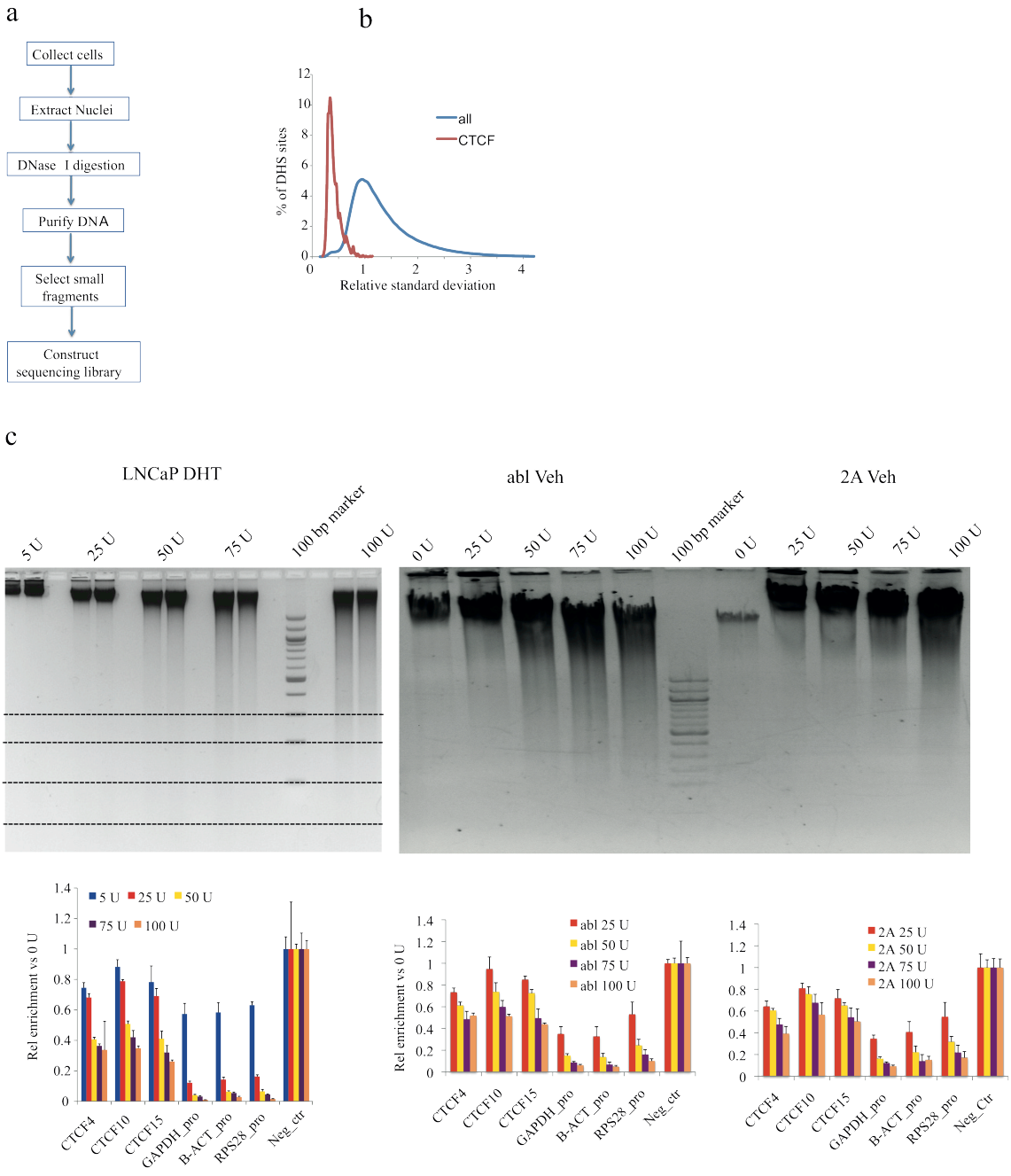
Supplementary materials for:

Analysis of optimized DNase-seq reveals intrinsic bias in transcription factor footprint identification

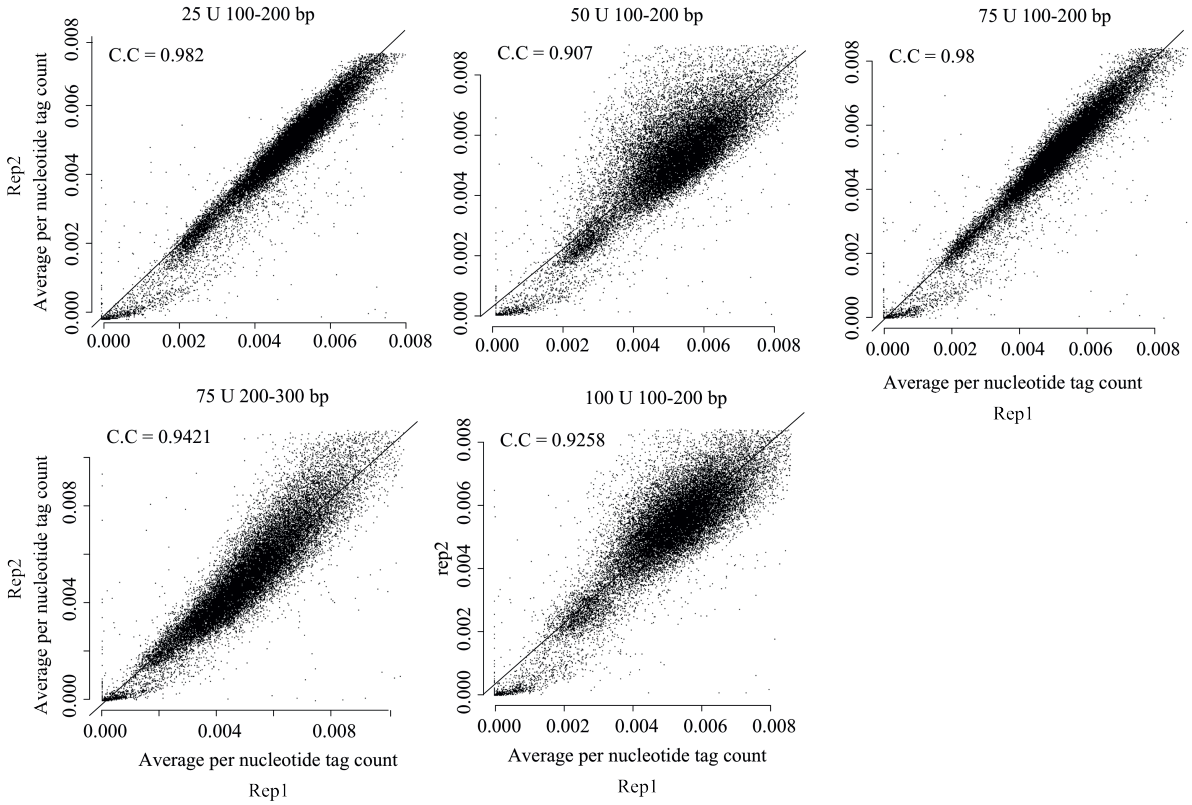
Housheng Hansen He[1,2,3,4,5,*], Clifford A. Meyer[1,3,*], Sheng'en Shawn Hu[3,6,*], Mei-Wei Chen[3], Chongzhi Zang[1,3], Yin Liu[3,6], Prakash K. Rao[3], Teng Fei[1,2,3], Han Xu[1,3], Henry Long[3,#], X. Shirley Liu[1,3,#] and Myles Brown[2,3,#]

[1] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, Massachusetts 02115, USA;
[2] Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, Massachusetts 02115, USA;
[3] Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA
[4] Ontario Cancer Institute, Princess Margaret Cancer Center/University Health Network, Toronto, Ontario, M5G1L7, Canada
[5] Department of Medical Biophysics, University of Toronto, Toronto, Ontario, M5G2M9, Canada
[6] Department of Bioinformatics, School of Life Science and Technology, Tongji University, Shanghai, 20092, China
*Equal contribution
#Correspondence

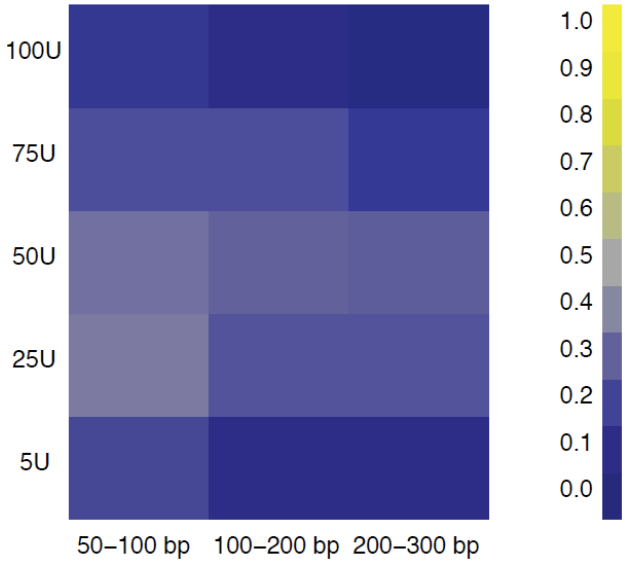
Supplementary Figure 1. DNase-seq experiment quality control. (a) Flow chart of the DNase-seq experimental process. **(b)** Variation in the relative DNase-seq tag count at non-promoter DNase hypersensitive regions across 74 ENCODE DNase-seq data sets. Constitutive CTCF sites, DHS sites that are present in all 74 ENCODE DNase-seq data sets and also overlap CTCF binding, are less variable than DHS sites in general. **(c)** Electrophoresis gel and qPCR quantification in LNCaP, abl and 2A cell lines. PCR primers spanning 3 constitutive CTCF binding sites and 3 housekeeping genes were used to quantify the relative DNA abundance over a range of DNase enzyme strength.



