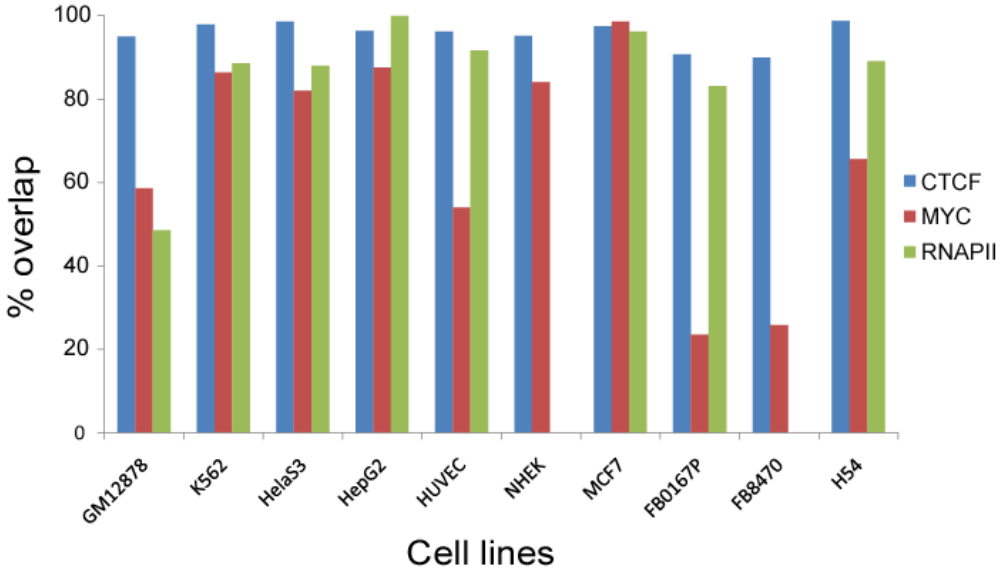
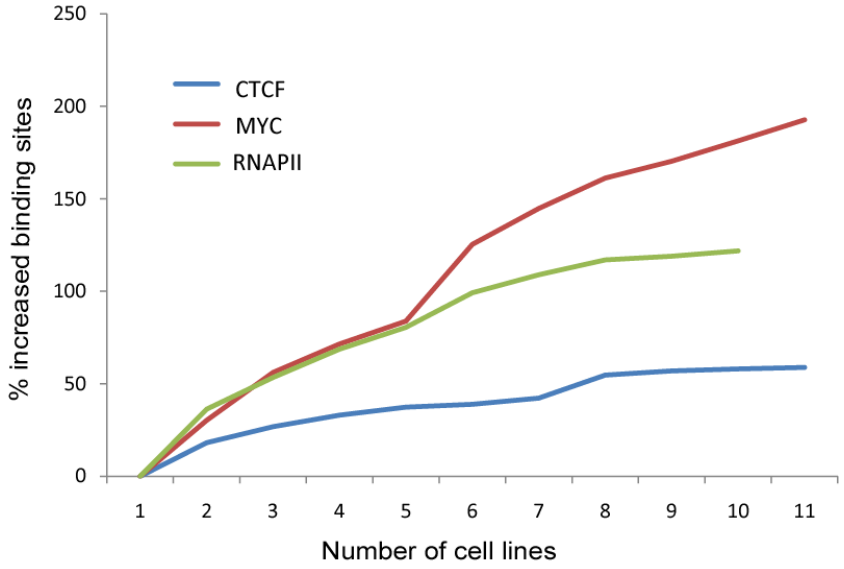
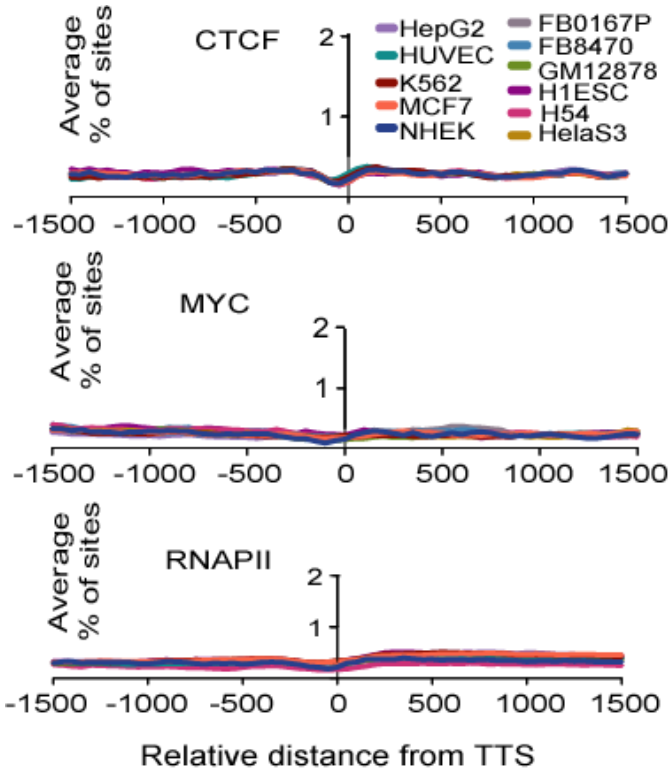
Supplemental data



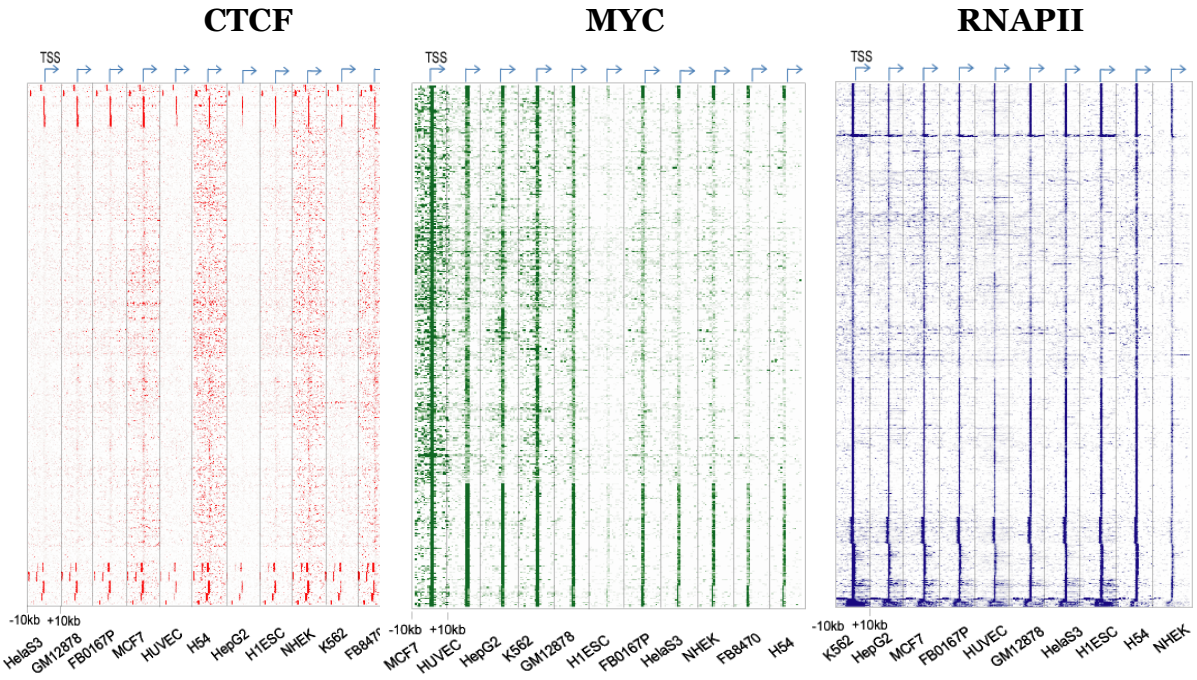
Supplemental Figure S1. Overlap analysis between two biological replicates of ChIP-seq data for each factor in each cell line. The top 50,000 CTCF, 15,000 MYC, and 30,000 Pol2 binding sites from one replicate were compared with the top 65,000, 30,000, and 50,000 binding sites from the other replicate, respectively. The Y-axis shows percent overlap between replicates for the cell types indicated on the X-axis. Overlap values for RNAPII are not shown for NHEK and FB8470 because only one replicate was used for NHEK, and no data was generated for RNAPII in FB8470. 2 cases of lower replicate overlap in the case of MYC was likely due to low MYC RNA expression levels as observed in array data (not shown). For consistent data processing we processed these two samples (MYC FB0167P and FB8470) in the same manner as the other data sets. Importantly, peak calling with these datasets with lower replicate overlap does not necessarily overestimate MYC targets because our peak calling program normalizes for total read depth internally.



Supplemental Figure S2. Increasing numbers of CTCF, MYC, and RNAPII binding sites in the human genome identified by sequencing additional cell types. For each factor, cell types were first sorted in order of highest to lowest number of binding sites. The number of new binding sites found with each additional cell type was determined using overlap analysis, and plotted as a percentage on the Y-axis.

