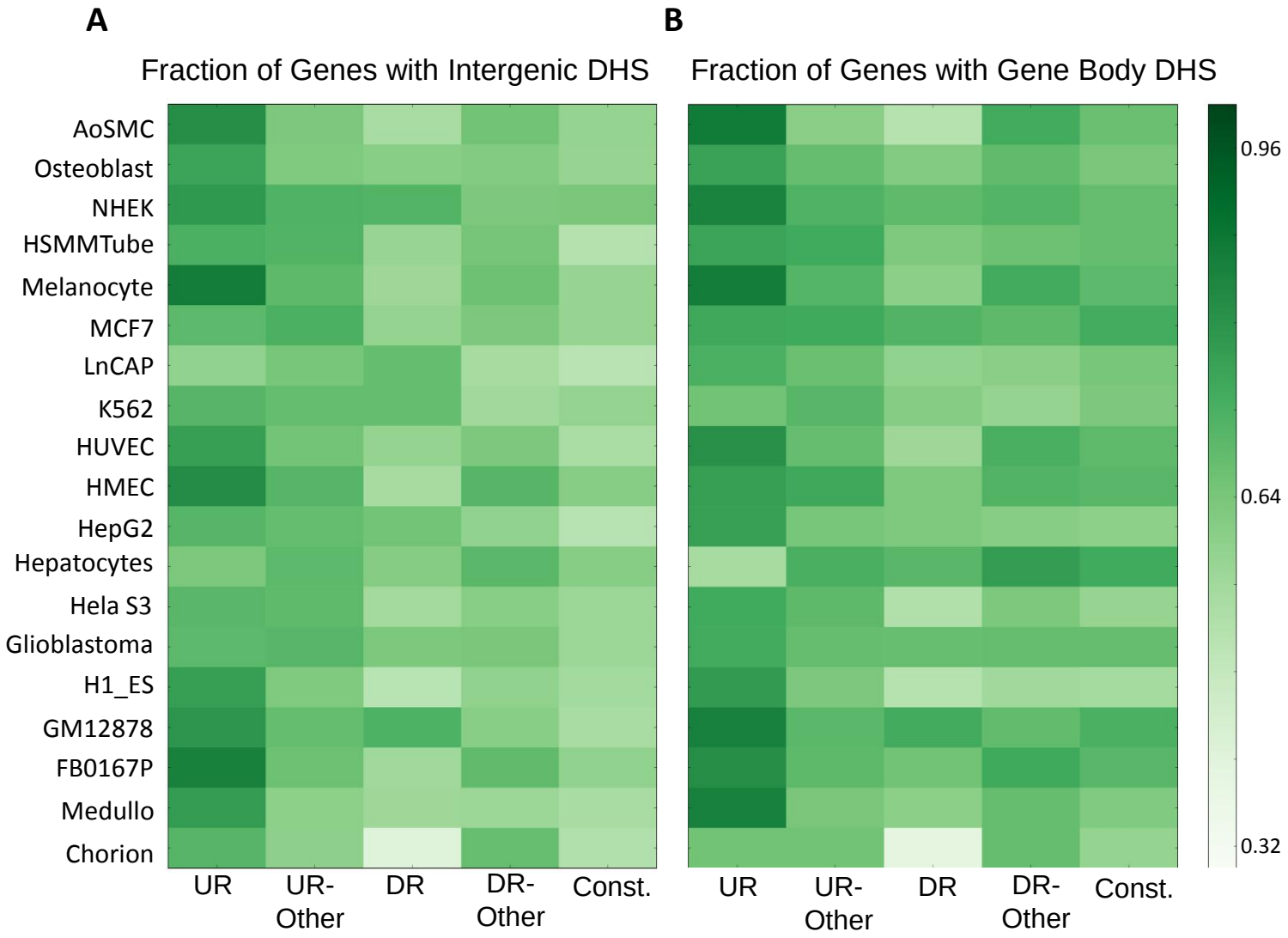
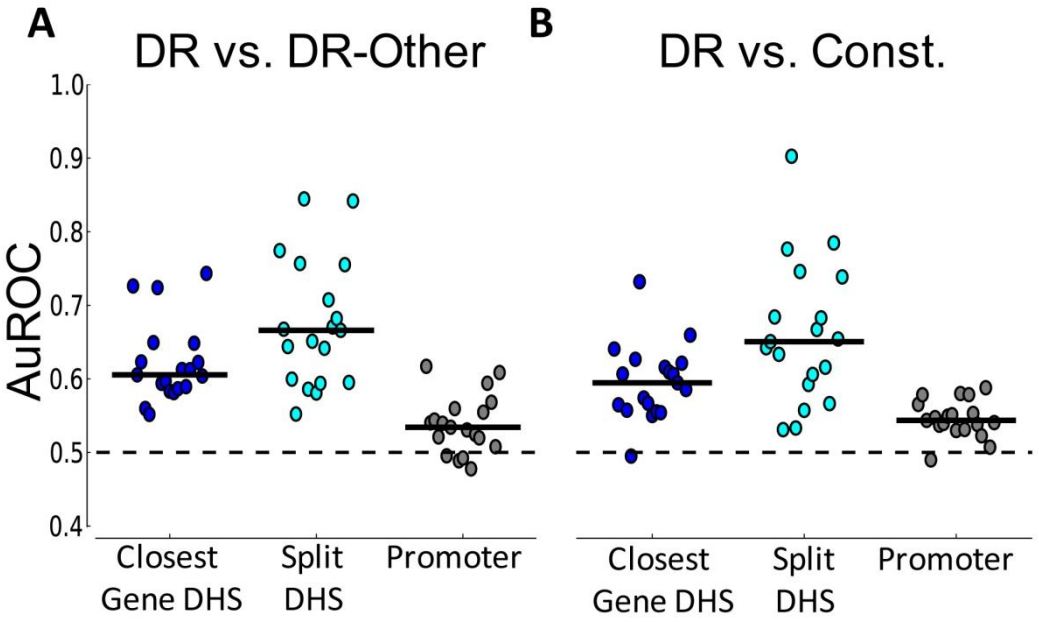
Supplementary Figure 1



