No gap and no flank nucleotides were allowed. For MEME, -mod zoops and -revcomp were specified. ChIPMunk was run under the peak model and the ZOOPS mode. For Weeder and MEME, the performance of prediction decreases as the length of the input sequences increases. For this reason, a 200-bp detection region is generally recommended for Weeder, MEME and DREME (34-35). We found, similarly, that the performance of MatrixREDUCE dropped when the detection window was enlarged to 1 kb (data not shown). Therefore, when running Weeder, MEME, DREME, and MatrixREDUCE, only the central 200-bp window of the 1-kb binding regions was used for motif discovery. For Weeder, MEME, and DREME, the input sequences were hard-masked for repeats using RepeatMasker (Smit, A.F.A., Hubley, R. and Green, P. RepeatMasker at http://repeatmasker.org), since that is the recommended procedure when using these algorithms. TherMos, MatrixREDUCE and ChIPMunk were run on unmasked sequence.

Experimental validation by Electrophoretic Mobility Shift Assays (EMSA)

Cloning of Esrrb/DBD Proteins

The DNA binding domain (DBD) of Esrrb was PCR-amplified from a mouse cDNA (IMAGE: 4030874; NCBI accession: BC132597) using primers containing consensus sequences to perform the GATEWAY BP reaction (Invitrogen) with a tobacco etch virus protease cleavage site at the 5' end.

Forward Primer:

5' GGGGACAAGT TTGTACAAAA AAGCAGGCTT CGAAAACCTG

TATTTTCAGGGC AACGCCATCCCCAAGCGC

Reverse Primer:

5' GGGGACCACTTTGTACAAGAAAGCTGGGTTTAGCTGTTCTCCGAATCCAGC

(gene-specific sequence underlined)

Purified PCR products were cloned into the pDONR221 vector (Invitrogen) by employing the GATEWAY BP technology to obtain the entry clone pENTR- *Esrrb/DBD*, spanning amino acid residues 96-194 of the full length Esrrb protein, which corresponds to the Esrr2 NMR construct 1LO1 except for C163A point mutation. Expression plasmids were generated using the GATEWAY LR cloning (Invitrogen) to obtain a pDEST-*HisMBP- Esrrb/DBD* construct encoding an N-terminal HisMBP ta*g*.

Purification of HisMBP-Esrrb/DBD Proteins

Expression plasmids were transformed into BL21-CodonPlus (DE3)-[RIPL] cells (Stratagene). Cell were grown in terrific broth (TB) supplemented with 100 ug/ml of ampicillin at 37 ^{o}C. When an OD_{600} of 0.6 was reached the temperature was lowered to 18 ^{o}C and protein expression was induced by adding 0.5mM isopropyl β-D-1-thiogalactopyranoside for 16 h. The bacteria cells were pelleted, resuspended in lysis buffer (10mM HEPES, pH7.3, 100mM NaCl) and lysed by sonification. The fusion protein was extracted from the soluble fraction at 4°C using Amylose Resin (New England Biolabs) following the manufacturer instructions. The fusion tag was removed by cleaving with the TEV protease overnight at 4 °C. The protein was further purified by cation-exchange chromatography using a Resource

S column (GE Healthcare) and eluted with a linear NaCl gradient. Fractions containing the EsrrbDBD proteins were pooled and desalted into a buffer containg 10mM Tris-HCl pH 8.0, 100mM NaCl., 2mM TCEP (tris(2-carboxyethyl)phosphine) using PD-10 Desalting Columns (GE Healthcare). The protein was then concentrated using Vivascience 3000 MWCO concentrators (Sartorius) and stored at − 80 °C.

Production of recombinant Klf4 protein

The DNA binding domain of mouse Klf4 (amino acids 398-483) was PCR amplified and cloned using the GATEWAY technology (Invitrogen), expressed in *Escherichia coli* BL21(DE3) cells with a cleavable N-terminal NusA-His6 tag and purified to homogeneity using procedures described elsewhere (36). The protein was quantified by measuring the absorbance at 280nm, frozen in liquid N_2 and stored in aliquots at -80°C until usage. Repeated freeze-thaw cycles were avoided.