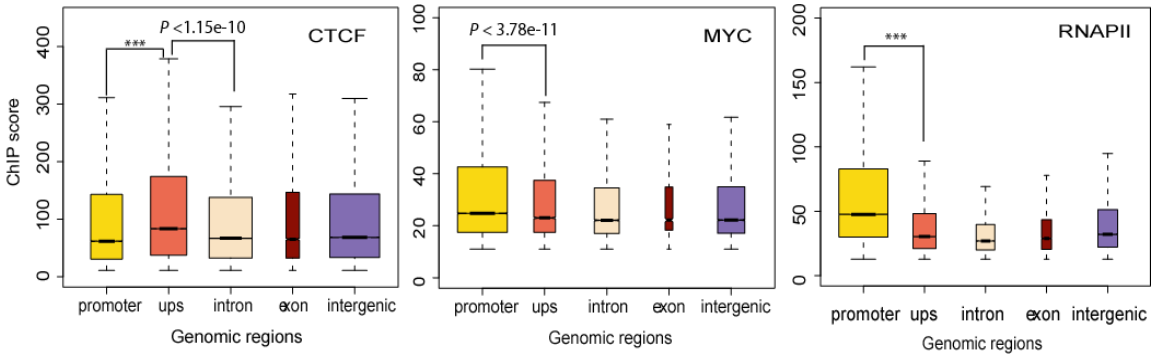


Supplemental Figure S3. The average genome-wide binding profiles of CTCF, MYC, and RNAPII around transcription termination sites (TTS). Average profiles were generated in the same manner as for Fig. 2A.

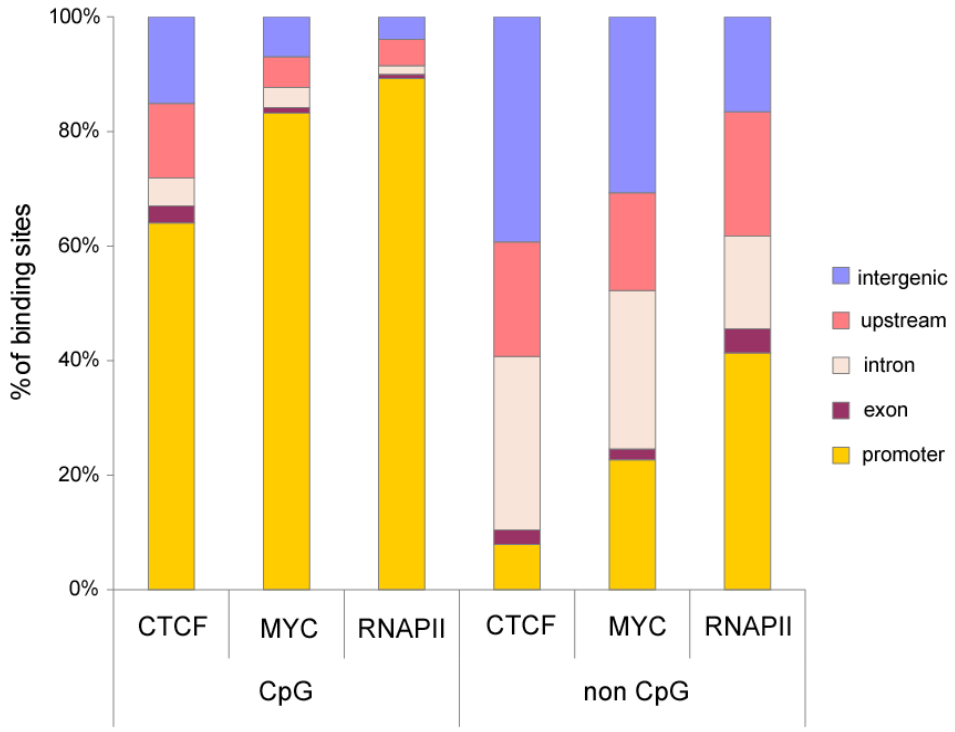


Supplemental Figure S4. Patterns of CTCF, MYC and RNAPII occupancy around genes. The heat map shows normalized occupancy scores for each gene with available data (rows) within ±10 kb from TSS (columns are 100 bp bins). Arrow indicates TSS and direction of transcription. For each factor, data in the first listed cell line was clustered using K-means clustering and data for the other cell lines are displayed in the same order. The ordering of genes is not the same for the 3 factors. The number of genes listed in the Y-axis is 13,610 for CTCF, 10,440 for MYC, and 13,000 for RNAPII.

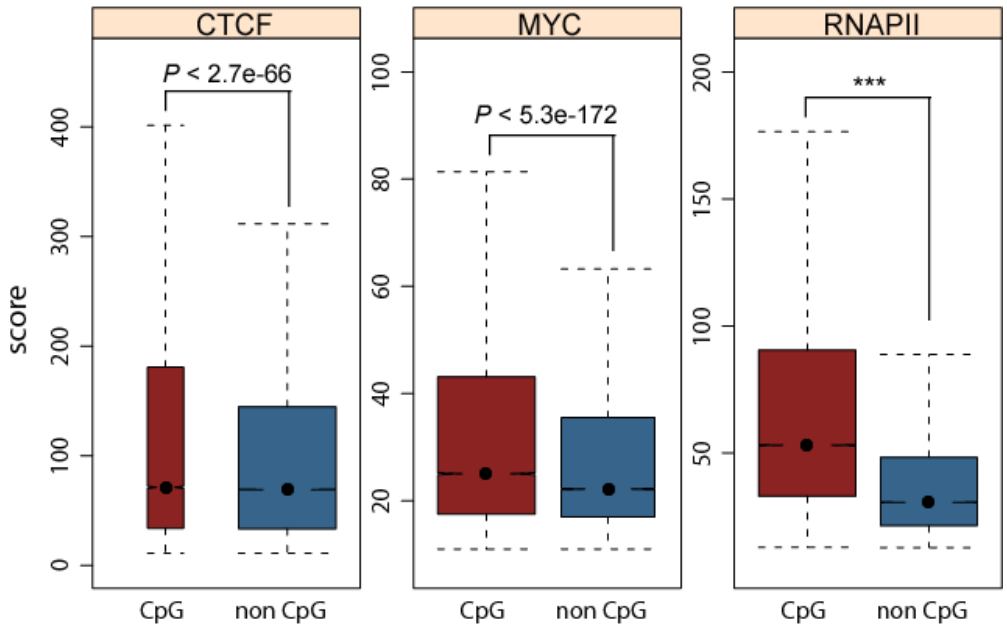
Supplemental Figure S5. Well-conserved binding sites of MYC in all cell types including H1ESCs.



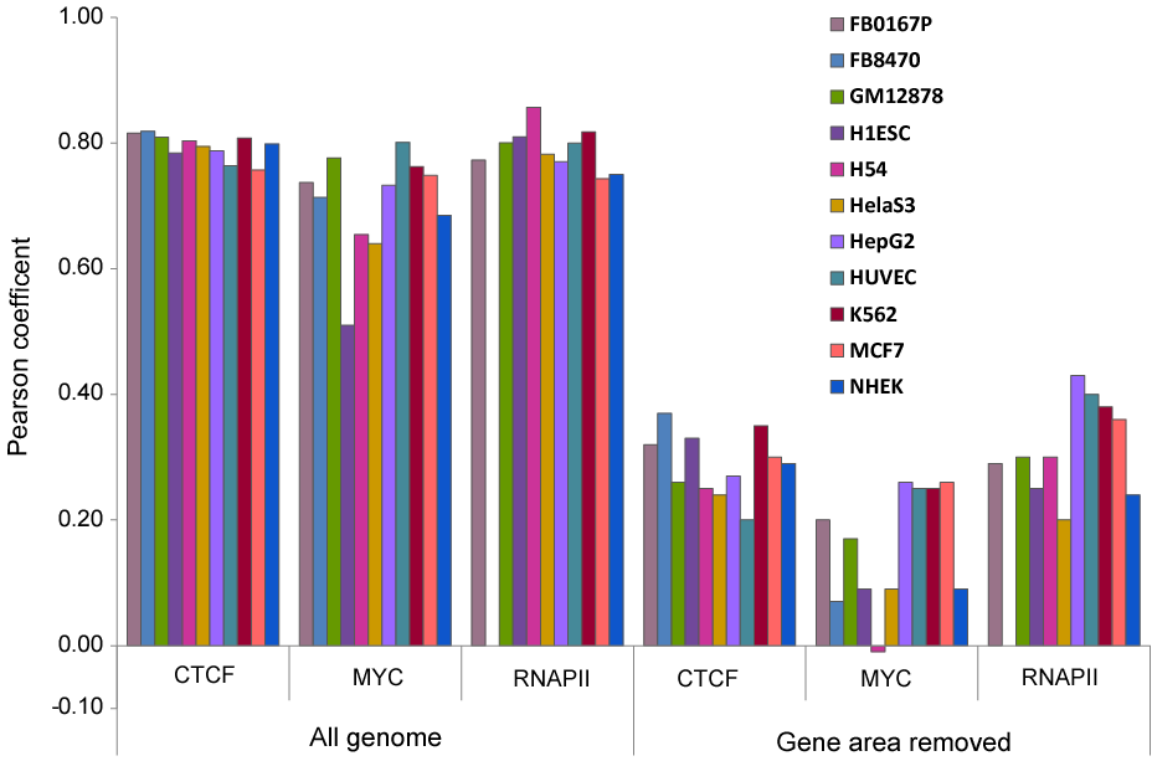
Supplemental Figure S6. Box plots showing ChIP-seq score distribution of CTCF, MYC, and RNAPII binding sites in 5 different genomic regions across all cell types. MYC and RNAPII binding sites in promoters had significantly higher ChIP scores than other genomic regions, whereas CTCF had higher ChIP scores in upstream regions. *P*-value was calculated by Wilcoxon rank sum test. Three asterisks (***) indicate a *P*-value of zero.



