Supplementary Materials for

Profiling the quantitative occupancy of myriad transcription factors across conditions by modeling chromatin accessibility data

Kaixuan Luo,[1,2,3,11] Jianling Zhong,[1,2,3] Alexias Safi,[2,4] Linda K. Hong,[2,4]
Alok K. Tewari,[5] Lingyun Song,[2,4] Timothy E. Reddy,[1,2,6,7,8] Li Ma,[1,9]
Gregory E. Crawford,[1,2,4] Alexander J. Hartemink[1,2,3,10,⋆]

[1]Computational Biology & Bioinformatics Graduate Program, Duke University,
[2]Center for Genomic & Computational Biology, Duke University,
[3]Department of Computer Science, Duke University,
Durham, NC 27708, USA
[4]Department of Pediatrics, Duke University Medical Center,
Durham, NC 27710, USA
[5]Department of Medical Oncology, Dana-Farber Cancer Institute,
Boston, MA 02215, USA
[6]Department of Biostatistics & Bioinformatics, Duke University Medical Center,
[7]Department of Molecular Genetics & Microbiology, Duke University Medical Center,
Durham, NC 27710, USA
[8]Department of Biomedical Engineering, Duke University,
[9]Department of Statistical Science, Duke University,
[10]Department of Biology, Duke University,
Durham, NC 27708, USA
[11]Department of Human Genetics, The University of Chicago,
Chicago, NC 60637, USA

[⋆]To whom correspondence should be addressed; E-mail: amink@cs.duke.edu.

Supplementary Tables

Supplementary Table S1 appears on the following page. Supplementary Tables S2, S3, and S4 are spreadsheets and have been uploaded separately.

- Supplementary Table S1: Different modeling frameworks for predicting TF binding using DNase-seq (or ATAC-seq) data.
- Supplementary Table S2: Combinations of TFs (motifs) and cell types used for training and testing TOP models with ATAC-seq and DNase-seq data as in Fig. 2.
- Supplementary Table S3: All motifs used for making predictions and screening as in Figs. 5 and 6.
- Supplementary Table S4: Pearson's correlations between predicted TF occupancy and TF expression as in Fig. 4C.

Supplementary Table S1. Different modeling frameworks for predicting TF binding using DNase-seq (or ATAC-seq) data.

	CENTIPEDE (2012)	MILLIPEDE (2013)	PIQ (2014)	BinDNase (2015)	msCentipede (2015)	TOP (this study)	GERV (2015)	Avocado (2020)
Has access to TF motifs as input?	Yes	Yes	Yes	Yes	Yes	Yes	No, uses k-mers instead	No
Genome-scale data used as input for TF prediction	One DNase-seq (or ATAC-seq) dataset	One DNase-seq (or ATAC-seq) dataset (ChIP-seq data used in training, but not used in prediction)	One DNase-seq (or ATAC-seq) dataset	One DNase-seq (or ATAC-seq) dataset (ChIP-seq data used in training, but not used in prediction)	One DNase-seq or ATAC-seq dataset (higher accuracy when multiple replicates are available)	One DNase-seq or ATAC-seq dataset (ChIP-seq data used in training, but not used in prediction)	Sequence variants (DNase-seq and ChIP-seq data used in training, but not used in prediction)	Trained on available ENCODE datasets including DNase-seq (or ATAC-seq), RNA-seq, and ChIP-seq both for TFs and for histone modifications. Transfer learning could be used for predicting new samples.
Motif/site-centric vs. genome-wide	Motif/site-centric	Motif/site-centric	Motifs modeled jointly	Motif/site-centric	Motif/site-centric	Motif/site-centric	Genome-wide	Genome-wide
Supervised vs. unsupervised	Unsupervised	Supervised	Unsupervised	Supervised	Unsupervised	Supervised	Supervised	Supervised
Independent model for each TF vs. joint model for TFs and/or cell types	Independent model for each TF	Independent model for each TF	Joint model for all TFs in each cell type	Independent model for each TF	Independent model for each TF	Joint model for all TFs in all cell types using a Bayesian hierarchical framework	Independent model for each TF	Joint model for all types of observations in all cell types
How DNase/ATAC data are used in the model	Nucleotide-resolution DNase digestion profile around motifs	Binned DNase digestion profile around motifs (same binning scheme for all TFs)	Nucleotide-resolution GP-smoothed DNase digestion profile around motifs	Binned DNase digestion profile around motifs (different binning scheme for each TF)	Multiscale nucleotide-resolution DNase/ ATAC digestion profile around motifs	Binned DNase/ATAC digestion profile around motifs (same binning scheme for all TFs)	Nucleotide-resolution binary (0/1) indicator variable denoting presence of a DNase read across genome	Multiscale binned DNase digestion profile across genome (at 25bp, 250bp, and 5000bp resolutions)
Output type and interpretation	TF binding probability	TF binding probability	TF binding probability	TF binding probability	TF binding probability	Quantitative ChIP-seq signal expressed as counts (in window ±100bp around motif) TF binding probability when using logistic version of TOP	Quantitative ChIP-seq signal expressed as counts (in window ±200bp around k-mer)	Quantitative ChIP-seq signal expressed as signal *p* value
Programming language or environment	R	R	R	R	Python	R	Python	Deep-learning using Keras with the Theano backend on GPUs

Supplementary Figures

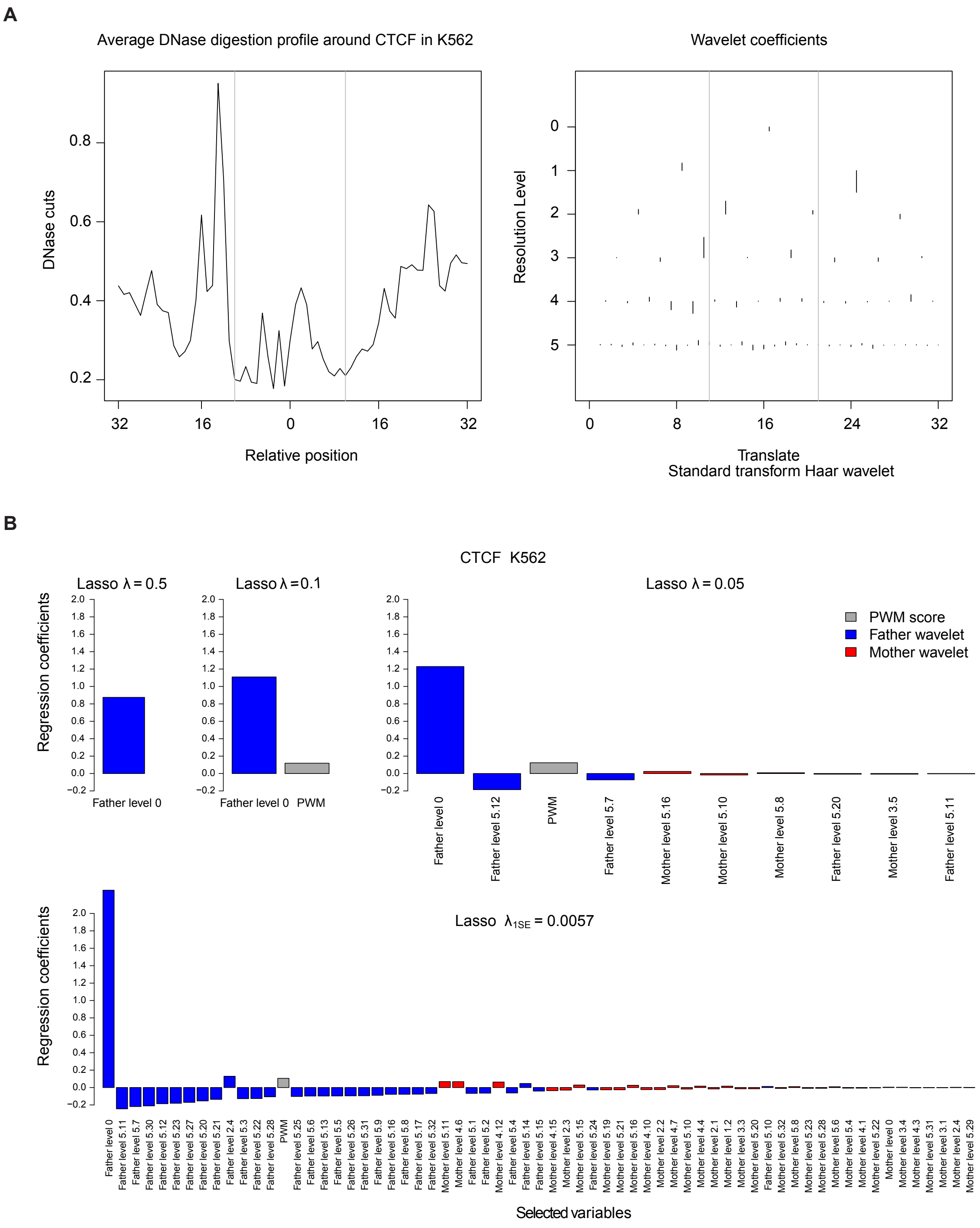

Fig. S1. Details of wavelet decomposition of CTCF DNase digestion profiles. (**A**) Example DNase digestion profile and Haar wavelet coefficients. (Left) Average DNase digestion profiles around CTCF motif matches in K562 cell type in a window of size 64 bp. (Right) Haar wavelet coefficients at different resolution levels for those DNase digestion profiles. (**B**) Selecting regression coefficients for CTCF in K562 cell line using Lasso with λ equals 0.5, 0.1, 0.05, or λ_{1SE} (largest value of λ such that mean cross-validated error is within one standard error of the minimum). Selected variables are ordered by the absolute value of coefficients. PWM score is in gray, mother wavelet coefficients are in red, and log2-transformed father wavelet coefficients are in blue. The names of the wavelet coefficients (bars) show the location indices of the wavelet coefficients (e.g., ‘Father level 5.12’ refers to the 12th father wavelet coefficient at resolution level 5).

