BinDNase: A discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data - Supplemental material

Juhani Kähärä, Harri Lähdesmäki

February 6, 2015

1 Location of the datasets

The datasets used in this work can be download from the ENCODE download interface http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeUwDgf. This site contains coordinates for the hotspots and the DNase I hypersensitivity measurements.

The ChIP-seq datasets can be found from the download interface at http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeAwgTfbsUniform.

2 Filenames of the ChIP-seq datasets

TF	Celltype	Filename
BACH1	K562	wgEncodeAwgTfbsSydhK562Bach1sc14700IggrabUniPk.narrowPeak
GATA1	K562	wgEncodeAwgTfbsSydhK562Gata1UcdUniPk.narrowPeak
MAFK	K562	wgEncodeAwgTfbsSydhK562Mafkab50322IggrabUniPk.narrowPeak
CEBPB	K562	wgEncodeAwgTfbsSydhK562CebpbIggrabUniPk.narrowPeak
E2F4	K562	wgEncodeAwgTfbsSydhK562E2f4UcdUniPk.narrowPeak
EGR1	K562	wgEncodeAwgTfbsHaibK562Egr1V0416101UniPk.narrowPeak
ELF1	K562	wgEncodeAwgTfbsHaibK562Elf1sc631V0416102UniPk.narrowPeak
ELK1	K562	wgEncodeAwgTfbsSydhK562Elk112771IggrabUniPk.narrowPeak
CTCFL	K562	wgEncodeAwgTfbsHaibK562Ctcflsc98982V0416101UniPk.narrowPeak
FOSL1	K562	wgEncodeAwgTfbsHaibK562Fosl1sc183V0416101UniPk.narrowPeak
FOS	K562	wgEncodeAwgTfbsSydhK562CfosUniPk.narrowPeak
MAFF	K562	wgEncodeAwgTfbsSydhK562MaffIggrabUniPk.narrowPeak
MXI1	K562	wgEncodeAwgTfbsSydhK562Mxi1af4185IggrabUniPk.narrowPeak
RFX5	K562	wgEncodeAwgTfbsSydhK562Rfx5IggrabUniPk.narrowPeak
SMC3	K562	wgEncodeAwgTfbsSydhK562Smc3ab9263IggrabUniPk.narrowPeak
RAD21	K562	wgEncodeAwgTfbsHaibK562Rad21V0416102UniPk.narrowPeak
STAT5	K562	wgEncodeAwgTfbsHaibK562Stat5asc74442V0422111UniPk.narrowPeak
SP2	K562	wgEncodeAwgTfbsHaibK562Sp2sc643V0416102UniPk.narrowPeak

TAF1	K562	wgEncodeAwgTfbsHaibK562Taf1V0416101UniPk.narrowPeak
USF2	K562	wgEncodeAwgTfbsSydhK562Usf2IggrabUniPk.narrowPeak
ZBTB33	K562	wgEncodeAwgTfbsHaibK562Zbtb33Pcr1xUniPk.narrowPeak
ZBTB7A	K562	wgEncodeAwgTfbsHaibK562Zbtb7asc34508V0416101UniPk.narrowPeak
ZNF143	K562	wgEncodeAwgTfbsSydhK562Znf143IggrabUniPk.narrowPeak
ATF3	K562	wgEncodeAwgTfbsHaibK562Atf3V0416101UniPk.narrowPeak
BDP1	K562	wgEncodeAwgTfbsSydhK562Bdp1UniPk.narrowPeak
BHLHE40	K562	wgEncodeAwgTfbsSydhK562Bhlhe40nb100IggrabUniPk.narrowPeak
GABPA	K562	wgEncodeAwgTfbsHaibK562GabpV0416101UniPk.narrowPeak
CTCF	K562	wgEncodeAwgTfbsBroadK562CtcfUniPk.narrowPeak
JUND	K562	wgEncodeAwgTfbsSydhK562JundIggrabUniPk.narrowPeak
NR2C2	K562	wgEncodeAwgTfbsSydhK562Tr4UcdUniPk.narrowPeak
NR2F2	K562	wgEncodeAwgTfbsHaibK562Nr2f2sc271940V0422111UniPk.narrowPeak
E2F6	K562	wgEncodeAwgTfbsHaibK562E2f6V0416102UniPk.narrowPeak
ETS1	K562	wgEncodeAwgTfbsHaibK562Ets1V0416101UniPk.narrowPeak
SP1	K562	wgEncodeAwgTfbsHaibK562Sp1Pcr1xUniPk.narrowPeak
USF1	K562	wgEncodeAwgTfbsHaibK562Usf1V0416101UniPk.narrowPeak
JUN	K562	wgEncodeAwgTfbsSydhK562CjunUniPk.narrowPeak
THAP1	K562	wgEncodeAwgTfbsHaibK562Thap1sc98174V0416101UniPk.narrowPeak
JUNB	K562	wgEncodeAwgTfbsUchicagoK562EjunbUniPk.narrowPeak
MAX	K562	wgEncodeAwgTfbsHaibK562MaxV0416102UniPk.narrowPeak
MEF2A	K562	wgEncodeAwgTfbsHaibK562Mef2aV0416101UniPk.narrowPeak
MYC	K562	wgEncodeAwgTfbsSydhK562CmycIggrabUniPk.narrowPeak
NFE2	K562	wgEncodeAwgTfbsSydhK562Nfe2UniPk.narrowPeak
REST	K562	wgEncodeAwgTfbsHaibK562NrsfV0416102UniPk.narrowPeak
NFYA	K562	wgEncodeAwgTfbsSydhK562NfyaUniPk.narrowPeak
NFYB	K562	wgEncodeAwgTfbsSydhK562NfybUniPk.narrowPeak
EP300	K562	wgEncodeAwgTfbsSydhK562P300IggrabUniPk.narrowPeak
SPI1	K562	wgEncodeAwgTfbsHaibK562Pu1Pcr1xUniPk.narrowPeak
SRF	K562	wgEncodeAwgTfbsHaibK562SrfV0416101UniPk.narrowPeak
TBP	K562	wgEncodeAwgTfbsSydhK562TbpIggmusUniPk.narrowPeak
TEAD4	K562	wgEncodeAwgTfbsHaibK562Tead4sc101184V0422111UniPk.narrowPeak
NRF1	K562	wgEncodeAwgTfbsSydhK562Nrf1IggrabUniPk.narrowPeak
YY1	K562	wgEncodeAwgTfbsHaibK562Yy1V0416102UniPk.narrowPeak
BRF1	K562	wgEncodeAwgTfbsSydhK562Brf1UniPk.narrowPeak
ZNF263	K562	wgEncodeAwgTfbsSydhK562Znf263UcdUniPk.narrowPeak
ATF1	K562	wgEncodeAwgTfbsSydhK562Atf106325UniPk.narrowPeak
ZNF274	K562	wgEncodeAwgTfbsSydhK562Znf274m01UcdUniPk.narrowPeak
GATA2	K562	wgEncodeAwgTfbsSydhK562Gata2UcdUniPk.narrowPeak
JUN	HepG2	wgEncodeAwgTfbsSydhHepg2CjunIggrabUniPk.narrowPeak
BHLHE40	HepG2	wgEncodeAwgTfbsSydhHepg2Bhlhe40cIggrabUniPk.narrowPeak
CEBPB	HepG2	wgEncodeAwgTfbsSydhHepg2CebpbIggrabUniPk.narrowPeak
ELF1	HepG2	wgEncodeAwgTfbsHaibHepg2Elf1sc631V0416101UniPk.narrowPeak
GABPA	HepG2	wgEncodeAwgTfbsHaibHepg2GabpPcr2xUniPk.narrowPeak
JUND	HepG2	wgEncodeAwgTfbsSydhHepg2JundIggrabUniPk.narrowPeak
MAFF	HepG2	wgEncodeAwgTfbsSydhHepg2Maffm8194IggrabUniPk.narrowPeak

SP2	HepG2	wgEncodeAwgTfbsHaibHepg2Sp2V0422111UniPk.narrowPeak
TAF1	HepG2	wgEncodeAwgTfbsHaibHepg2Taf1Pcr2xUniPk.narrowPeak
USF1	HepG2	wgEncodeAwgTfbsHaibHepg2Usf1Pcr1xUniPk.narrowPeak
YY1	HepG2	wgEncodeAwgTfbsHaibHepg2Yy1sc281V0416101UniPk.narrowPeak
SRF	HepG2	wgEncodeAwgTfbsHaibHepg2SrfV0416101UniPk.narrowPeak
EP300	HepG2	wgEncodeAwgTfbsHaibHepg2P300V0416101UniPk.narrowPeak
CTCF	HepG2	wgEncodeAwgTfbsHaibHepg2Ctcfsc5916V0416101UniPk.narrowPeak
ZNF274	HepG2	wgEncodeAwgTfbsSydhHepg2Znf274UcdUniPk.narrowPeak
ATF3	HepG2	wgEncodeAwgTfbsHaibHepg2Atf3V0416101UniPk.narrowPeak
MAFK	HepG2	wgEncodeAwgTfbsSydhHepg2Mafkab50322IggrabUniPk.narrowPeak
RFX5	HepG2	wgEncodeAwgTfbsSydhHepg2Rfx5200401194IggrabUniPk.narrowPeak
SP1	HepG2	wgEncodeAwgTfbsHaibHepg2Sp1Pcr1xUniPk.narrowPeak
USF2	HepG2	wgEncodeAwgTfbsSydhHepg2Usf2IggrabUniPk.narrowPeak
MXI1	HepG2	wgEncodeAwgTfbsSydhHepg2Mxi1UniPk.narrowPeak
MYC	HepG2	wgEncodeAwgTfbsUtaHepg2CmycUniPk.narrowPeak
RAD21	HepG2	wgEncodeAwgTfbsHaibHepg2Rad21V0416101UniPk.narrowPeak
REST	HepG2	wgEncodeAwgTfbsHaibHepg2NrsfV0416101UniPk.narrowPeak
MAX	HepG2	wgEncodeAwgTfbsSydhHepg2MaxIggrabUniPk.narrowPeak
SMC3	HepG2	wgEncodeAwgTfbsSydhHepg2Smc3ab9263IggrabUniPk.narrowPeak
NRF1	HepG2	wgEncodeAwgTfbsSydhHepg2Nrf1IggrabUniPk.narrowPeak
ZBTB7A	HepG2	wgEncodeAwgTfbsHaibHepg2Zbtb7aV0416101UniPk.narrowPeak
TBP	HepG2	wgEncodeAwgTfbsSydhHepg2TbpIggrabUniPk.narrowPeak
TEAD4	HepG2	wgEncodeAwgTfbsHaibHepg2Tead4sc101184V0422111UniPk.narrowPeak
NR2C2	HepG2	wgEncodeAwgTfbsSydhHepg2Tr4UcdUniPk.narrowPeak

3 High-resolution analysis improves predictions

The performance difference between the DNase activity predictor (the binding score is the total number of DNase cuts within 50bp window) and BinDNase is plotted against the total number of DNase cuts in the window. The DNase activity predictor works well for TFs whose binding sites are located within the high DNase activity regions, especially for negative set 1. There are however many TFs which bind to sites with lower DNase activity when only DNase hotspots are considered. For some of the TFs the DNase activity predictor fails completely.