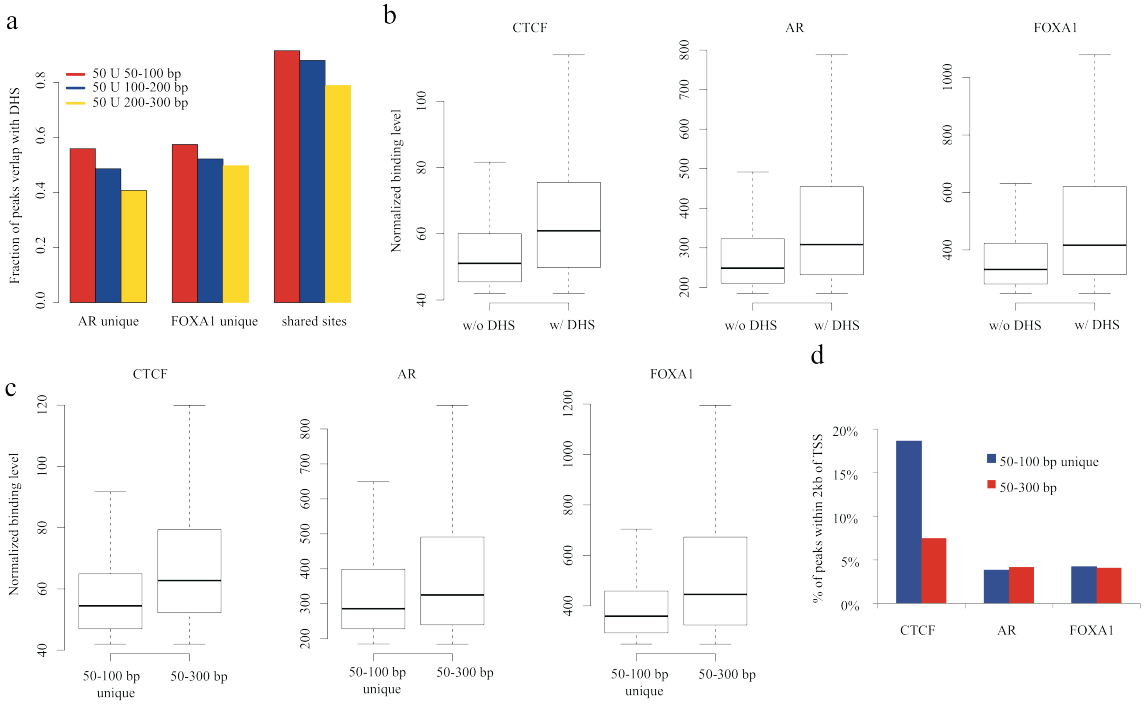
Supplementary Figure 2. Correlation between DNase-seq biological replicates.
Scatter plots show the genome wide correlations between two biological replicates under five different conditions. From each DNase-seq dataset 15M reads were sampled randomly and the average per nucleotide tag count in every 100kb genomic region was calculated along with Pearson correlation coefficients (C.C.). Each point in these scatter plots corresponds to one 100kb genomic interval.



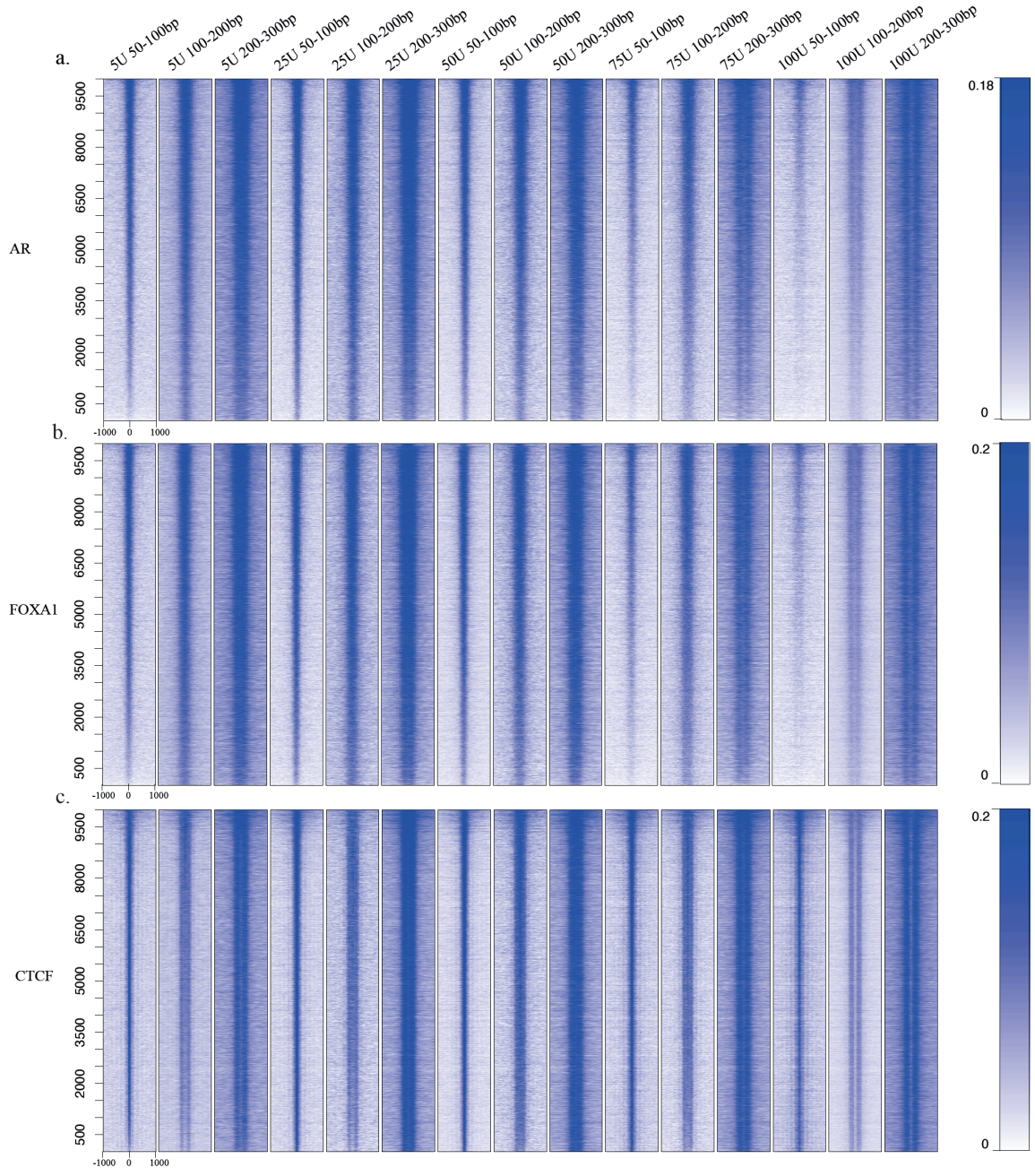
Supplementary Figure 3. Effect of digestion level and fragment size on H3K4me2 peak recovery. Proportion of H3K4me2 ChIP-seq regions discovered as DNaseI hypersensitive sites in LNCaP cells. 15M reads were sampled from each experimental condition. Rows correspond to the DNaseI enzyme strength and columns represent fragment sizes. Colors represent the proportion of H3K4me2 sites detected by DNase-seq.



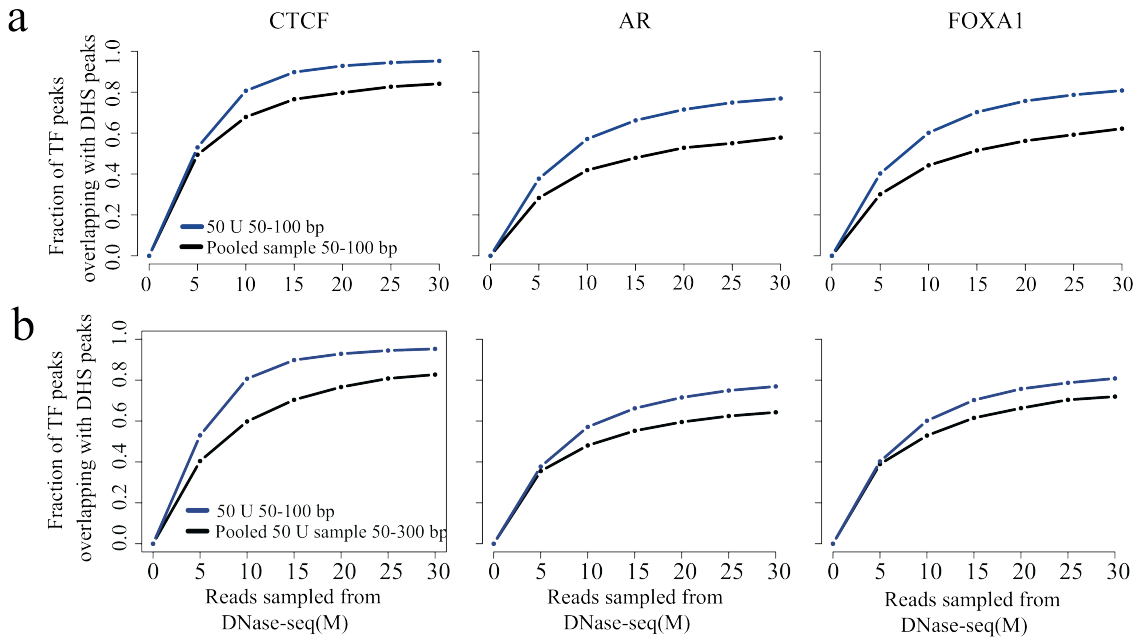
Supplementary Figure 4. Effect of fragment size on recovery of known transcription factor binding sites. (a) Proportion of ChIP-seq enriched regions discovered as DNaseI hypersensitive (DHS) sites for AR unique, FOXA1 unique and AR and FOXA1 shared sites in LNCaP cells. 15M randomly sampled reads were used to call DHS sites. Relative to unique sites, shared AR and FOXA1 sites are more likely to be DHS. **(b)** CTCF, AR and FOXA1 binding level as measured by MACS ChIP-seq analysis score for sites overlapping with DHS sites (w/ DHS) and sites not overlapping with DHS sites (w/o DHS). **(c)** CTCF, AR and FOXA1 binding levels as measured by MACS score. ChIP-seq sites overlapping with the intersection of DHS sites discovered from 50-100bp, 100-200bp and 200-300bp tags (50-300bp) have higher binding levels than DHS sites identified from 50-100bp tags alone (50-100bp unique). **(d)** Percentage of 50-100bp unique and 50-300bp shared CTCF, AR and FOXA1 in proximal promoter regions (2kb of TSS).