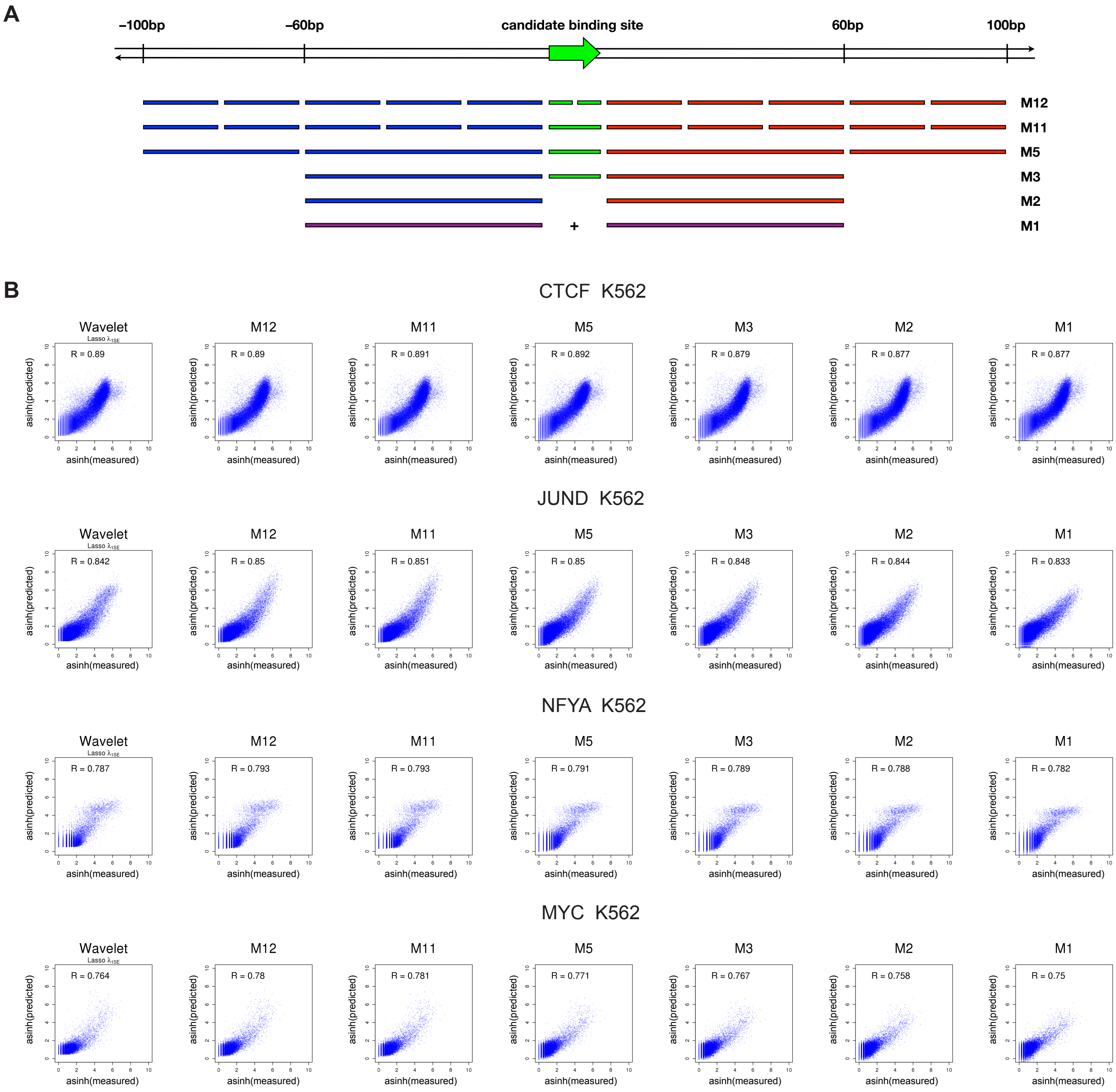

Fig. S2. Details of MILLIPEDE binning and performance in comparison with wavelet models. (**A**) Understanding the relationships between bins in various MILLIPEDE models. All bins are defined relative to the strand orientation of the candidate binding site: green bins are within the binding site, blue bins are upstream, red bins are downstream, and purple bins in the M1 model sum together cleavage events from the M2 model's blue and red bins. Models are arranged from most to least complex (Luo and Hartemink, 2013). In this work, we use model M5, which has two upstream bins, a bin spanning the motif site, and two downstream bins. (**B**) Examples of prediction performances in 5-fold cross-validation using DNase wavelet coefficients and different MILLIPEDE bins. The wavelet models used variables selected using Lasso with λ_{1SE} (largest value of λ such that mean cross-validated error is within one standard error of the minimum). The M5 model outperforms the wavelet model in all these examples.

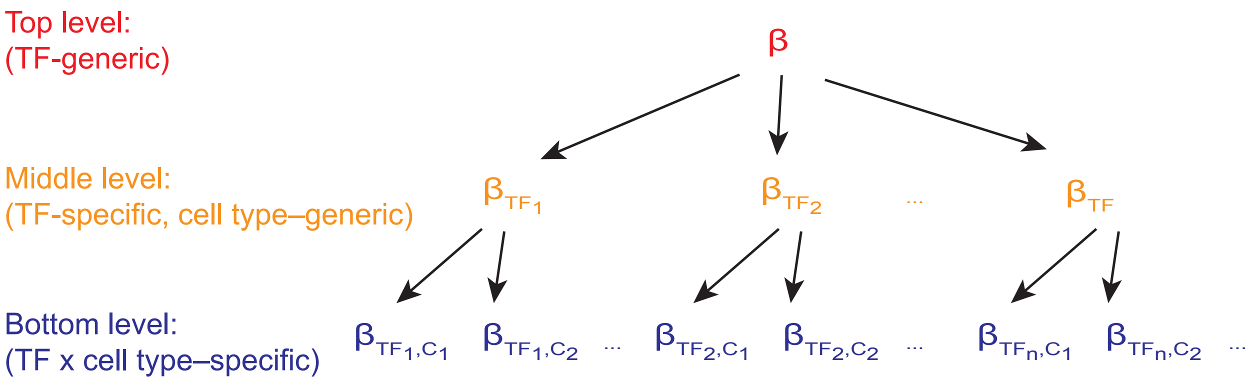

Fig. S3. Bayesian hierarchical model structure. Our Bayesian hierarchical model aims to jointly estimate multiple parameters from multiple TFs across multiple cell types. Instead of learning them separately by fitting one TF in one cell type separately as in our earlier work (Luo and Hartemink, 2013), we regard the parameters to be related by the hierarchical structure as shown in the figure. A key feature of Bayesian hierarchical model is that by leveraging the grouping structure, model parameters are organized in a hierarchy to allow for borrowing or sharing of information, and the observed data can be used to estimate the population distribution even though the population level parameters may not be directly observed. Intuitively, from our biological knowledge and empirical observations, a TF would have similar DNA binding signatures or footprints among different cell types, therefore, the parameters of a TF in different cell types are very related (cell type–generic). Similarly, different TFs also share similar DNA binding footprint profiles (TF-generic), as we often observe a depletion in the motif region surrounded by evaluated DNase- or ATAC-seq signals in the nearby flanking regions. In our hierarchical model structure, at the bottom level are regression parameters specific to a particular TF × cell-type combination. For each TF, the bottom level parameters are themselves drawn from a shared distribution for the TF at the middle level. Likewise, the parameters associated with each TF's distribution at the middle level are themselves drawn from a single shared distribution for all TFs at the top level.

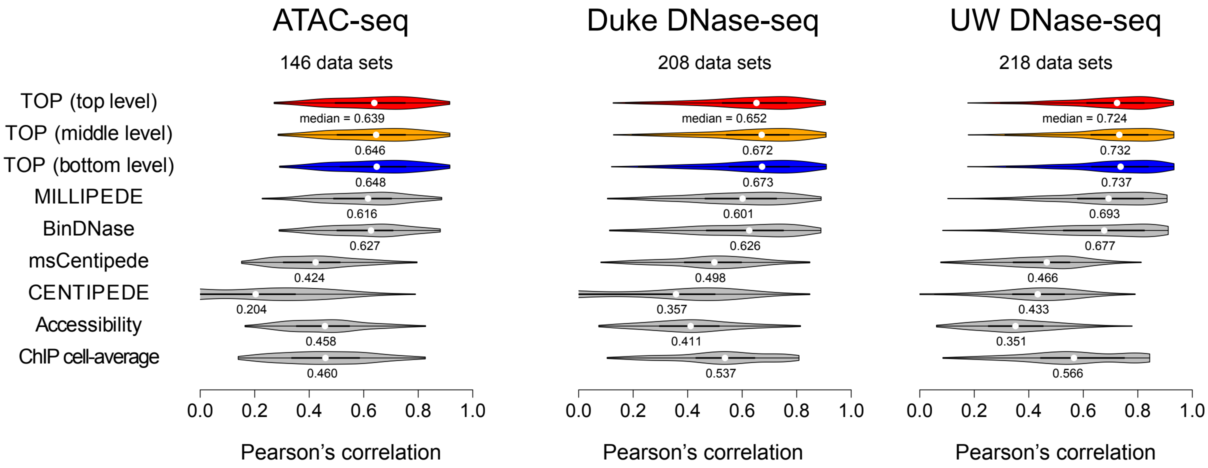

Fig. S4. Prediction performance evaluated with ChIP cell-average occupancy baseline. Prediction performance of different methods was evaluated using quantitative TF occupancy at motif matches, similar to Fig. 2B. Here we added into our benchmark a ChIP cell-average occupancy baseline, where the occupancy of a TF in a cell type was represented by the measured ChIP occupancy of this TF in the other cell types. To obtain the average ChIP occupancy from multiple cell types, we only included TFs with available ChIP-seq data in at least three cell types (Supp. Table S2), thus, the TF × cell-type combinations included here are slightly fewer than those included in Fig. 2B.