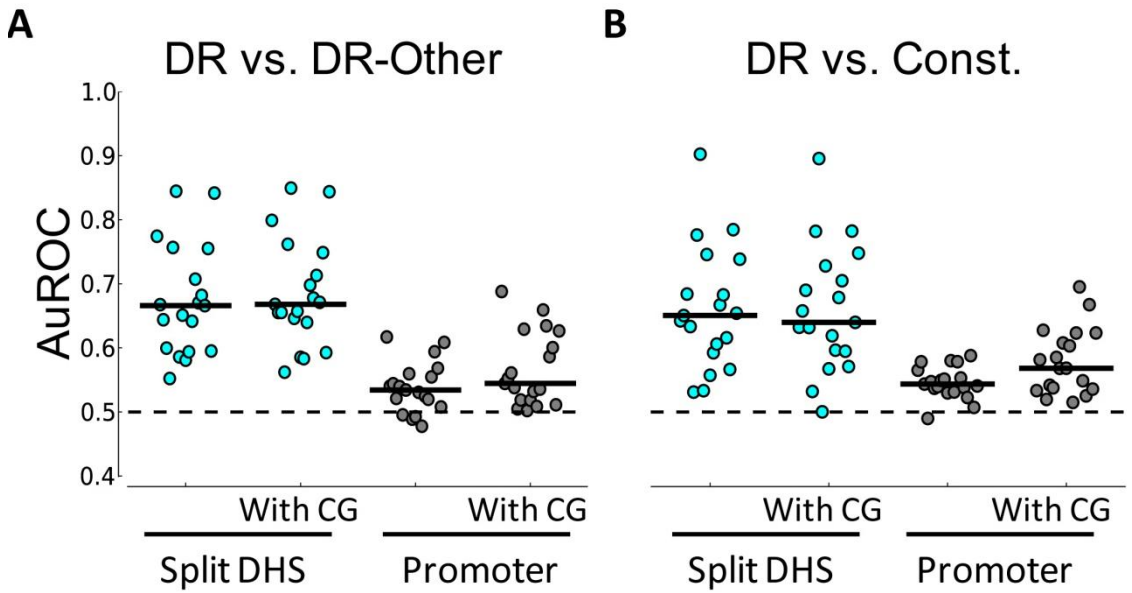
The fraction of genes in each gene set that had Intergenic DHSs **(A)** and Gene Body DHS **(B). (A and B)** share the same color map.