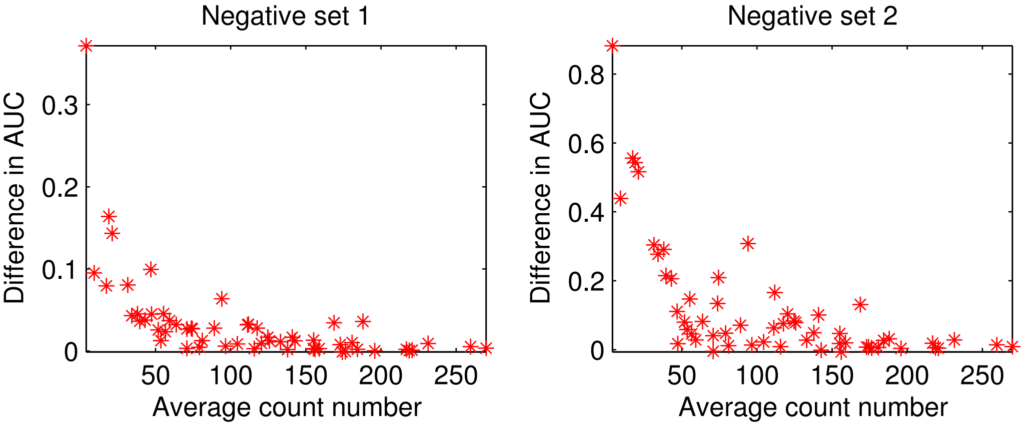

Figure 1: BinDNase improves the predictions the most for TFs who bind to low DNase I activity sites.

4 Models for K562 (left) and HepG2 (right)

Figures in this section show the logistic regression models for different TFs. In each figure, the left panel shows the model trained on cell line K562 and the right panel shows the model trained on cell line HepG2. In each panel, the upper panels show the average DNase I cleavage centered at the TF binding motifs. The coloured bars indicate the optimised feature selection and their coefficients in the logistic regression model. Red (resp. blue) colour indicates positive (resp. negative) coefficient.