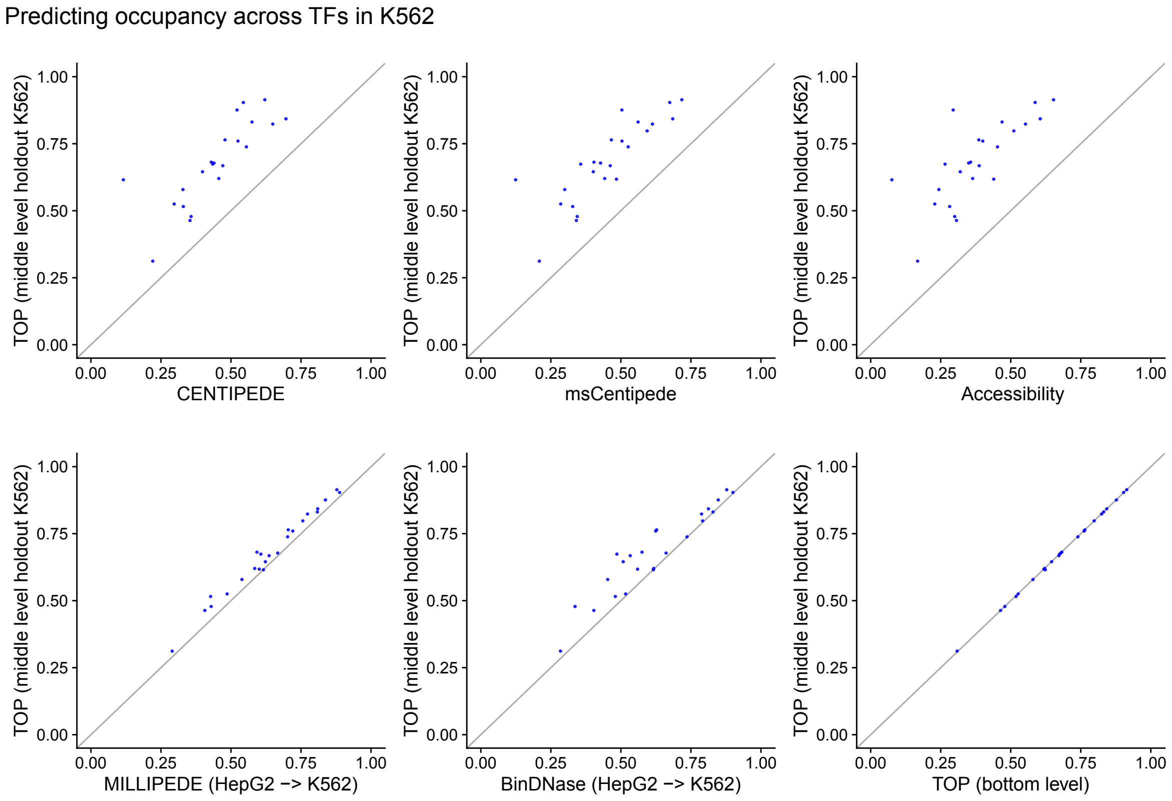

Fig. S5. Prediction performance for cross-cell type predictions. Comparing the performance of TOP with other methods in cross-cell type prediction setting. Similar to Fig. 2C, a 'held-out' version of the TOP model was trained after holding out the K562 cell type from training set. Predictions were then made for all TFs in K562 in the test chromosomes using TOP model (middle level parameters) trained with the held-out training set. Results were compared with predictions in K562 made by CENTIPEDE, msCentipede, MILLIPEDE, and BinDNase trained with HepG2 data, total chromatin accessibility, as well as TOP model trained using the full training set. Each dot represents the Pearson's correlations between predicted and measured occupancy for one TF. Shown are results from UW DNase-seq data. Duke DNase-seq and ATAC-seq results are similar.

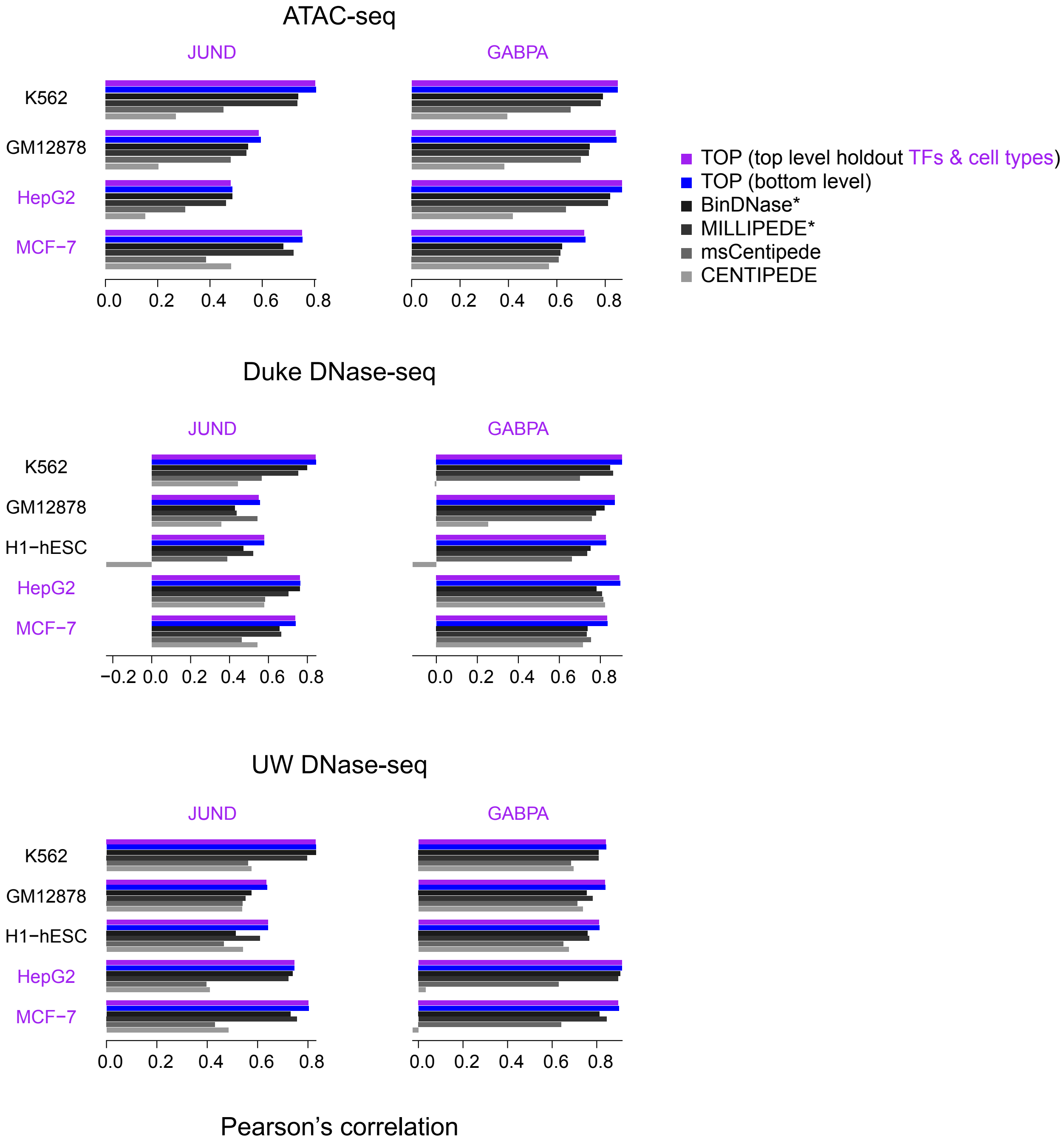

Fig. S6. Prediction performance for held-out TFs and cell types. Similar to Figs. 2C,D, 'held-out' versions of the TOP model were trained after holding out a subset of TFs (JUND and GABPA, along with all related TF family members with similar motifs) and cell types (MCF-7 and HepG2) from training sets. Then predictions were made for the TFs and cell types in the test chromosomes using the held-out models. Results were compared with TOP model (bottom level) trained using full training set, BinDNase, and MILLIPEDE trained with these TFs and cell types, as well as CENTIPEDE and msCentipede. The TFs and cell types holding out from training set are in purple. Note, BinDNase and MILLIPEDE (by default) do not make predictions for held-out TFs. They were include here simply as 'upper bound' references when training data of the exact same TFs and cell types of interest are available.