Supplementary Figure 2



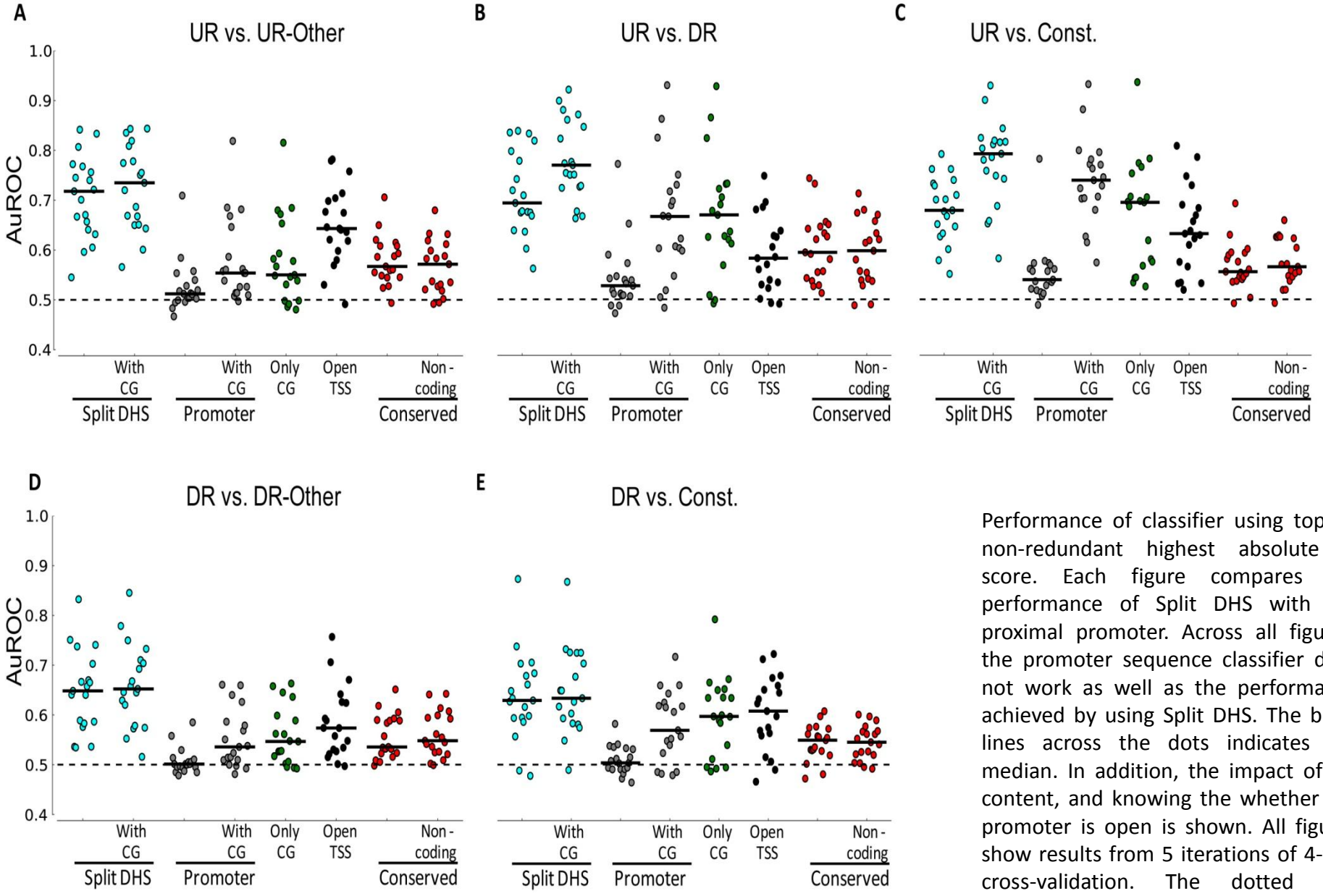
Classifier performance for discriminating DR genes from other classes. **(A, B)** Performance of the classifier using all PWMs. Each figure compares the performance of 2 methods of associating DHSs to genes (Closest Gene DHS and Split DHS) with the proximal promoter. Across all figures, the promoter sequence classifier does not work as well as the performance achieved by using closest gene DHS and Split DHS and is significant at the 0.05 level (t-test). The black lines across the dots indicates the median. All figures show results from 5 iterations of 4-fold cross-validation. The dotted line indicates an AuROC of 0.5 which is the performance of a random classifier.