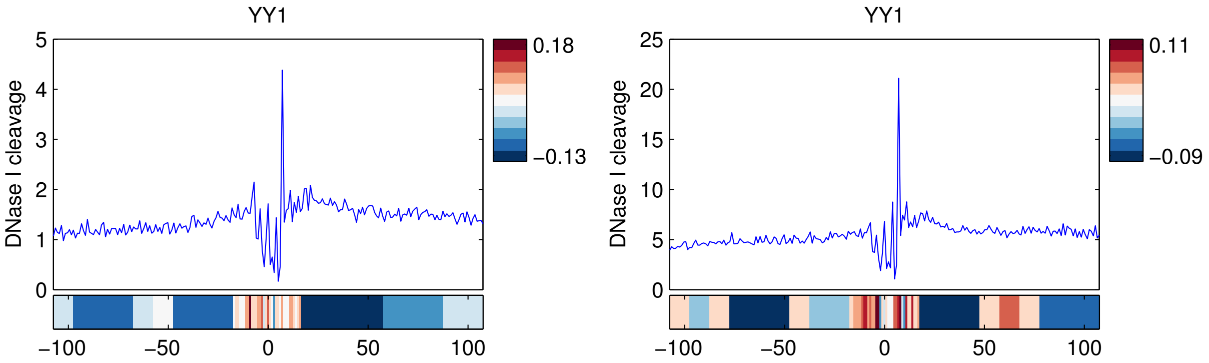

Figure 2: BinDNase models for YY1

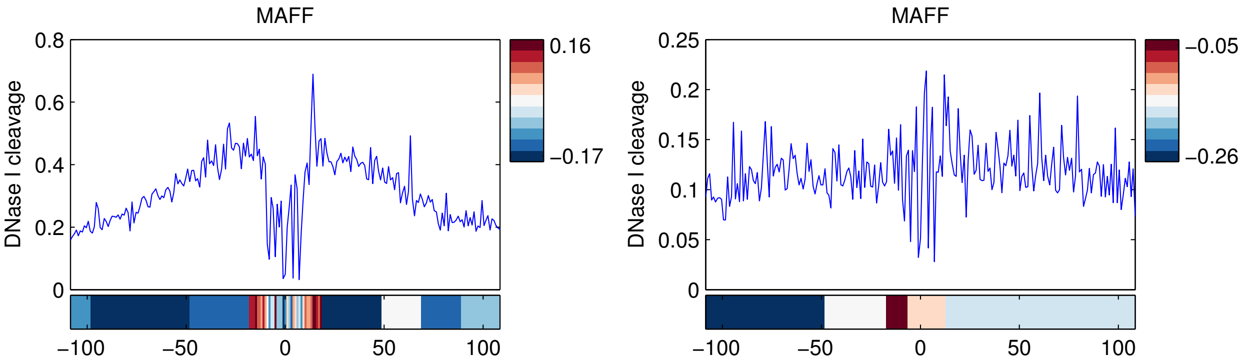

Figure 3: BinDNase models for MAFF

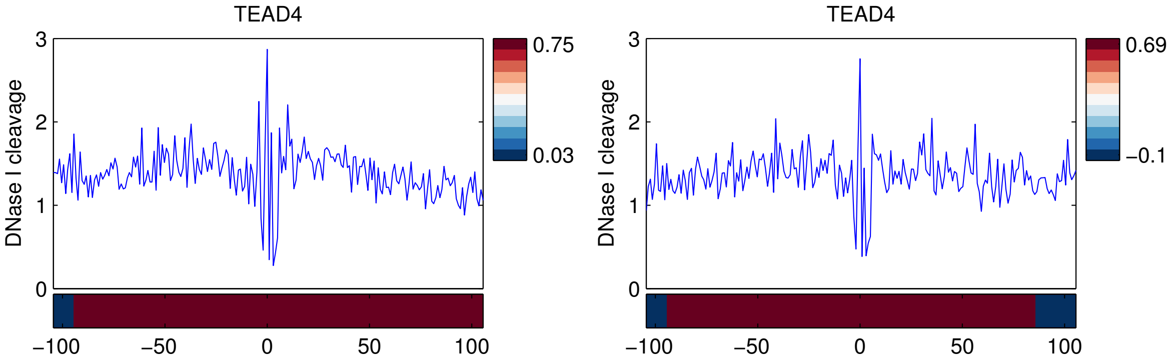

Figure 4: BinDNase models for TEAD4

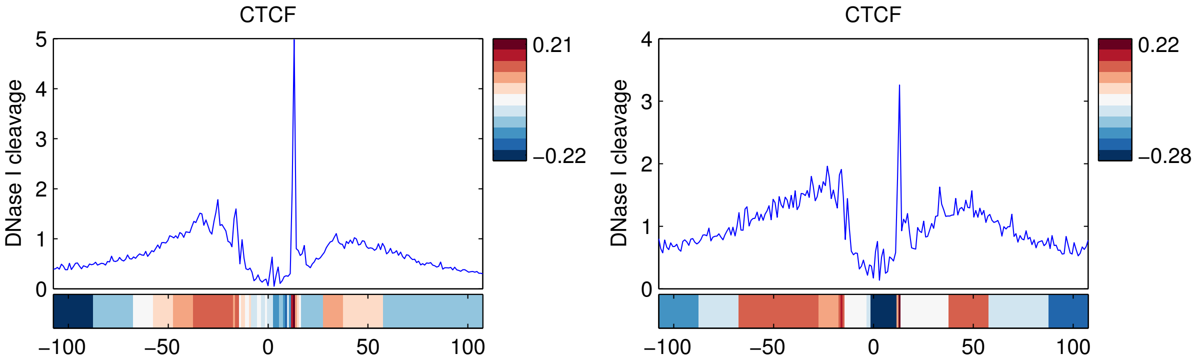

Figure 5: BinDNase models for CTCF

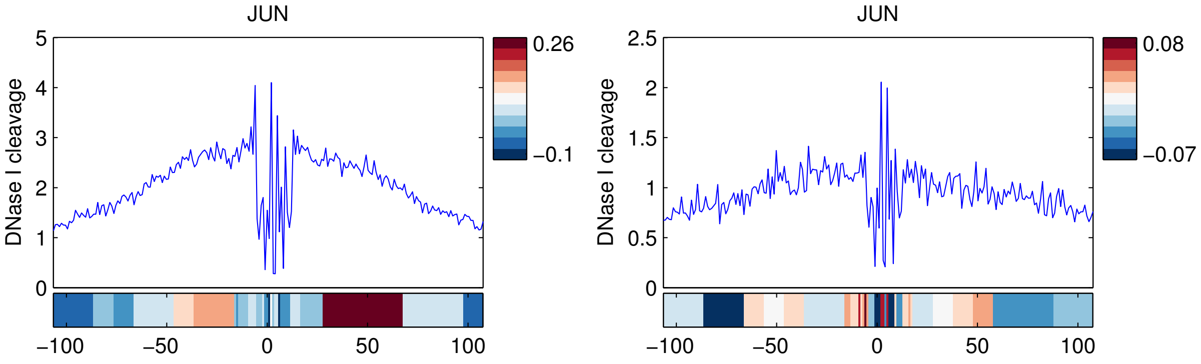

Figure 6: BinDNase models for JUN

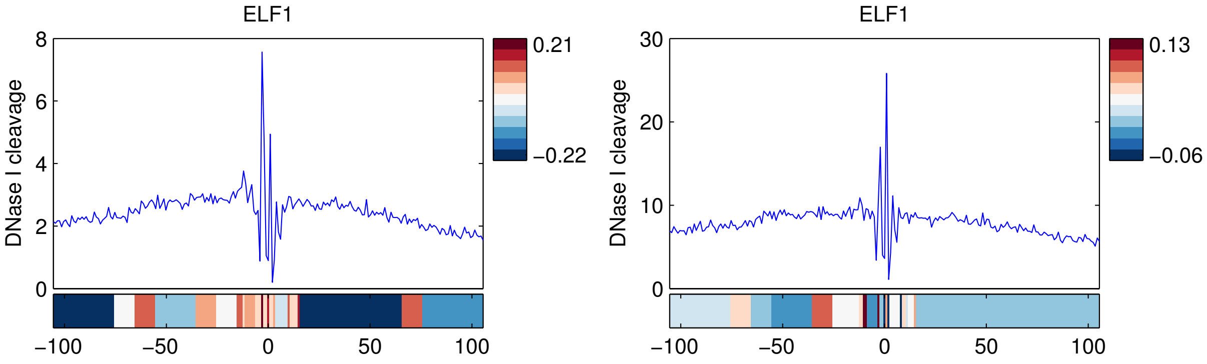

Figure 7: BinDNase models for ELF1

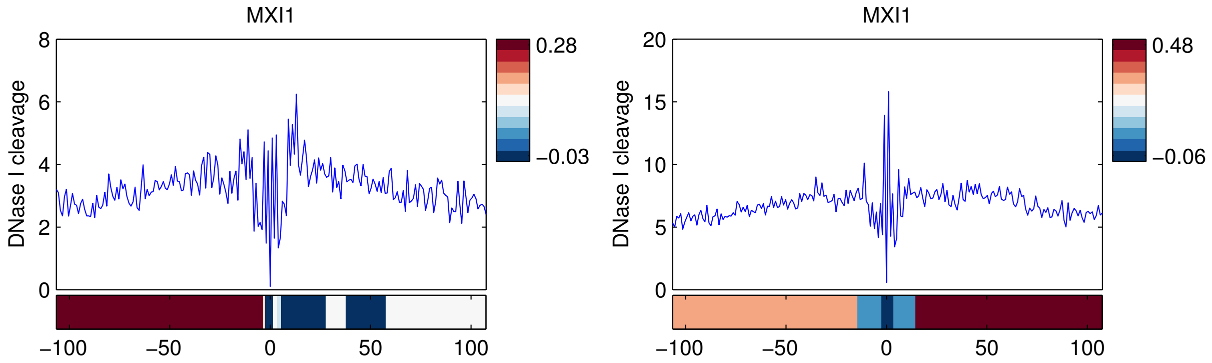

Figure 8: BinDNase models for MXI1

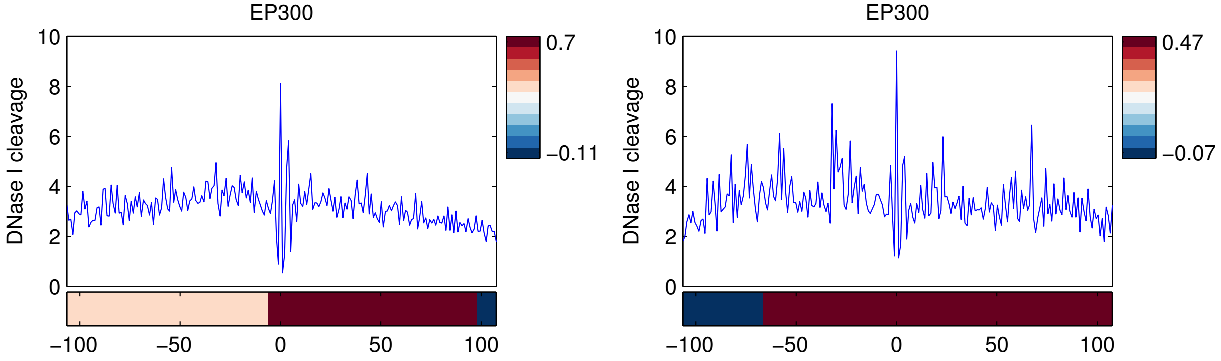

Figure 9: BinDNase models for EP300

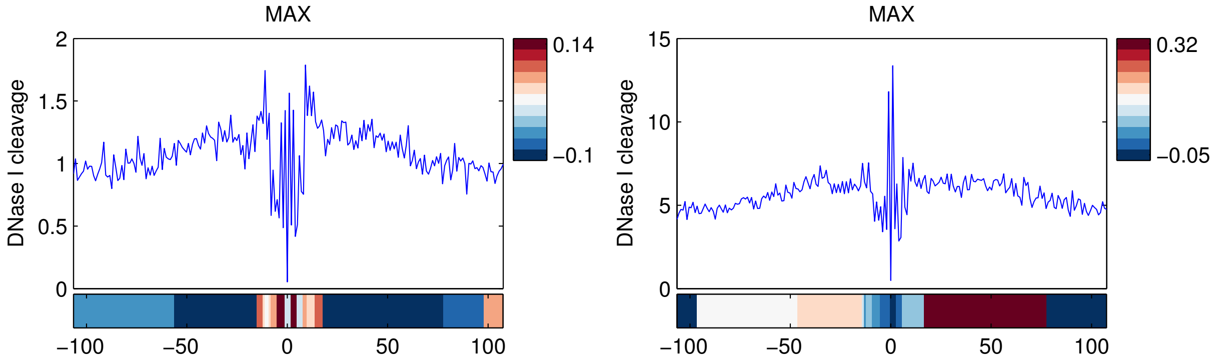

Figure 10: BinDNase models for MAX

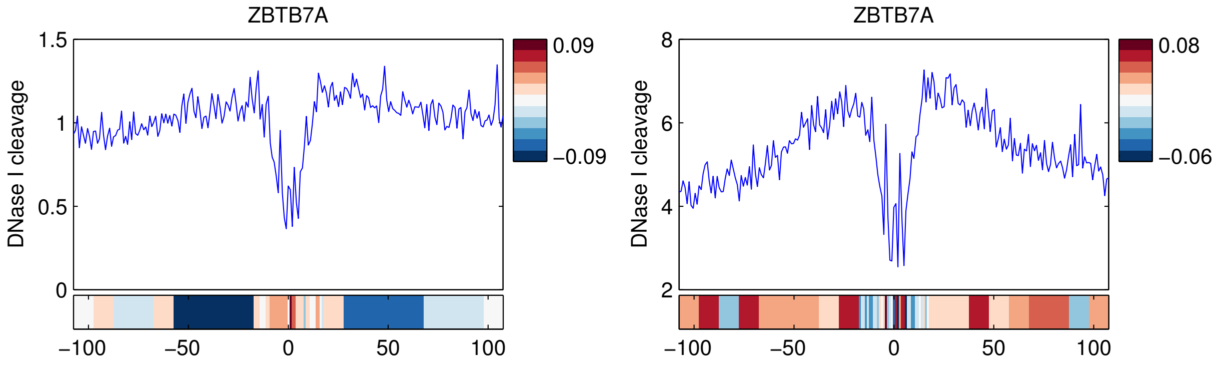

Figure 11: BinDNase models for ZBTB7A1

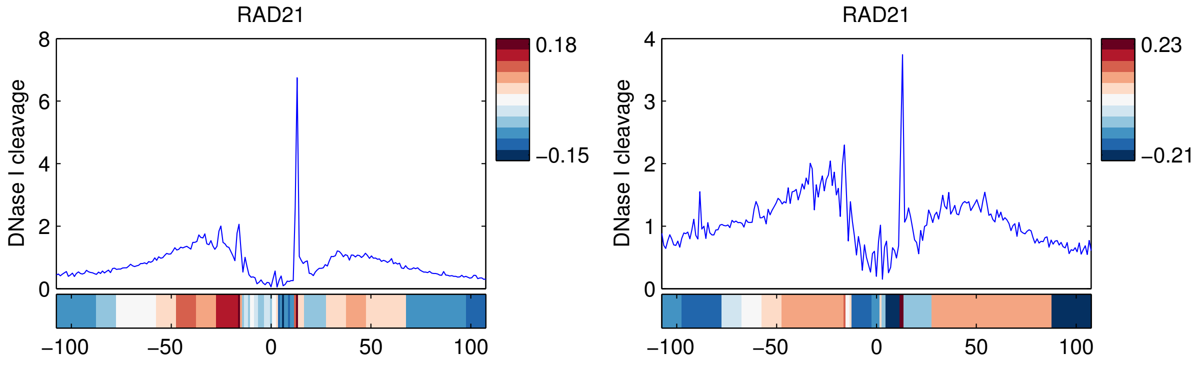

