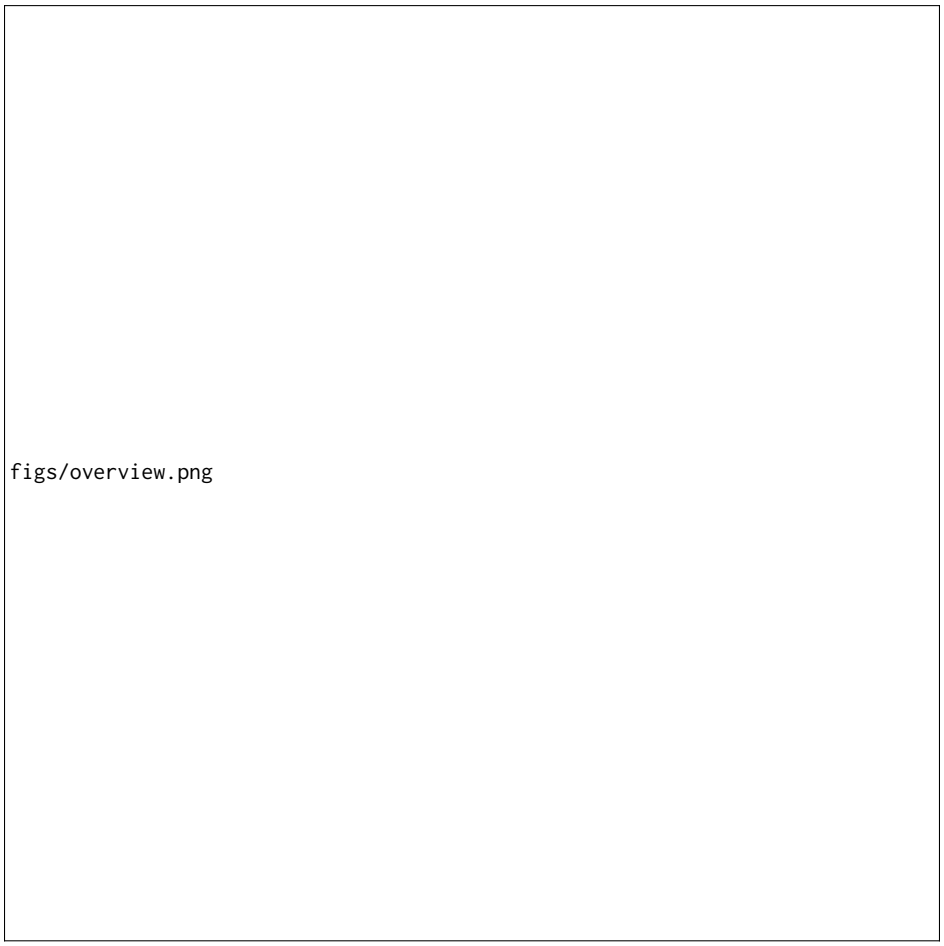
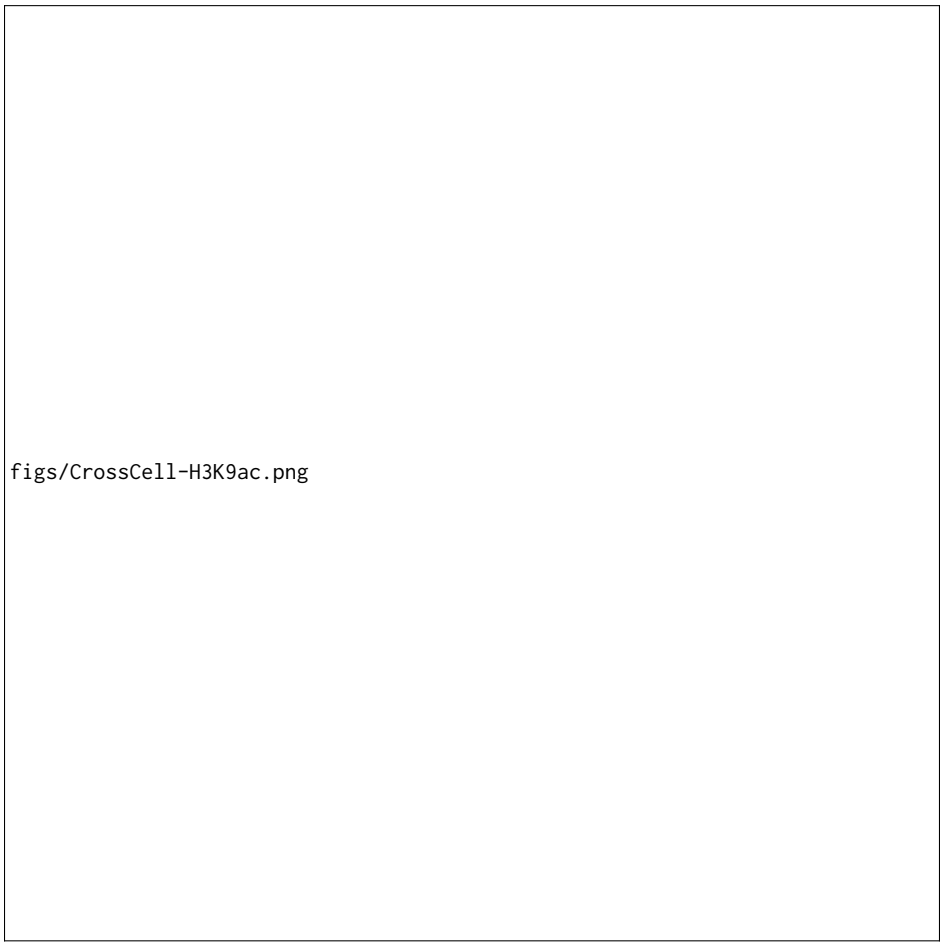
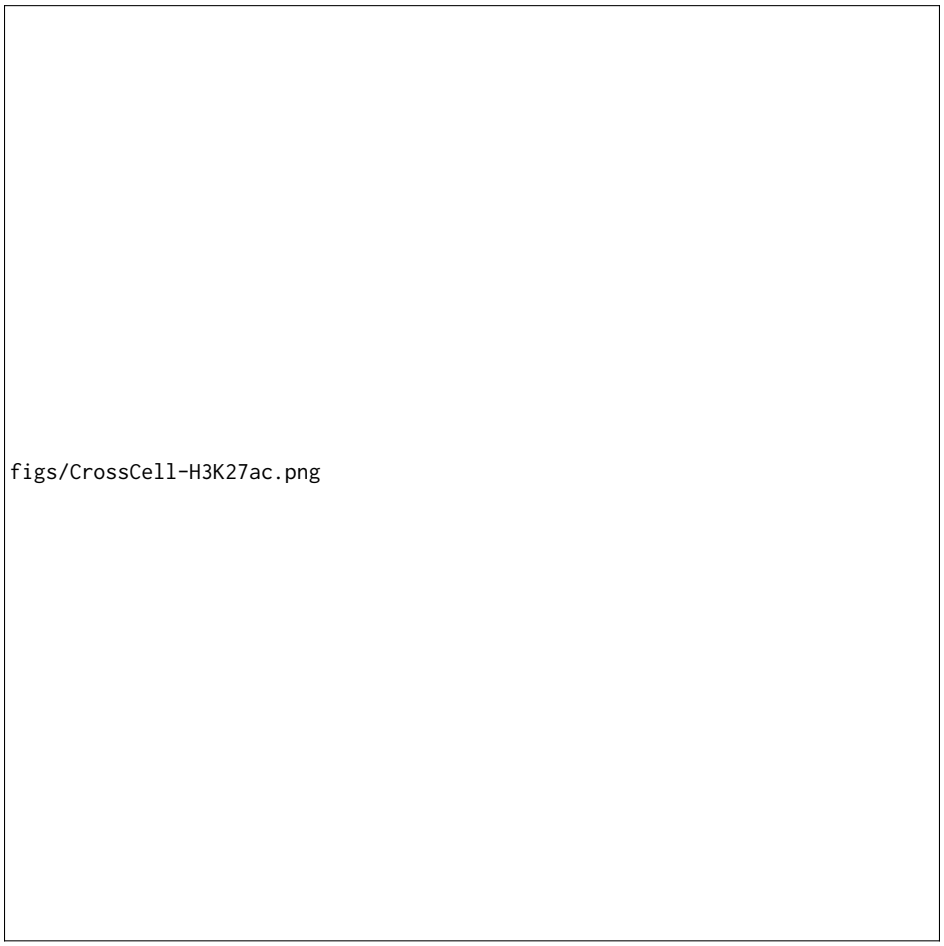
SUPPLEMENTARY FIGURES