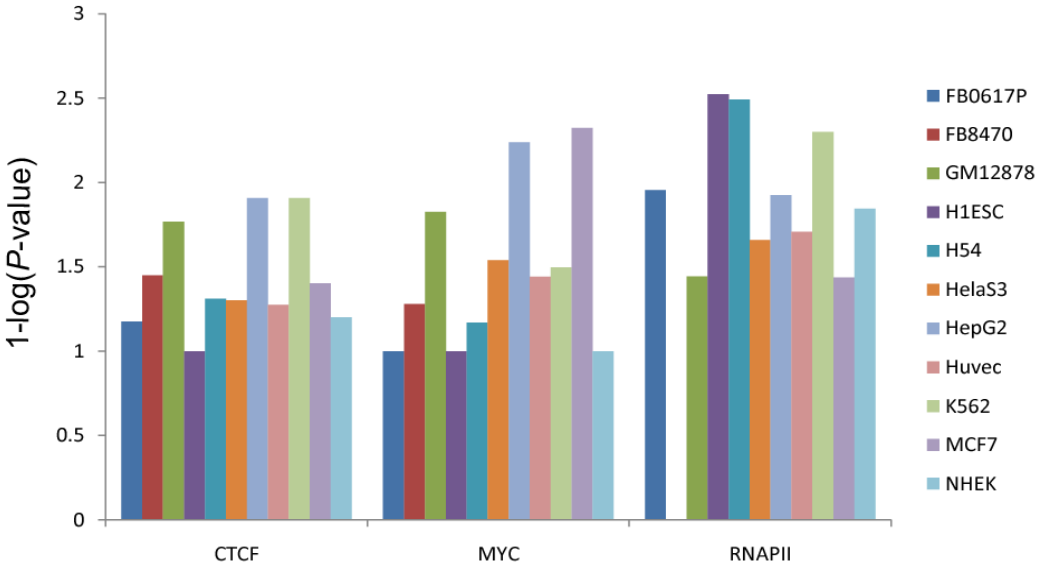
Supplemental Figure S7 The distribution of CpG and non-CpG binding sites of CTCF, MYC, and RNAPII in 5 different genomic regions.



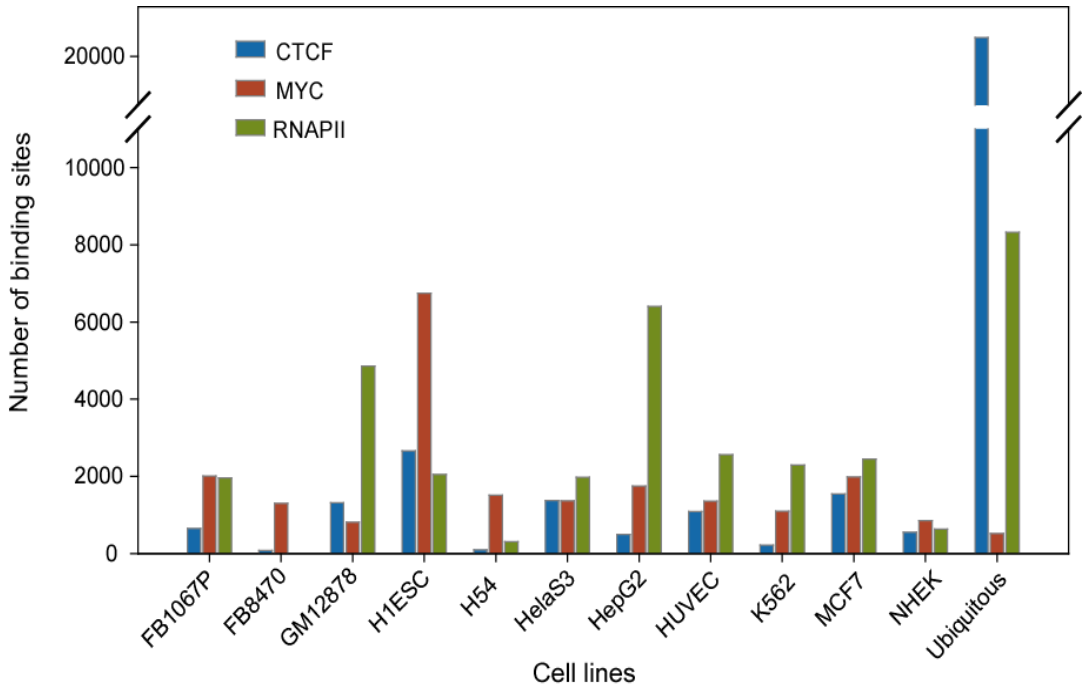
Supplemental Figure S8. MYC and RNAPII binding sites show significantly higher occupancy scores in CpG islands than in non-CpG regions. *P*-value was calculated by the Wilcoxon rank sum test. Three asterisks (***) indicate a *P*-value of zero.



Supplemental Figure S9. Pearson correlation coefficients between the number of CTCF, MYC, and RNAPII binding sites and gene density (plotted on the Y-axis). Removing all binding sites of CTCF, MYC, and RNAPII up to 20 kb away from genes still showed a positive correlation. A negative control dataset generated by random permutation of data showed no positive correlation (all values at zero, not shown).



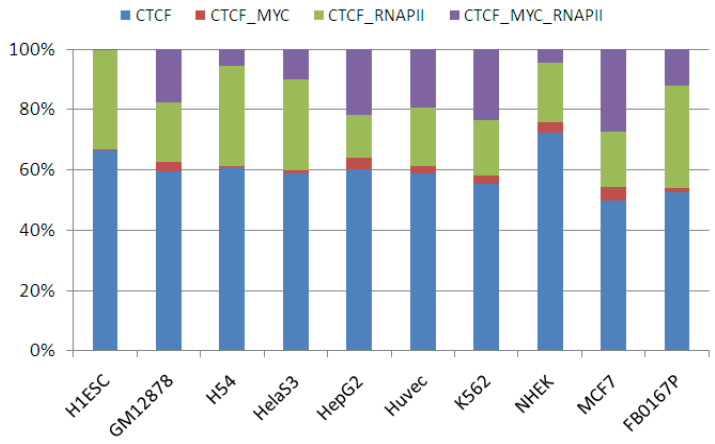
Supplemental Figure S10. TF binding in bidirectional promoters activates both genes equally, regardless of its binding distance from transcription start site of genes. A value of 3 on the Y-axis indicates *P*-value of 0.01, with lower values indicating higher *P*-values. *P*-values were calculated by the binomial distribution function. The null hypothesis was that the binding of a factor in a bidirectional promoter is equally likely to regulate both genes to the same extent, resulting in similar expression levels. The high *P*-values fail to reject the null hypothesis.



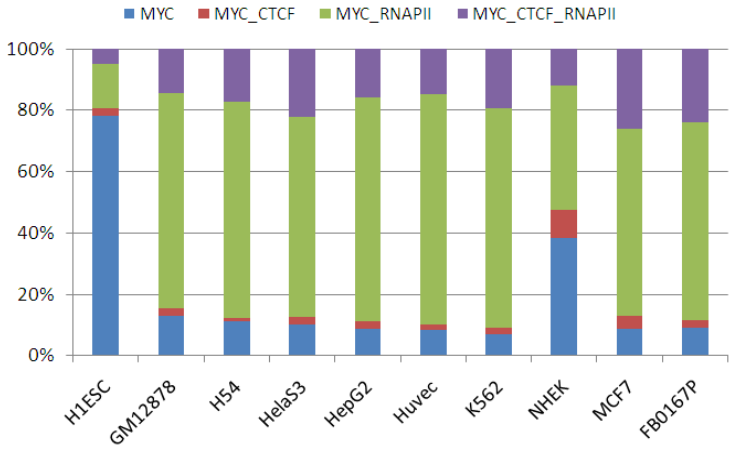
Supplemental Figure S11. The number of tissue-specific binding sites of CTCF, MYC, and RNAPII across the genome in different cell types.

Supplemental Figure S12. Distribution of single and combinatorial binding sites of CTCF, MYC and RNAPII in different cell types. X-axis shows cell lines and Y-axis represents % binding sites. Y-axis represents % binding sites.