Supplementary Figure 3



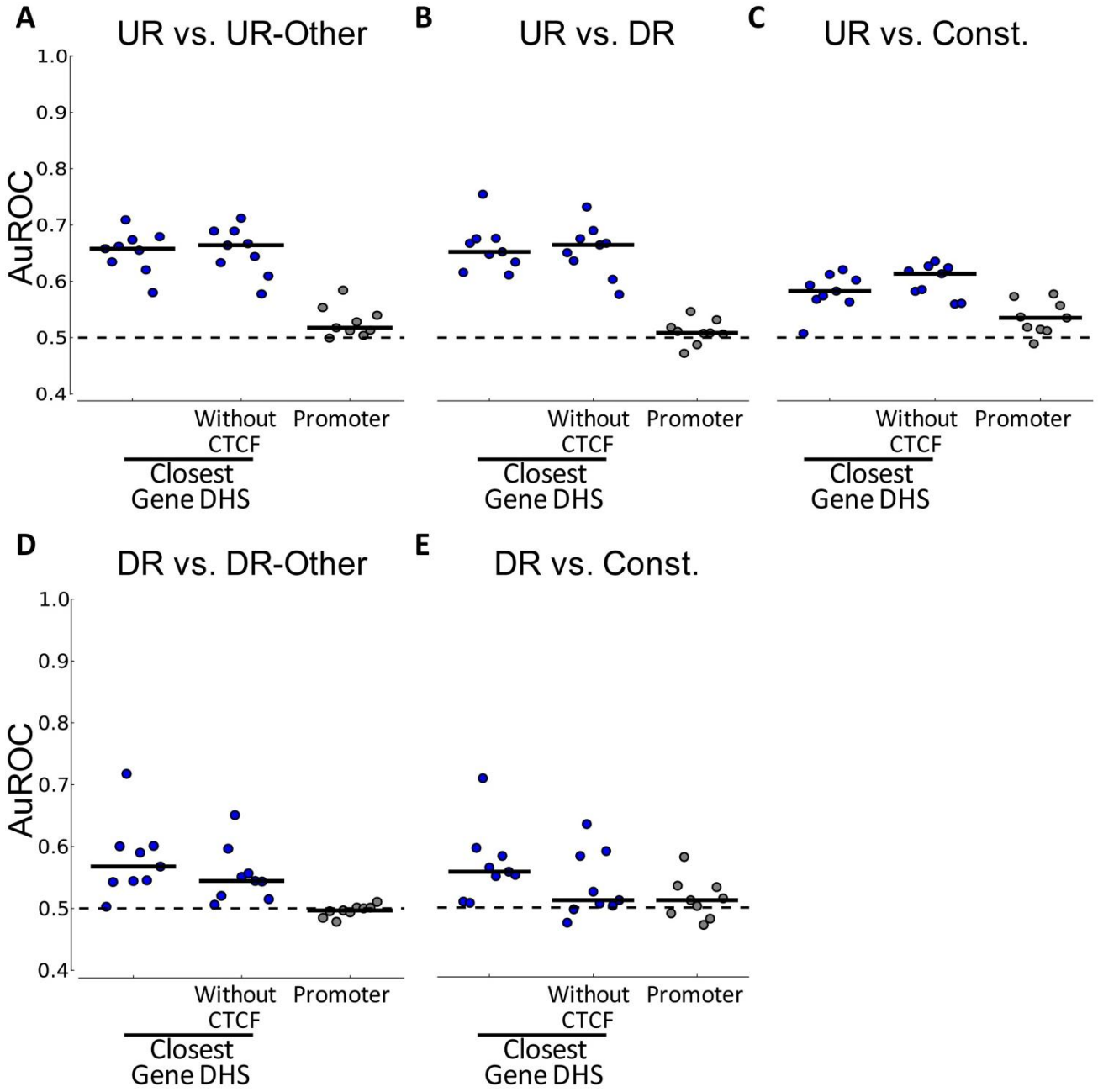
Impact of normalized CG dinucleotide content on classifier performance for discriminating DR genes. **(A, B)** Results using the Split DHS and Promoter sequence are shown. The dotted line indicates an AuROC of 0.5 which is the performance of a random classifier. CG content is shown to help for the proximal promoter where other factors are not informative.