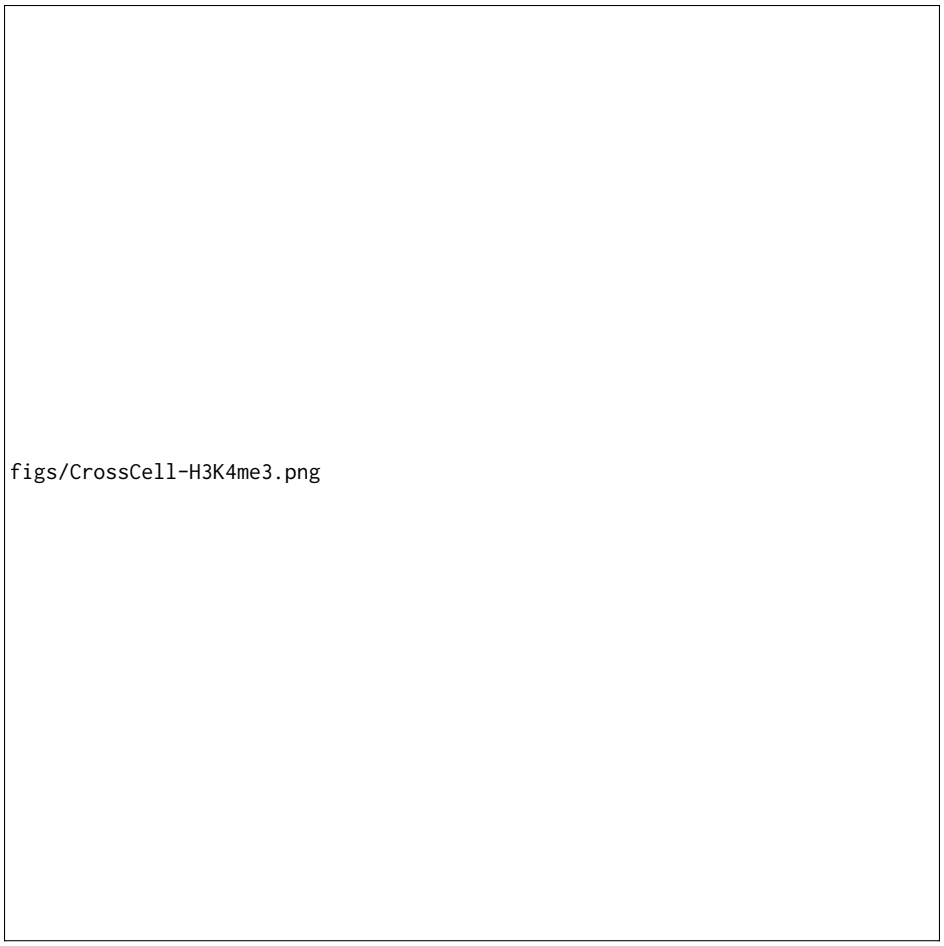
Supplementary Figure 1: Overview of the proposed method to compute the minimal feature sets