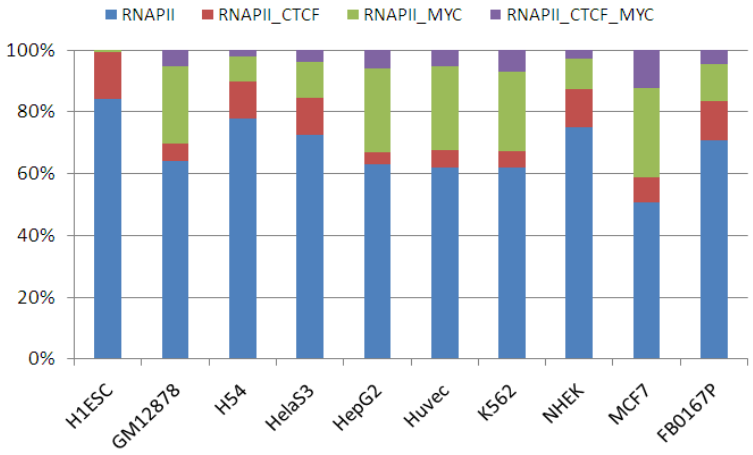
CTCF



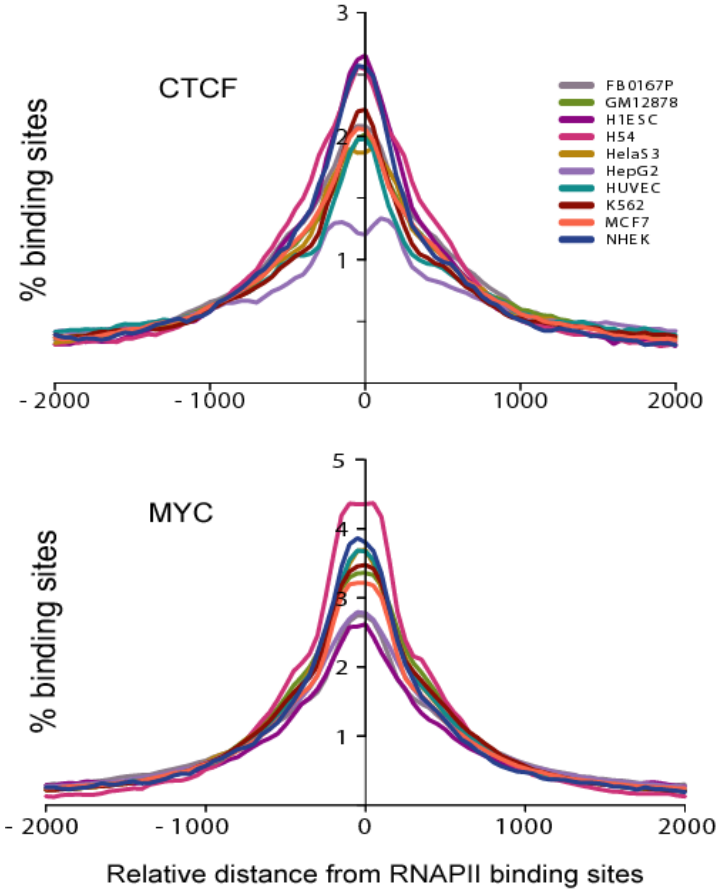
MYC



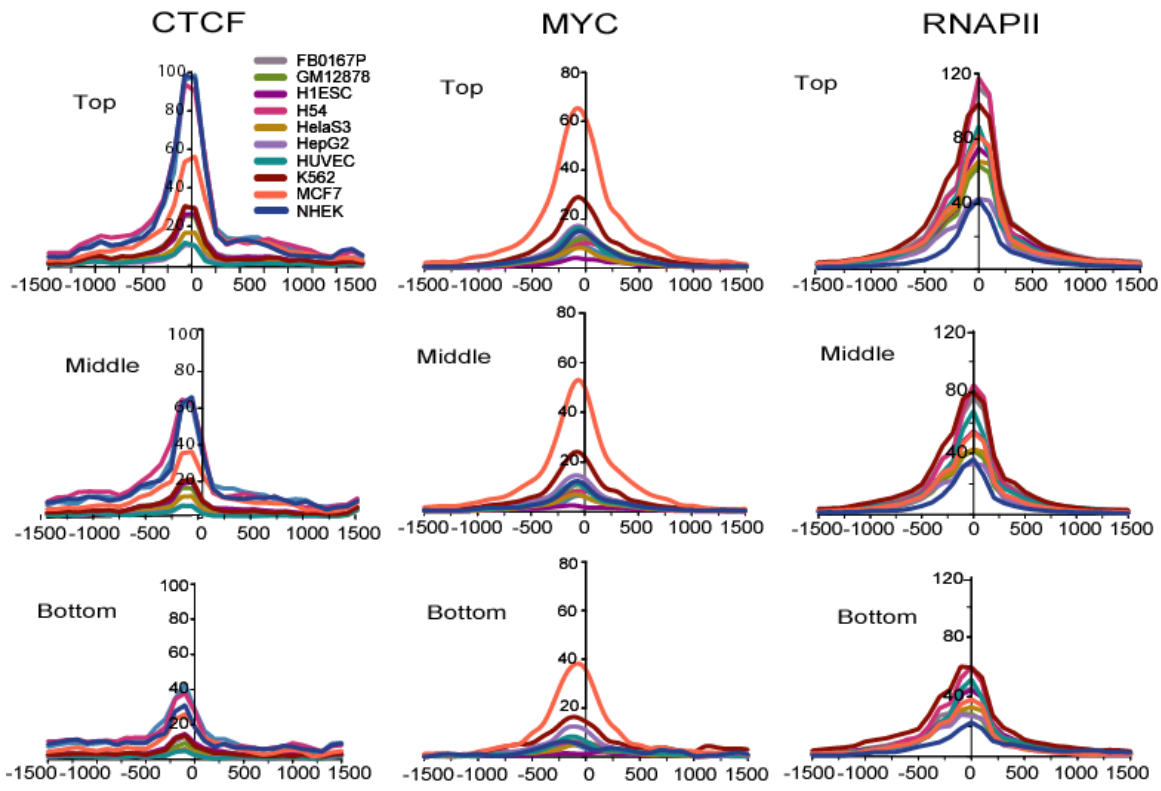
RNAPII



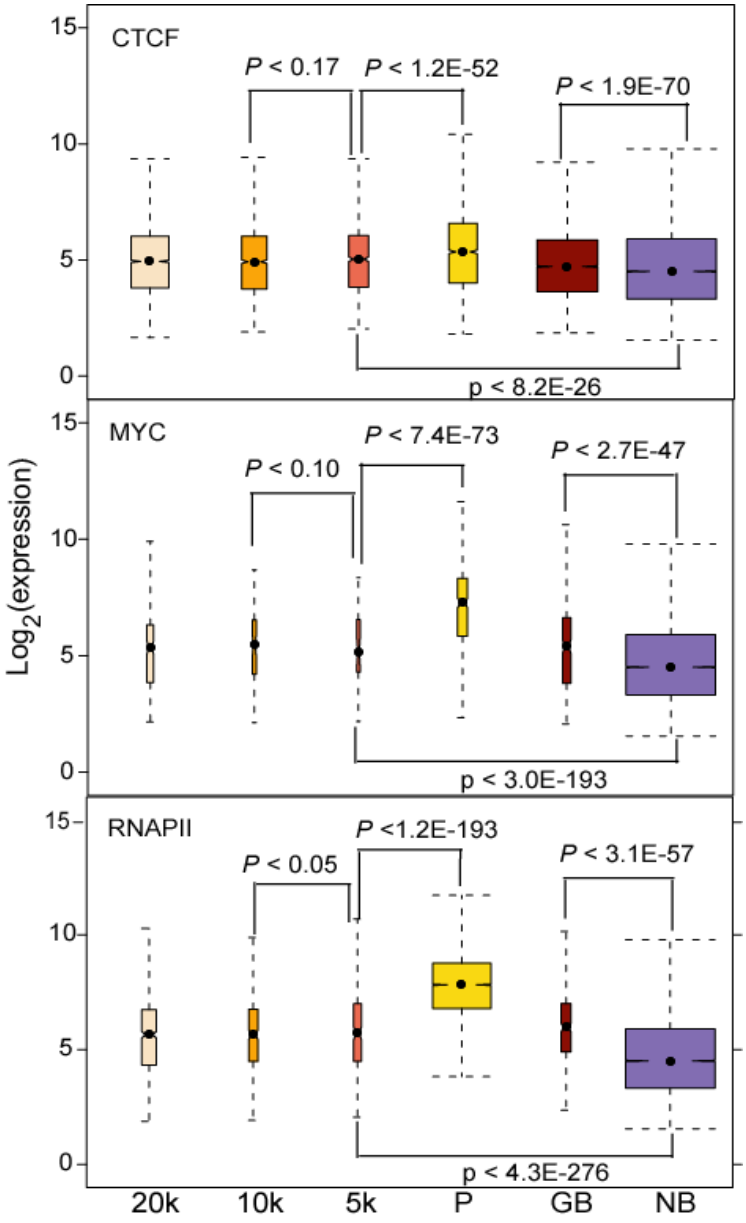
Supplemental Figure S13. Distribution of single and combinatorial target genes of CTCF, MYC and RNAPII in different cell types. X-axis shows cell lines and Y-axis represents % target genes. Y-axis indicates % target genes.