Figure 12: BinDNase models for RAD1

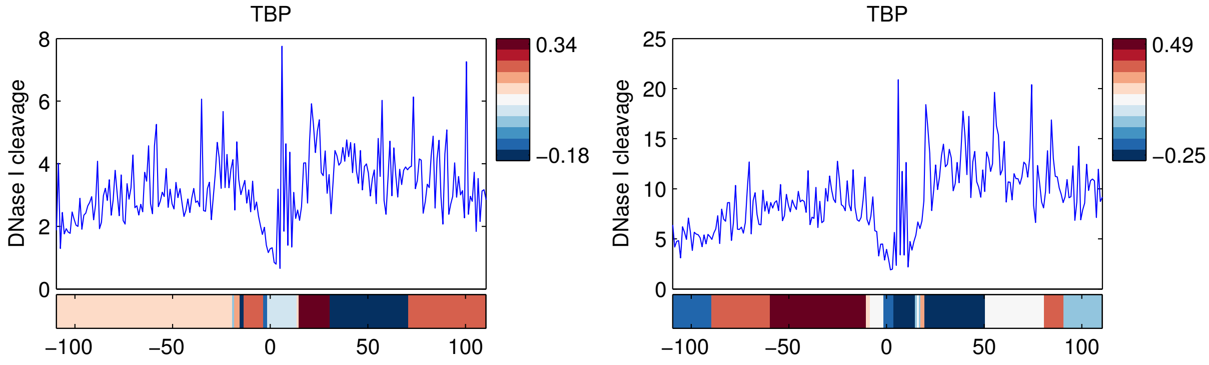

Figure 13: BinDNase models for TBP1

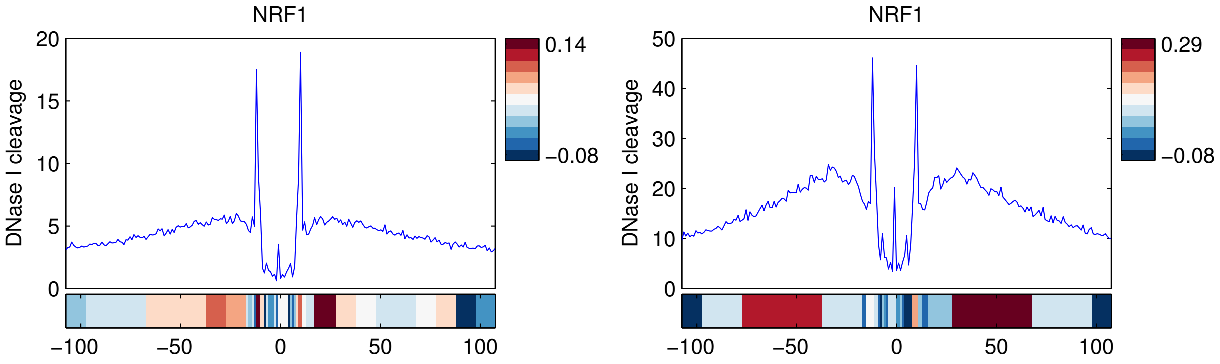

Figure 14: BinDNase models for NRF1

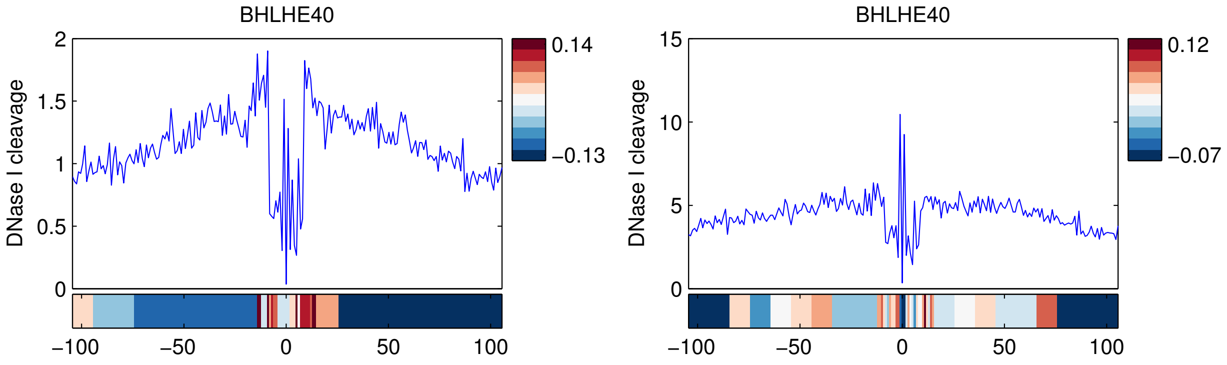

Figure 15: BinDNase models for BHLHE40

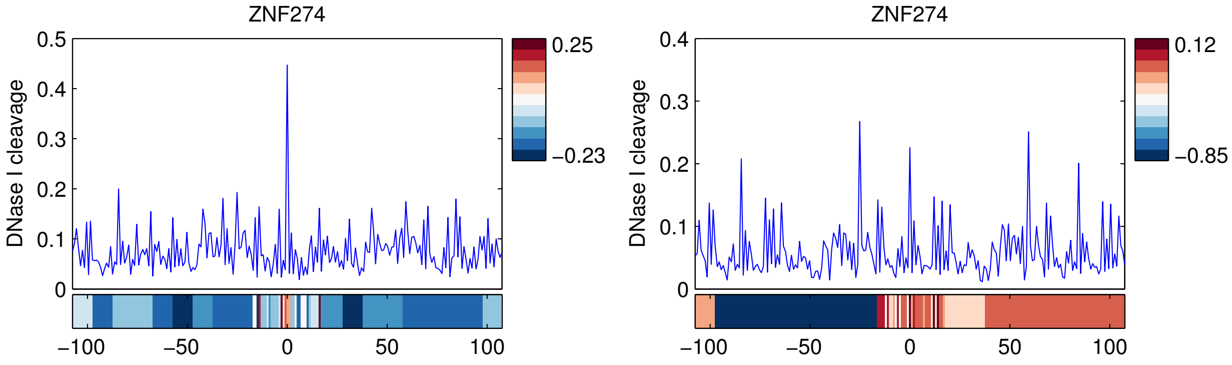

Figure 16: BinDNase models for ZNF274

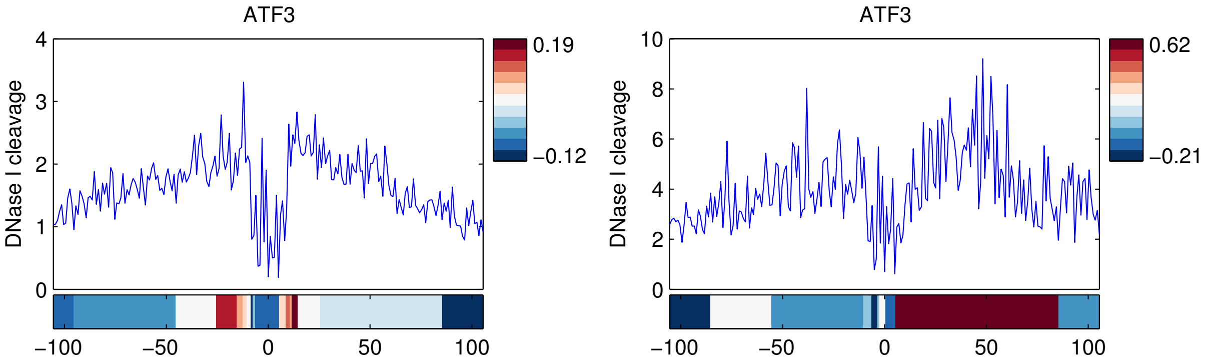

Figure 17: BinDNase models for ATF3

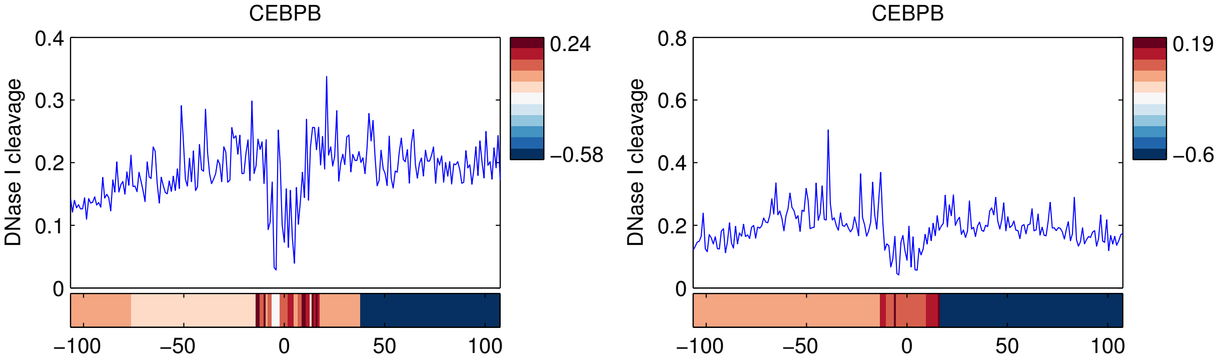

Figure 18: BinDNase models for CEBPB

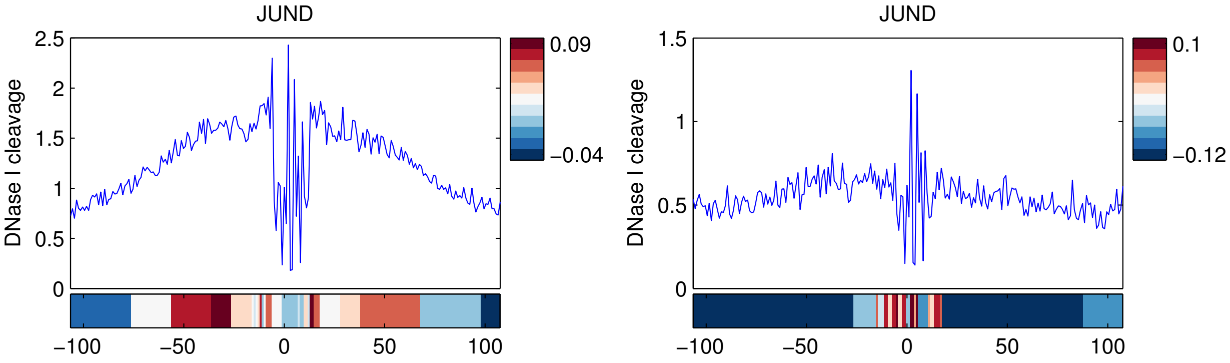

Figure 19: BinDNase models for JUND

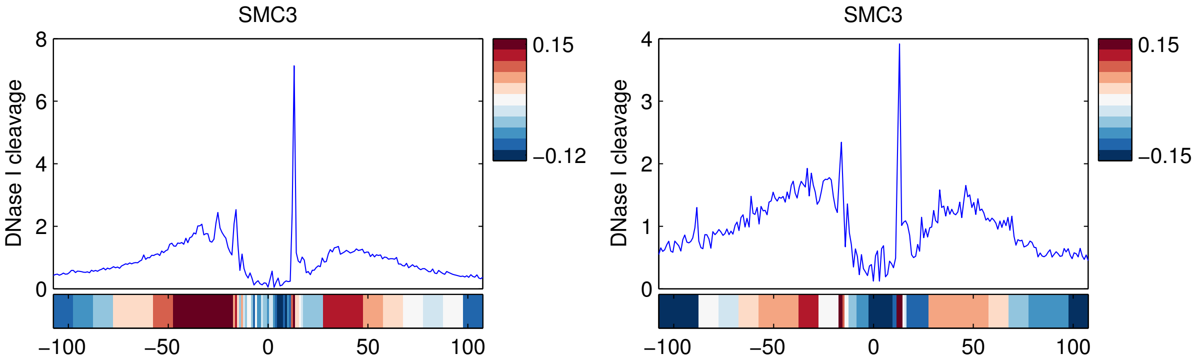

Figure 20: BinDNase models for SMC3

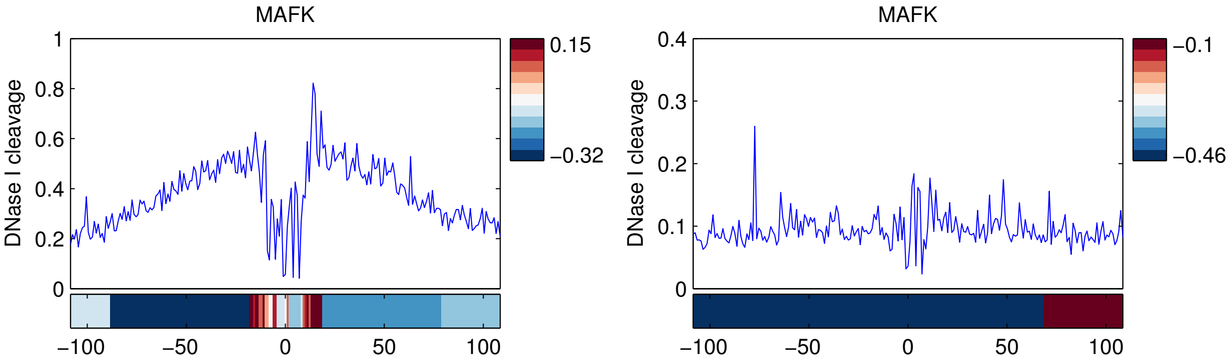

Figure 21: BinDNase models for MAFK

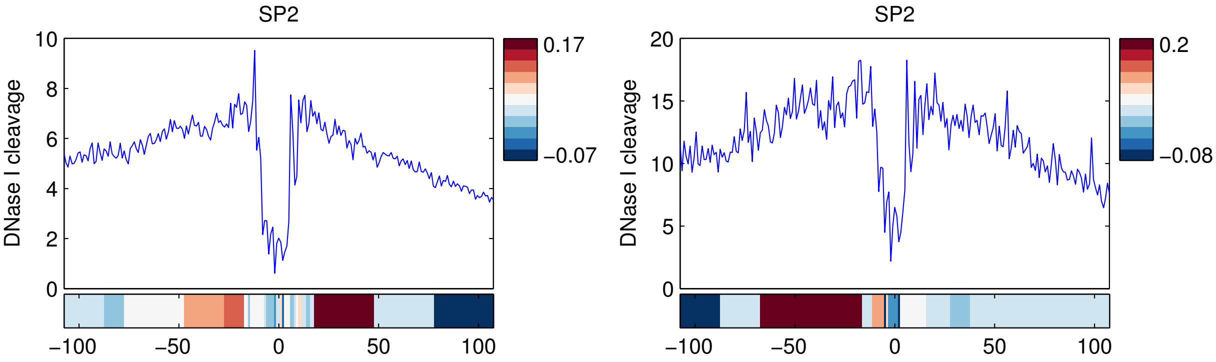

Figure 22: BinDNase models for SP2