Supplementary Figure 2: Prediction of histone PTMs across cell lines using TF-binding data; shown are AUPR curves for DeepHistone for H3K9ac; deepHistone is trained on the cell line indicated by the row and tested on each of the three cell lines (indicated by the column); the AUC is given on the plot in each case



Supplementary Figure 3: Prediction of histone PTMs across cell lines using TF-binding data; shown are AUPR curves for DeepHistone for H3K27ac; DeepHistone is trained on the cell line indicated by the row and tested on each of the three cell lines (indicated by the column); the AUC is given on the plot in each case



Supplementary Figure 4: Prediction of histone PTMs across cell lines using TF-binding data; shown are AUPR curves for DeepHistone for H3K4me3; DeepHistone is trained on the cell line indicated by the row and tested on each of the three cell lines (indicated by the column); the AUC is given on the plot in each case

Supplementary Figure 5: Two-dimensional projection using t-SNE for H1-H3K4me3, K562-H3K9ac and GM12878-H3K27ac (left to right) on DNA sequence data; red points are negative examples; greens points are positive examples.

Supplementary Figure 6: Two-dimensional projection using t-SNE for K562-H3K9ac, GM12878-H3K4me3, GM12878-H3K9ac (left to right) on TF ChIP-Seq data; red points are negative examples; greens points are positive examples.