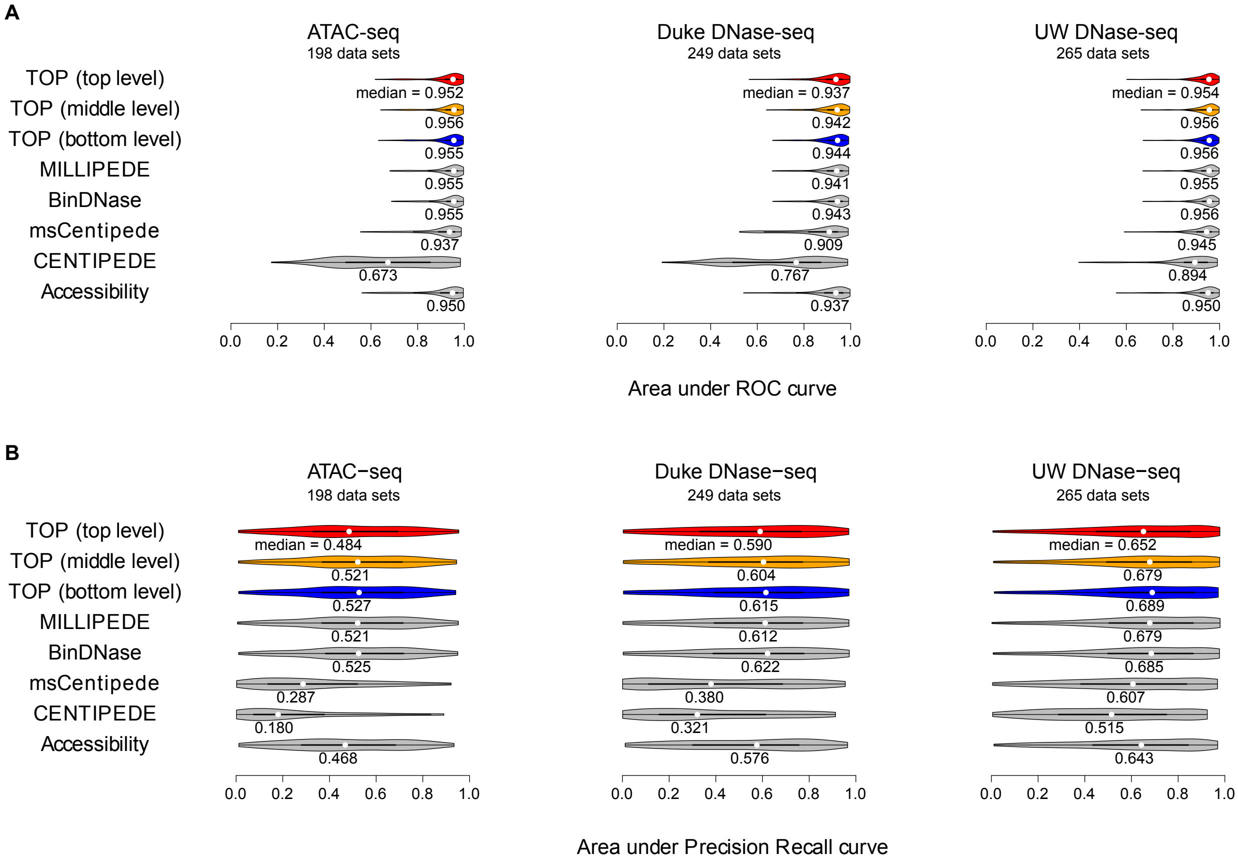

Fig. S7. Prediction performance in the binary TF binding setting. Though the main focus of the paper is on quantitative occupancy predictions, a logistic version of the TOP model was trained for binary TF binding prediction. Specifically, candidate sites were labeled as bound if they overlapped with ChIP-seq peaks, and unbound otherwise. A logistic version of the TOP model was trained using binary ChIP labels. Predictions of different methods in the test chromosomes were evaluated with binary ChIP labels using the area under ROC curve (AUROC) in panel **A** and the area under precision recall curve (AUPR) in panel **B**.

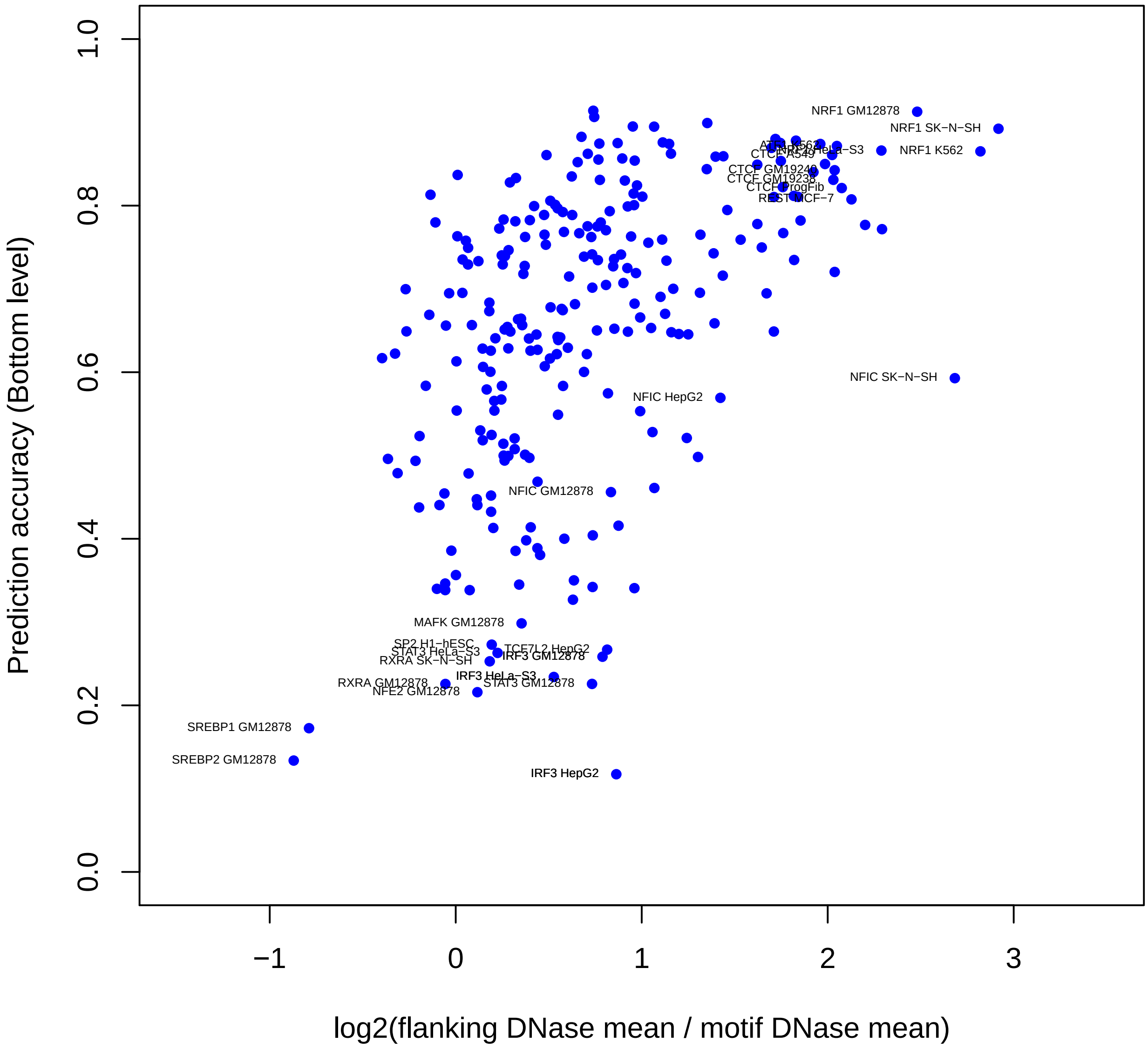

Fig. S8. Prediction accuracy (bottom level) as a function of DNase depletion ratio. DNase depletion ratio was calculated as the log ratio of the average number of DNase cleavage events in the 60 bp proximal flanking regions divided by the average number of DNase cleavage events within the motif itself. Each dot represents one TF × cell-type combination in the Duke DNase data.