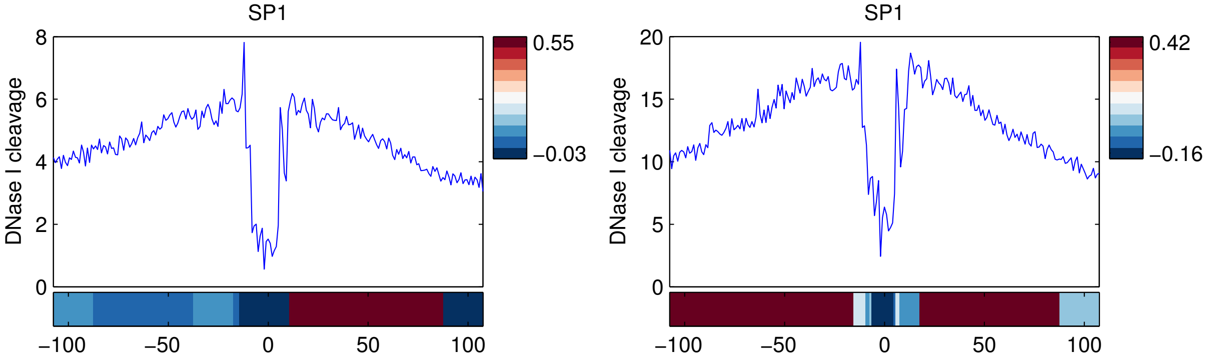

Figure 23: BinDNase models for SP1

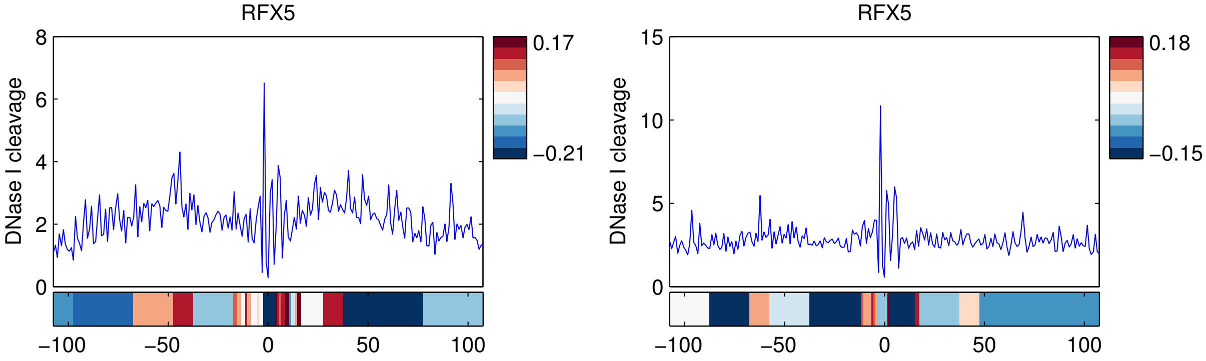

Figure 24: BinDNase models for RFX5

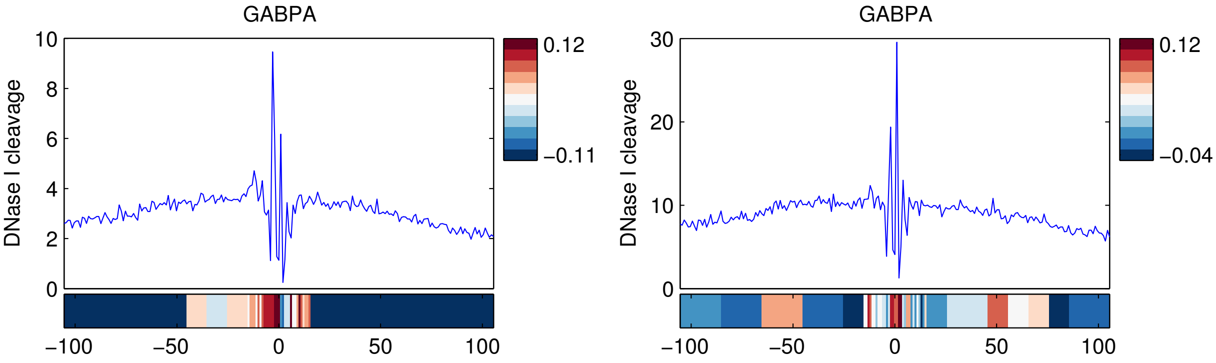

Figure 25: BinDNase models for GABPA

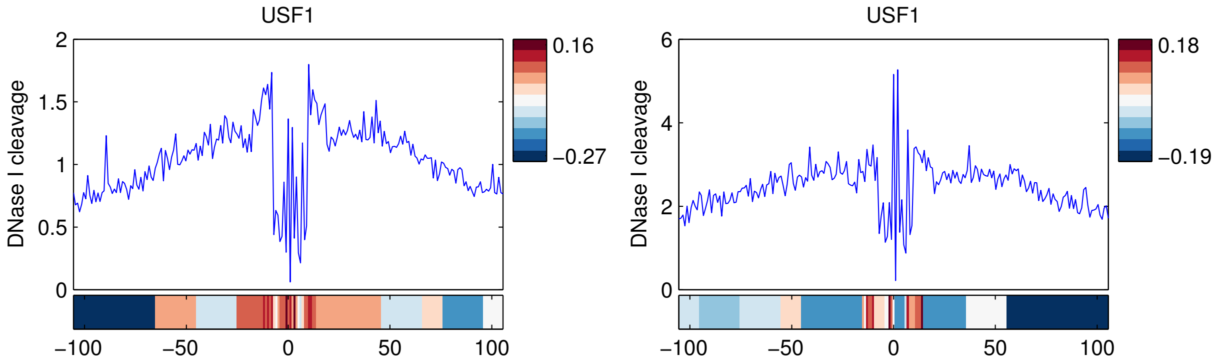

Figure 26: BinDNase models for USF1

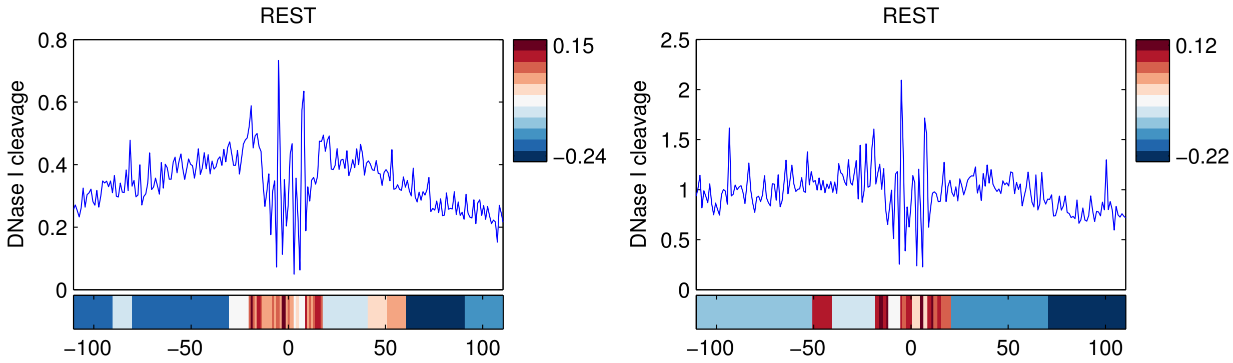

Figure 27: BinDNase models for REST

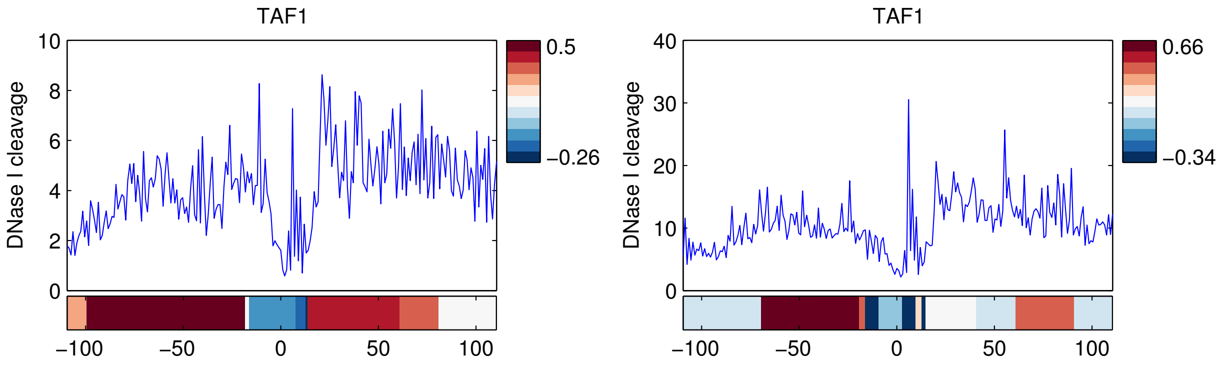

Figure 28: BinDNase models for TAF1

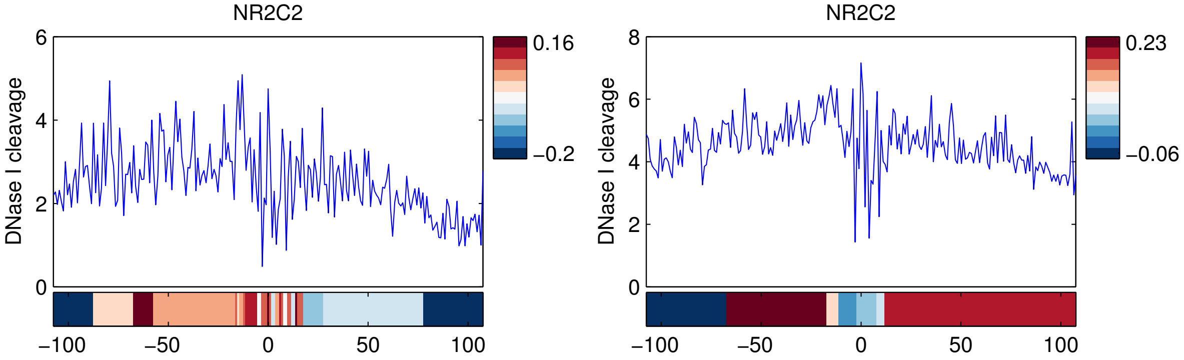

Figure 29: NR2C2

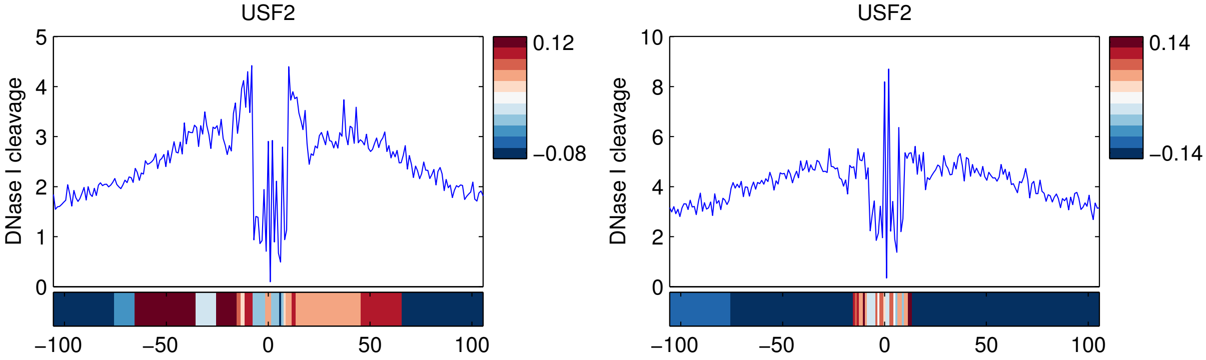

Figure 30: BinDNase models for USF2

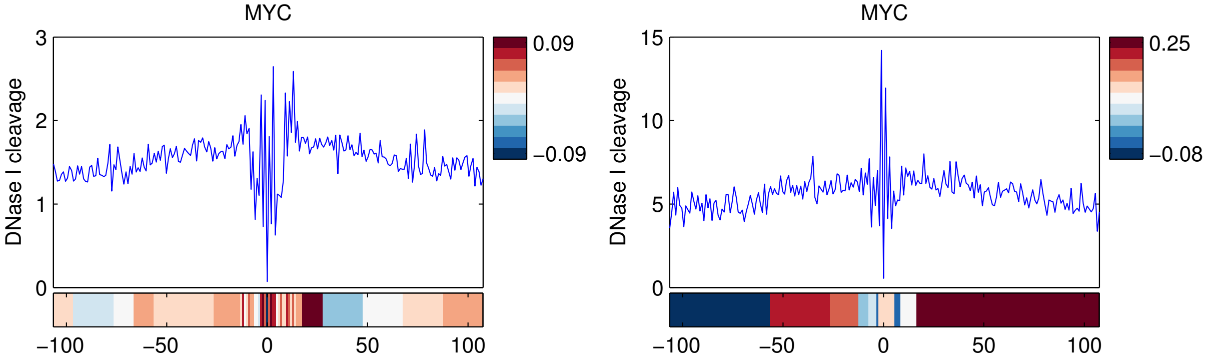

Figure 31: BinDNase models for MYC

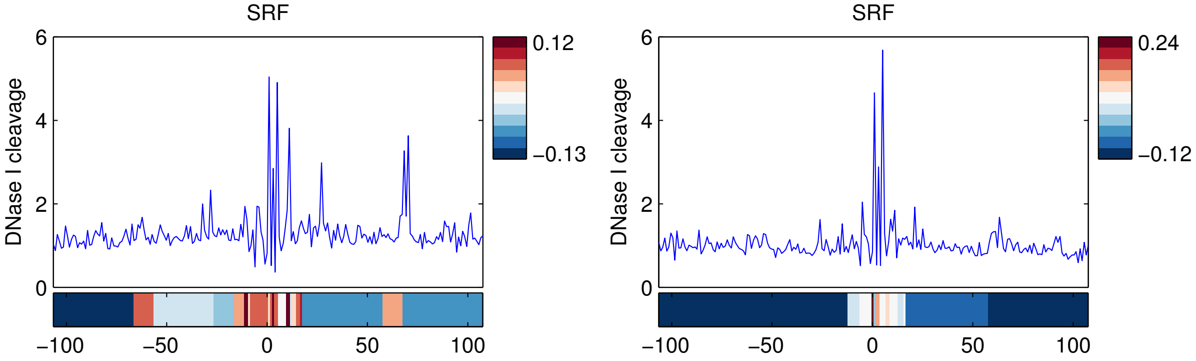

Figure 32: SRF

5 K562 only

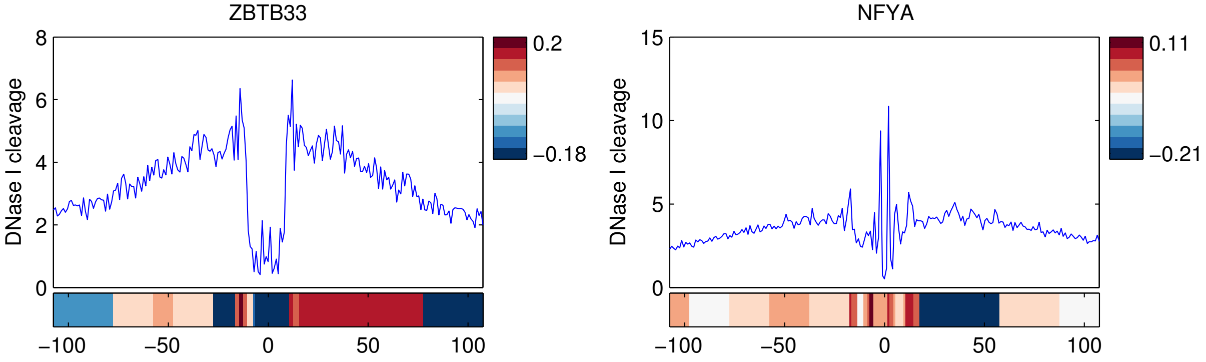

Figure 33: BinDNase models for ZBTB33 and NFYA