SUPPLEMENTARY TABLES

PTM	Positive %	Positives/total
H3K4me3	39.8%	11,865/29,828
H3K9ac	24.1%	7,199/29,828
H3K27ac	13.8%	4,123/29,828
H3K27me3	8.3%	2,470/29,828

Supplementary Table 1: Positive/negative ratio for cell line H1 (before NCL)

PTM	Positive (%)	Positives/total	Additional positives
H3K4me3	44.1%	11,865/26,858	4.4%
H3K9ac	28.4%	7,199/25,342	4.3%
H3K27ac	16.1%	4,123/25,552	2.3%
H3K27me3	9.2%	2,470/26,569	1.0%

Supplementary Table 2: Positive/negative ratio after NCL on the TF datasets for cell line H1

PTM	Positive (%)	Positives/total	Additional positives
H3K4me3	44.4%	11,865/26,698	4.7%
H3K9ac	28.3%	7,199/25,440	4.2%
H3K27ac	16.0%	4,123/25,740	2.2%
H3K27me3	9.0%	2,470/26,316	0.98%

Supplementary Table 3: Positive/negative ratio after NCL on the sequence datasets for cell line H1

		AUPR		AUROC		Accuracy	
Cell Line	PTM	LR_{Clean}	LR_{Raw}	LR_{Clean}	LR_{Raw}	LR_{Clean}	LR_{Raw}
H1	H3K4me3	0.067	0.848	0.908	0.918	0.910	0.856
H1	H3K9ac	0.860	0.649	0.941	0.867	0.882	0.825
H1	H3K27ac	0.629	0.447	0.895	0.828	0.673	0.861
H1	H3K27me3	0.320	0.332	0.804	0.806	0.885	0.902

Supplementary Table 4: Comparative analysis of Logistic Regression on cleaned and raw data for sequence (H1 cell line). LR_{Clean} indicates performance of logistic regression on cleaned data and LR_{Raw} indicates performance of logistic regression on raw data

		AUPR		AUROC		Accuracy	
Cell Line	PTM	DeepHistone$_{Clean}$	DeepHistone$_{Raw}$	DeepHistone$_{Clean}$	DeepHistone$_{Raw}$	DeepHistone$_{Clean}$	DeepHistone$_{Raw}$
H1	H3K4me3	0.982	0.455	0.985	0.894	0.951	0.923
H1	H3K9ac	0.922	0.725	0.966	0.915	0.908	0.851
H1	H3K27ac	0.730	0.723	0.927	0.874	0.896	0.830
H1	H3K27me3	0.614	0.447	0.912	0.894	0.927	0.920

Supplementary Table 5: Comparative analysis of DeepHistone on cleaned and raw data for sequence (H1 cell line). DeepHistone$_{Clean}$ indicates performance of DeepHistone on cleaned data and DeepHistone$_{Raw}$ indicates performance of DeepHistone on raw data

		AUPR		AUROC		Accuracy	
Cell Line	PTM	LR$_{Clean}$	DeepHistone$_{Clean}$	LR$_{Clean}$	DeepHistone$_{Clean}$	LR$_{Clean}$	DeepHistone$_{Clean}$
H1	H3K4me3	0.067	0.982	0.908	0.918	0.910	0.951
H1	H3K9ac	0.860	0.922	0.941	0.966	0.882	0.908
H1	H3K27ac	0.629	0.730	0.895	0.927	0.673	0.896
H1	H3K27me3	0.320	0.614	0.804	0.912	0.885	0.927

Supplementary Table 6: Comparative analysis of Logistic Regression on cleaned and raw data for sequence (H1 cell line). LR$_{Clean}$ indicates performance of logistic regression on cleaned data and DeepHistone$_{Clean}$ indicates performance of DeepHistone on cleaned data

		AUPR		AUROC		Accuracy	
Cell Line	PTM	DeepHistone$_{Aug}$	DeepHistone$_{Clean}$	DeepHistone$_{Aug}$	DeepHistone$_{Clean}$	DeepHistone$_{Aug}$	DeepHistone$_{Clean}$
H1	H3K4me3	0.973	0.982	0.977	0.918	0.925	0.951
H1	H3K9ac	0.907	0.922	0.965	0.966	0.909	0.908
H1	H3K27ac	0.623	0.730	0.913	0.927	0.970	0.896
H1	H3K27me3	0.422	0.614	0.848	0.912	0.901	0.927