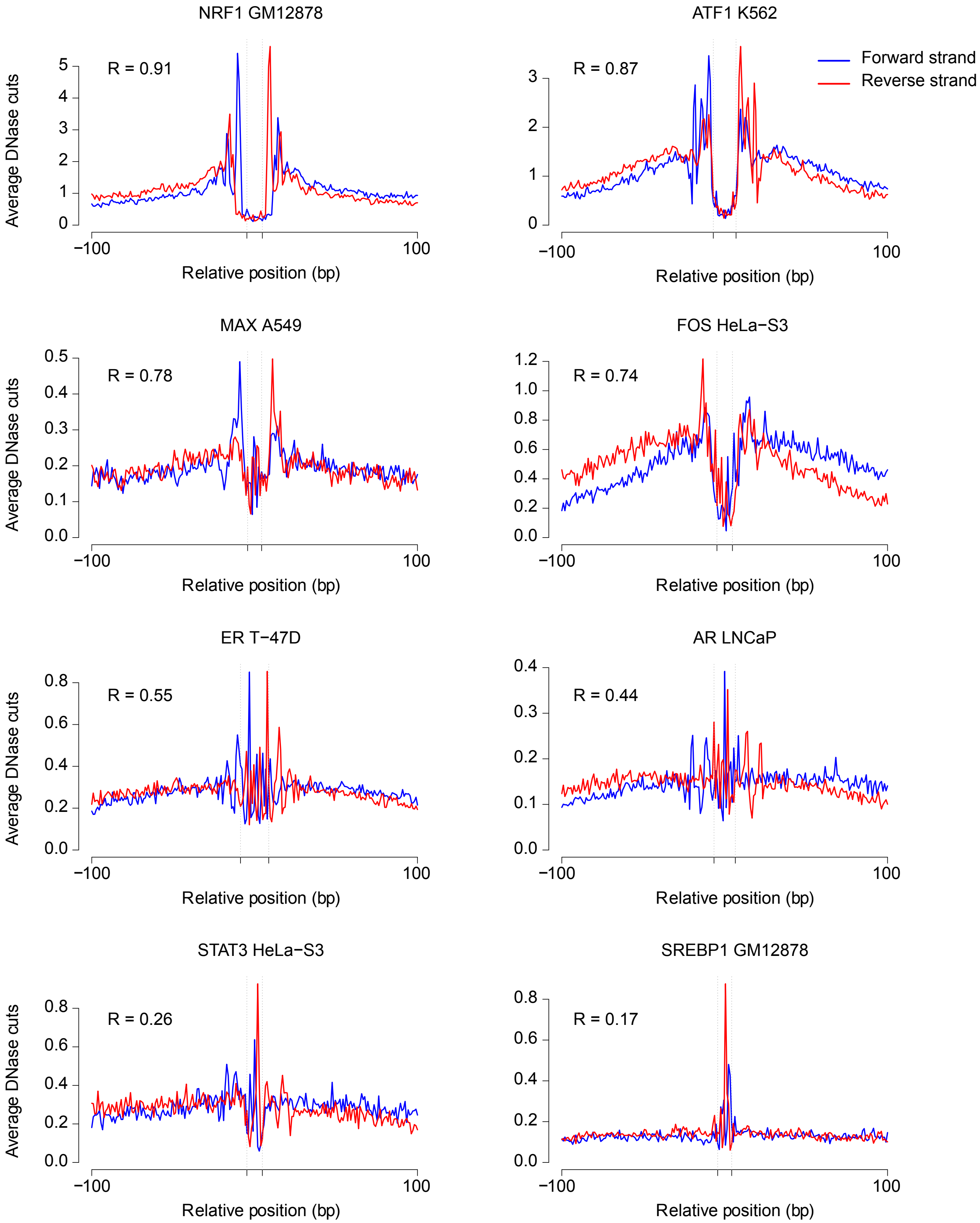

Fig. S9. Examples of DNase digestion profiles on both strands around a motif, showing TF × cell-type combinations with high to low prediction accuracy. DNase profiles were averaged over 1000 binding sites with the highest occupancy measured by ChIP-seq. The TF × cell-type combinations that achieved higher prediction accuracy tend to show much clearer DNase depletion patterns (depletion of DNase cleavage events within the motif region, coupled with elevation in the proximal flanking regions).

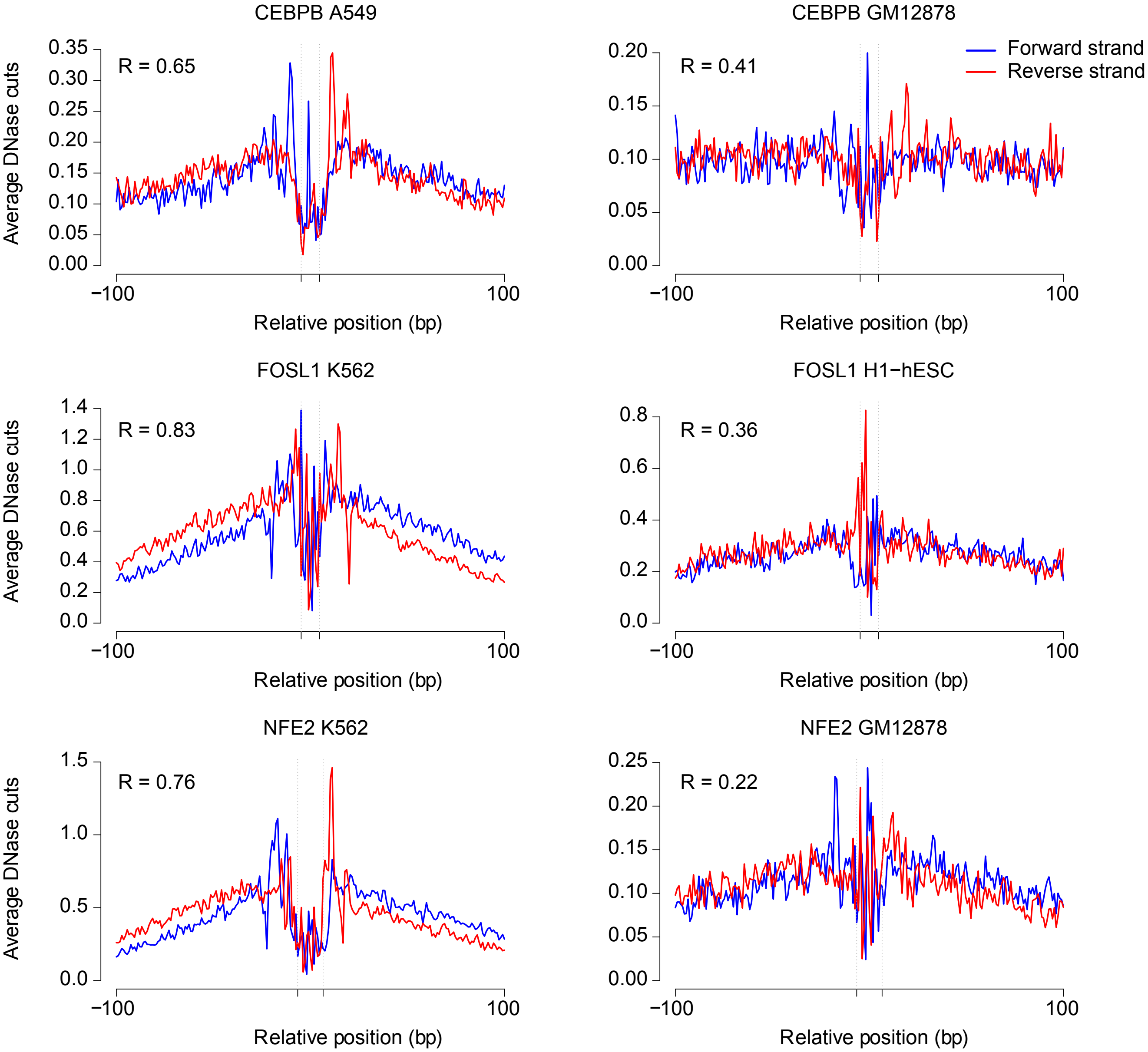

Fig. S10. Examples of DNase digestion profiles for TF × cell-type combinations with higher prediction accuracy in one cell type (left) but lower prediction accuracy in a different cell type (right). Note that scales on left and right differ. Consistent with Fig. S9, cell types with higher prediction accuracy often show much clearer DNase depletion patterns.

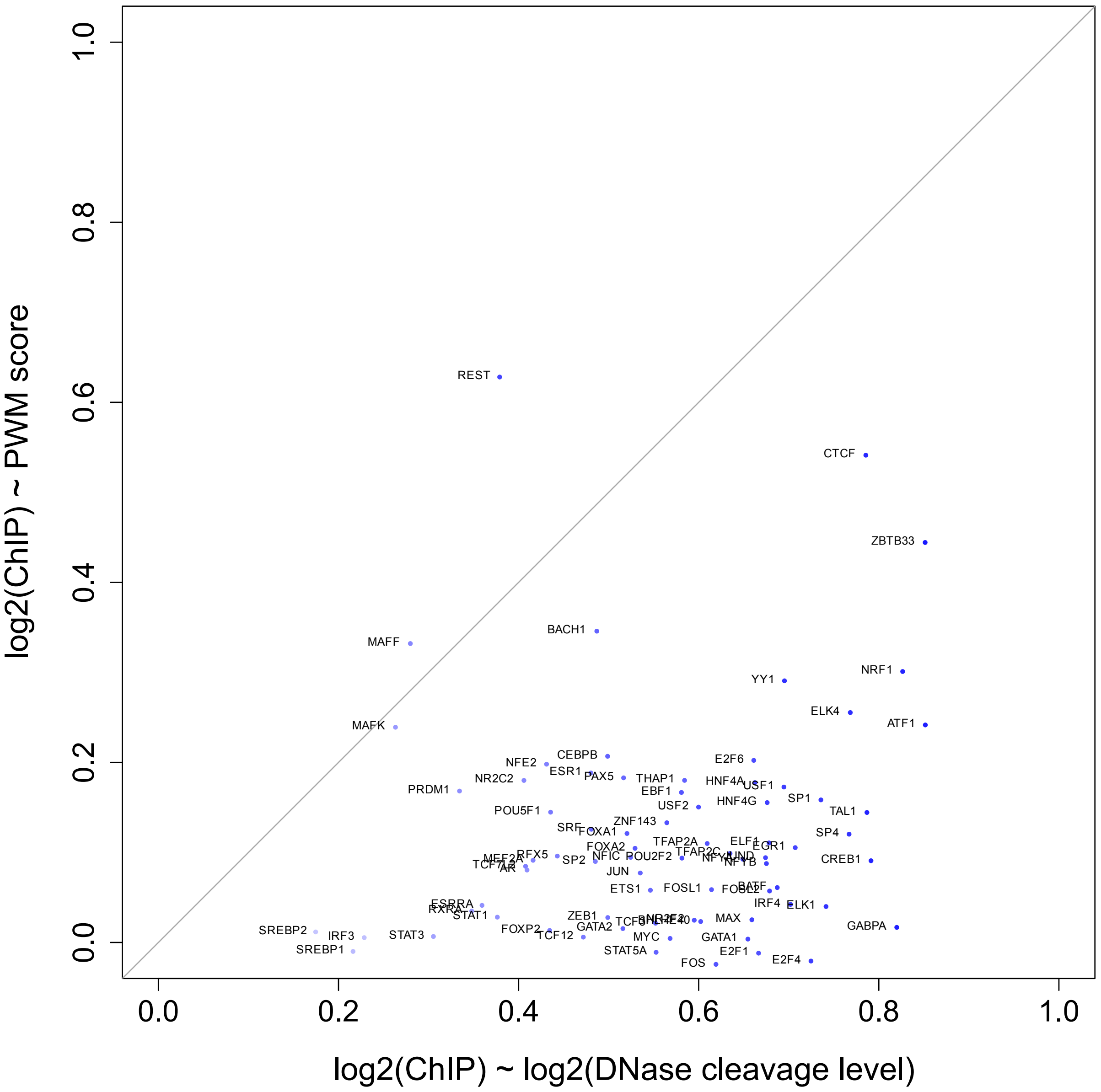

Fig. S11. Scatter plot comparing the correlation of measured TF occupancy with total number of nearby DNase cleavage events vs. the correlation of measured TF occupancy with PWM score. X-axis shows Pearson's correlations of log2(measured TF occupancy) with log2(total number of nearby DNase cleavage events). Y-axis shows Pearson's correlations of log2(measured TF occupancy) with PWM score. Overall level of nearby DNase cleavage is more correlated with measured TF occupancy than PWM score is—and typically markedly so—for nearly all TFs, with REST and MAFF being exceptions.