EMSA for Esrrb and Klf4

EMSA studies were performed as reported previously (37) with the following modifications. For competition assays a master mix containing the 14bp dsCy5-labeled Esrrb element *(5'-*AGCCAAGGTCACCA *-3')* derived from the exon of the *Tmem91* gene and the EsrrbDBD protein was mixed with an excess of unlabeled dsDNA competitor. Final concentrations were 1nM cy5-labeled DNA, either 20nM. 200nM or 500nM competitor DNA, and 15.625nM protein in a reaction buffer containing 2mM β-mercaptoethanol, 10mM Tris-HCl pH 8.0, 100mM KCl, 50µM $ZnCl_2$, 2mM $MgCl_2$ 10%Glycerol, 0.1%NP-40 and 0.1mg/mL BSA (New England Biolabs). In the absence of the unlabeled dsDNA competitor, the reactions were first incubated for 1 hour on ice and were further incubated for another 30mins after the addition of the competitor. The reactions were then electrophoresed at 4˚C in the dark on a pre-run tris-glycine (TG) 12% native polyacrylamide gel at 200V for 30mins using 1X TG running buffer. The gel was imaged using a Typhoon 9140 PhosphorImager (GE Healthcare) and bound and unbound samples were quantified using the ImageQuant TL software (GE Healthcare).

For single mutation competition assays a master mix containing the 15bp dsCy5-labeled Klf4 element *(5'- CATAGGGTGTGGTCA -3')* derived from the control region downstream of the *Uck2* gene and the Klf4 protein was mixed with an excess of unlabeled dsDNA competitor. Final concentrations were 1nM cy5-labeled DNA, either 10nM, 500nM or 1000nM competitor DNA and 10nM protein in a reaction buffer containing 2mM β-mercaptoethanol, 10mM Tris-HCl pH 8.0, 100mM KCl, 50μM $ZnCl_2$, 2mM $MgCl_2$ 10%Glycerol, 0.1%NP-40 and 0.1mg/mL BSA (New England Biolabs). The reactions were incubated for 1 hour on ice and electrophoresed at 4˚C in the dark on a pre-run tris-glycine (TG) 12% native polyacrylamide gel at 200V for one hour using 1X TG running buffer. The gel was imaged using a Typhoon 9140 PhosphorImager (GE Healthcare) and bound and unbound samples were quantified using the ImageQuant TL software (GE Healthcare). For multiple mutation competition assays, a master mix containing the 14bp dsCy5-labeled Klf4 element (*5'-AGAGGGCGTATTCAG-3'*) and the Klf4 protein was mixed with an either 10-fold or 30-fold excess of unlabeled dsDNA competitor.

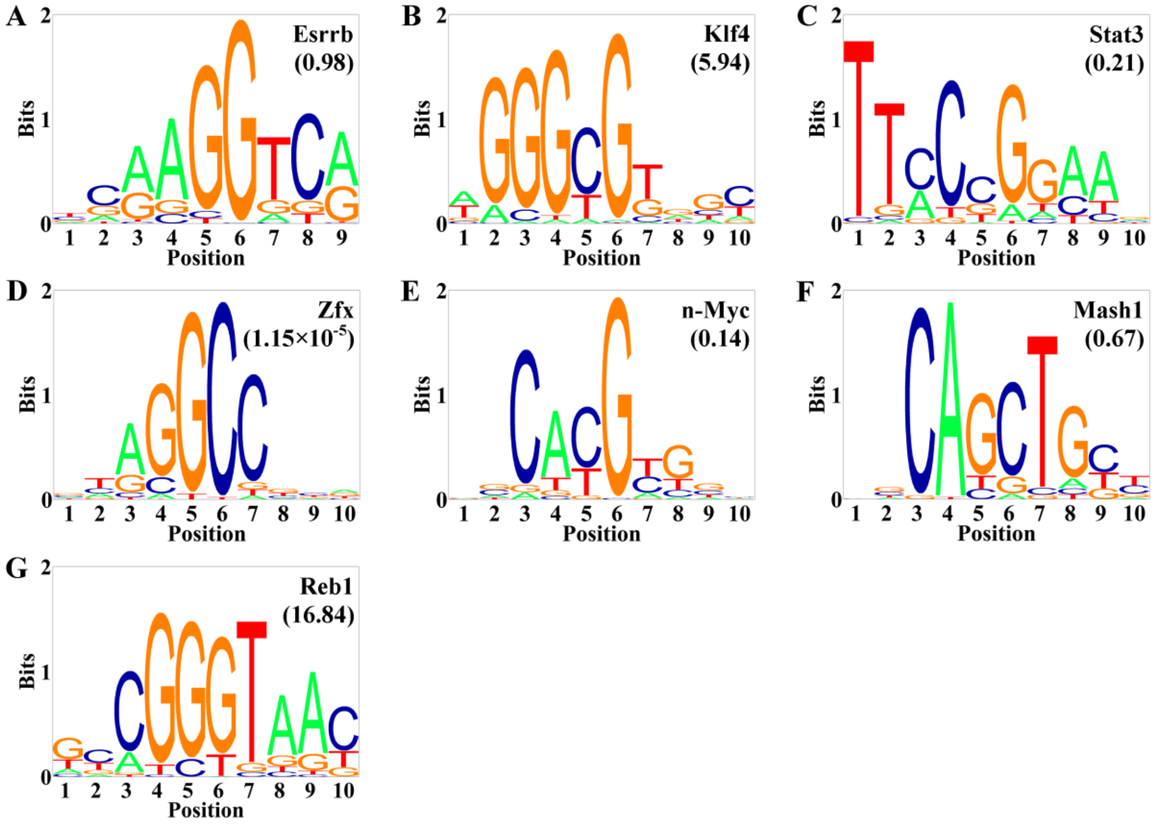
Comparison of PSEMs in Euclidean distance