Supplementary Table 7: Comparative analysis of DeepHistone on cleaned and augmented data for sequence (H1 cell line). DeepHistone$_{Aug}$ indicates performance of DeepHistone on augmented data and DeepHistone$_{Clean}$ indicates performance of DeepHistone on cleaned data

		AUPR		AUROC		Accuracy	
Cell Line	PTM	LR$_{Clean}$	LR$_{Raw}$	LR$_{Clean}$	LR$_{Raw}$	LR$_{Clean}$	LR$_{Raw}$
H1	H3K4me3	0.944	0.885	0.993	0.950	0.968	0.888
H1	H3K9ac	0.923	0.729	0.981	0.921	0.936	0.844
H1	H3K27ac	0.820	0.545	0.963	0.909	0.920	0.878
H1	H3K27me3	0.601	0.437	0.909	0.877	0.927	0.922

Supplementary Table 8: Comparative analysis of Logistic Regression on cleaned and raw data for TF-ChIP-Seq (H1 cell line). LR$_{Clean}$ indicates performance of logistic regression on cleaned data and LR$_{Raw}$ indicates performance of logistic regression on raw data

		AUPR		AUROC		Accuracy	
Cell Line	PTM	DeepHistone$_{Clean}$	DeepHistone$_{Raw}$	DeepHistone$_{Clean}$	DeepHistone$_{Raw}$	DeepHistone$_{Clean}$	DeepHistone$_{Raw}$
H1	H3K4me3	0.996	0.880	0.996	0.949	0.977	0.896
H1	H3K9ac	0.970	0.781	0.966	0.988	0.953	0.863
H1	H3K27ac	0.901	0.605	0.978	0.916	0.940	0.885
H1	H3K27me3	0.810	0.560	0.957	0.915	0.954	0.930

Supplementary Table 9: Comparative analysis of DeepHistone on cleaned and raw data for TF-ChIP-Seq (H1 cell line). DeepHistone$_{Clean}$ indicates performance of DeepHistone on cleaned data and DeepHistone$_{Raw}$ indicates performance of DeepHistone on raw data

		AUPR		AUROC		Accuracy	
Cell Line	PTM	DeepHistone$_{Clean}$	LR$_{Clean}$	DeepHistone$_{Clean}$	LR$_{Clean}$	DeepHistone$_{Clean}$	LR$_{Clean}$
H1	H3K4me3	0.996	0.944	0.996	0.993	0.977	0.968
H1	H3K9ac	0.970	0.923	0.966	0.981	0.953	0.936
H1	H3K27ac	0.901	0.820	0.978	0.963	0.940	0.920
H1	H3K27me3	0.810	0.601	0.957	0.909	0.954	0.927

Supplementary Table 10: Comparative analysis of DeepHistone and logistic regression on cleaned data for TF-ChIP-Seq (H1 cell line). DeepHistone$_{Clean}$ indicates performance of DeepHistone on cleaned data and LR$_{Clean}$ indicates performance of logistic regression on cleaned data

		AUPR		AUROC		Accuracy	
Cell Line	PTM	DeepHistone$_{Clean}$	DeepHistone$_{Aug}$	DeepHistone$_{Clean}$	DeepHistone$_{Aug}$	DeepHistone$_{Clean}$	DeepHistone$_{Aug}$
H1	H3K4me3	0.996	0.989	0.996	0.992	0.977	0.962
H1	H3K9ac	0.970	0.959	0.966	0.983	0.953	0.937
H1	H3K27ac	0.901	0.833	0.978	0.957	0.940	0.909
H1	H3K27me3	0.810	0.615	0.957	0.916	0.954	0.905

Supplementary Table 11: Comparative analysis of DeepHistone and logistic regression on cleaned data for TF ChIP-Seq (H1 cell line). DeepHistone$_{Clean}$ indicates performance of DeepHistone on cleaned data and DeepHistone$_{Aug}$ indicates performance of DeepHistone on augmented data

	Test on H1			Test on K562			Test on GM12878		
	DeepHistone		LR	DeepHistone		LR	DeepHistone		LR
	Test Set with NCL	Test Set w/o NCL	w/o NCL	Test Set with NCL	Test Set w/o NCL	w/o NCL	Test Set with NCL	Test Set w/o NCL	w/o NCL
Train on H1	0.97	0.851	0.681	0.95	0.85	0.7638	0.84	0.72	0.661
Train on K562	0.90	0.801	0.606	0.99	0.90	0.8793	0.68	0.582	0.453
Train on GM12878	0.86	0.75	0.596	0.95	0.852	0.799	0.98	0.88	0.80