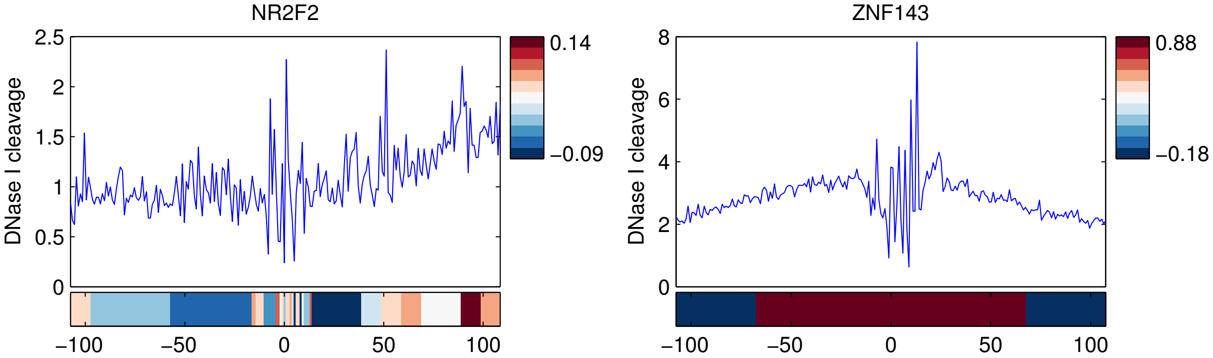

Figure 34: BinDNase models for NR2F2 and ZNF143

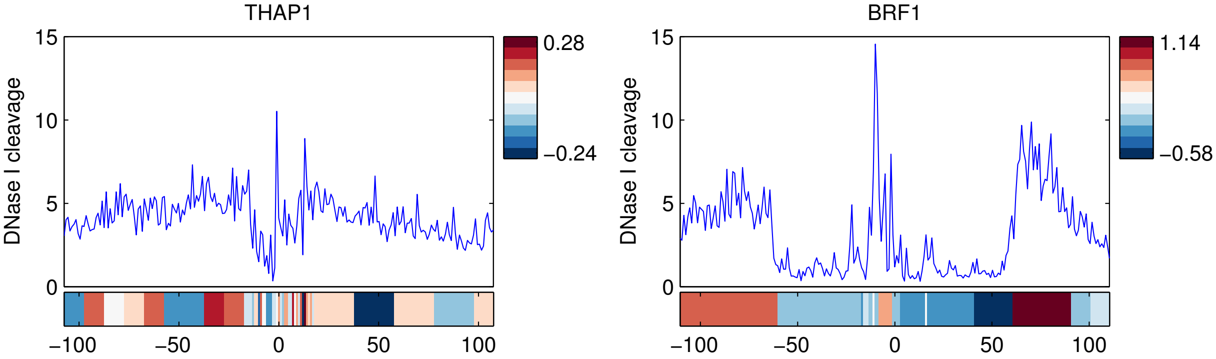

Figure 35: BinDNase models for THAP1 and BRF1

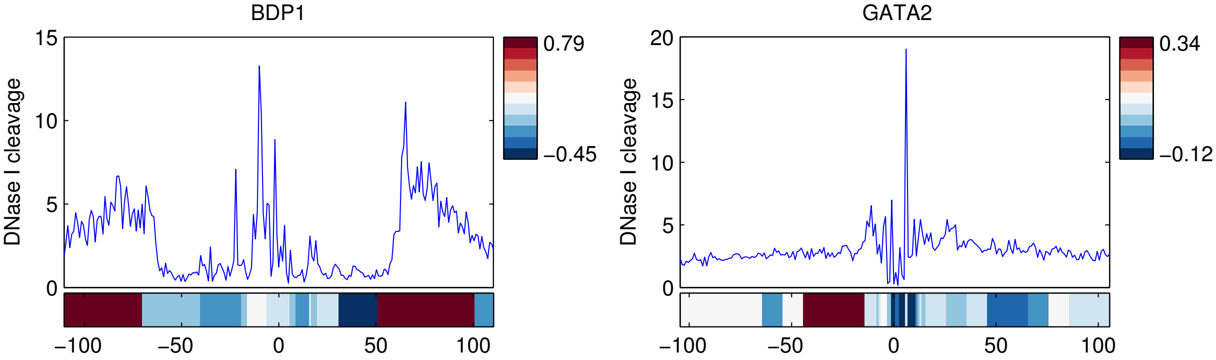

Figure 36: BDP1 and GATA2

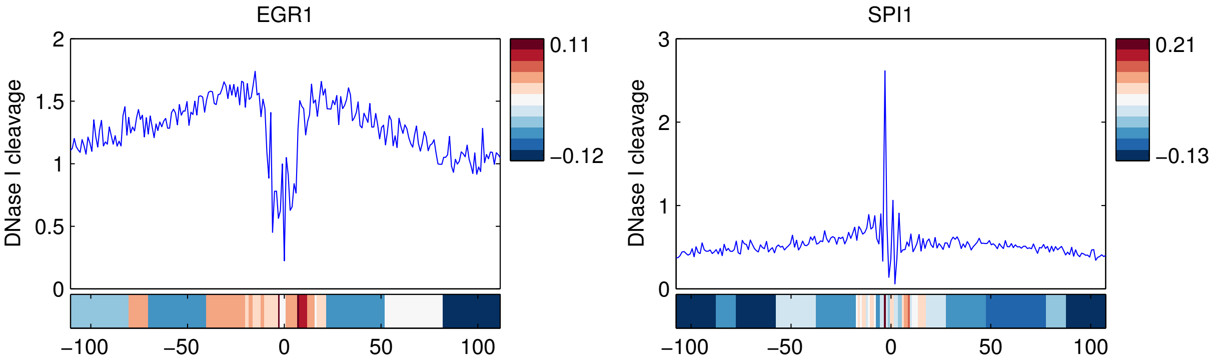

Figure 37: BinDNase models for EGR1 and SPI1

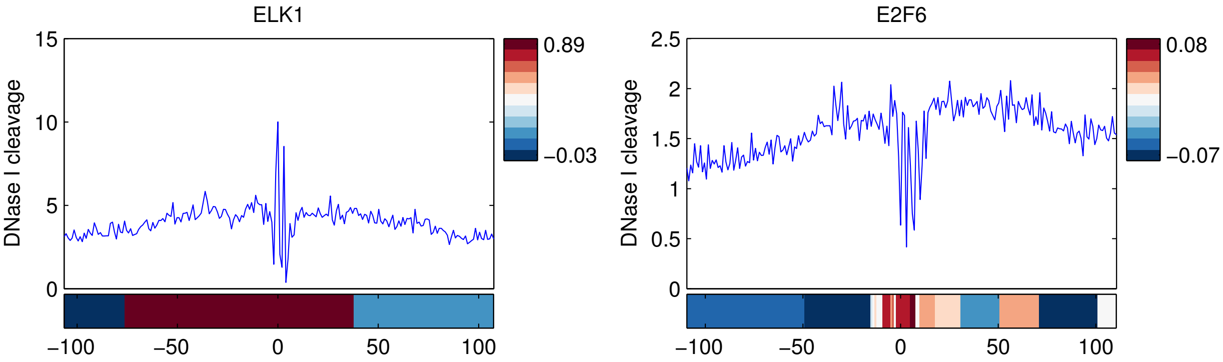

Figure 38: BinDNase models for EL1 and E2F6

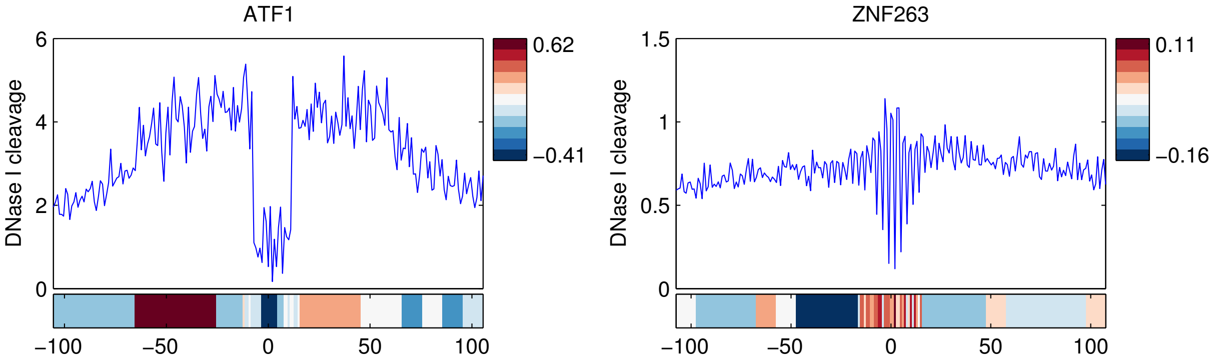

Figure 39: BinDNase models for ATF1 and ZNF263

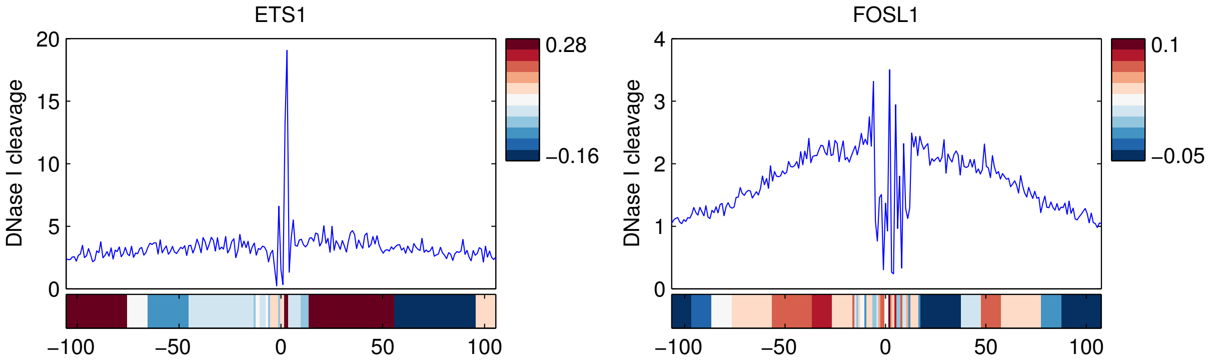

Figure 40: BinDNase models for ETS1 and FOSL1

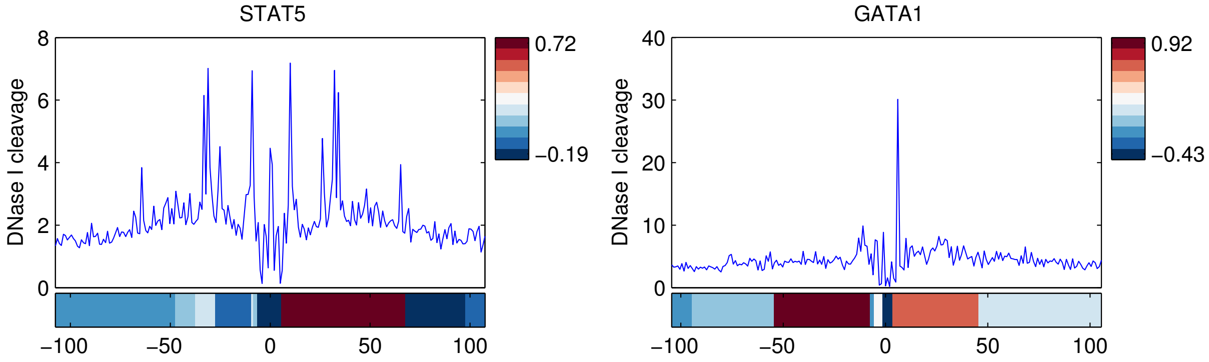

Figure 41: BinDNase models for STAT5 and GATA1

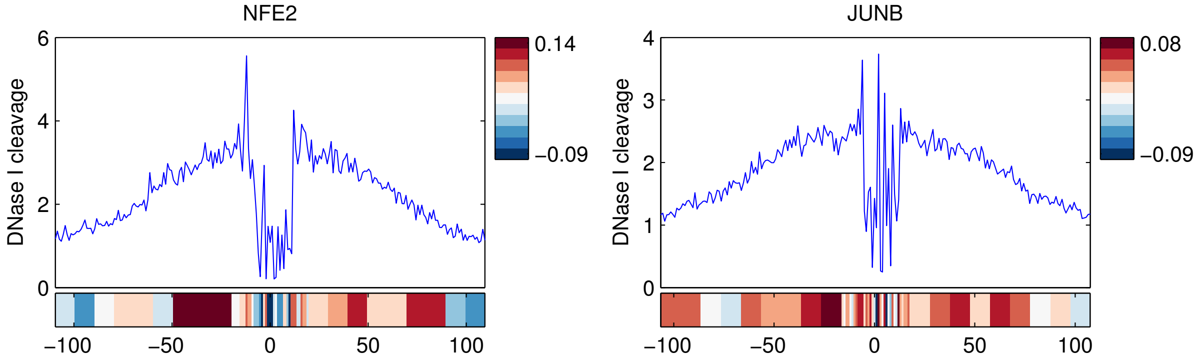

Figure 42: BinDNase models for NFE2 and JUNB

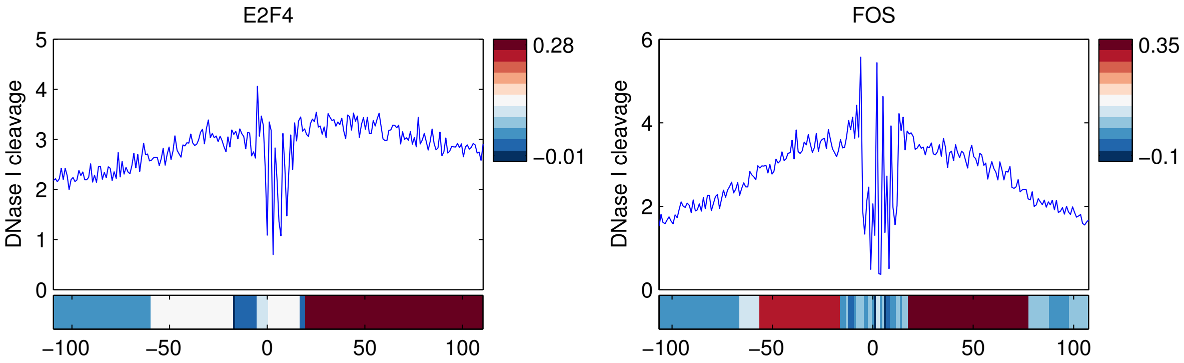