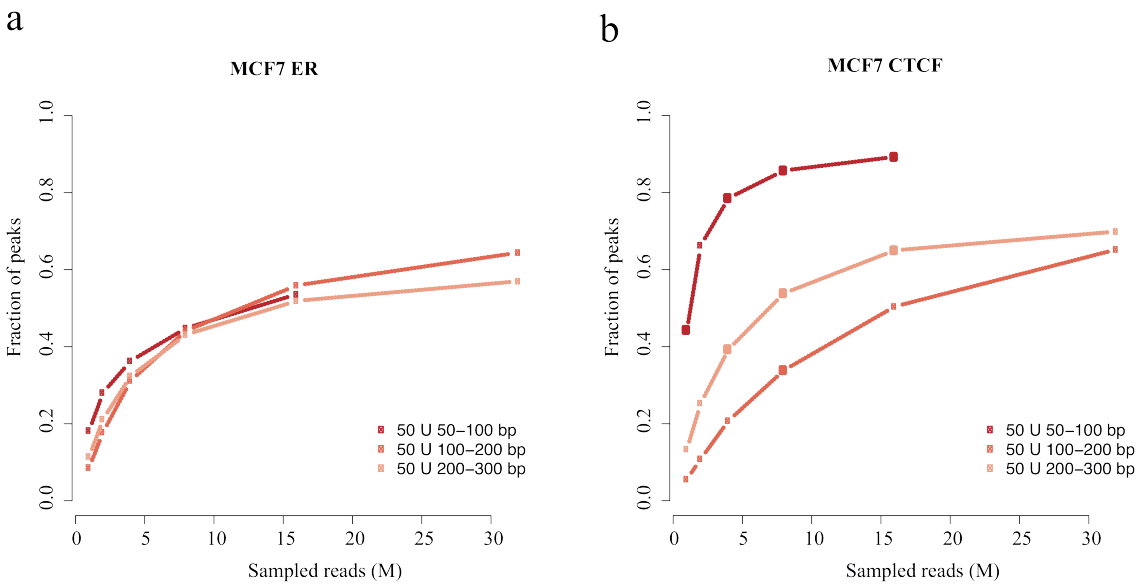
Supplementary Figure 5. DNase-seq tag count densities at AR, FOXA1 and CTCF ChIP-seq sites.. Each row in each of the heatmaps represents a genomic locus centered on a ChIP-seq peak center. Sites in the heatmaps are ordered by 5U 100-200bp DNase-seq tag count. The colors in the heatmaps represent 50bp segment averages of ChIP-seq signal (normalized by macs2 to 1M reads).



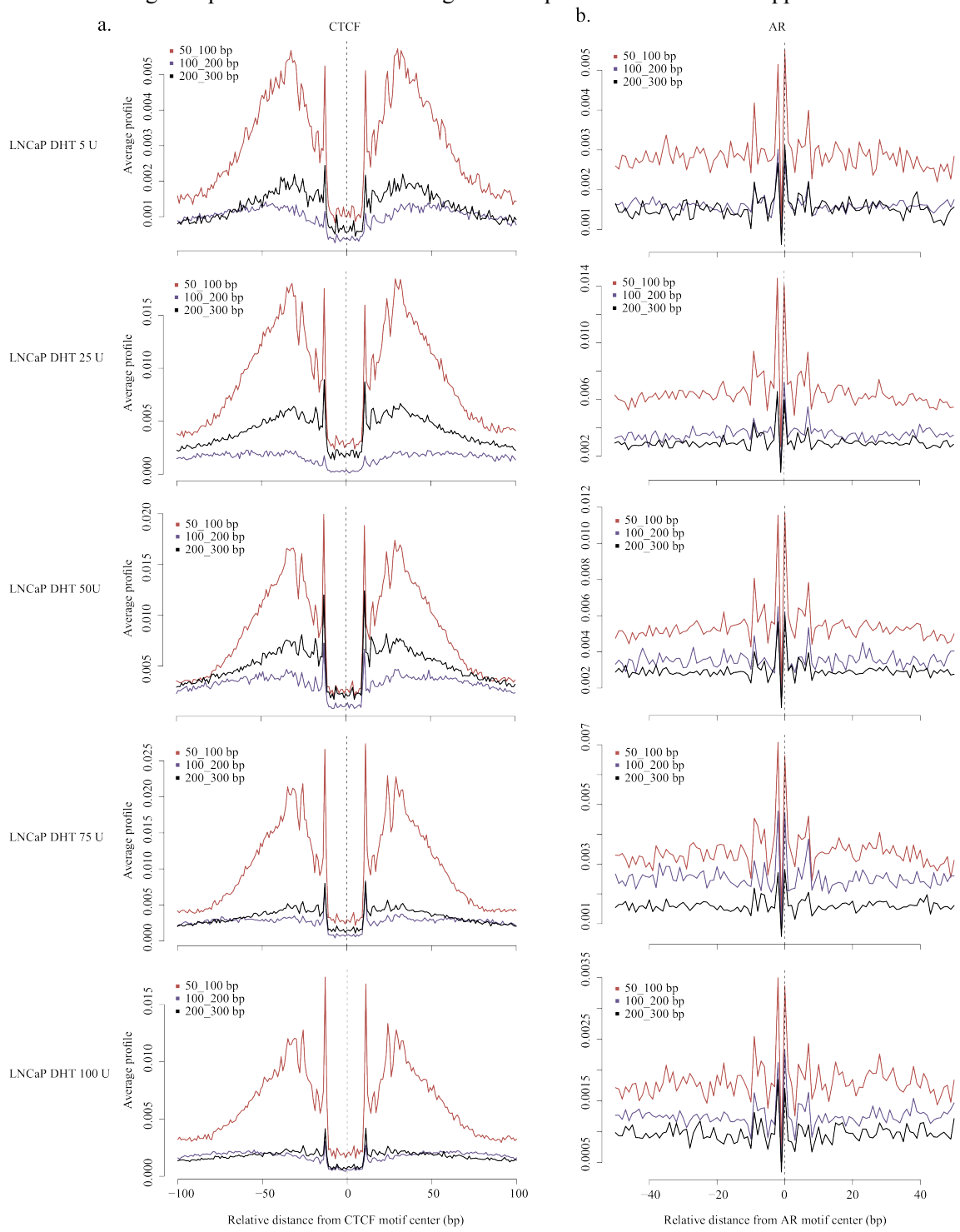
Supplementary Figure 6. Effect of pooling digestion levels and fragment sizes on recoveries of CTCF, AR and FOXA1 binding sites. These plots represent, as a function of read depth, the proportion of CTCF, AR and FOXA1 ChIP-seq regions recovered as DNaseI hypersensitive sites in LNCaP cells. **(a)** Pooling different digestion levels, 5U, 25U, 50U, 75U and 100U, using the single 50-100bp fragment size range is less efficient for TF binding site recovery than using the single 50U digestion level with 50-100bp fragments. **(b)** At the 50U digestion level, pooling different fragment size ranges, 50-100bp, 100-200bp and 200-300bp, is less efficient than using the single 50-100bp range.



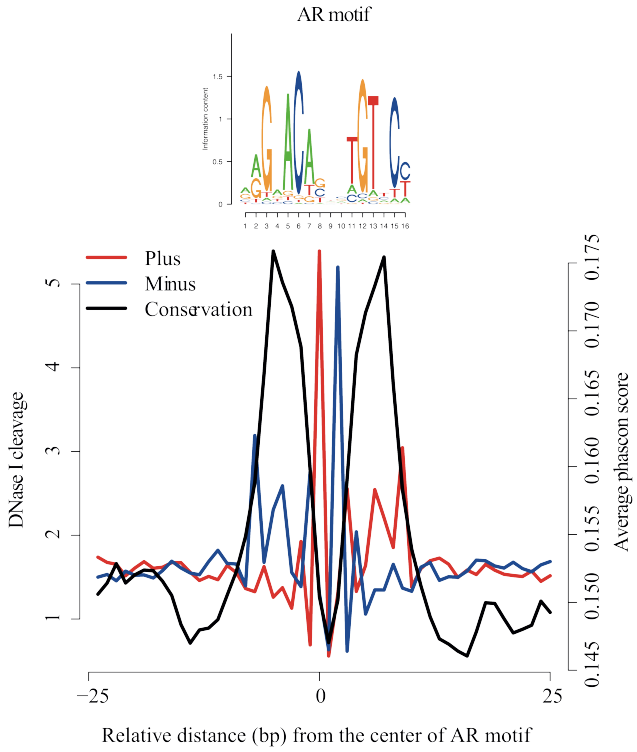
Supplementary Figure 7. Effect of fragment size on retrieval of (a) ER and (b) CTCF binding sites in MCF7 cells.