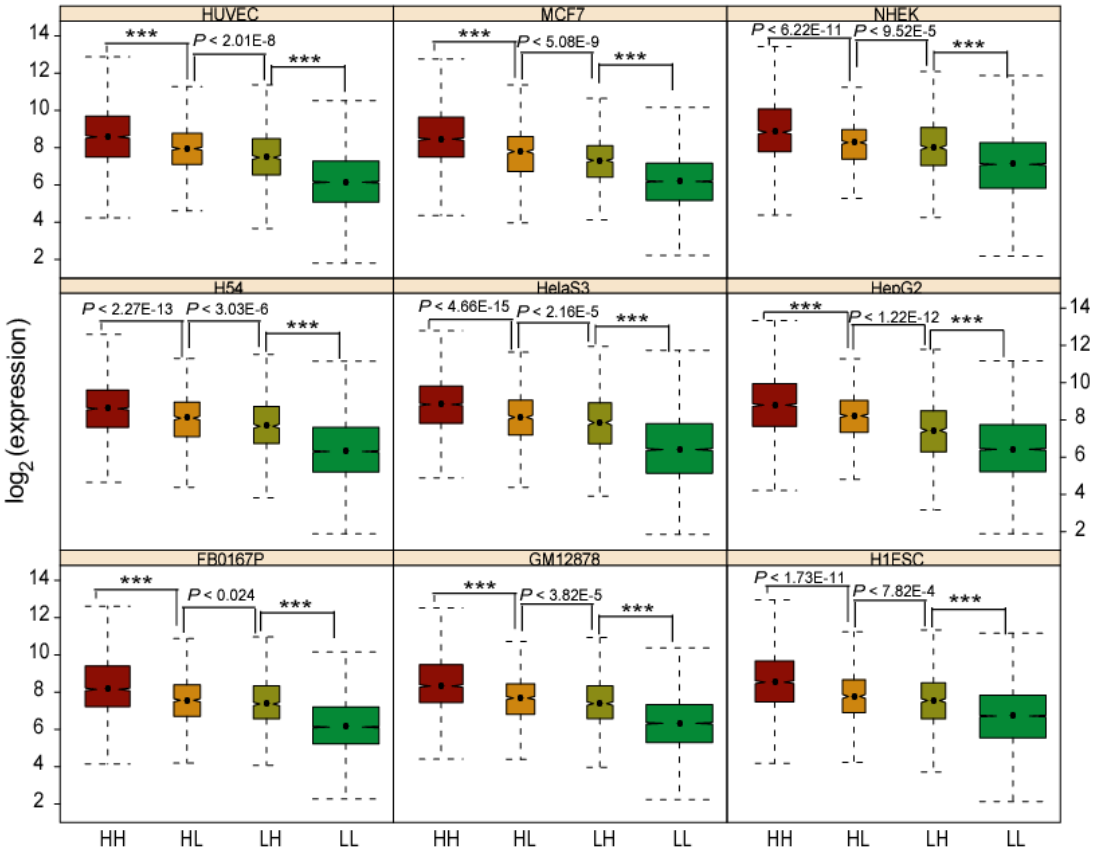
Supplemental Figure S14. Average CTCF and MYC binding profiles around RNAPII binding sites.



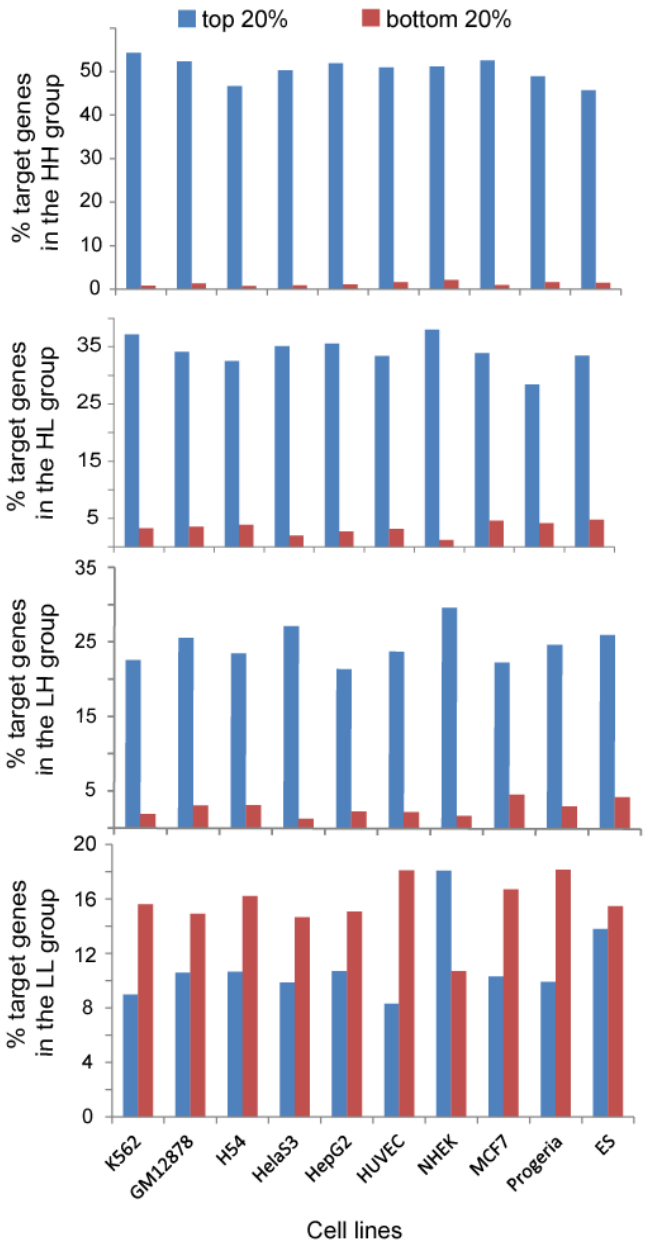
Supplemental Figure S15. Binding profiles of CTCF, MYC, and RNAPII in 3 classes of expression groups around TSS. Genes were classified into 3 different expression groups containing the top 33%, middle 33% and bottom 33% based on their expression level. X-axis represents relative distance from TSS. The Y-axis indicates the normalized, input-corrected occupancy score and is therefore comparable across the three expression groups.



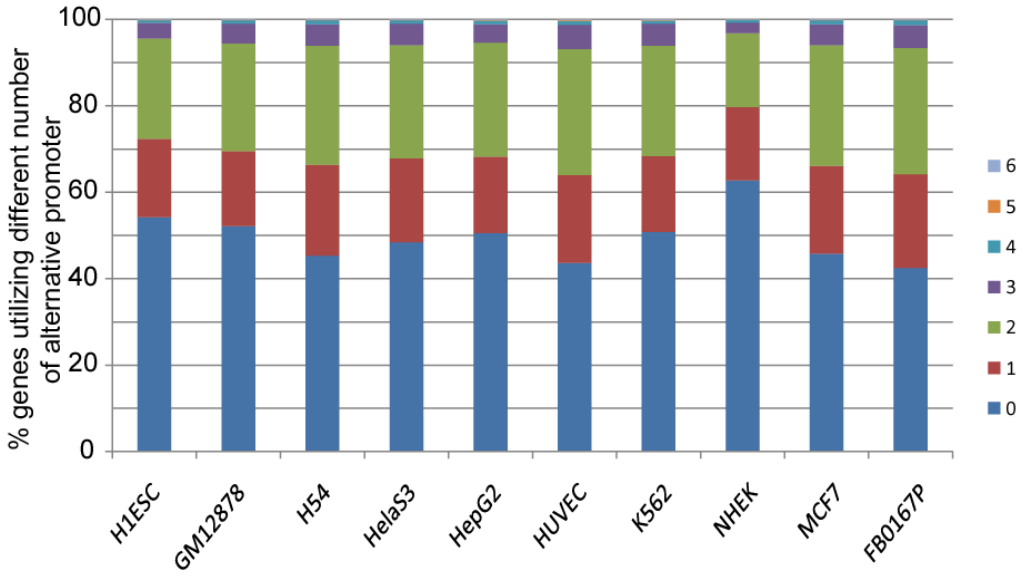
Supplemental Figure S16. Location effect of CTCF, MYC, and RNAPII sites on expression of target genes. X-axis represents binding position. 5/10/20K, GB, and NB denote 5, 10 or 20 Kb upstream from TSS, gene body, and non-binding, respectively. Y-axis indicates $\log_2$ expression. *P*-values were calculated by Wilcoxon rank sum test.