Figure 43: BinDNase models for E2F4 and FOS

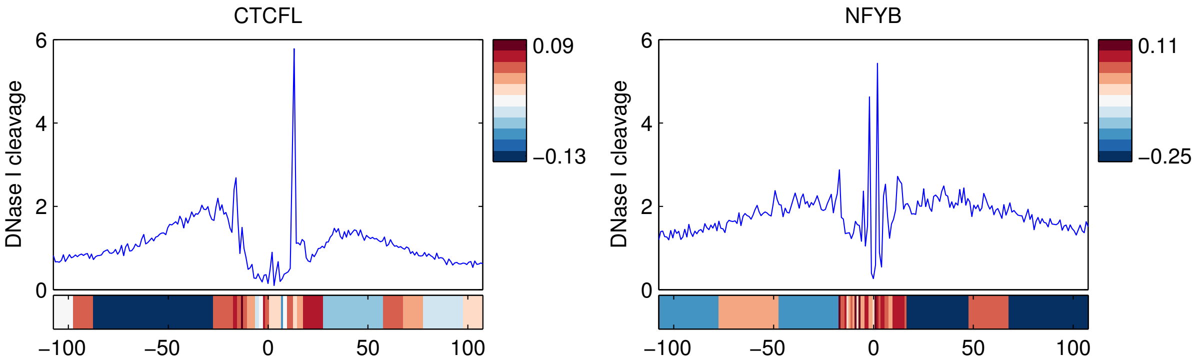

Figure 44: BinDNase models for CTCFL and NFYB

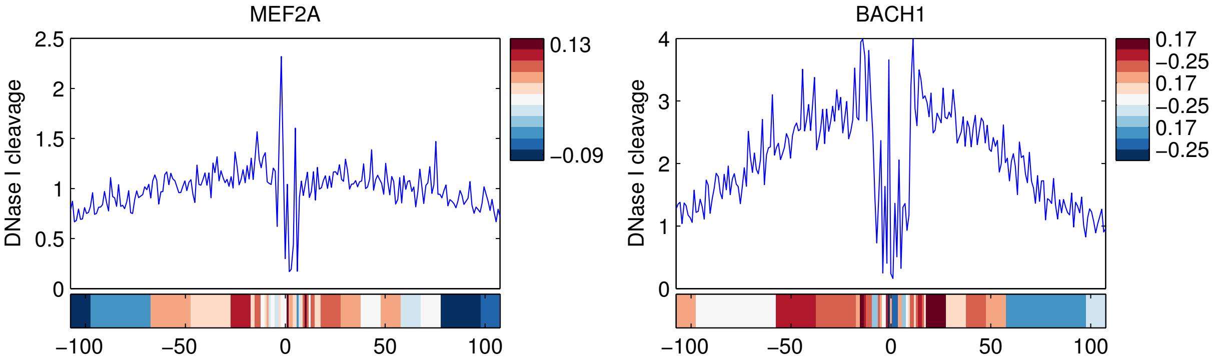

Figure 45: BinDNase models for MEF2A and BACH1

6 The number of candidate binding sites in the training and testing sets has little effect on model accuracy

We tried to vary the number of instances in training and testing sets for transcription factor CTCF. The protein was chosen because it is one of factors that has lots of binding sites. This test was conducted with negative set 2.

Number of bound sites	Number of unbound sites	AUC
500	500	0.856
1000	1000	0.861
2500	2500	0.869
5000	5000	0.868

Table 2: The effect of variation in the number of sites in model training, when the number of sites in the testing data is kept constant.