Supplementary Table 12: DeepHistone's predictive performance (AUPR) across cell lines for H3K9ac

	Test on H1			Test on K562			Test on GM12878		
	DeepHistone		LR	DeepHistone		LR	DeepHistone		LR
	Test Set with NCL	Test Set w/o NCL	w/o NCL	Test Set with NCL	Test Set w/o NCL	w/o NCL	Test Set with NCL	Test Set w/o NCL	w/o NCL
Train on H1	0.87	0.68	0.48	0.93	0.81	0.719	0.80	0.67	0.604
Train on K562	0.76	0.587	0.416	0.99	0.901	0.85	0.63	0.53	0.423
Train on GM12878	0.61	0.48	0.376	0.93	0.823	0.7579	0.98	0.863	0.735

Supplementary Table 13: DeepHistone's predictive performance (AUPR) across cell lines for H3K27ac

		AUPR		
Cell Line	PTM	500bp	1000bp	2000bp
H1	H3K4me3	0.997	0.996	0.996
H1	H3K9ac	0.984	0.989	0.991
H1	H3K27ac	0.939	0.956	0.961
H1	H3K27me3	0.844	0.876	0.888
GM12878	H3K4me3	0.976	0.981	0.965
GM12878	H3K9ac	0.979	0.980	0.960
GM12878	H3K27ac	0.979	0.975	0.967

Supplementary Table 14: DeepHistone's prediction performance (AUPR) for different choices of window sizes around the TSS

TF	Gini Importance	TF	Gini Importance
SIN3A	0.23238437	USF2	0.02192348
MAX	0.12652680	EGR1	0.01872315
TCF12	0.06496394	ATF2	0.01860226
YY1	0.05945533	CJUN	0.01346616
SP4	0.04706157	POU5F1	0.01159581
E2F6	0.04343570	SRF	0.01150319
CTCF	0.04228849	SIX5	0.01068433
CREB1	0.03942244	USF1	0.01055930
TEAD4	0.03397813	CEBPB	0.01050230
NRSF	0.03321331	FOSL1	0.01029178
ZNF274	0.03080357	ATF3	0.00918099
GABP	0.02630001	BACH1	0.00752202
NANOG	0.02315774	CMYC	0.00599580
SP1	0.02284828	RFX5	0.00590655
USF2	0.02192348	JUND	0.00466759

Supplementary Table 15: Gini importance for the TFs of H3K4me3, cell line H1

TF	Gini Importance	TF	Gini Importance
SIN3A	0.15565900	SP1	0.02237557
TCF12	0.08263545	MAFK	0.01608166
NRSF	0.08120345	USF1	0.01551900
CREB1	0.08098141	USF2	0.01506276
YY1	0.06765905	TEAD4	0.01139045
SP4	0.05682064	CJUN	0.01122649
SIX5	0.05230694	BACH1	0.00921277
GABP	0.04594152	EGR1	0.00714014
MAX	0.04300521	FOSL1	0.00713903
ATF2	0.04032396	POU5F1	0.00687461
CMYC	0.03481369	SRF	0.00674068
CTCF	0.03060233	RFX5	0.00667074
E2F6	0.02800676	JUND	0.00506986
NANOG	0.02695069	CEBPB	0.00414045
ZNF274	0.02615859	ATF3	0.00228711

Supplementary Table 16: Gini importance for the TFs of H3K9ac, cell line H1

TF	Gini Importance	TF	Gini Importance
YY1	0.086448	ZNF274	0.02237557
SIN3A	0.083164	JUND	0.01608166
ATF2	0.069840	BACH1	0.01551900
CREB1	0.066675	ATF3	0.01506276
TCF12	0.061500	USF2	0.01139045
SP4	0.061451	TEAD4	0.01122649
NRSF	0.057449	MAFK	0.00921277
SP1	0.051684	NANOG	0.00714014
CTCF	0.044918	CJUN	0.00713903
GABP	0.039888	FOSL1	0.00687461
CMYC	0.036264	EGR1	0.00674068
SIX5	0.036003	SRF	0.00667074
USF1	0.032884	POU5F1	0.00506986
MAX	0.032866	CEBPB	0.00414045
E2F6	0.028427	RFX5	0.00745000