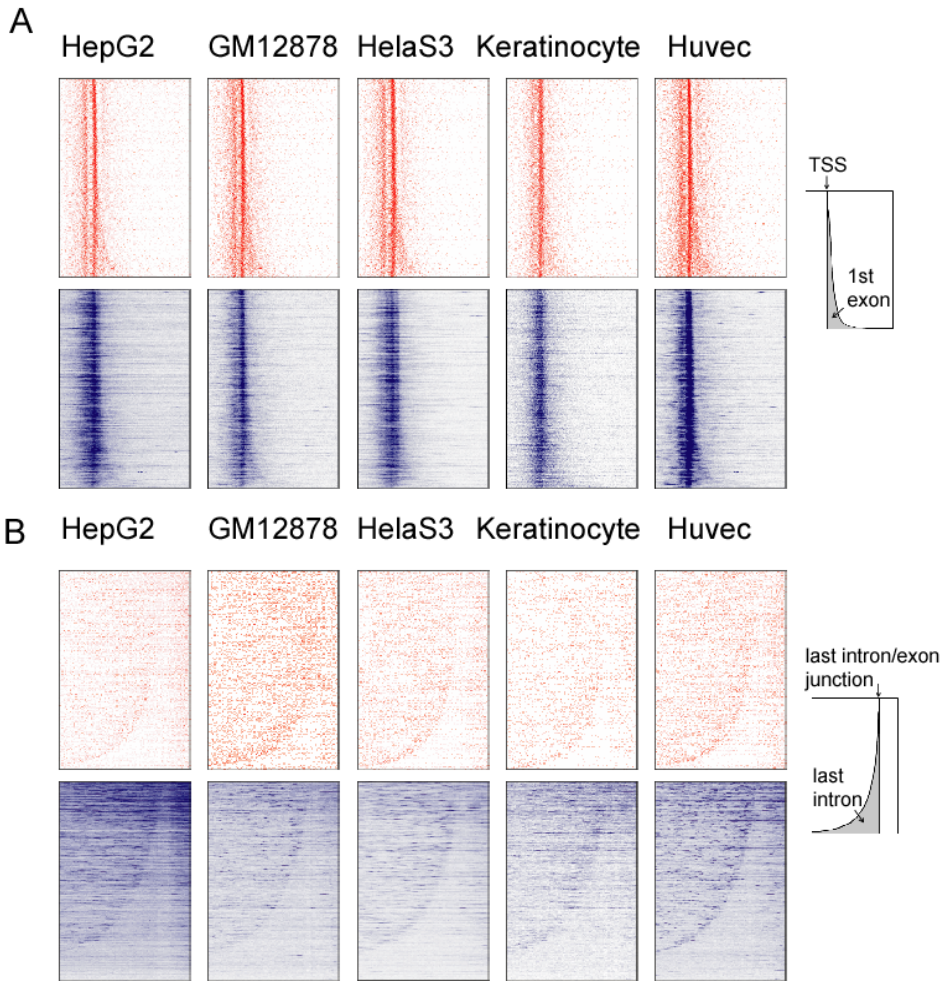
Supplemental Figure S17. Box plots showing distribution of gene expression in the 4 RNAPII binding groups for additional cell lines (see Fig. 6B). *P*-values were calculated by Wilcoxon rank sum test. Asterisks indicate a calculated *P*-value of zero.



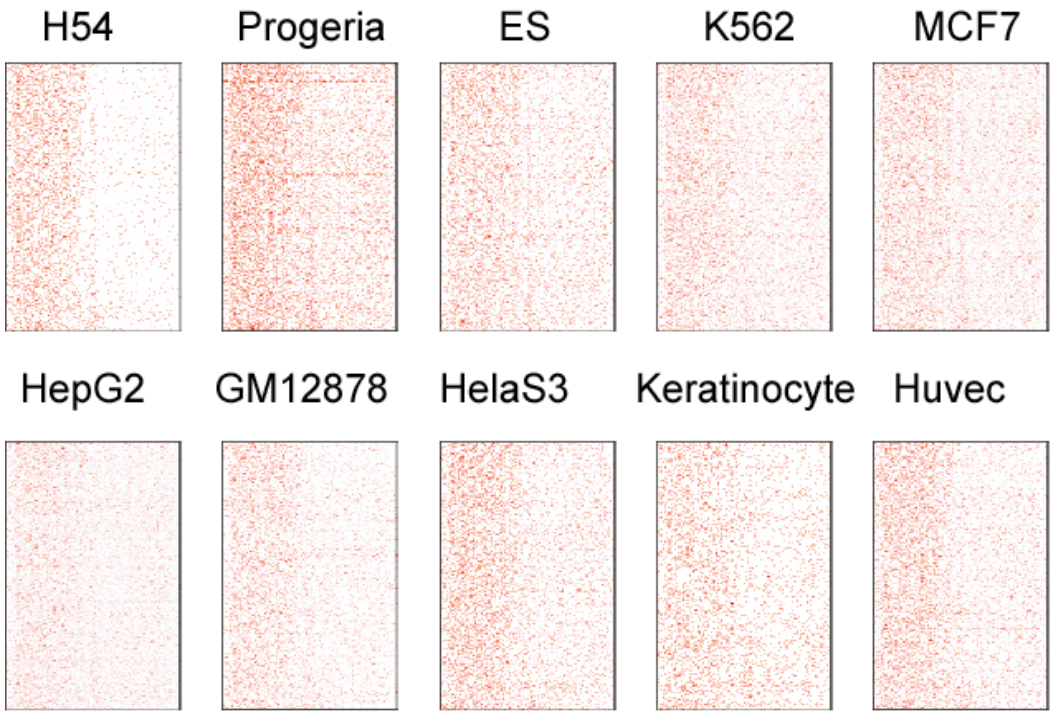
Supplemental Figure S18. Proportions of genes in the 4 RNAPII binding groups that are highly expressed (top 20%) as well as low-expressed (bottom 20 %). Y-axis shows the percentage of genes in each group.



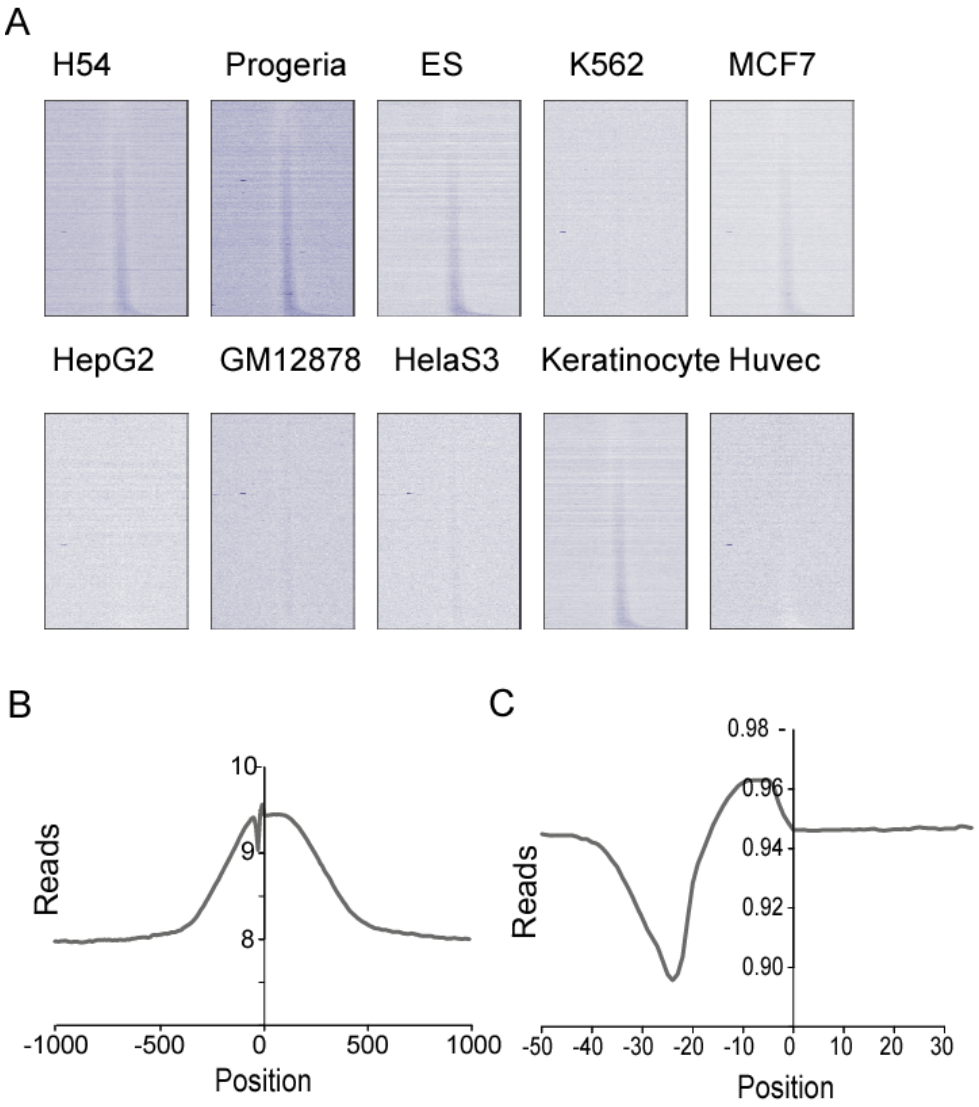
Supplemental Figure S19. Percentage of genes utilizing the indicated numbers of alternative promoters among genes containing at least 2 alternative promoters. Each color indicates number of alternative promoters utilized.