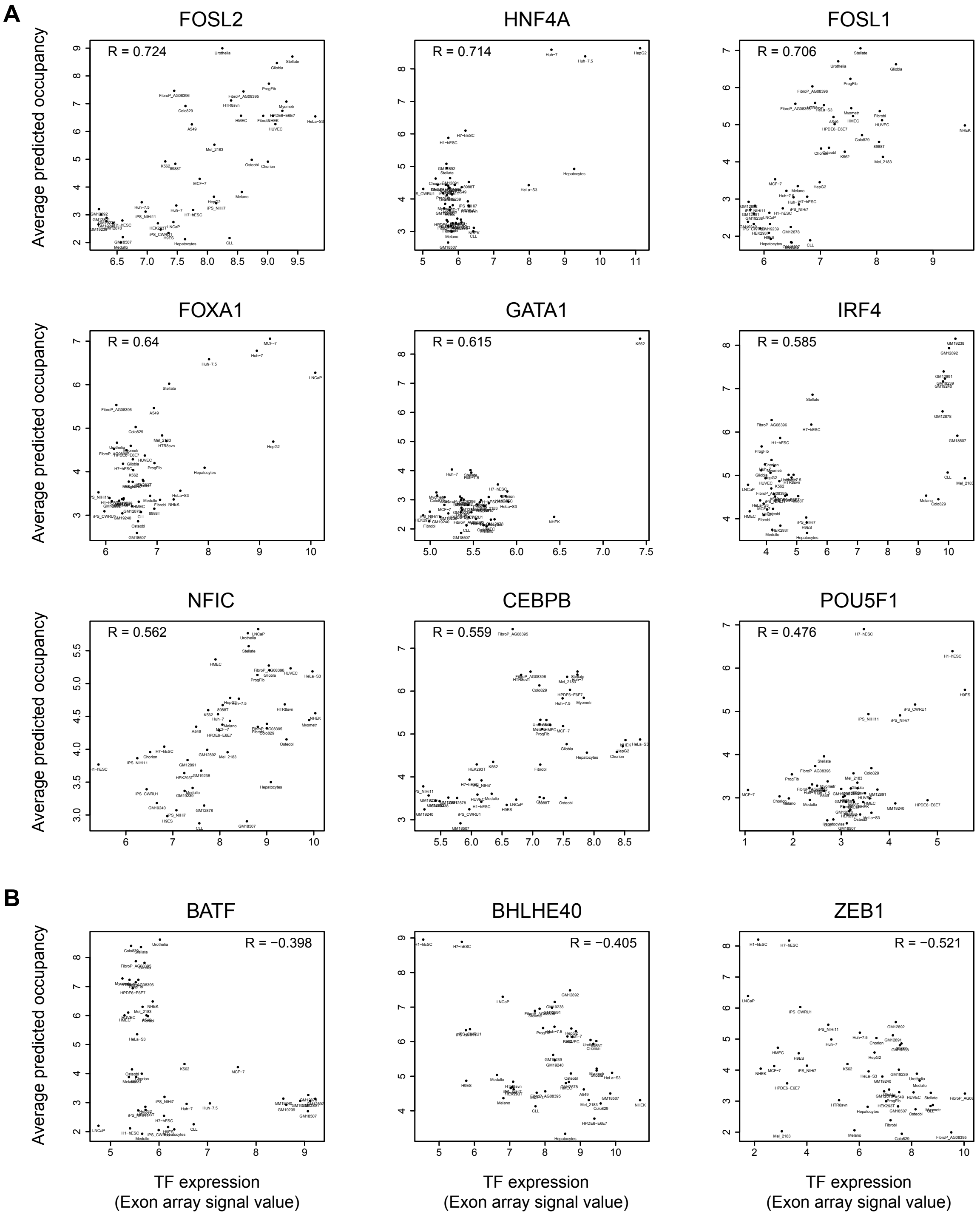

Fig. S12. Examples showing strong relationships between TF occupancy and TF expression in different cell types. (**A**) The nine TFs exhibiting the strongest positive correlation between predicted TF occupancy and measured TF expression level (which serves as a rough but imperfect proxy for active nuclear TF concentration). X-axis shows TF expression signal values (normalized gene expression value) from ENCODE Affymetrix Exon Array data generated by Crawford group. Y-axis shows predicted occupancy averaged across candidate sites along the genome. (**B**) The three TFs exhibiting statistically significant negative correlation between predicted TF occupancy and measured TF expression level.

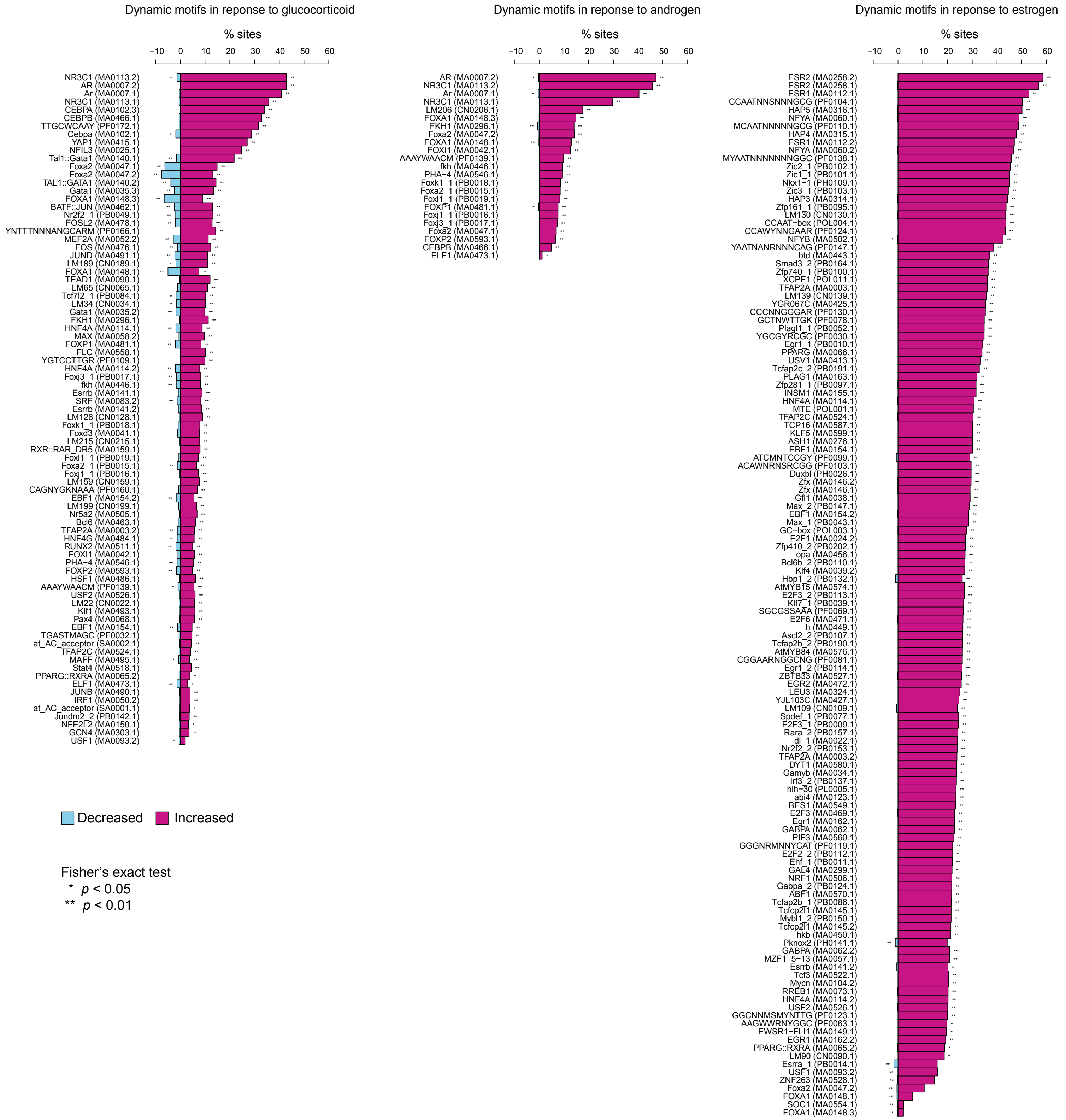

Fig. S13. TF occupancy dynamics in response to hormone stimulation. The three panels are analogous to the ones shown in Fig. 5, depicting the same analysis but here conducted at the level of individual TF motifs rather than RSAT TF motif clusters. Motif identifiers are displayed in parentheses next to motif names.