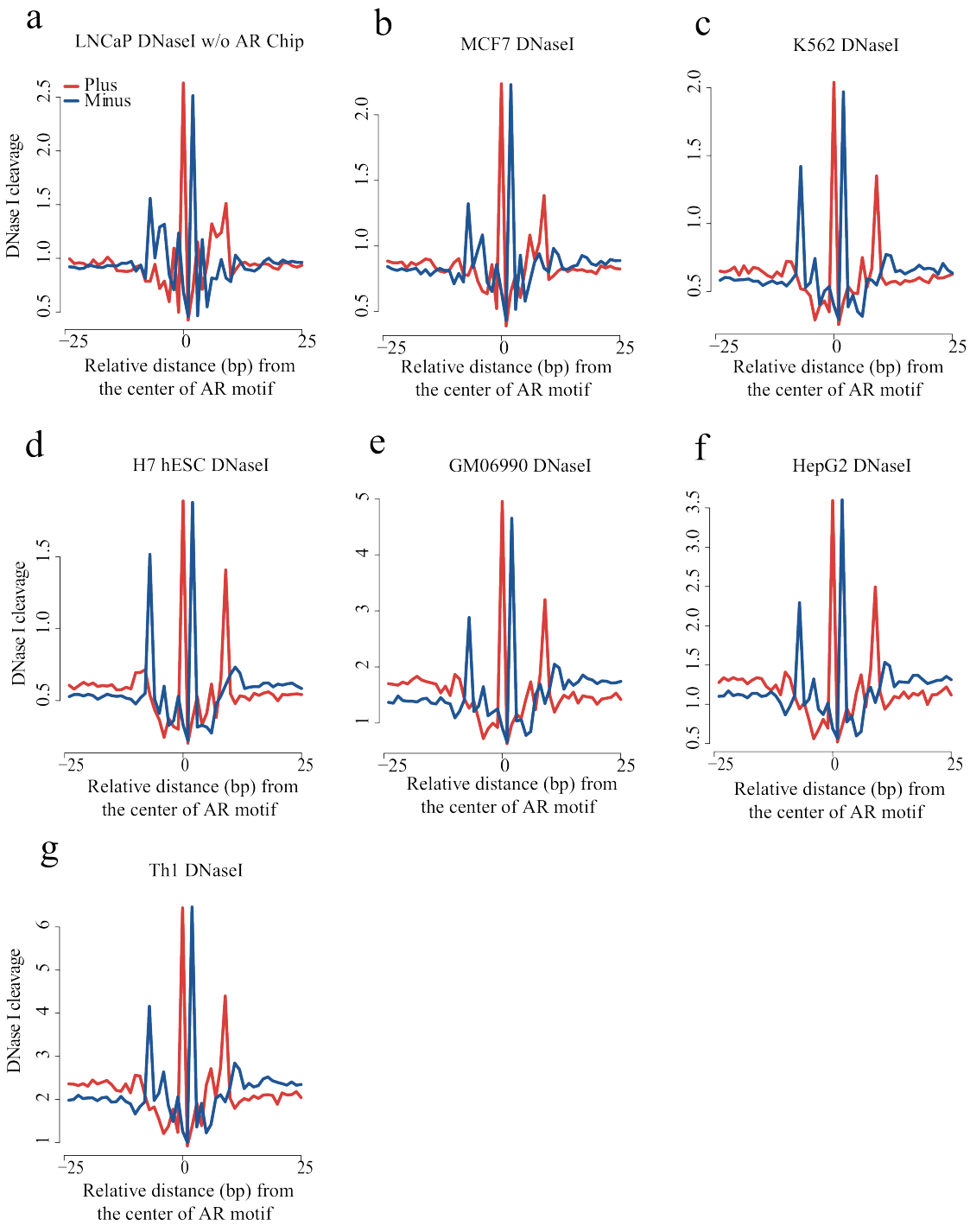
Supplementary Figure 8. Effect of digestion level and fragment size on CTCF and AR footprint. **(a)** Nucleotide resolution DNaseI cleavage frequencies across CTCF recognition sequences at CTCF ChIP-seq peaks in LNCaP. DNase-seq signals were normalized to 1M reads and 5' ends of reads counted in a non-strand specific manner. Short 50-100bp fragments produce clearer cleavage signals than 100-200bp or 200-300bp fragments across all different digestion levels. **(b)** Nucleotide resolution DNaseI cleavage frequencies across AR recognition sequences at AR ChIP-seq peaks in LNCaP.



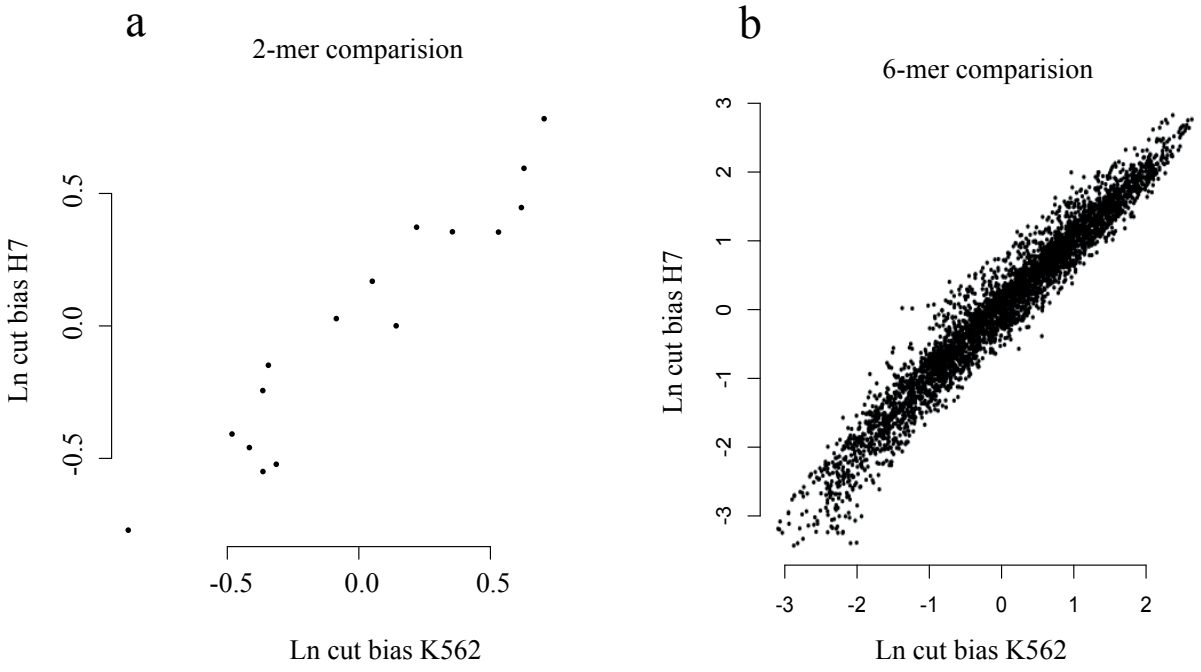
Supplementary Figure 9. PhastCons score evolutionary conservation of DNA sequence at AR motifs. The AR motif is a palindrome composed of an androgen response element half-site and its reverse complement, separated by a gap of 3 non-informative nucleotides. When AR binds to DNA the contacts between AR and DNA occur at the half sites, not in the gap. The three gap nucleotides are less well conserved than the half-sites themselves. DNaseI cleavage is highest in the gap consistent with the regions of contact between AR and DNA. The DNase cleavage pattern is, however, also seen in naked DNA suggesting that either the evolutionary conservation pattern is coincidental or that the accessibility of DNaseI has something in common with the AR DNA interaction.