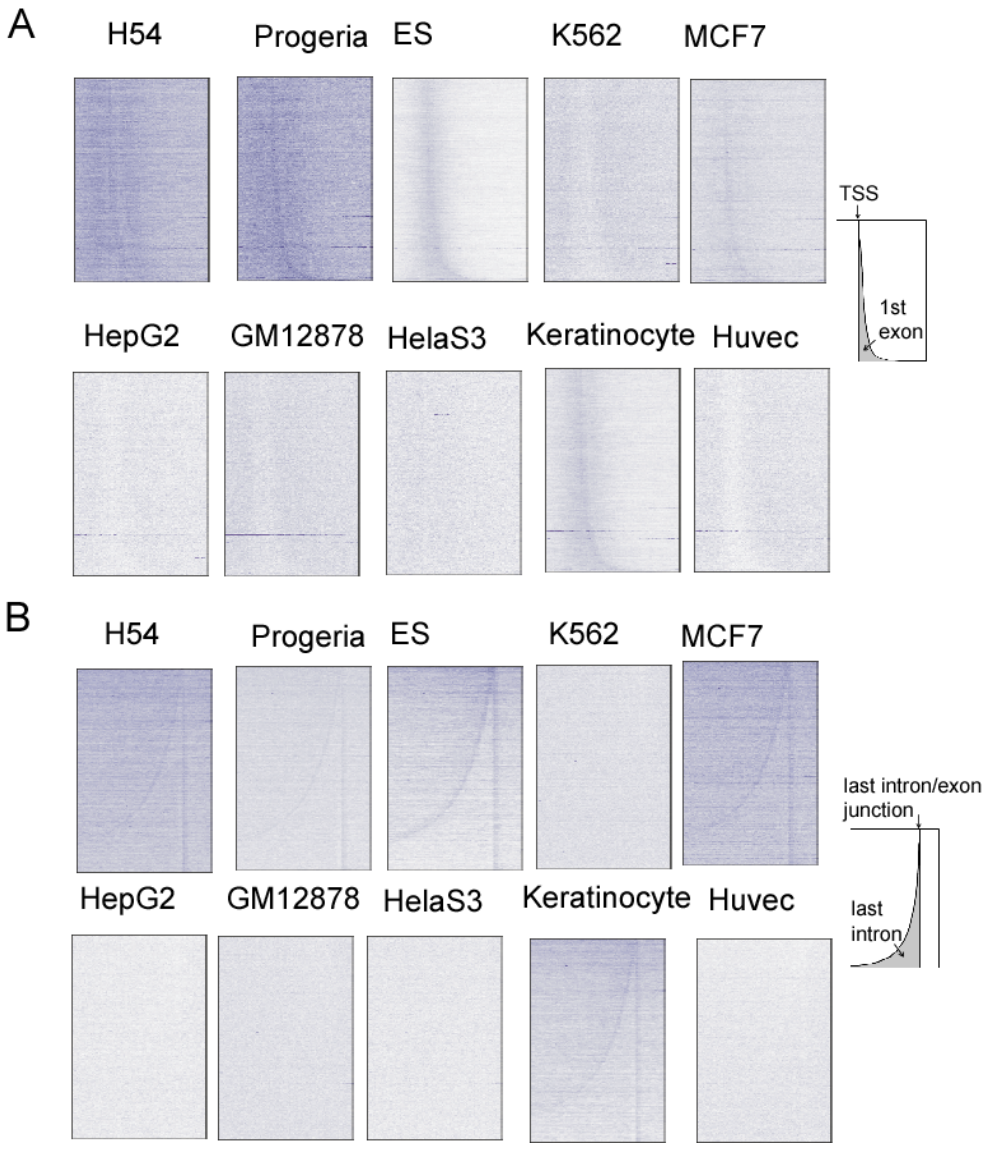
Supplemental Figure S20. Pol2 is enriched at first and last exons. (A and B) Pol2 signals at the first and last exon/intron junctions are plotted as in Figure 8.



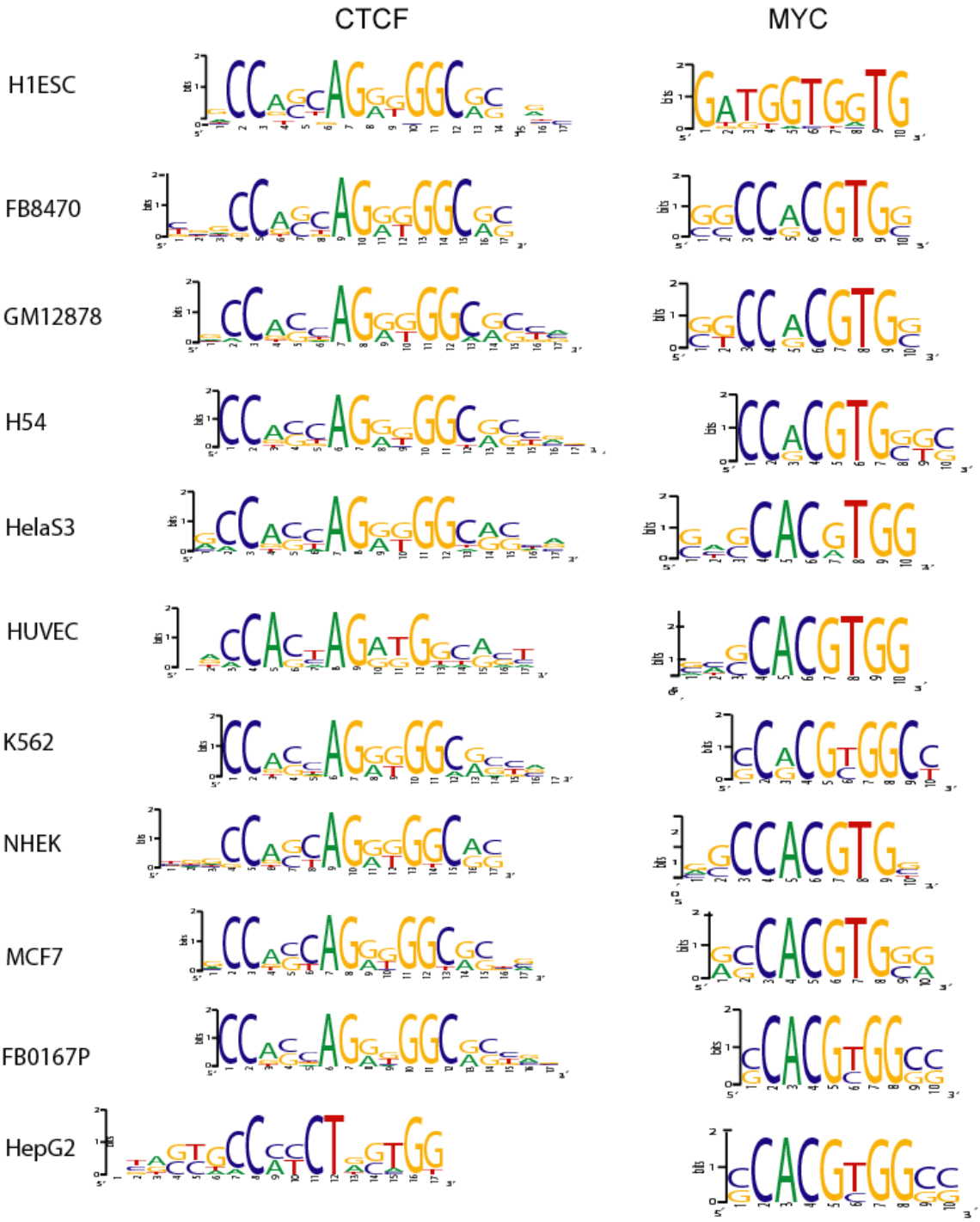
Supplemental Figure S21. Pol2 ChIP peaks at constitutive internal exons. Pol2 ChIP peaks (normalized and input-corrected) in 10 cell lines were plotted as in Figure 8.



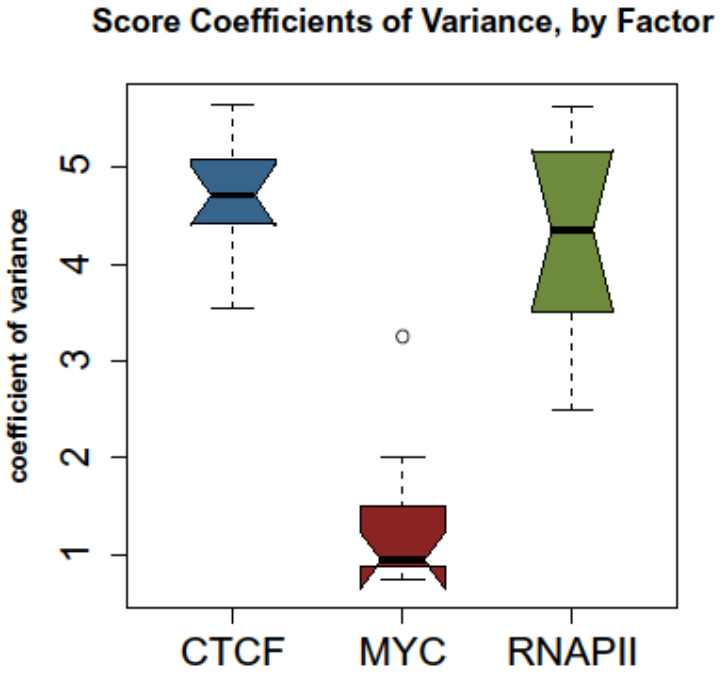
Supplemental Figure S22. Input bias at constitutive internal exons. (A) Input reads are plotted around the internal exons as in Figure 8. (B) Exons and their immediate surroundings are more uniquely mappable than regions farther away. All internal exons were aligned to their start sites and mappability scores were assigned into 10 bp bins surrounding the exon start sites. Average of the mappability scores is plotted. (C) Higher resolution view of mappability in the 30 bp region around the exon start sites. Internal exons are aligned as in B and mappability scores are assigned into 1 bp bins.