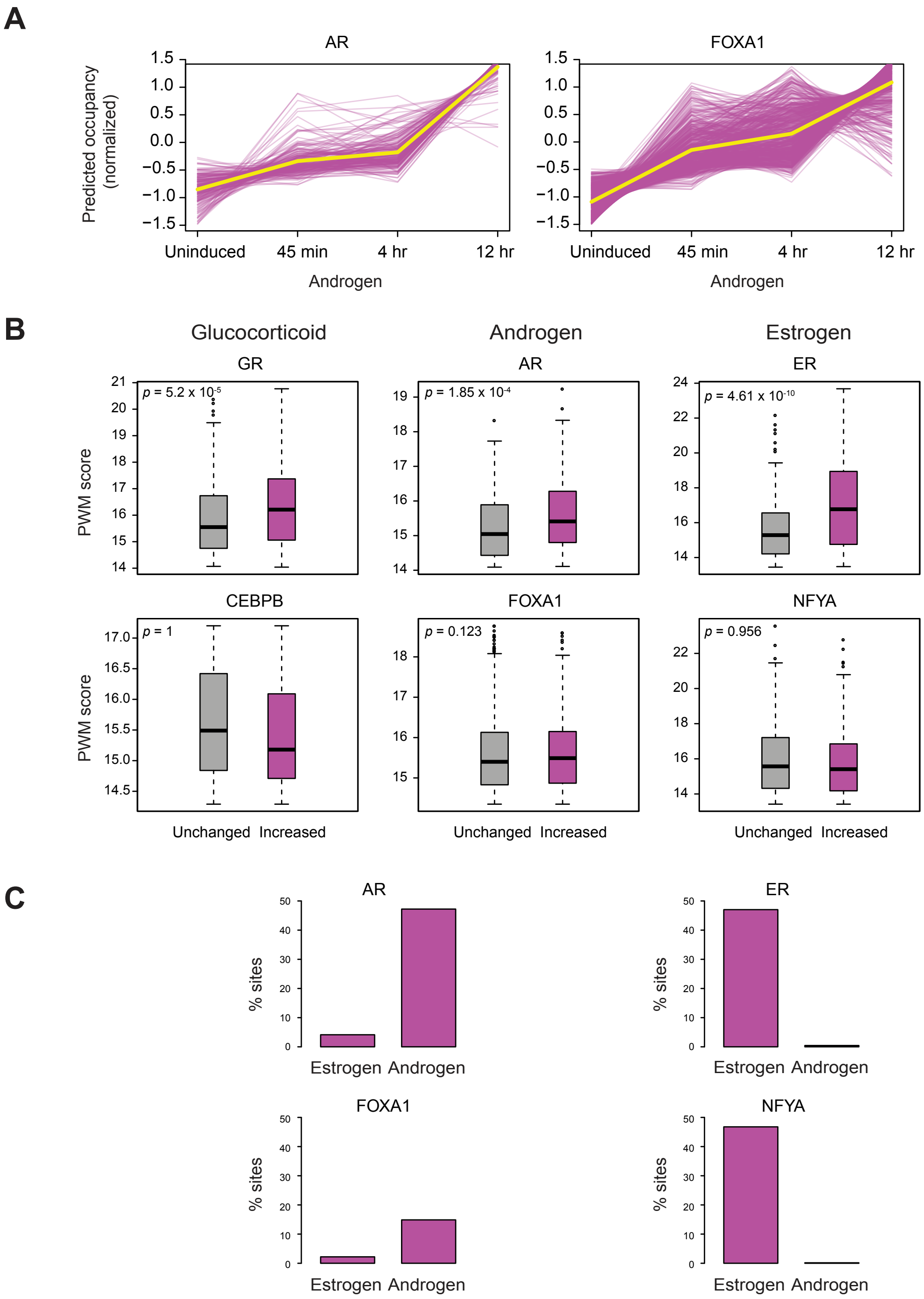

Fig. S14. Further details of TF occupancy dynamics in response to hormone stimulation. (**A**) Under androgen stimulation, AR and FOXA1 sites increase in occupancy gradually over the time course, revealing the importance of a quantitative perspective on occupancy. Yellow lines highlight the average trends of those sites. (**B**) For GR in glucocorticoid stimulation, AR in androgen stimulation, and ER in estrogen stimulation, sites with significantly increased occupancy possess significantly higher average PWM scores than sites with unchanged occupancy. This is not true for CEBPB, FOXA1, or NFYA, the second-most responsive TFs in the respective treatment conditions. (**C**) Specificity of increased occupancy. Bar plots on the left show the response of AR and FOXA1, revealing that their increased occupancy is highly specific to androgen induction. Bar plots on the right show the response of ER and NFYA, revealing that their increased occupancy is highly specific to estrogen induction.

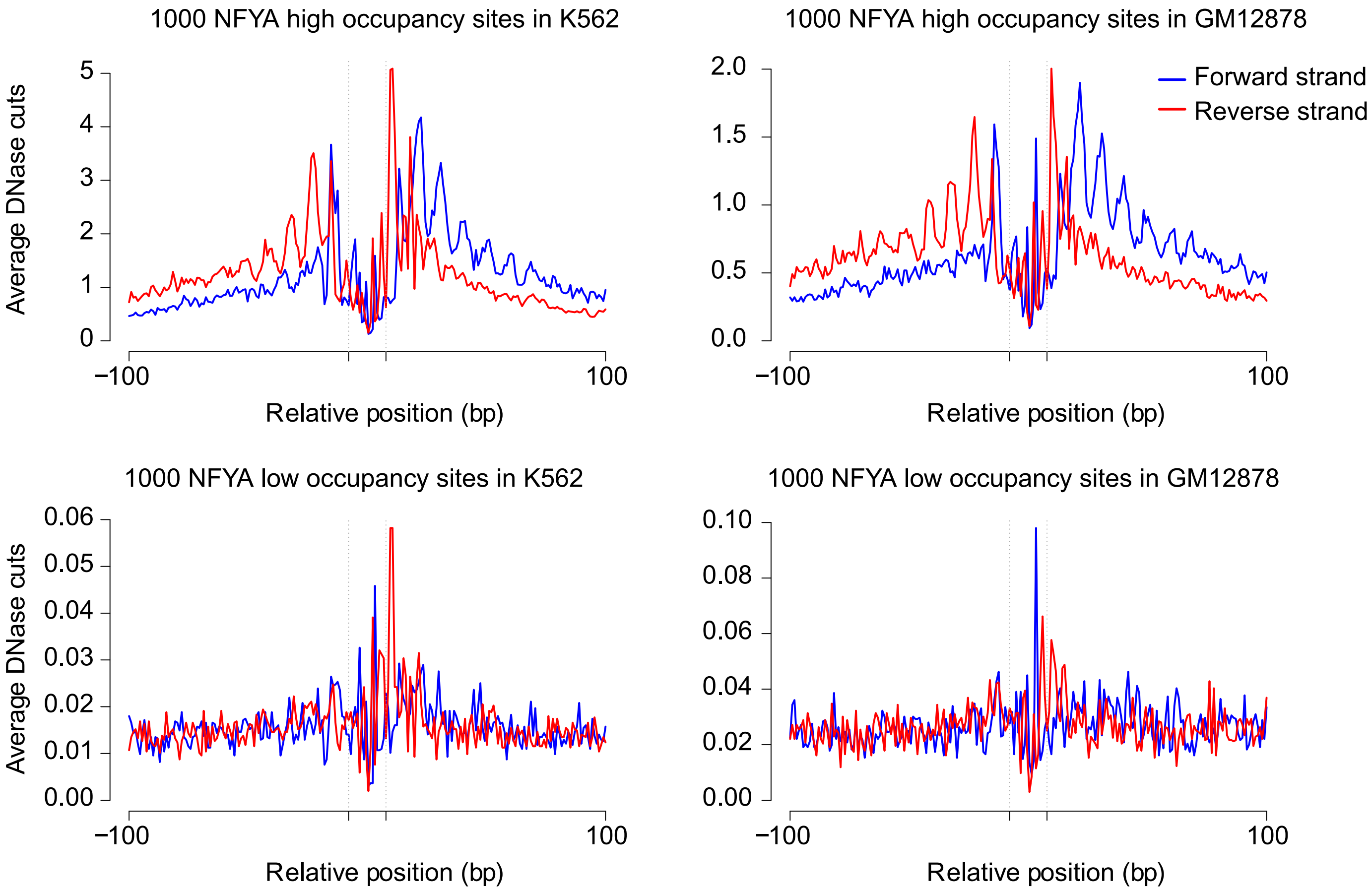

Fig. S15. Average DNase digestion profiles around 1000 high occupancy and 1000 low occupancy NFYA sites in K562 and GM12878 cell types. In both cell types, the oscillation patterns of DNase cleavage events in the flanking regions of NFYA high occupancy sites (top row) are similar to the DNase oscillation patterns previously observed within nucleosomes (Zhong et al., 2016), suggesting that NFYA is perhaps more likely to bind flanked by nucleosomes.