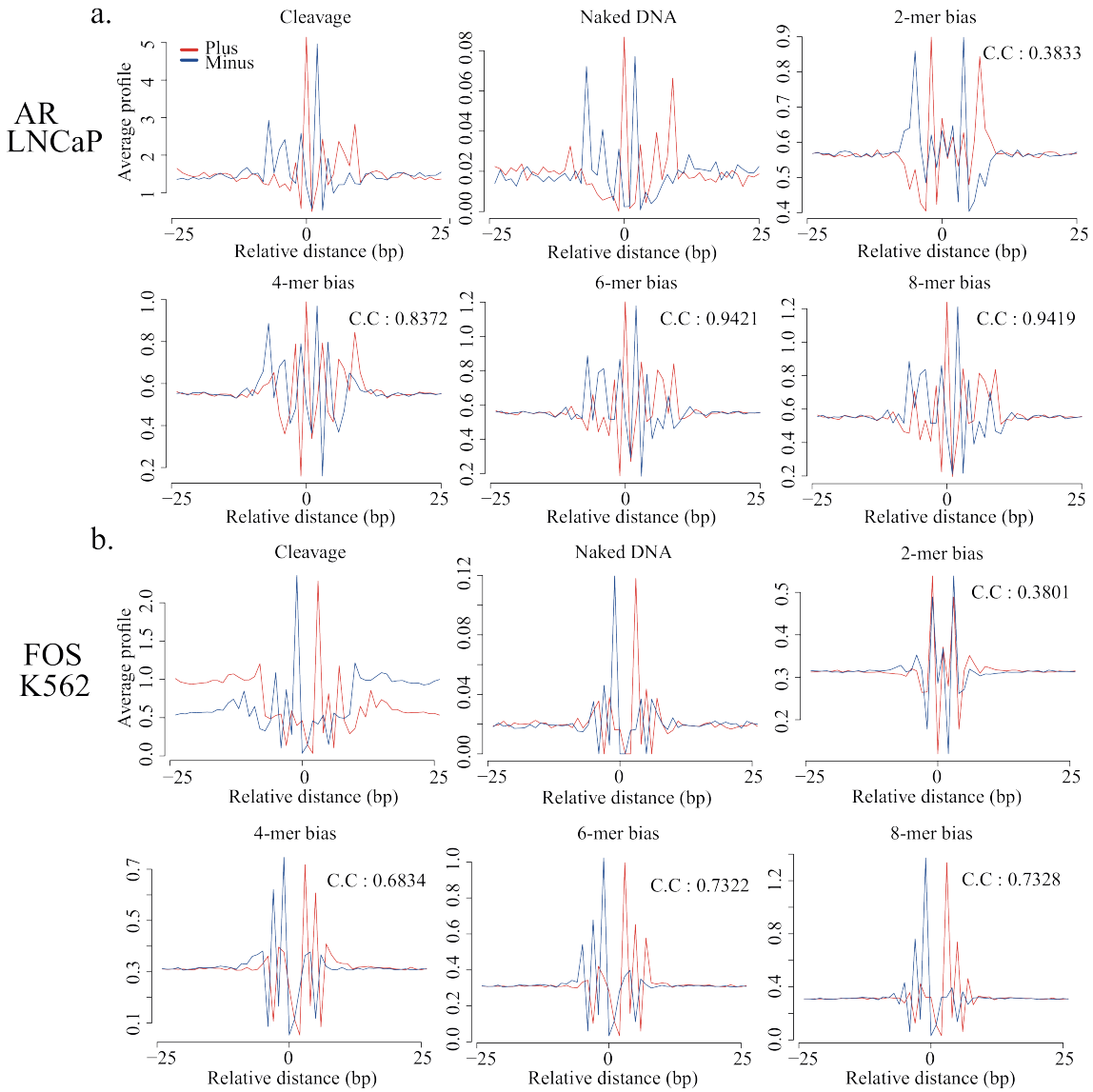
Supplementary Figure 10. DNaseI cleavage at AR motifs in different cell lines, **(a)** LNCaP, **(b)** MCF7, **(c)** K562, **(d)** H7, **(e)** GM06990, **(f)** HepG2, **(g)** Th1. The y-axis represents average counts of the 5' end of DNase-seq tags. Differences in the scale of the y-axis are due to differences in read depth.



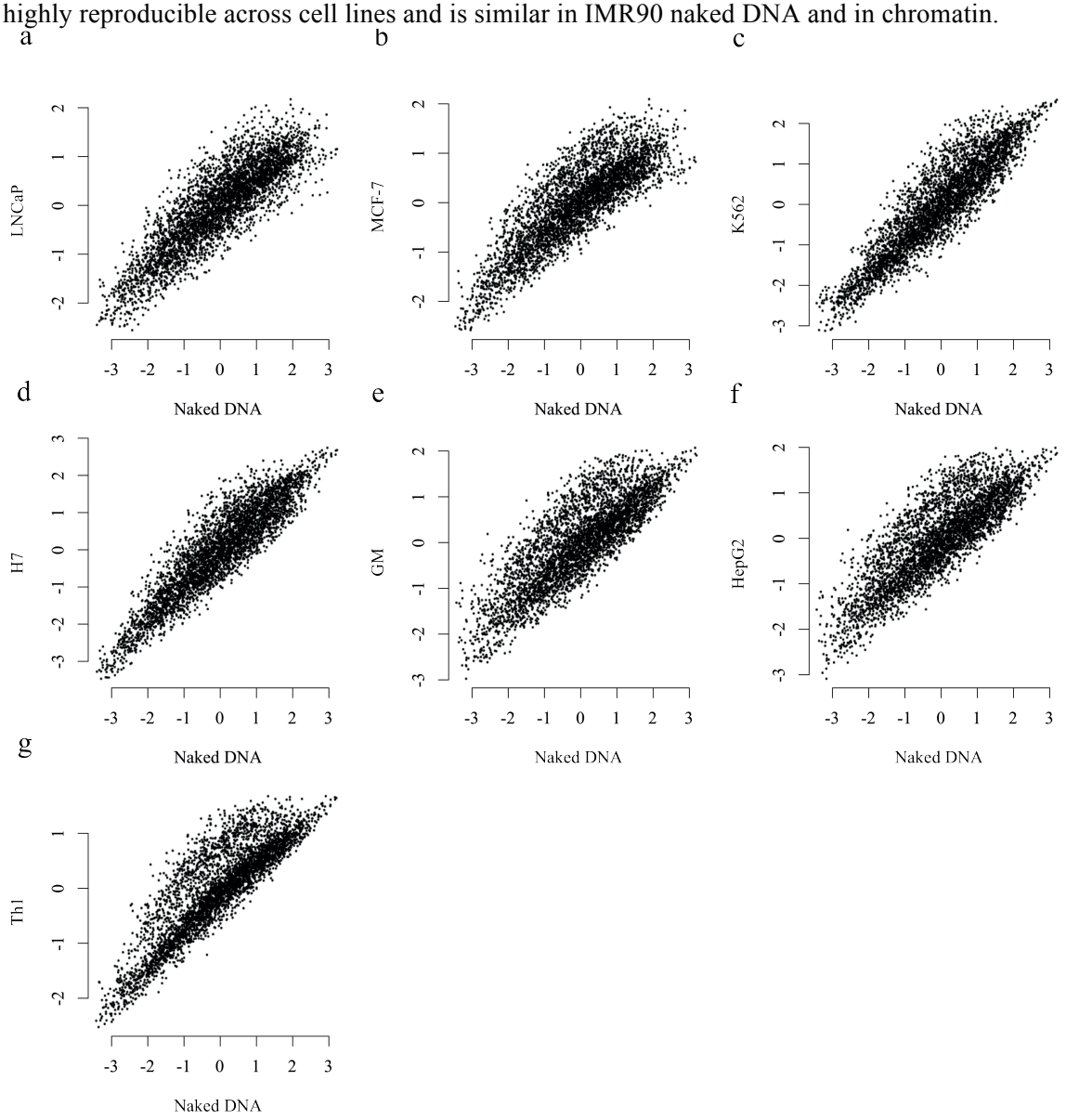
Supplementary Figure 11. Comparison of DNase cleavage bias in K562 and H7 cells. **(a)** DNase cleavage bias calculated based on 2-mer model. **(b)** DNase cutting bias calculated based on 6-mer model. Whereas in the 2-mer case, the highest bias value is approximately 5 fold that of the lowest, for 6-mers this ratio is greater than 400.



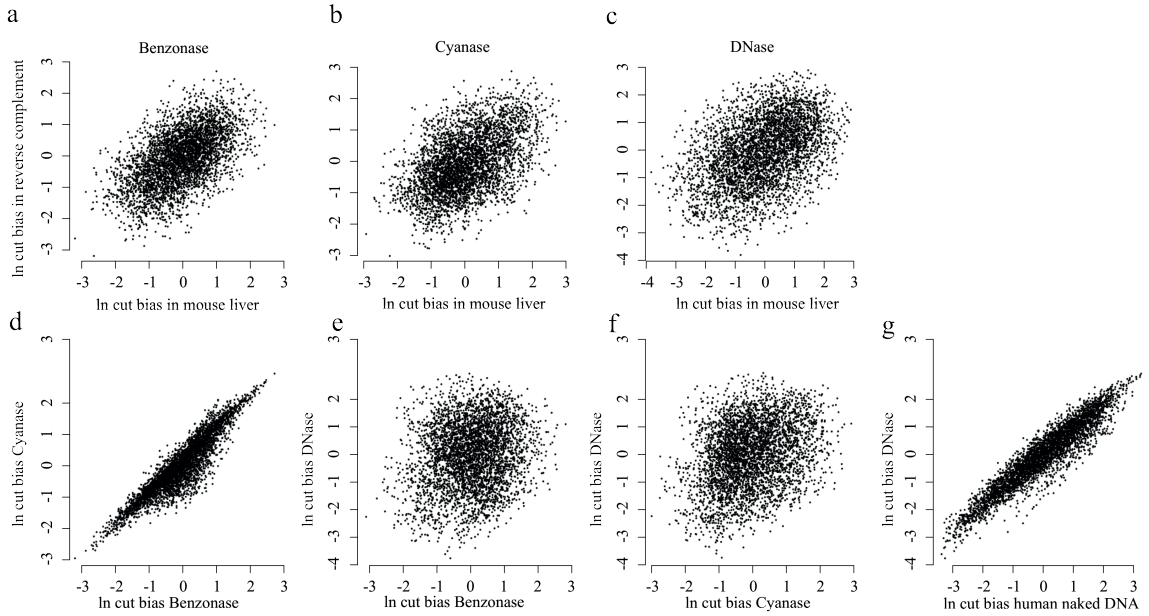
Supplementary Figure 12. Observed, intrinsic and 2-,4-,6- and 8-mer model predicted DNaseI cleavage at AR and FOS binding sites. **(a)** DNaseI cleavage in chromatin, naked DNA, and model predicted bias for AR in LNCaP cells. **(b)** DNaseI cleavage in chromatin, naked DNA and model predicted bias for FOS in K562 cells. TF binding sites are centered on TF binding motifs within ChIP-seq peak regions. The correlation between the observed cleavage pattern and model predicted cleavage patterns are similar for 6-mer and 8-mer models. The 6-mer model predicts cleavage bias patterns more accurately than the 2-mer and 4-mer models.