For an n-mer sequence, the Euclidean distance can be calculated as

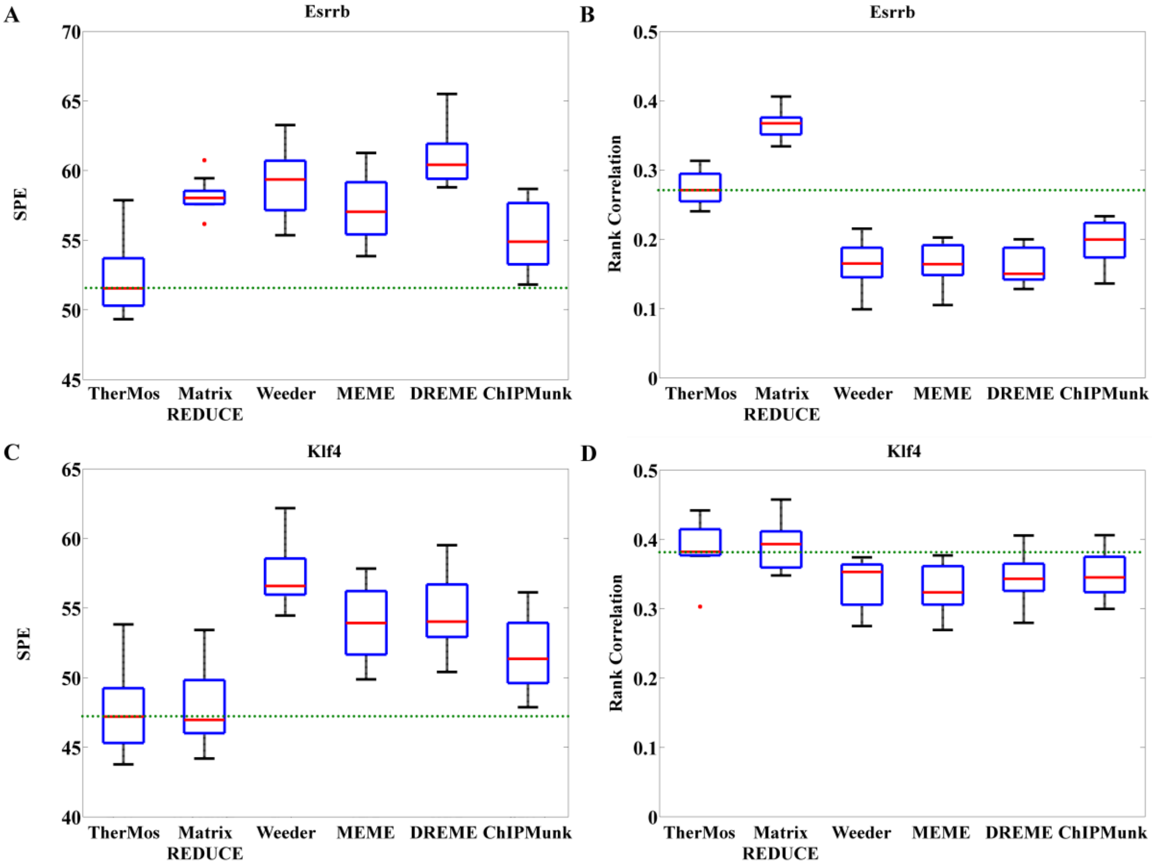
$$\mathrm{ED} = \sqrt{\frac{\sum_i \sum_j (G_{ij}\,(\mathrm{Predicted}) - G_{ij}\,(\mathrm{EMSA}))^2}{n}} \qquad \{22\}$$

where G_{ij} is the entry at the row i and column j in the position frequency matrix.