Supplementary Figure 4



Performance of classifier using top 10 non-redundant highest absolute z-score. Each figure compares the performance of Split DHS with the proximal promoter. Across all figures, the promoter sequence classifier does not work as well as the performance achieved by using Split DHS. The black lines across the dots indicates the median. In addition, the impact of CG content, and knowing the whether the promoter is open is shown. All figures show results from 5 iterations of 4-fold cross-validation. The dotted line indicates an AuROC of 0.5 which is the performance of a random classifier.

Supplementary Figure 5



Performance of classifier using top 10 non-redundant highest absolute z-score. Each figure compares the performance of 2 methods of associating DHSs to genes using Closest Gene DHS and the inclusion and exclusion of DHS with CTCF ChIP-seq peaks and with the proximal promoter. Note that only 9 cell lines had CTCF ChIP-seq data. All figures show results from 5 iterations of 4-fold cross-validation. The dotted line indicates an AuROC of 0.5 which is the performance of a random classifier.

Supplemental Table 1

Cell Line	Number of DHS	Number of bases in DHS (% of genome size)
Chorion	140042	3.55
Medulloblastoma	151858	2.44
FB0167P	126789	1.91
GM12878	124321	1.96
H1_ES	129061	2.83
Glioblastoma	118222	1.71
Hela S3	141165	2.04
Hepatocytes	171897	3.48
HepG2	116018	1.72
HMEC	158764	2.37
HUVEC	126695	2.08
K562	133372	1.87
LnCAP	144070	2.82
MCF7	119828	1.58
Melanocyte	130073	1.50
HSMMTube	161083	2.68
NHEK	140520	1.80
Osteoblast	151381	2.45
AoSMC	121731	1.62

DHS statistics by cell line. Number of DHS per cell line and the total number of bases covered by the DHS peaks

Supplemental Table 2

Number of Cell Lines TSS Open in	Number of genes
1	976
2	581
3	455
4	362
5	306
6	345
7	292
8	290
9	311
10	307
11	348
12	408
13	417
14	457
15	536
16	614
17	863
18	1798
19	8393

Number of genes in each box plot in Figure 2C. For example, 8393 TSS are in regions of open chromatin in all cell lines.