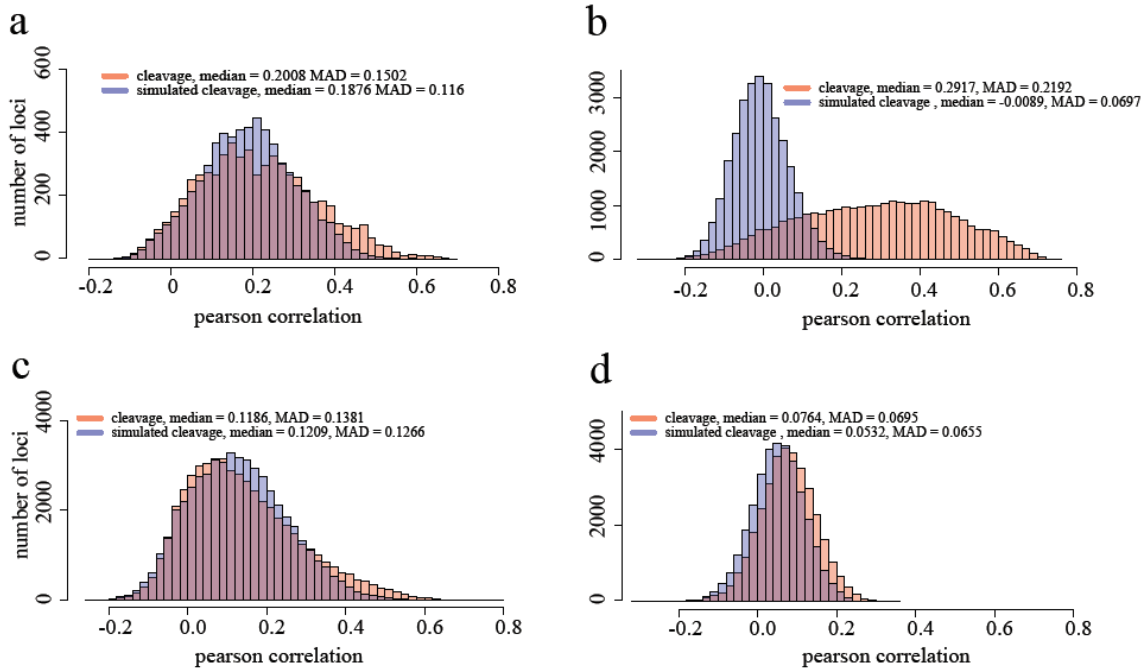
Supplementary Figure 13. DNaseI cleavage bias in naked DNA and 7 different cell lines, **(a)** LNCaP, **(b)** MCF7, **(c)** K562, **(d)** H7, **(e)** GM06990, **(f)** HepG2, **(g)** Th1. DNaseI cleavage bias is highly reproducible across cell lines and is similar in IMR90 naked DNA and in chromatin.



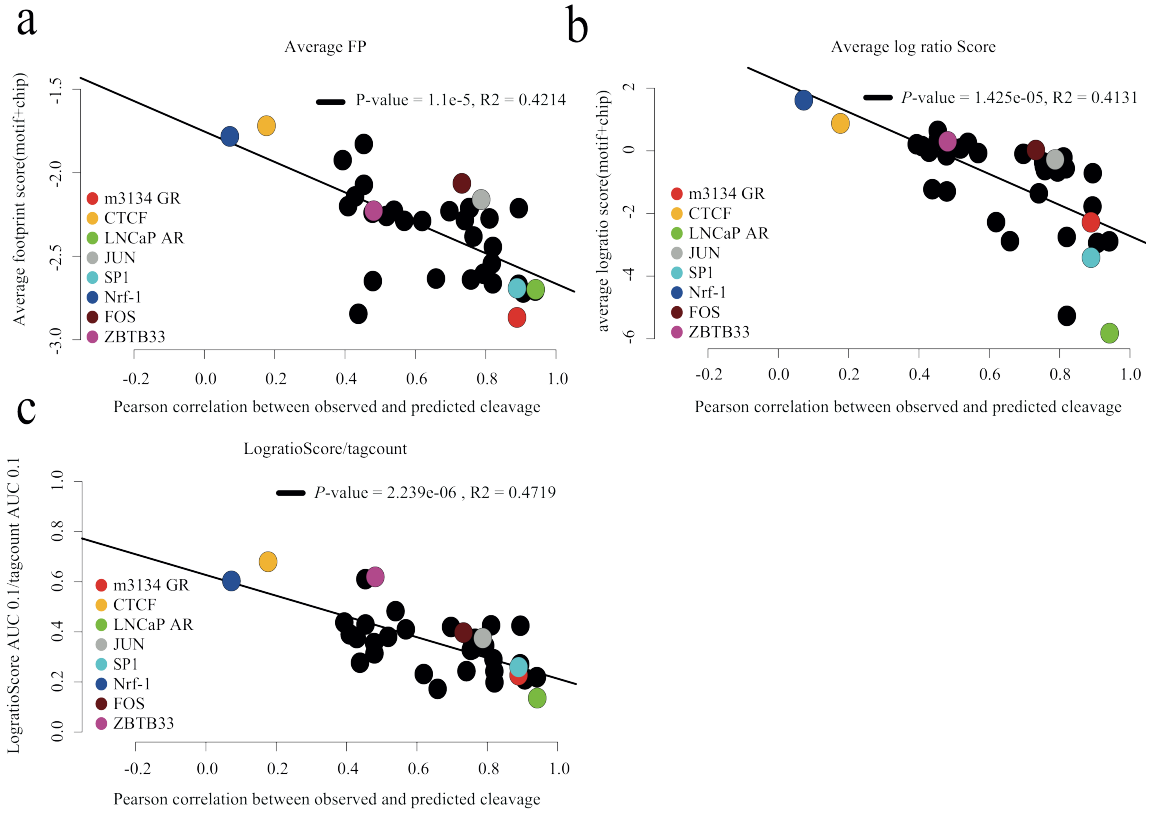
Supplementary Figure 14. Benzonase, cyanase and DNaseI cleavage biases. Comparison of cleavage bias in **(a)** benzonase, **(b)** cyanase, **(c)** DNaseI in mouse liver. Comparison of **(d)** cyanase with benzonase, **(e)** DNaseI with benzonase, **(f)** DNaseI with cyanase, **(g)** DNaseI in mouse liver with DNaseI in human IMR90 naked DNA. All three nucleases exhibit strong 6-mer DNA cleavage biases. The biases of benzonase and cyanase are similar to each other but distinct from that of DNaseI.



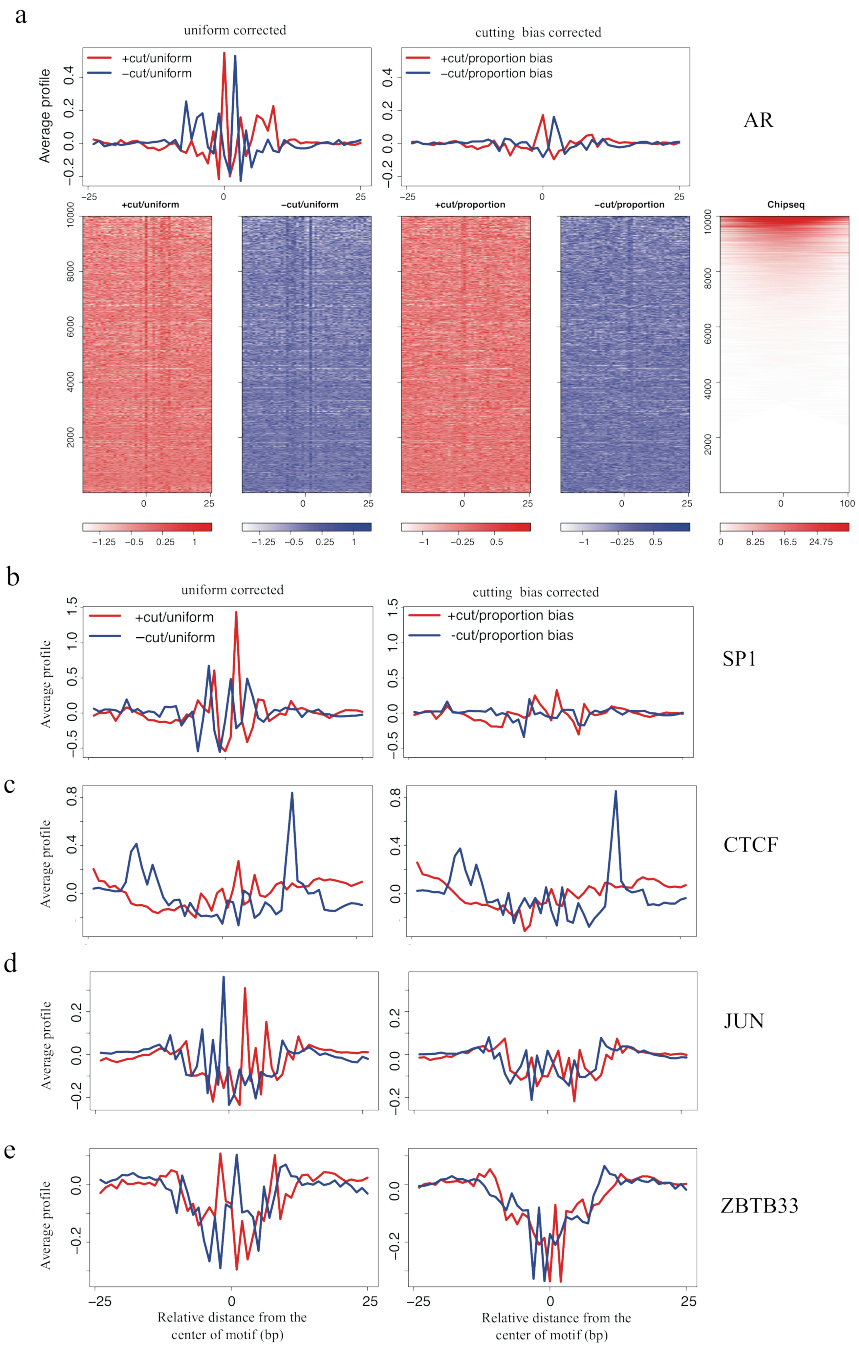
Supplementary Figure 15. The sequence bias contribution to DNaseI cleavage patterns in CTCF and AR. Pearson correlation coefficients were calculated between observed locus specific cleavage patterns (red) and the mean observed cleavage patterns derived from DNaseI cuts at **(a)** AR and **(b)** CTCF motifs in ChIP-seq identified binding sites. To show the contribution of sequence bias, Pearson correlation coefficients were also calculated between 6-mer model predicted cleavage patterns (blue) and the mean observed cleavage patterns. In the case of AR **(a)** there is an almost complete overlap between distributions for observed and model predicted cases. In sharp contrast, the CTCF **(b)** distributions are clearly different. Examining sites that are DNaseI hypersensitive, contain the respective AR **(c)** and CTCF **(d)** binding motifs, but are not enriched in ChIP-seq signal for the respective factors we see for AR **(c)** the model predicted and AR distributions are similar, as before. In the CTCF case **(d)** the observed distribution is now more similar to the predicted one. MAD: median absolute deviation.