Number of bound sites	Number of unbound sites	AUC
200	1000	0.856
500	1000	0.880
1000	1000	0.875
200	2000	0.867
500	2000	0.871
1000	2000	0.861
1500	2000	0.866
2000	2000	0.868

Table 3: The effect of variation in the number of sites in model testing, when the number of sites in the training data is kept constant

7 The effect of PWM score in the modeling

BinDNase uses the motif match score (PWM score) as a variable in the modeling. Excluding the motif match score decreases the prediction results for some TFs but generally BinDNase performs well even without using the PWM score.

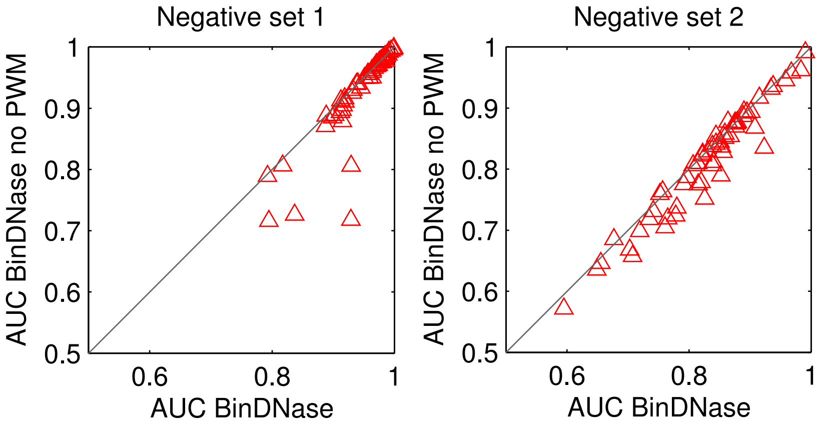

Figure 46: The prediction accuracies of BinDNase with and without PWM score in the model for negative set 1 (left) and negative set 2 (right).

8 Bias correction

A simple DNase I sequence bias correction method was developed. The corrected counts in each genomic position can be calculated with the following equation.

$$D_{corrected} = D_{data} - D_{bias} = D_{data} - c * p \tag{1}$$

,where D_{data} is the number of DNase cuts in the data, D_{bias} is the number of cuts that can be attributed to the sequence bias, p is a probability of having a bias induced cut in the current position and c is a scaling constant, which equals to the number of DNase cuts in the genomic window used in modeling. Probabilities p were given in: Sung, M.-H. et al. (2014). Dnase footprint signatures are dictated by factor dynamics and dna sequence. Molecular Cell, 56(2), 275285.

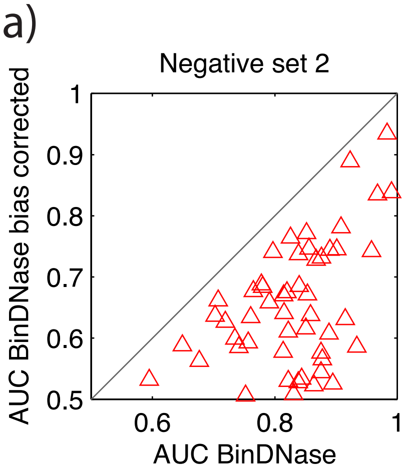

Figure 47: The prediction accuracies obtained with the original BinDNase model and with the BinDNase model including a DNase bias correction.

9 Strandspecific BinDNase

We also implemented a strandspecific version of the BinDNase algorithm. This scheme treats the DNase induced cuts separately for each strand. The prediction performance is almost identical to the basic BinDNase model for negative set 1, but the strandspecific version of BinDNase slightly improves predictions with negative set 2.

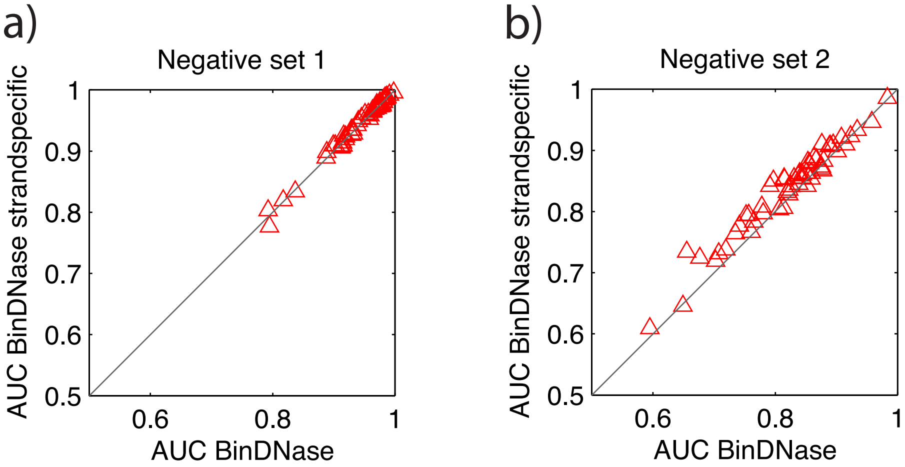

Figure 48: The prediction accuracies obtained with a) the original BinDNase model and with b) the BinDNase model using strand-specific DNase cut counts

10 Examples of distortion caused by non-uniform preprocessing

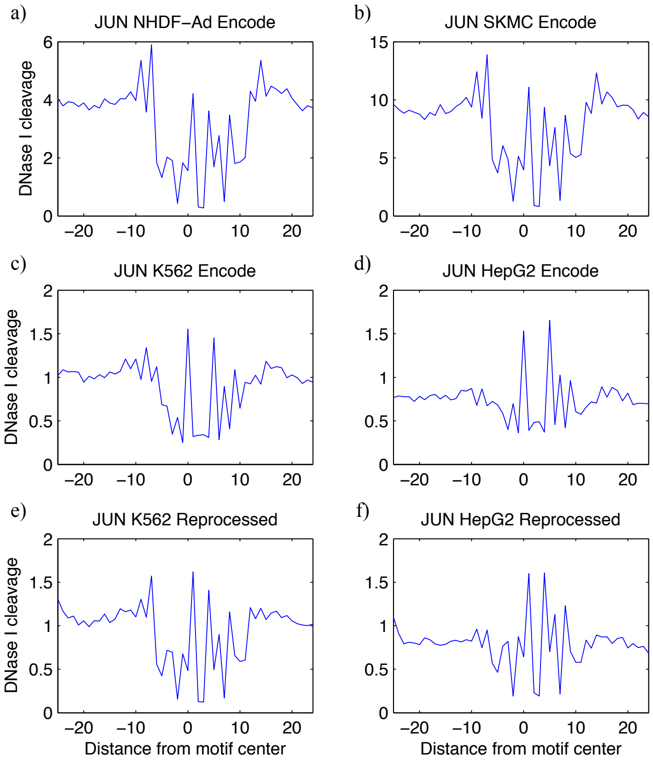

Figure 49: Average DNase I cleavage profiles for protein JUN in four ENCODE cell types: a) NHDF-Ad, b) SKMC, c) K562, d) HepG2, e) K562 reprocessed, and f) HepG2 reprocessed. For each cell type, the average DNase-seq signal at nucleotide resolution centered at JUN motif overlapping DNase-seq hotspot is shown. The celltypes K562 and HepG2 exhibit clearly distinct average pattern in c) and d). Careful preprocessing of the data makes these unexpected cell type specific differences disappear as shown in e) and f) and is essential for nucleotide resolution analysis of the DNase I data.

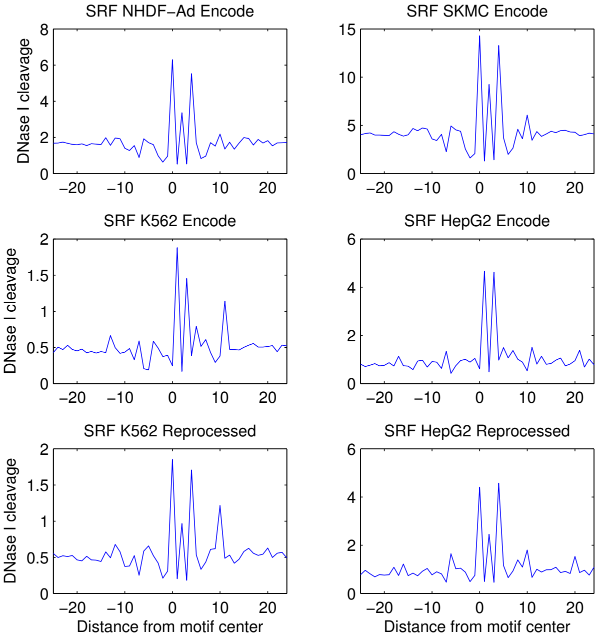

Figure 50: Transcription factor SRF serves as a clear example of distortion caused by non-uniform DNase-seq data preprocessing.

11 Generalisation and protein family.

The Supplementary Table 4 lists the prediction accuracies for TFs in both cell lines used in this study. The cell line used in model training is indicated with brackets. The protein families are taken from ENCODE related website www.factorbook.org.

TF	AUC K562 (K562)	AUC K562 (HepG2)	AUC HepG2 (K562)	AUC HepG2 (HepG2)	Protein family
JUN	0.844	0.732	0.573	0.601	leucine zipper
JUND	0.822	0.482	0.380	0.688	leucine zipper
MAFF	0.840	0.795	0.893	0.935	leucine zipper
CEBPB	0.797	0.811	0.859	0.856	leucine zipper
ATF3	0.780	0.725	0.816	0.843	leucine zipper
MAFK	0.839	0.739	0.895	0.931	leucine zipper
BHLHE40	0.655	0.648	0.741	0.745	helix-loop-helix
USF2	0.853	0.776	0.610	0.750	helix-loop-helix
MXI1	0.878	0.872	0.899	0.898	helix-loop-helix
MYC	0.814	0.806	0.926	0.932	helix-loop-helix
USF1	0.765	0.735	0.742	0.749	helix-loop-helix
MAX	0.677	0.611	0.824	0.876	helix-loop-helix
ELF1	0.791	0.757	0.855	0.847	winged helix
RFX5	0.761	0.717	0.754	0.738	winged helix
RAD21	0.814	0.815	0.773	0.832	winged helix
GABPA	0.850	0.849	0.920	0.918	winged helix
YY1	0.826	0.784	0.890	0.903	beta-beta-alpha zinc finger
SP2	0.875	0.859	0.785	0.790	beta-beta-alpha zinc finger
SP1	0.876	0.866	0.886	0.894	beta-beta-alpha zinc finger
CTCF	0.876	0.852	0.853	0.851	beta-beta-alpha zinc finger
ZNF274	0.983	0.983	0.986	0.986	beta-beta-alpha zinc finger
ZBTB7A	0.595	0.588	0.749	0.806	beta-beta-alpha zinc finger
REST	0.923	0.930	0.938	0.930	beta-beta-alpha zinc finger
NR2C2	0.852	0.844	0.829	0.853	zinc finger
TAF1	0.938	0.903	0.950	0.947	TAF(II)230 TBP-binding fragment
SRF	0.820	0.785	0.739	0.744	SRF-like
EP300	0.859	0.865	0.798	0.787	TAZ domain
SMC3	0.867	0.813	0.763	0.814	Smc hinge domain
NRF1	0.901	0.905	0.901	0.908	unknown
TBP	0.844	0.832	0.918	0.933	TBP-like
TEAD4	0.807	0.797	0.719	0.736	TEA/ATTS

Table 4: The prediction accuracies for TFs when the cell line used in model training is varied.