Supplementary Table 17: Gini importance for the TFs of H3K27ac, cell line H1

TF	Gini Importance	TF	Gini Importance
TCF12	0.122498	RFX5	0.019292
SIN3A	0.109399	ZNF274	0.015939
E2F6	0.104137	USF1	0.014193
ATF2	0.083487	FOSL1	0.011417
GABP	0.082123	EGR1	0.010195
NRSF	0.062741	SRF	0.010066
SIX5	0.049258	USF2	0.008650
TEAD4	0.041044	CEBPB	0.006137
SP4	0.039053	CJUN	0.006085
MAX	0.036303	ATF3	0.005618
BACH1	0.035637	NANOG	0.004659
SP1	0.030465	POU5F1	0.003487
CMYC	0.029646	CTCF	0.002622
CREB1	0.028472	JUND	0.002566
YY1	0.024804	MAFK	6E-06

Supplementary Table 18: Gini importance for the TFs of H3K27me3, cell line H1

TF	Gini Importance	TF	Gini Importance
E2F6	0.196256	NFYB	0.016176
MAX	0.138325	FOSL1	0.015863
TEAD4	0.040534	CEBPD	0.013123
ATF3	0.039185	BACH1	0.011328
GATA2	0.038282	ETS1	0.010107
ZNF263	0.037788	RFX5	0.009204
CREB1	0.037324	BHLHE40	0.006487
E2F4	0.035283	CEBPB	0.006333
SP1	0.031753	STAT5A	0.006216
CMYC	0.029631	NR2F2	0.006073
TAL1	0.027513	NFYA	0.005686
GATA1	0.025325	MEF2A	0.005414
ZNF274	0.025263	GABP	0.003524
CTCF	0.021088	YY1	0.002995
CFOS	0.020187	SIX5	0.002594
CJUN	0.019749	MAFK	0.002314
JUND	0.019427	ELK1	0.002304
NFE2	0.019186	ATF1	0.001717
HCFC1	0.019131	MAFF	0.001508
USF2	0.017150	SRF	0
EGR1	0.016452	USF1	0
TR4	0.016201		

Supplementary Table 19: Gini importance for the TFs of H3K4me3, cell line K562

TF	Gini Importance	TF	Gini Importance
E2F6	0.196256	NFYB	0.016176
MAX	0.138325	FOSL1	0.015863
TEAD4	0.040534	CEBPD	0.013123
ATF3	0.039185	BACH1	0.011328
GATA2	0.038282	ETS1	0.010107
ZNF263	0.037788	RFX5	0.009204
CREB1	0.037324	BHLHE40	0.006487
E2F4	0.035283	CEBPB	0.006333
SP1	0.031753	STAT5A	0.006216
CMYC	0.029631	NR2F2	0.006073
TAL1	0.027513	NFYA	0.005686
GATA1	0.025325	MEF2A	0.005414
ZNF274	0.025263	GABP	0.003524
CTCF	0.021088	YY1	0.002995
CFOS	0.020187	SIX5	0.002594
CJUN	0.019749	MAFK	0.002314
JUND	0.019427	ELK1	0.002304
NFE2	0.019186	ATF1	0.001717
HCFC1	0.019131	MAFF	0.001508
USF2	0.017150	SRF	0
EGR1	0.016452	USF1	0
TR4	0.016201		

Supplementary Table 20: Gini importance for the TFs of H3K9ac, cell line K562

TF	Gini Importance	TF	Gini Importance
MAX	0.118014	USF2	0.015532
CMYC	0.070880	FOSL1	0.013067
TEAD4	0.070726	BACH1	0.012797
E2F6	0.067724	CJUN	0.010579
CTCF	0.065102	NFYB	0.009546
HCFC1	0.055546	GABP	0.009454
ATF3	0.039058	BHLHE40	0.008547
ZNF263	0.039007	USF1	0.006958
JUND	0.038854	MAFK	0.006584
ZNF274	0.031512	YY1	0.005852
NR2F2	0.031452	CEBPD	0.005728
E2F4	0.030254	NFYA	0.005106
GATA1	0.028832	RFX5	0.003683
ETS1	0.025718	MAFF	0.003592
SP1	0.025559	ELK1	0.003136
GATA2	0.024006	STAT5A	0.002921
CREB1	0.022697	SIX5	0.001919
GABP	0.019951	ATF1	0.001875
FOSL1	0.017570	MEF2A	0.001597
CEBPB	0.016846	EGR1	0
USF2	0.016359	SRF	0
CFOS	0.015860		