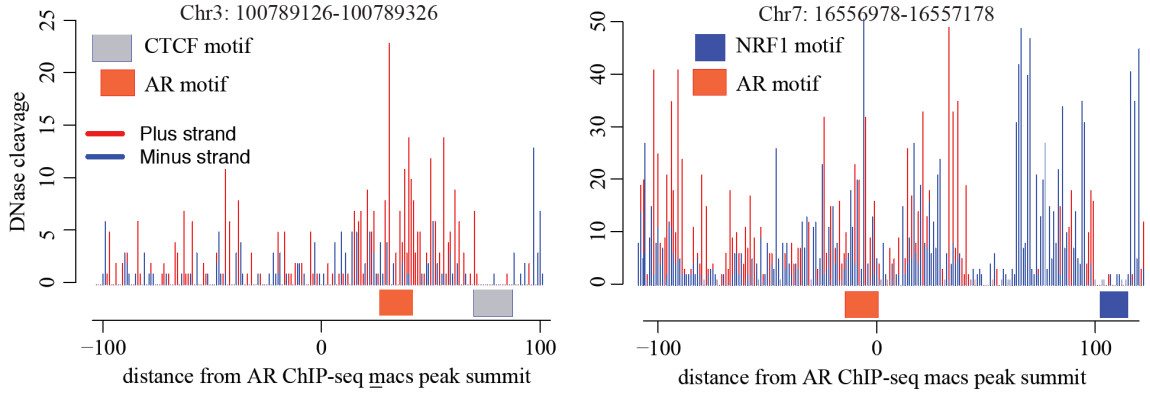
Supplementary Figure 16. Predicting transcription factor binding from bias normalized DHS. **(a)** Average footprint scores relative to Pearson correlation coefficients between cleavage and 6-mer model predicted cleavage for 36 transcription factors. The correlation between observed cleavage patterns and predicted cleavage bias is inversely related to the strength of the trough-like footprint pattern. **(b)** Average sequence bias normalized footprint score relative to the correlation between observed cleavage and predicted cleavage. The sequence bias normalized footprint score is the sum of the normalized DNase-seq sensitivity values in the central region, spanning the TF binding motif, subtracted from the sum of normalized DNaseI sensitivity values in the regions flanking the motif: $f^{\text{seq-norm}} = \sum_{s\in\{+,-\}}(\sum_{i\in\text{flank}} z_i^s - \sum_{i\in\text{center}} z_i^s)$ (see Online Methods for the definition of z). **(c)** The relative prediction power of $f^{\text{seq-norm}}$ in relation to the correlation between observed cleavage to bias correlation. The relative prediction power is calculated as the ratio of areas under the ROC curve.



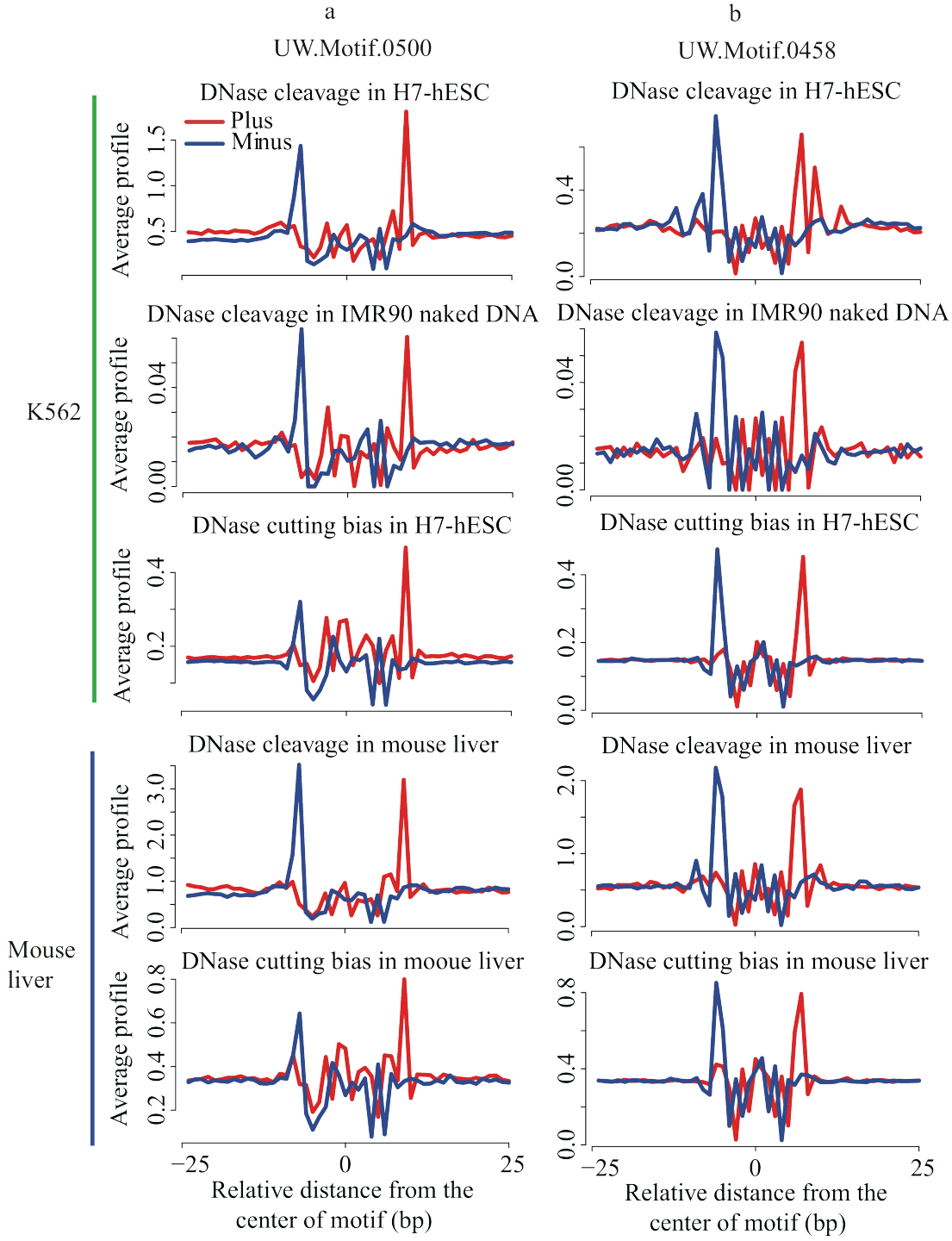
Supplementary Figure 17. Uniform and sequence bias normalizations of DNase-seq. We calculate two normalizations of DNaseI cleavage counts, a *sequence bias* normalization that takes into account the intrinsic cleavage biases of all nucleotides in a 50bp window, and a *uniform* normalization that assumes that all 6-mers are cut with the same frequency (see Online Methods). These two normalizations differ only in the sequence bias parameters, allowing data from these normalizations to be compared to each other on the same scale. (a) Uniform (left) and sequence bias (right) normalizations of AR. Heatmaps represent strand specific, nucleotide resolution normalized 5' tag counts relative to the center of the AR motif, with rows ordered by ChIP-seq tag count for AR in LNCaP. (b) Uniform (left) and sequence bias normalization (right) for SP1 (c) CTCF (d) JUN and (e) ZBTB33.