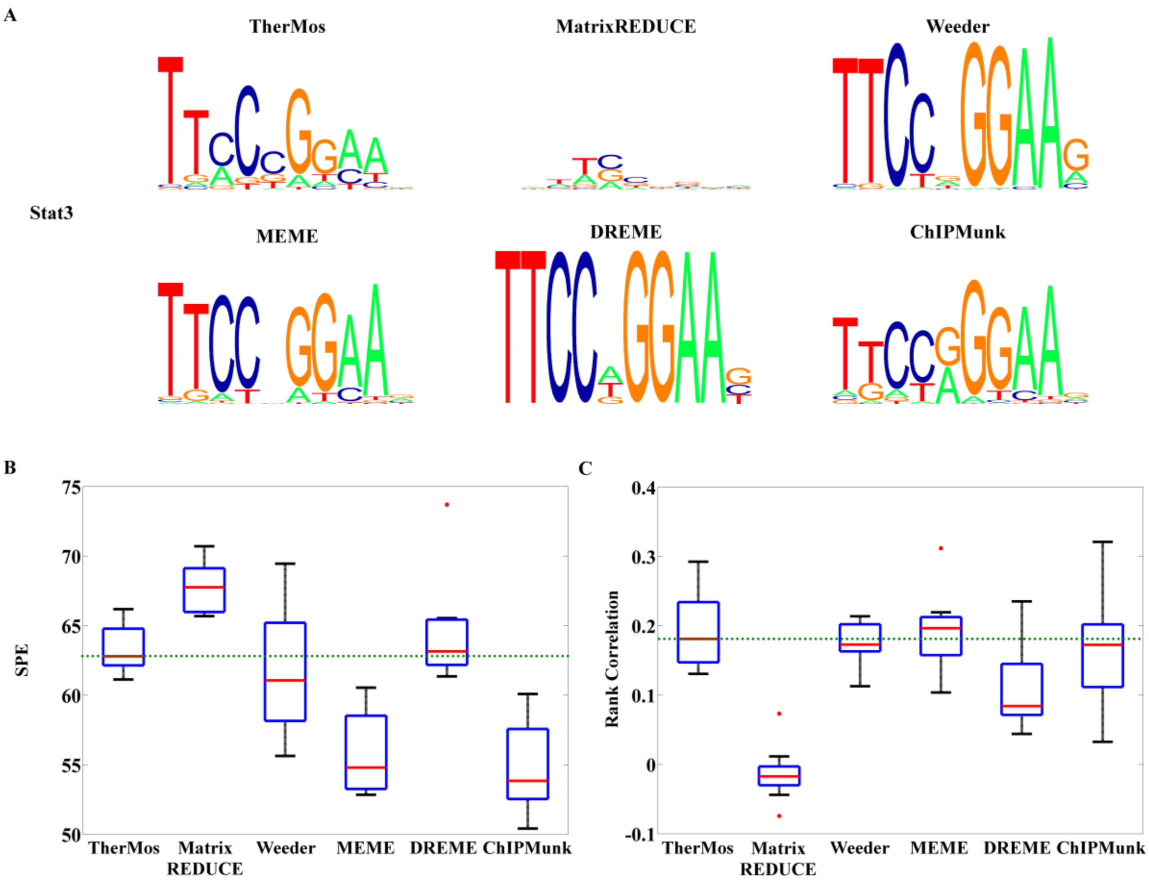
Supplementary Figures



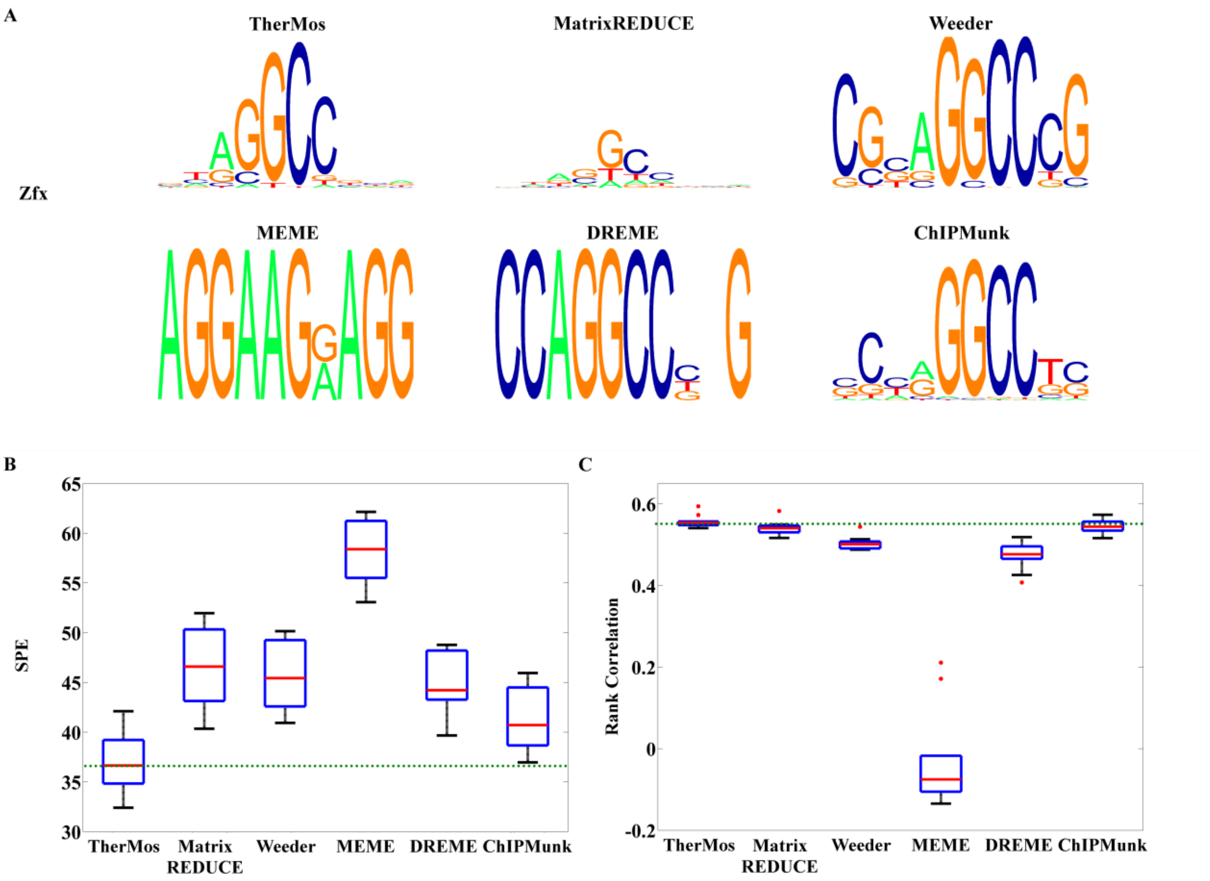
S1 Sequence logos of the PSEMs predicted by TherMos. [*TF*]/K_d*(ref)* values are shown in parentheses. The reference sequence is the consensus. (**A**) Esrrb. (**B**) Klf4. (**C**) Stat3. (**D**) Zfx. (**E**) n-Myc. (**F**) Mash1. (**G**) Reb1.