Supplemental Table 3

Cell Type	UR genes	UR Other Genes	DR genes	DR Other genes	Constitutive genes
Chorion	200	2340	200	857	168
Medulloblastoma	200	2303	200	1531	168
FB0167P	200	2191	200	1763	168
GM12878	200	2610	200	1347	168
H1_ES	200	2152	200	1600	168
Glioblastoma	200	2411	200	2218	168
Hela S3	200	1807	200	2654	168
Hepatocytes	200	1876	200	1263	168
HepG2	200	1563	200	2688	168
HMEC	200	2503	200	1354	168
HUVEC	200	1866	200	2230	168
K562	200	2493	200	1792	168
LnCAP	200	2574	200	1590	168
MCF7	200	2419	200	1667	168
Melanocyte	200	2423	200	1315	168
HSMMTube	200	2358	200	1523	168
NHEK	200	1276	200	2955	168
Osteoblast	200	2268	200	1876	168
AoSMC	200	2421	200	1734	168

Number of genes in each category for each cell line. Expression values were first z-score transformed. UR and DR genes were the top and bottom 200 genes, respectively. UR Other genes were genes up-regulated in a different cell line with expression z-score < 0 in the current cell line. Constitutive genes were picked as genes in neither UR or DR sets across all cell lines, were expressed above background in all cell lines and had a maximum absolute z-score of expression less than 1.7 across all cell lines

Supplemental Table 4

Cell Line	GO Categories of UR genes
Chorion	Placenta, inflammatory response, extracellular region, cytokine binding
Medulloblastoma	Retina, Brain, visual perception, ion channel complex, gated channel activity
B0167P	Skeletal system development, pattern specification process, extracellular region part, intrinsic to plasma membrane, growth factor activity
GM12878	B-cell, spleen, immune response, cell activation, MHC class II receptor activity
H1_ES	Brain, ion transport, synapse, plasma membrane, gated channel activity
Glioblastoma	Regulation of transcription, DNA-dependent, anterior-posterior pattern formation, DNA binding, zinc ion binding
Hela S3	
Hepatocytes	Liver, acute inflammatory response, complement activation, oxygen binding
HepG2	Liver, lipid homeostasis, cholesterol metabolic process, sterol homeostasis, extracellular space
HMEC	Keratinocyte, ectoderm development, epidermis development, epithelial cell differentiation
HUVEC	Umbilical Vein Endothelial cell, angiogenesis, vasculature development, plasma membrane part, cell adhesion
K562	Blood, platelet, hemopoiesis, intrinsic to plasma membrane
LnCAP	Prostate, Prostatic carcinoma, synaptic transmission
MCF7	
Melanocyte	Skin, Melanoma, pigmentation during development, melanocyte differentiation, melanosome
HSMMTube	Skeletal muscle, heart, muscle system process, muscle tissue development, structural constituent of muscle
NHEK	Keratinocyte, keratinocyte differentiation, epithelial cell differentiation, desmosome
Osteoblast	Fibroblast, Osteoblast, skeletal system development, extracellular structure organization
AoSMC	Fibroblast, response to wounding, cell adhesion, extracellular region part, chemokine activity

GO analysis for UR genes in each cell line. The UP_Tissue entries from DAVID were used to identify the similarity of expression to known tissue types

Supplemental Table 5

Cell Line	Number of DHS with CTCF (% of total)	Number of TSS DHS with CTCF (% of total TSS DHS)	Number of Gene Body DHS with CTCF (% of total Gene Body DHS)	Number of Intergenic DHS with CTCF (% of total Intergenic DHS)
GM12878	34065(27.4%)	4696(37.5%)	12853(22.7%)	16516(29.9%)
H1_ES	40379(31.3%)	7056(47.9%)	15597(26.7%)	17726(31.6%)
Glioblastoma	30508(25.8%)	4131(34.6%)	11166(22.0%)	15211(27.4%)
Hela S3	37811(26.8%)	4772(39.8%)	14590(22.9%)	18449(28.2%)
HepG2	36767(31.7%)	4058(30.4%)	14321(28.8%)	18388(34.7%)
HUVEC	27394(21.6%)	3587(26.3%)	10344(17.4%)	13463(25.1%)
K562	35335(26.5%)	4610(35.5%)	13772(21.7%)	16953(29.7%)
MCF7	39226(32.7%)	5670(45.8%)	14837(29.3%)	18719(32.9%)
NHEK	33416(23.8%)	5564(42.1%)	12415(20.1%)	15437(23.6%)

Number and percentage of DHS with CTCF ChIP-seq peaks. Across all cell lines, a median of 27.4% of all DHS have a CTCF ChIP-seq peak. BEDTools (Quinlan and Hall 2010) was used to compute overlap.