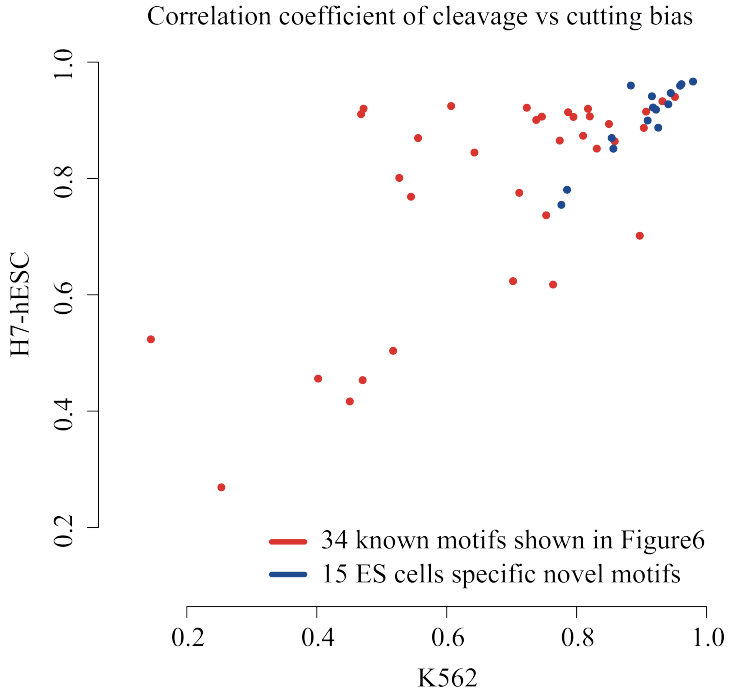
Supplementary Figure 18. Examples of DNaseI cleavage patterns at AR ChIP-seq peaks. The DNaseI troughs we see in these regions are not at the AR motif but are instead associated with the motifs of factors that have stronger DNaseI footprints.



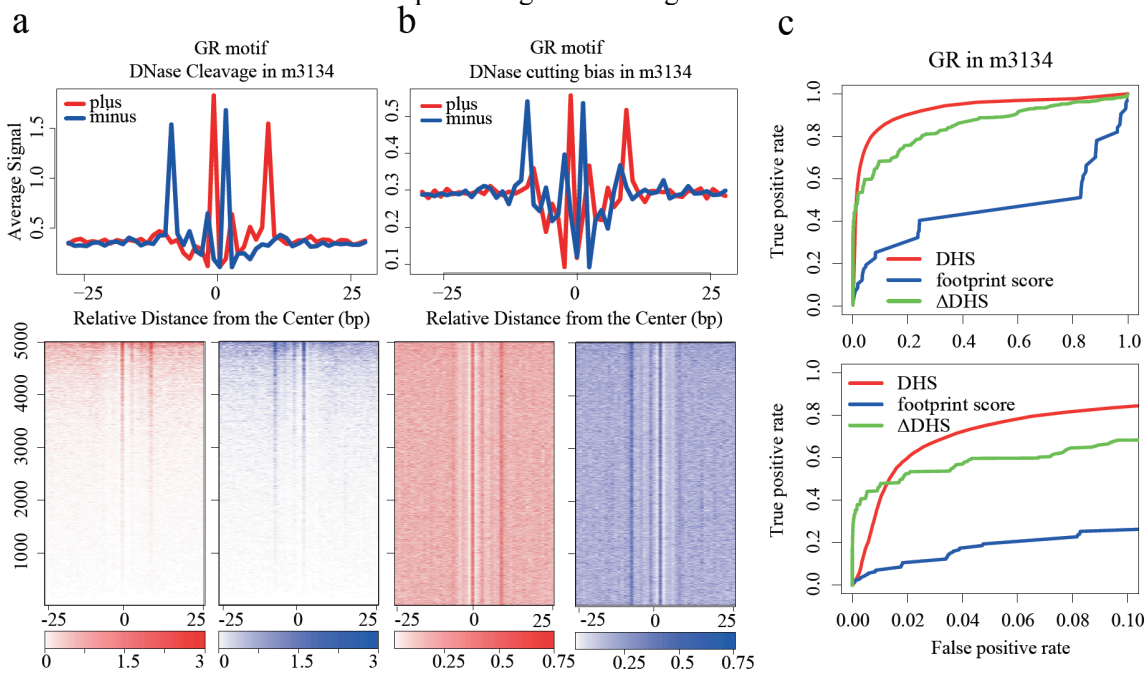
Supplementary Figure 19. Comparison of observed, predicted and naked DNA cleavage bias in *de novo* motifs UW.Motif.0500 and UW.Motif.0458 and UW.Motif.0423. Observed and predicted DNaseI cleavage patterns in DNaseI peaks centered on motif hits for **(a)** UW.Motif.0500 and **(b)** UW.Motif.0458. The rows, are as follows: observed DNaseI cleavage pattern in K562, observed cleavage in IMR90 naked DNA at K562 loci, K562 6-mer bias predicted cleavage at K562 loci, observed DNaseI cleavage in mouse liver, mouse liver 6-mer predicted cleavage at mouse liver loci. In each case the plots are based on the 5000 top scoring motif matches within K562 or mouse liver DNase-seq peaks. In the human data only mappable regions are included.



Supplementary Figure 20. Summary of comparison of observed and predicted cleavage bias in known and *de novo* motifs from Neph *et al* (2012). In this analysis, Pearson correlation coefficients are summarized for 15 ES cell specific *de novo* motifs from Neph et al (2012) along with the 34 known motifs in Supplementary Table 1 and Figure 6b (AR and GR excluded). In this analysis, motifs are scanned in DNase-seq peak regions, while in the main Figure 6b and Supplementary Table 2, motifs are further filtered using ChIP-seq data, resulting in some differences between the correlation coefficients for the 34 known motifs in this figure and main Figure 6b and Supplementary Table 2.



Supplementary Figure 21. DNaseI cleavage at the GR motif. **(a)** DNaseI cleavage in m3134 cells at GR motifs, showing the average cleavage, and in the heatmaps, site specific cleavage for the plus and minus strand. **(b)** Cleavage pattern predicted for GR based on the 6-mer DNaseI cutting bias. **(c)** Performance of the absolute DNase-seq tag count (DHS), the DNaseI footprint score, and the differential DNaseI score ΔDHS in predicting GR binding in m3134 cells.



Supplementary Table 1. Ten least sensitive and ten most sensitive 6-mers in K562 DNase-seq data.

6-mer	Bias estimate
CAGATA	0.0086
CAGATT	0.0087
CAGATC	0.0088
CAGATG	0.0091
GAGATA	0.0099
GAGATG	0.0099
GAGATT	0.0100
GAGATC	0.0104
CACATA	0.0106
CAGGTG	0.0107
TCTTAA	2.2376
GCTTGT	2.2541
GCTTAC	2.2874
TCTTGT	2.3062
TCTTAC	2.3557
ACTTAA	2.3871
ACTTGA	2.4169
ACTTGC	2.4565
ACTTAC	2.4978
ACTTGT	2.5958

Supplementary Table 2. Performance of footprint scores relative to total DNase-seq tag counts (DHS) in the recovery of ChIP-seq determined transcription factor binding sites.

Transcription Factor	Pearson Correlation	Tag Count AUC	Footprint AUC	Footprint $AUC_{FPR<0.1}$/ Tag-count $AUC_{FPR<0.1}$
ATF3	0.54	0.92	0.56	0.34
CEBPB	0.81	0.71	0.52	0.41
CTCF	0.18	0.88	0.67	0.74
E2F6	0.41	0.92	0.57	0.36
EGR1	0.46	0.89	0.58	0.36
ELF1	0.76	0.90	0.48	0.20
ETS1	0.89	0.93	0.47	0.21
FOS	0.74	0.96	0.60	0.28
FOSL1	0.75	0.92	0.56	0.32
GABPA	0.79	0.88	0.48	0.20
GATA1	0.39	0.92	0.65	0.42
GATA2	0.43	0.95	0.60	0.35
IRF1	0.62	0.98	0.53	0.22
JUN	0.79	0.94	0.58	0.30
JUND	0.74	0.91	0.54	0.35
MAX	0.57	0.83	0.54	0.34
MEF2A	0.66	0.93	0.48	0.15
MYC	0.52	0.97	0.57	0.31
NFE2	0.46	0.92	0.66	0.54
Nrf-1	0.07	0.95	0.69	0.56
NRSF	0.90	0.85	0.56	0.36
PU.1	0.82	0.83	0.48	0.31
SIX5	0.82	0.93	0.50	0.20
SP1	0.89	0.97	0.47	0.14
SP2	0.90	0.87	0.44	0.17
SRF	0.82	0.87	0.47	0.16
STAT1	0.94	0.93	0.47	0.19
TAL1	0.48	0.96	0.48	0.17
USF1	0.70	0.78	0.53	0.39
USF2	0.44	0.90	0.43	0.26
YY1	0.75	0.85	0.54	0.31
ZBTB33	0.48	0.83	0.55	0.38
ZBTB7A	0.48	0.87	0.54	0.33
ZNF263	0.77	0.75	0.49	0.35
AR	0.94	0.94	0.49	0.02
GR	0.87	0.92	0.46	0.25

In this table $AUC_{FPR<0.1}$ refers to the area under the ROC curve for false positive rates between 0.0 and 0.1. AUC refers to the area under the full ROC curve.

Supplementary Table 3. Primers used in this work.

	Forward	Reverse
CTCF4	CCCCAGAGAGTAGGGAACAG	GGCACGCAAAGACATACTGA
CTCF10	AGAGCACCCCCTACTGGCTAA	TAAGAAGCTGTGCGCGATGAC
CTCF15	CTTAGGGGACCTTTTCTACAGGA	GAGCACTTGTAAACTCGTCTGCT
GAPDH_pro	AAAAGCGGGGAGAAAGTAGG	GCTGCGGGCTCAATTTATAG
B-ACT_pro	TCGAGCCATAAAAGGCAACT	TCTCCCTCCTCCTCTTCCTC
RPS28_pro	CGGCAGCTGACACGTAAGTC	CAATGCAGAGCGACACTCAC