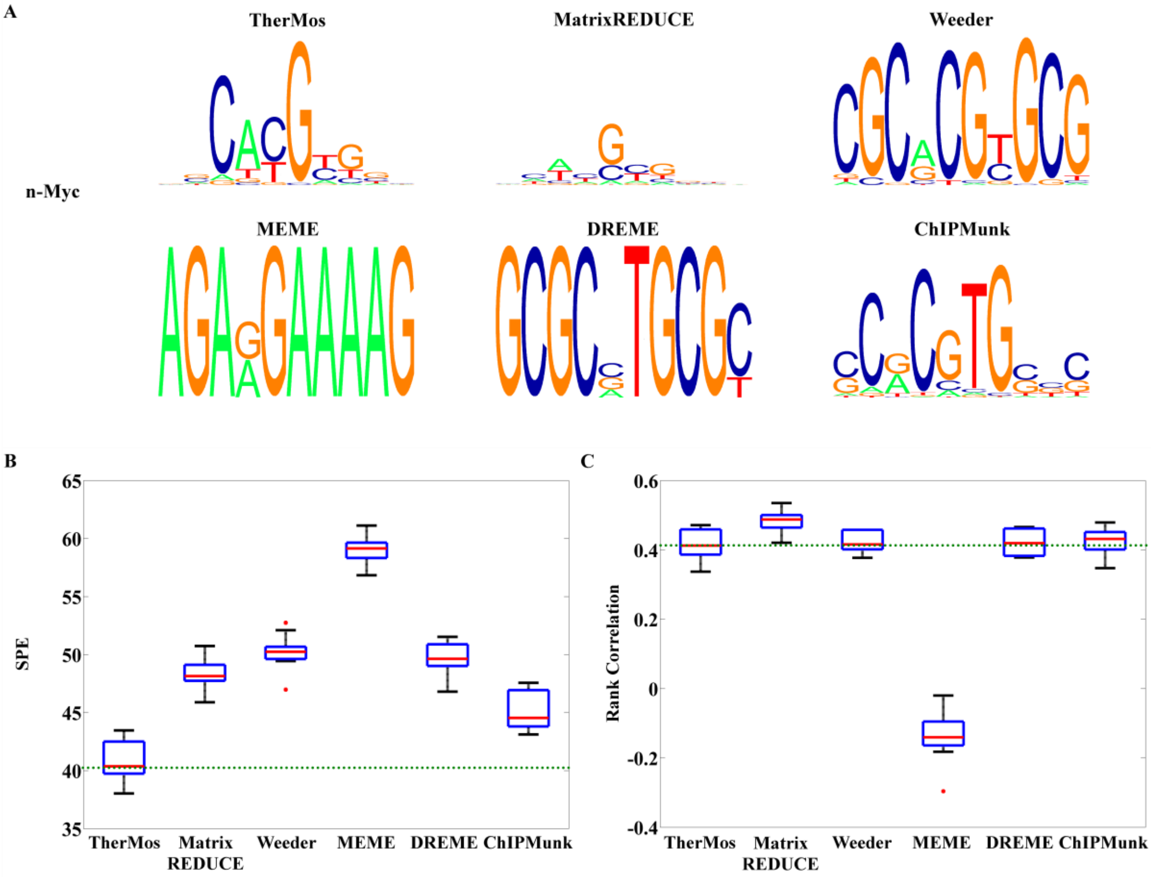
S2 10-fold cross-validation tests for Esrrb and Klf4. SPE is calculated between predicted (motif) and observed (experimental data) ChIP-seq binding profile. Rank correlation is calculated between predicted (motif) and observed (experimental data) ChIP-seq tag counts. (**A**) SPEs of the cross-validation test for Esrrb predicted by TherMos, and five other algorithms. (**B**) Rank correlation coefficients of the cross-validation test for Esrrb predicted by TherMos, and five other algorithms. (**C**) SPEs of the cross-validation test for Klf4 predicted by TherMos, and five other algorithms. (**D**) Rank correlation coefficients of the cross-validation test for Klf4 predicted by TherMos, and five other algorithms.