Supplemental Figure S23. Input read density at the first and last exon/intron junctions. (A and B) Input reads are plotted at the first and last exon/intron junctions as in Figure 8.



Supplemental Figure S24. Motifs discovered from strong binding sites de novo using DME on ChIP-seq peaks. Virtually identical motifs were found in weak and moderate sites (not shown)



Supplemental Figure S25. Distribution of coefficients of variance (standard deviation/mean occupancy score) for input corrected and filtered binding sites before applying target cutoff thresholds for CTCF, MYC and RNAPII (11, 11 and 10 cell lines respectively). The qualitative impression that MYC binding sites exhibit generally weaker signal is reflected in the lower CoVs for most MYC cell lines.

Supplemental Table S1. The number of independent biological replicates and aligned reads for all ChIP-seq as well as input data sets.

	Input		CTCF		MYC		RNAPII	
	Replicates	# Aligned Reads	Replicates	# Aligned Reads	Replicates	# Aligned Reads	Replicates	# Aligned Reads
FB0167P	1	31,510,516	2	46,874,291	2	25,230,535	2	49,082,997
FB8470	1	40,979,563	2	39,266,345	2	10,955,174		
GM12878	1	16,568,390	3	30,830,198	2	27,762,943	2	41,965,165
H1ESC	2	21,110,780	1	14,563,540	1	19,025,332	1	22,033,234
H54	1	51,821,256	2	104,978,274	2	58,004,041	2	36,683,532
HelaS3	1	11,798,703	2	28,322,770	2	14,228,304	2	18,851,779
HepG2	1	11,929,226	2	14,063,412	3	25,203,575	4	36,930,350
HUVEC	1	14,898,732	2	22,204,482	2	30,953,678	2	28,353,668
K562	1	16,493,115	3	27,270,533	3	34,160,398	2	40,791,878
MCF7	1	29,404,931	2	45,850,125	2	80,691,282	2	58,363,979
NHEK	1	26,350,143	2	26,316,271	1	13,234,802	1	20,058,797
Average		**24,805,941**		**36,412,749**		**30,859,097**		**35,311,538**
Total		**241,354,839**		**353,665,950**		**314,219,529**		**304,032,382**
				Average		**31,109,556**		
				Total		**1,213,272,700**		

Supplemental Table S2. The Number of CTCF, MYC, and RNAPII binding sites at the selected score threshold in diverse cell lines

	CTCF			RNAPII			MYC		
	Top %	Score	# Sites	Top %	Score	# Sites	Top %	Score	# Sites
FB0167P	3.5	10.97	**46,424**	2	18.46	**36,325**	0.5	17.72	**7,100**
FB8470	4.2	33.22	**35,652**				0.5	12.65	**4,661**
GM12878	3.7	10.97	**50,366**	2	20.06	**31,118**	0.39	10.97	**8,134**
H1ESC	4	21.25	**40,890**	2	25.94	**25,089**	0.5	12.83	**9,222**
H54	4	29.68	**43,501**	2	20.78	**23,069**	0.26	10.97	**3,938**
HelaS3	3.8	10.97	**53,731**	2	16.15	**29,249**	0.48	10.97	**6,949**
HepG2	4	16.66	**43,234**	2	13.64	**42,550**	0.5	13.14	**11,623**
HUVEC	3.6	10.97	**45,175**	2	21.82	**33,924**	0.5	12.82	**11,862**
K562	4	30.54	**37,289**	2	28.36	**31,560**	0.5	19.91	**9,685**
MCF7	4	22.98	**45,705**	2	12.73	**41,224**	0.92	33.22	**15,435**
NHEK	7.2	33.22	**38,009**	2	26.28	**13,511**	0.5	16.85	**4,697**
Average			**43,355**			**30,762**			**8,482**
Total			**479,976**			**307,619**			**93,306**

Supplemental Table S3. Functional categories enriched among the ubiquitous binding sites of CTCF, MYC, and RNAPII, reported by the DAVID web-based functional annotation program. Enrichment and FDR Q-Values based on the hypergeometric distribution analyzing overlap between genes in the GO Biological Process ontology term and the list of genes bound by RNAPII (8167), MYC (185) or CTCF (3321) across all cell types. Genes bound by CTCF in all cell types showed enrichment in a variety of lipid metabolic processes, while genes ubiquitously bound by RNAPII or MYC showed enrichment in constitutive housekeeping functions such as translation and protein degradation.

		RNAPII			MYC			CTCF		
GO Biological Process Term	Total Genes	Gene Hits	Fold Enrich	FDR Q-Val	Gene Hits	Fold Enrich	FDR Q-Val	Gene Hits	Fold Enrich	FDR Q-Val
translational elongation	105	92	1.8	0.0%	32	7.7	0.0%			
ribosome biogenesis	135	113	1.7	0.0%	22	5.4	0.0%			
nuclear export	62	55	1.8	0.0%						
protein folding	176	131	1.5	0.0%						
regulation of ubiquitin-protein ligase activity	81	68	1.7	0.0%						
RNA processing	582				80	3.5	0.0%			
ubiquitin-dependent protein catabolic process	273				14	4.6	0.1%			
RNA elongation from RNA polymerase II promoter	49	46	2.0	0.0%						
cellular alkene metabolic process	25							25	1.4	1.1%
unsaturated fatty acid biosynthetic process	38							35	1.3	6.4%
icosanoid biosynthetic process	35							32	1.2	10.7%
leukotriene metabolic process	24							24	1.4	1.4%
glutathione biosynthetic process	12							12	1.4	23.7%
release of cytochrome c from mitochondria	22							20	1.2	34.8%
neutral amino acid transport	20							18	1.2	45.3%

Supplemental Table S4. Functional categories enriched among target genes occupied by combinations of CTCF, MYC, and RNAPII, as reported by the DAVID web-based functional annotation program. Values are fold enrichment; hypergeometric distribution-based FDR Q-Values for all enrichments are < 5%.

		CTCF + MYC					CTCF + Pol II					Pol II + MYC								CTCF + MYC + Pol II								
GO Biological Process term	# gene	HeLaS3	HepG2	HUVEC	K562	MCF7	GM12878	H1ESC	HepG2	MCF7	NHEK	FB0167P	GM12878	H54	HeLaS3	HepG2	HUVEC	K562	MCF7	NHEK	FB0167P	GM12878	H54	HeLaS3	HepG2	HUVEC	K562	MCF7	NHEK
translation	337							1.7				2.3	2.2	4.2	2.4	2.0	1.9	2.1	2.0	4.0	2.4	2.8	4.6	2.9	2.7	2.3	2.6	2.0	3.5
ribosome biogenesis	135											2.4	2.3	3.3	2.4	2.1	2.1	2.1	2.1	3.3	2.4	3.4	5.8	3.4	2.1	2.4	2.4	1.8	4.5
ncRNA metabolic process	241											2.3	2.1	3.0	2.3	2.0	1.8	2.1	2.1	2.8		2.3	2.9	2.2		1.9			
RNA splicing	301								1.8				2.0	2.6	1.8	1.8		1.8	1.9	2.6									
tRNA metabolic process	118											2.4	2.2	3.4	2.5	2.2	1.9	2.2	2.1										
tRNA processing	78											2.5			2.5		2.0	2.3	2.3	3.7									
tRNA modification	20													6.3	2.9		2.5												
tRNA aminoacylation for protein translation	45																									3.3			
protein folding	176											1.8	1.8	2.3	1.6		1.5	1.6	1.8										
'de novo' posttranslational protein folding	12					3.9																							
regulation of protein ubiquitination	123																								2.1				
positive regulation of ubiquitin-protein ligase activity	70																		1.7										
chromatin silencing	23																								4.3				
histone acetylation	53																										3.0		
histone H2A acetylation	12																									5.6			
protein amino acid acetylation	58							2.1																			2.9		
gas homeostasis	8																									7.0			
spindle organization	51																1.8												
NADP metabolic process	19																								4.6				
NADPH regeneration	12																								5.5				
glucose catabolic process	57											2.1																	
monosaccharide catabolic process	71									2.1																			
negative regulation of gene expression, epigenetic	26																										3.7		
DNA damage checkpoint	61																										3.1		
DNA integrity checkpoint	65																										2.9		
Golgi vesicle transport	138											1.8																	
post-Golgi vesicle-mediated transport	60											2.0																	
regulation of cell migration	233	2.5																											
apoptotic mitochondrial changes	30						4.0			2.7																			
regulation of transcription factor activity	144				2.7																								
regulation of binding	213	2.4			2.6																								
response to inorganic substance	270	2.3																											
actin cytoskeleton organization	257		2.1																										
icosanoid biosynthetic process	35			5.3																									
unsaturated fatty acid biosynthetic process	38			4.9																									
leukotriene biosynthetic process	20			5.9																									
cellular alkene metabolic process	25			4.7																									
positive regulation of signal transduction	292			2.2																									
SMAD protein nuclear translocation	8					5.1																							
regulation of I-kappaB kinase/NF-kappaB cascade	131						2.1																						
hemidesmosome assembly	11										6.6								6.9										
cell-substrate junction assembly	32										3.5																		