Supplementary Table 21: Gini importance for the TFs of H3K27ac, cell line K562

TF	Gini Importance	TF	Gini Importance	TF	Gini Importance
ATF2	0.148595	MTA3	0.018221	ATF3	0.005581
BATF	0.077526	NFATC1	0.017614	CEBPB	0.005491
BCL3	0.071042	NFE2	0.015291	CREB1	0.004933
BHLHE40	0.059730	NFIC	0.014777	CTCF	0.00466
CFOS	0.053180	NFYA	0.014490	EGR1	0.004477
E2F4	0.042806	NFYB	0.012906	GABP	0.004387
EBF1	0.041229	PBX3	0.012512	JUND	0.004289
ELF1	0.037559	POU2F2	0.011695	MAFK	0.00273
ELK1	0.031329	RUNX3	0.010129	MAX	0.002344
ERRA	0.029913	RXRA	0.008446	RFX5	0.001959
ETS1	0.029511	STAT1	0.007964	SIX5	0.001724
FOXM1	0.028797	STAT3	0.007670	SP1	0.001551
IKZF1	0.028649	STAT5A	0.007262	SRF	0.000179
IRF3	0.026735	TCF12	0.006995	USF1	0
IRF4	0.023174	TCF3	0.006167	USF2	0
MEF2A	0.023124	TR4	0.005887	YY1	0
MEF2C	0.018895	ZEB1	0.005872	ZNF274	0

Supplementary Table 22: Gini importance for the TFs of H3K4me3, cell line GM12878

TF	Gini Importance	TF	Gini Importance	TF	Gini Importance
SP1	0.120071	SRF	0.018063	TCF12	0.007442
NFATC1	0.118832	YY1	0.017558	FOXM1	0.007308
CREB1	0.091158	CTCF	0.016291	NFIC	0.006543
TCF3	0.042794	ZNF274	0.015206	RXRA	0.006089
STAT3	0.042297	MTA3	0.014740	MAFK	0.005623
BATF	0.037999	PBX3	0.014655	ELK1	0.004361
ELF1	0.035553	NFYA	0.013748	USF1	0.003899
RUNX3	0.033227	STAT5A	0.012509	EGR1	0.003817
POU2F2	0.032332	CEBPB	0.011997	MEF2C	0.003125
BCL3	0.029621	ATF2	0.011925	JUND	0.002580
MEF2A	0.025376	EBF1	0.010829	ERRA	0.002329
BHLHE40	0.021213	SIX5	0.009894	ATF3	0.001071
ZEB1	0.021043	ETS1	0.009459	E2F4	0.001017
NFE2	0.02052	CFOS	0.008964	IRF3	0.000491
STAT1	0.020475	TR4	0.008896	RFX5	0.000471
MAX	0.020419	NFYB	0.008684	USF2	0.000135
IRF4	0.018718	GABP	0.008632	IKZF1	0

Supplementary Table 23: Gini importance for the TFs of H3K9ac, cell line GM12878

TF	Gini Importance	TF	Gini Importance	TF	Gini Importance
NFATC1	0.22886	BATF	0.019819	YY1	0.005152
SP1	0.067851	MAX	0.019670	MEF2A	0.004452
CREB1	0.058443	PBX3	0.017083	IKZF1	0.004394
ELF1	0.052343	CFOS	0.014929	JUND	0.003402
CTCF	0.045843	FOXM1	0.014667	IRF3	0.003067
TCF3	0.045590	STAT1	0.013192	E2F4	0.002980
BCL3	0.032479	MTA3	0.013037	MEF2C	0.002809
BHLHE40	0.030360	SIX5	0.011125	NFYB	0.002335
POU2F2	0.030256	NFYA	0.009903	USF1	0.001537
NFE2	0.027149	ZEB1	0.008742	STAT5A	0.001408
IRF4	0.026063	EBF1	0.008388	USF2	0.001331
STAT3	0.025837	TR4	0.006440	ERRA	0.000546
RUNX3	0.024066	RXRA	0.006403	ETS1	0.000474
ZNF274	0.022131	NFIC	0.006221	GABP	0.000241
SRF	0.020747	ATF2	0.005951	ELK1	0
CEBPB	0.020596	TCF12	0.005788	ATF3	0
EGR1	0.020289	MAFK	0.005614	RFX5	0

Supplementary Table 24: Gini importance for the TFs of H3K27ac, cell line GM12878