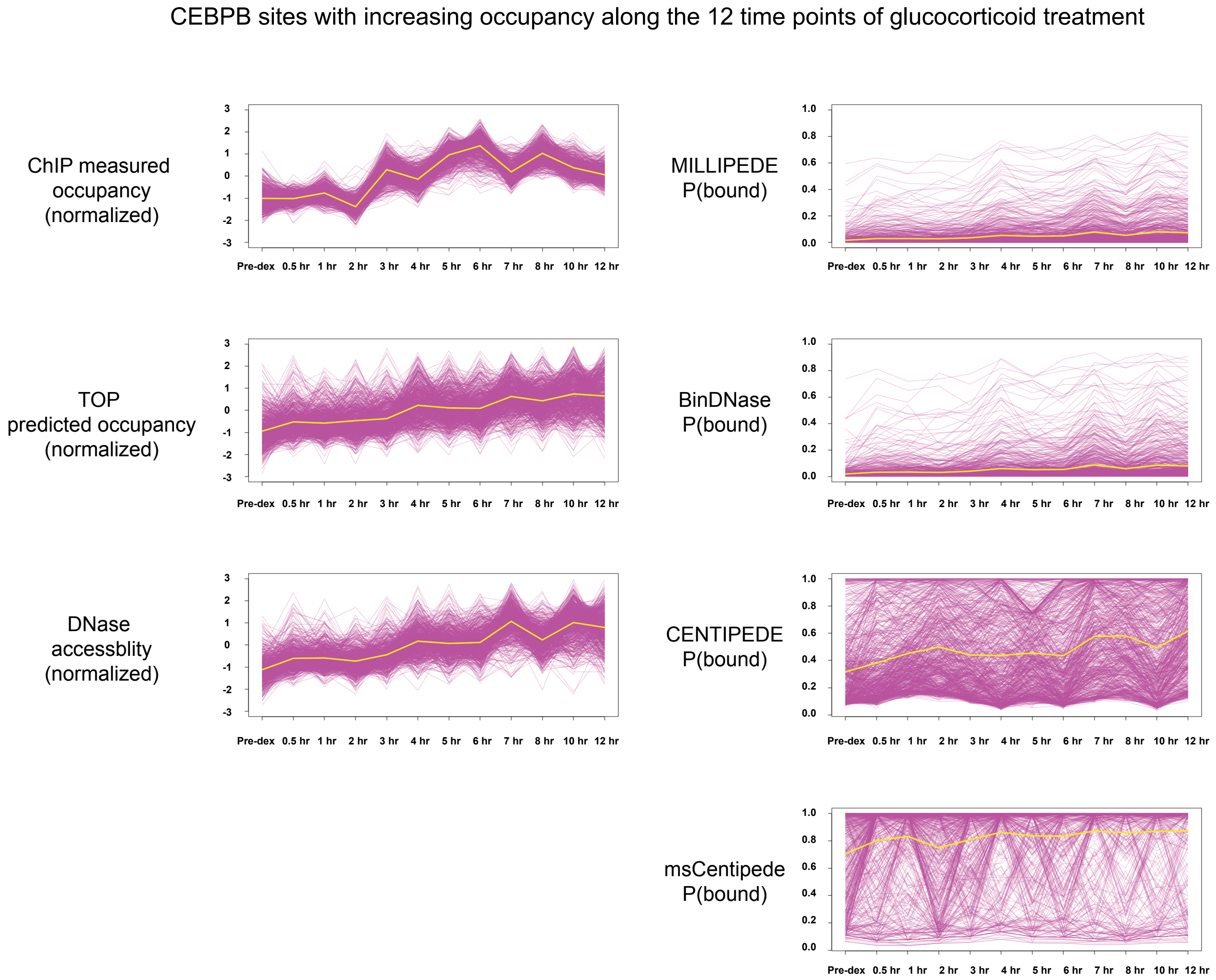

Fig. S16. Measured and predicted occupancy or binding probabilities for CEBPB sites with increasing occupancy along the 12 time points of glucocorticoid (GC) treatment. Following glucocorticoid (GC) treatment, CEBPB sites display gradual increasing trends of occupancy over the time course. Left panel shows measured occupancy, TOP (bottom level model) predicted occupancy, and total DNase accessibility along the 12 time points of GC treatment. Right panel shows CEBPB binding probabilities from four other computational methods, which predict TF binding from a binary perspective. Yellow lines highlight the average trends of those sites.

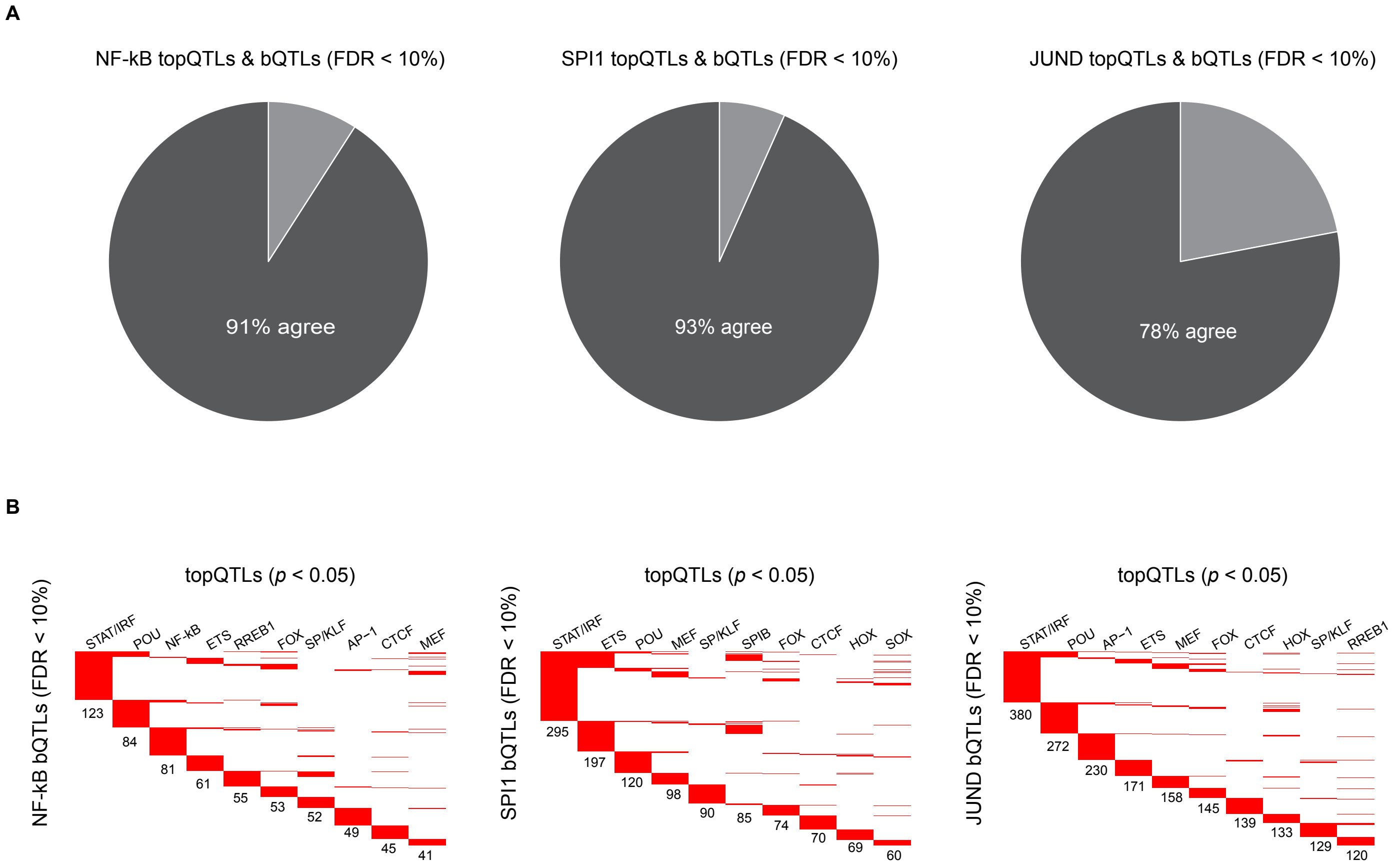

Fig. S17. Agreement between topQTLs and bQTLs for NF-kB, SPI1, and JUND. (**A**) Among SNPs that are both significant topQTLs (FDR < 10%) and bQTLs (FDR < 10%), a large percentage of them show directionality agreement on high occupancy alleles. STAT1 and POU2F1 were not included because very few SNPs are both significant topQTLs and bQTLs when using a 10% FDR cutoff. (**B**) Many bQTLs are observed to be topQTLs for different motifs; only the most frequent ten topQTL motifs are shown here, but more than 10% of bQTLs correspond to topQTLs ($p < 0.05$) for different motifs.

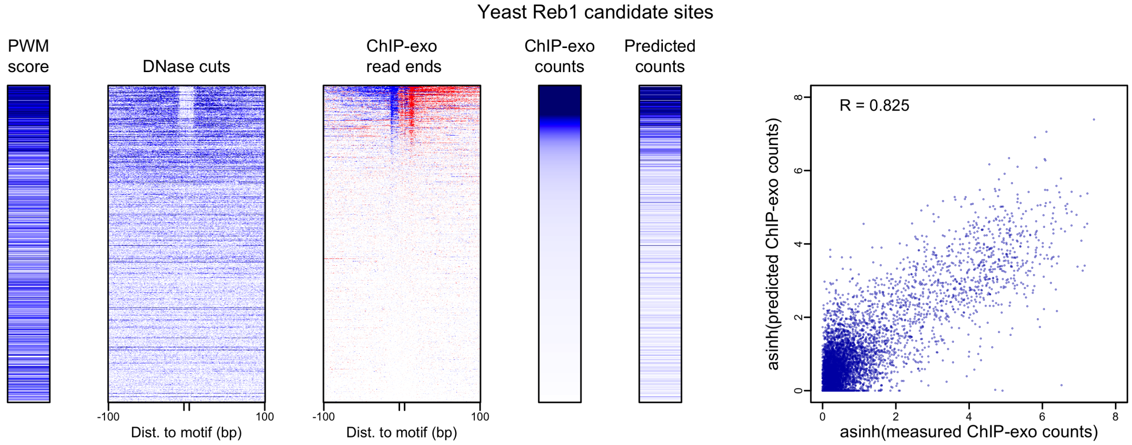

Fig. S18. Predicting quantitative Reb1 occupancy in yeast. (Non-hierarchical) regression model was trained on ChIP-exo read counts around Reb1 candidate binding sites in yeast. Rows in the left panels correspond to candidate binding sites and were ordered by the measured number of ChIP-exo read counts (column 4). Darker colors mean higher PWM score, higher number of DNase cleavage events, or higher occupancy (ChIP-exo read counts).

References

Luo, K. and Hartemink, A. J., 2013. Using DNase digestion data to accurately identify transcription factor binding sites. In *Pac. Symp. Biocomputing*, pages 80–91. World Scientific, Hackensack, NJ.

Zhong, J., Luo, K., Winter, P. S., Crawford, G. E., Iversen, E. S., and Hartemink, A. J., 2016. Mapping nucleosome positions using DNase-seq. *Genome Res.*, **26**(3):351–364.