Supplemental Table 6

Cell Type	UR – UR Other Genes			UR – DR genes		
	AuROC	TFs - Positive Coefficient	TFs - Negative Coefficient	AuROC	Positive Coefficient	Negative Coefficient
Chorion	0.54		*ZNF143*	0.87	*HBP1*	*E2F1, NFYA, GABPA*
Medulloblastoma	0.79	*PDX1, CRX, REST*		0.80	*MEF2A, CRX, REST*	
FB0167P	0.74	*DMRT1, JDP2*		0.73	*POU2F1*	
GM12878	0.79	*YY1, NFE2L1-MAFG, SPI1, IRF8*	*E2F3, E2F4-TFDP2, E2F1*	0.76	*INSM1, IRF8*	*AHR-ARNT, HINFP*
H1_ES	0.66			0.77	*NANOG*	*TFAP2A, GABPA, ELK4*
Glioblastoma	0.82	*ZNF143*		0.74	*STAT5B*	
Hela S3	0.84	*USF1, GABPA*		0.84		
Hepatocytes	0.70		*RFX1, ZFP161, FOXN1*	0.80	*NR2F2, RXRA-NR1H2*	*E2F1, FOXN1*
HepG2	0.77	*ZEB1*		0.68	*HNF4A*	
HMEC	0.62			0.72		*ELK4*
HUVEC	0.70	*ETS1, SPI1*		0.74		*ELK4, E2F*
K562	0.69	*YY1, LMO2 bound to TAL1, TCF3 and GATA1, ETS1*		0.60		
LnCAP	0.66			0.65	*SOX5*	
MCF7	0.76	*GATA6, ZFX, TCF3, HINFP*		0.68	*RBPJ*	
Melanocyte	0.78	*SREBF1, MEF2A, AHR-ARNT*		0.84	*MYCN, ELK4*	*GABPA*
HSMMtube	0.64	*MEF2A, PKNOX2*	*ZNF423, PAX6, PAX3*	0.79	*BACH2*	*NFYA, GABPA*
NHEK	0.73	*ZNF410*		0.69	*MAF, MTF1*	*FOXD1*
Osteoblast	0.62			0.61	*MYB*	*FOXN1*
AoSMC	0.87	*DMRT1, CEBPB, PPARG*		0.88	*CEBPB, PATZ1*	*NFYA*

Results for each cell line using Split DHS from All TFs for the UR – UR Other and UR – DR classification task. TFs with positive and negative coefficients are shown for both sets of TFs used.