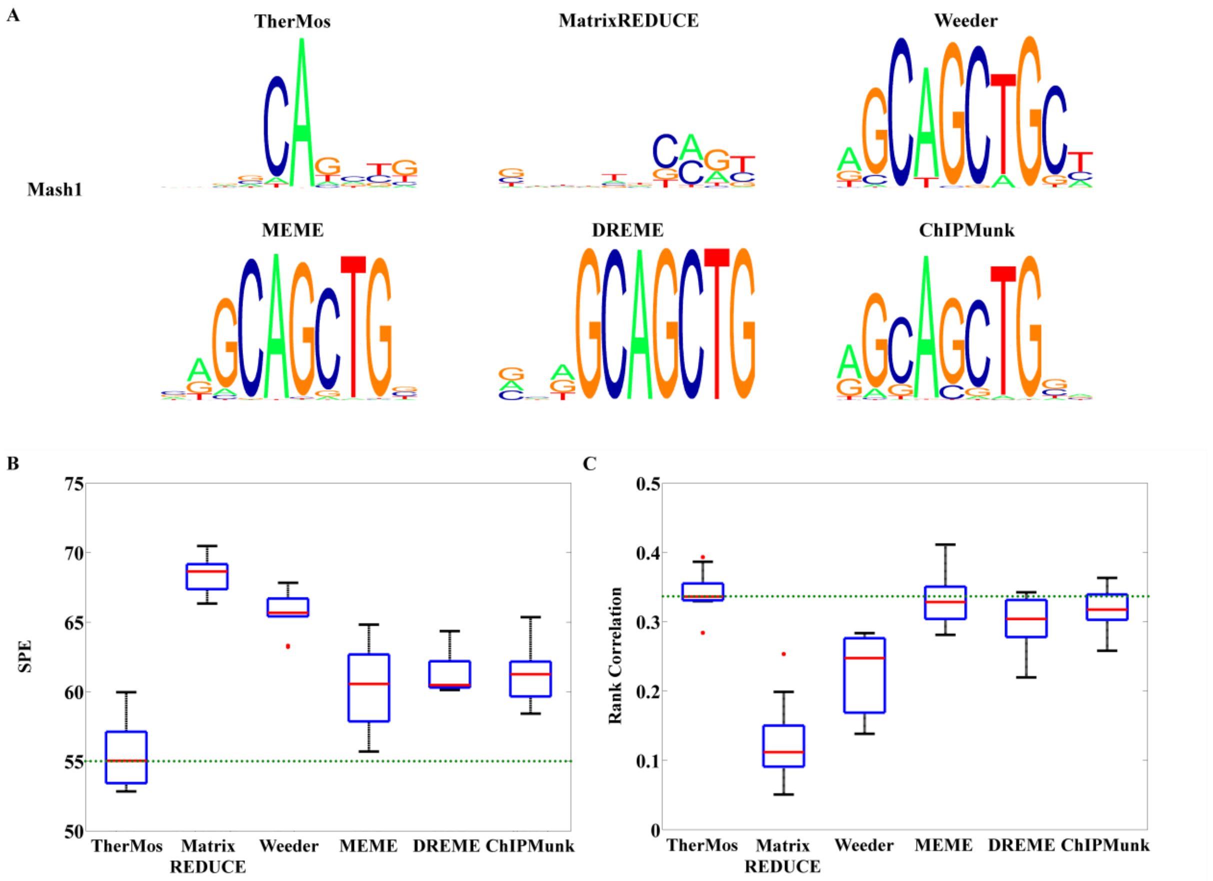
S3 10-fold cross-validation tests for Stat3. SPE is calculated between predicted (motif) and observed (experimental data) ChIP-seq binding profile. Rank correlation is calculated between predicted (motif) and observed (experimental data) ChIP-seq tag counts. (**A**) Sequence logos of the motifs predicted by TherMos, MatrixREDUCE, Weeder, MEME, DREME and ChIPMunk for the test which predicts the lowest SPE. (**B**) SPEs of the cross-validation test predicted by TherMos, and five other algorithms. (**C**) Rank correlation coefficients of the cross-validation test predicted by TherMos, and five other algorithms.