Supplemental Table 7

Cell Line	Cell-Type Specifically Up-Regulated TFs	Cell-Type Specifically Down-Regulated TFs
Chorion	*ASCL2, EGR1, OSR2, GCM1, DLX5*	*IRF3, E2F3, ZFP161, USF1, ELK1*
Medulloblastoma	*CRX, INSM1, NHLH1, HLX, SIX3, SOX11*	*REST, SMAD3, NR2F2, SP100*
FB0167P	*ZBTB12, STAT1, LHX9, ZIC1, OSR1, HOXC11*	*AIRE, RFX3, IRF5,BACH2*
GM12878	*HIC1, ARID5A, IRF4, SPIB, EGR2, POU2F2*	*FOXP1, GLIS2, TCF7L2, BCL6*
H1_ES	*ZBTB3, MYCN, SOX21, SOX11, ZIC3, SOX2, OTX2*	*MEIS1, NR2F2, HOXC9*
Glioblastoma	*IRF3, HOXD10, HOXB5, NKX3-2, PAX6, ZIC1, PITX2, HOXD11*	*AHR, STAT5A*
Hela S3	*ESRRA,E2F2, PAX6, ARNT, SP1, MAFK, ELK1, FOXF2, ATF1, MEOX1*	
Hepatocytes	*BACH2, HNF4A, RXRA, AR, STAT3*	*FOXJ3, E2F3, RFX7, SIX4, HOXA6*
HepG2	*CEBPA, HNF1A, GFI1, HNF4A, HOXD1, NFYA, FOXA2, HOXA3, TCF7, SOX9*	
HMEC	*STAT4, IRF6, EGR3, OSR1, HOXA5*	*ZBTB3, DR1, HOXA11, SIX1, HOXA5, PAX2*
HUVEC	*SOX18, GBX2, HIC1, ARID5A, SOX17, HOXA3, BCL6B, HOXA9*	*ZBTB6, BACH1*
K562	*WT1, HOXB9, GFI1B, ESRRB, STAT5A, LEF1, MYB, GATA1*	*ARX, KLF7*
LnCAP	*ZBTB7B, HOXC6, AR, NKX3-1, MAFB, ELF5, HOXB13*	*FOXJ1, STAT6, KLF7*
MCF7	*ESR1, SPDEF, LMX1B, IRX5, GSC, MSX2, GATA3*	*SOX14, POU6F1, FOXI1*
Melanocyte	*MAF, IRF4, PAX3, TBX5, IRX6, LEF1*	*NKX3-1, BBX, GABPA, CUX1*
HSMMTube	*MYF6, SOX11, SIX1, PITX2*	*ZBTB3, ZBTB12, HOXD13, GATA6, STRA13, HLXB9*
NHEK	*FOXJ2, VDR, MTF1, SOX15, EHF, SOX8, HOXA1, MAF, IRX4, GATA5*	
Osteoblast	*STAT4, STAT1, EGR2, BACH1, SIX1, HOXA11, GLIS2, BARX1, PROP1*	*SOX21*
AoSMC	*OSR1, MEIS1, HBP1, CUX1,*	*POU3F2, HOXC11, PAX4, SOX8, PITX3, SOX7*

Top 10 non-redundant highest absolute expression z-score in each cell line.

Supplemental Table 8

Cell Line	UR – UR Other Genes	UR – DR Genes
Chorion	*SPI1*	*FOXP1, SP1*
Medulloblastoma	*TAL1-GATA1*	*EWSR1-FLI1*
FB0167P	*SPI1*	*SPI1*
GM12878	***SPIB*** *, ***FOXP1***	***SPIB*** *, ***FOXP1***
H1_ES	***ZIC3*** *, *EWSR1-FLI1, FOXP1*	*EWSR1-FLI1,SP1, FOXP1*
Glioblastoma	*EWSR1-FLI1, NFE2L2*	*IRF*
Hela S3	*EWSR1-FLI1, TAL1-GATA1,* ***FOXF2****	***SP1, FOXF2****
Hepatocytes	***STAT3****	***FOXJ3****
HepG2	*SP1, ZNF219,* ***FOXA2***	*SP1,* ***FOXA2****
HMEC	*SPI1*	*ETS2*
HUVEC	*ETS2, FOXP1, ZBTB7B*	*EWSR1-FLI1, ZBTB7B*
K562	*FOXP1, SP1,* ***WT1***	***WT1,*** *FOXP1*
LnCAP	***ELF5***	***ELF5***
MCF7	***FOXI1***	***GATA3****
Melanocyte	*IRF, SP1*	*IRF,* ***IRX6***
HSMMTube	*IRF*	*IRF*
NHEK	*ZNF219,* ***FOXJ2****, SP1*	*SP1,****FOXJ2***
Osteoblast	*IRF, SP1*	*IRF*
AoSMC	*EWSR1-FLI1, ZNF281, NFE2L2, FOXP1*	***PAX4, SOX7****, IRF*

Matches to motifs identified using MEME. Motifs were first compared to the top10 non-redundant TFs using STAMP. The matches found to that list are shown in bold. * indicates that the TF also was identified in the classifier as being predictive.