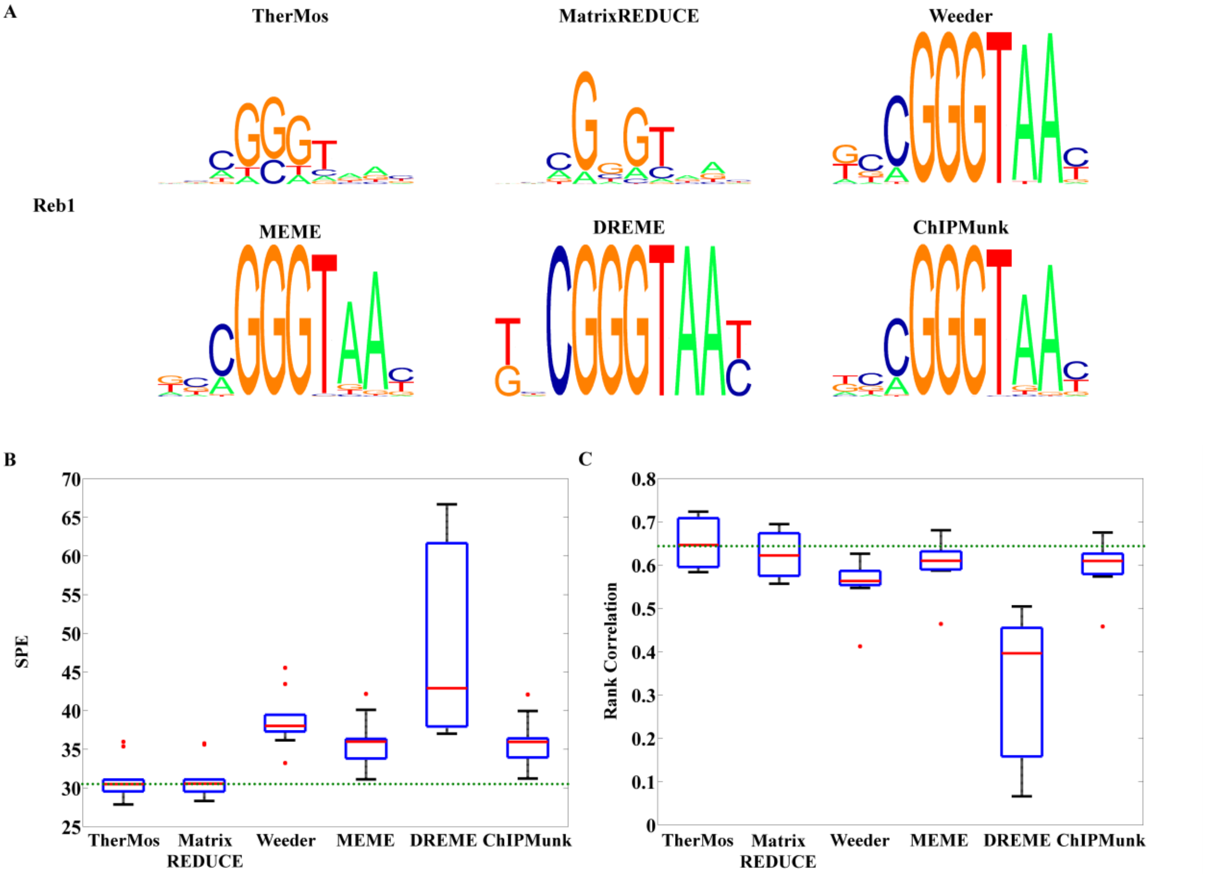
S4 10-fold cross-validation tests for Zfx. SPE is calculated between predicted (motif) and observed (experimental data) ChIP-seq binding profile. Rank correlation is calculated between predicted (motif) and observed (experimental data) ChIP-seq tag counts. (**A**) Sequence logos of the motifs predicted by TherMos, MatrixREDUCE, Weeder, MEME, DREME and ChIPMunk for the test which predicts the lowest SPE. (**B**) SPEs of the cross-validation test predicted by TherMos, and five other algorithms. (**C**) Rank correlation coefficients of the cross-validation test predicted by TherMos, and five other algorithms.



S5 10-fold cross-validation tests for n-Myc. SPE is calculated between predicted (motif) and observed (experimental data) ChIP-seq binding profile. Rank correlation is calculated between predicted (motif) and observed (experimental data) ChIP-seq tag counts. (**A**) Sequence logos of the motifs predicted by TherMos, MatrixREDUCE, Weeder, MEME, DREME and ChIPMunk for the test which predicts the lowest SPE. (**B**) SPEs of the cross-validation test predicted by TherMos, and five other algorithms. (**C**) Rank correlation coefficients of the cross-validation test predicted by TherMos, and five other algorithms.



S6 10-fold cross-validation tests for Mash1. SPE is calculated between predicted (motif) and observed (experimental data) ChIP-seq binding profile. Rank correlation is calculated between predicted (motif) and observed (experimental data) ChIP-seq tag counts. (**A**) Sequence logos of the motifs predicted by TherMos, MatrixREDUCE, Weeder, MEME, DREME and ChIPMunk for the test which predicts the lowest SPE. (**B**) SPEs of the cross-validation test predicted by TherMos, and five other algorithms. (**C**) Rank correlation coefficients of the cross-validation test predicted by TherMos, and five other algorithms.



S7 10-fold cross-validation tests for Reb1. SPE is calculated between predicted (motif) and observed (experimental data) ChIP-seq binding profile. Rank correlation is calculated between predicted (motif) and observed (experimental data) ChIP-seq tag counts. (**A**) Sequence logos of the motifs predicted by TherMos, MatrixREDUCE, Weeder, MEME, DREME and ChIPMunk for the test which predicts the lowest SPE. (**B**) SPEs of the cross-validation test predicted by TherMos, and five other algorithms. (**C**) Rank correlation coefficients of the cross-validation test predicted by TherMos, and five other algorithms.

Supplementary References

33. Hillier, L.W., Marth, G.T., Quinlan, A.R., Dooling, D., Fewell, G., Barnett, D., Fox, P., Glasscock, J.I., Hickenbotham, M. and Huang, W. et al. (2008) Whole-genome sequencing and variant discovery in C. elegans. Nat. Methods, 5, 183-188.

34. Hu, J., Li, B. and Kihara, D. (2005) Limitations and potentials of current motif discovery algorithms. Nucleic Acids Res., 33, 4899-4913.

35. Valen, E., Sandelin, A., Winther, O. and Krogh, A. (2009) Discovery of regulatory elements is improved by a discriminatory approach. PLoS Comput. Biol., 5, e1000562. (10.1371/journal.pcbi.1000562)

36. Ng, C.K.L., Palasingamn, P., Venkatachalam, R., Baburajendran, N., Cheng, J., Jauch, R. and Kolatkar, P.R. (2008) Purification, crystallization and preliminary X-ray diffraction analysis of the HMG domain of the Sox17 in complex with DNA. Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun., 64, 1184-1187.

37. Jauch, R., Ng, C.K.L., Saikatendu, K.S., Stevens, R.C. and Kolatkar, P.R. (2008) Crystal structure and DNA binding of the homeodomain of the stem cell transcription factor Nanog. J. Mol. Biol., 376, 758-770.