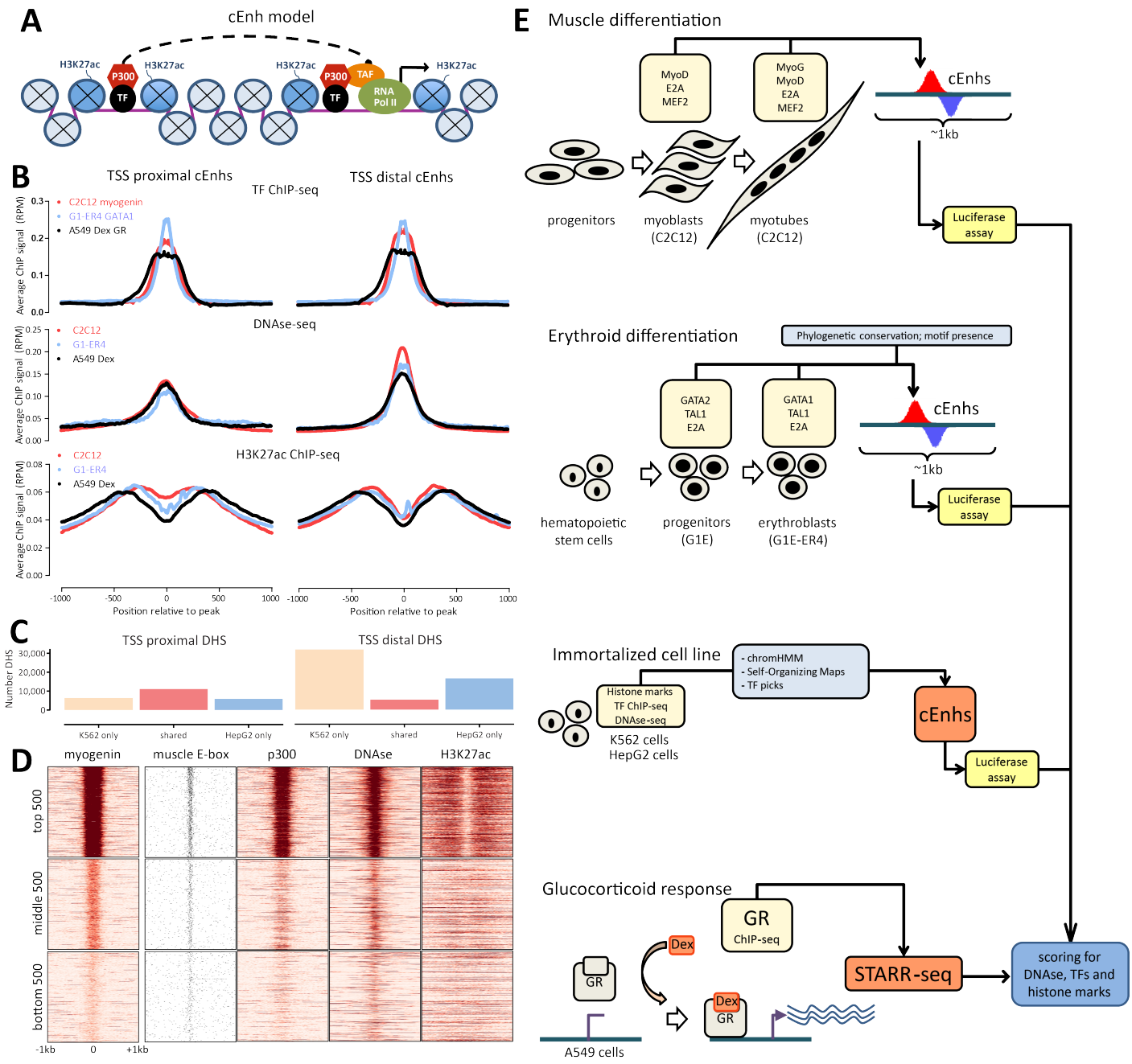


Figure 1: Biochemical signatures and functional testing of candidate enhancer elements (cEnhs) in mammalian genomes. (A) Biochemical signatures of cEnhs and promoters. Active enhancers are characterized by DNAse hypersensitivity due to nucleosome depletion, by p300 occupancy and by H3K27ac, as well H3K4me1 (not shown). Promoter elements share some of these features, but also associate with components of the transcription and transcription initiation machineries, and are marked by H3K4me3 (not shown); (B) Genome-wide commonalities and differences between the biochemical signatures of enhancers and promoters. Shown is the average signal profile around TSS distal (right; defined as regions more than 1kb away from an annotated TSS) and TSS proximal (left) cEnhs (defined as statistically significant peaks in the respective datasets; see the Methods section for further detail) in mouse and human cells for TFs (myogenin in differentiating mouse muscle cell, GATA1 in erythroid mouse cells, and the glucocorticoid receptor upon Dexamethasone stimulation of human A549 cells), DNAse hypersensitivity and H3K27ac; (C) The distribution of biochemical signal strength varies over a large continuum. Shown are the signal distribution for myogenin, p300, DNAse-seq, and H3K27ac relative to the summits of the top 500, middle 500 and bottom 500 reproducible myogenin ChIP-seq sites (total $n = 32,278$) in differentiated C2C12 muscle cells, as well as the distribution of the cognate myogenin TF binding

(legend continued on next page)

motif. (D) Different cell types share a small fraction of their distal cEnh elements, in contrast to promoter elements. Shown are the common and cell-type specific TSS proximal (within 1kb of an annotated TSS) and TSS distal DHSs between the human erythroid K562 and hepatocyte HepG2 immortalized cell lines; E) Outline of cENH selection approaches, biological systems, experimental design and functional assays used in this study. Sets of cEnhs for functional testing were compiled based on: TF ChIP-Seq occupancy measurements (of the master regulators of muscle differentiation, MyoD and myogenin) in differentiating mouse C2C12 cells; phylogenetic conservation patterns and TF occupancy measurements (of the regulators of erythropoesis GATA1 and TAL1) in differentiating mouse G1E-ER4 cells; TF occupancy (multiple TFs) in immortalized K562 cells; machine learning methods (Self-Organizing Maps, chromHMM and Segway) defining integrated chromatin states over multiple histone modification, DNAse and TF occupancy measurements. These cEnhs were tested using luciferase assays. In addition, DNA fragments from GR ChIP-seq experiments in Dex-stimulated A549 cells were cloned and assayed for activity using the STARR-seq assay. Active elements identified using these methods were then evaluated for the presence and distribution of various biochemical signatures.

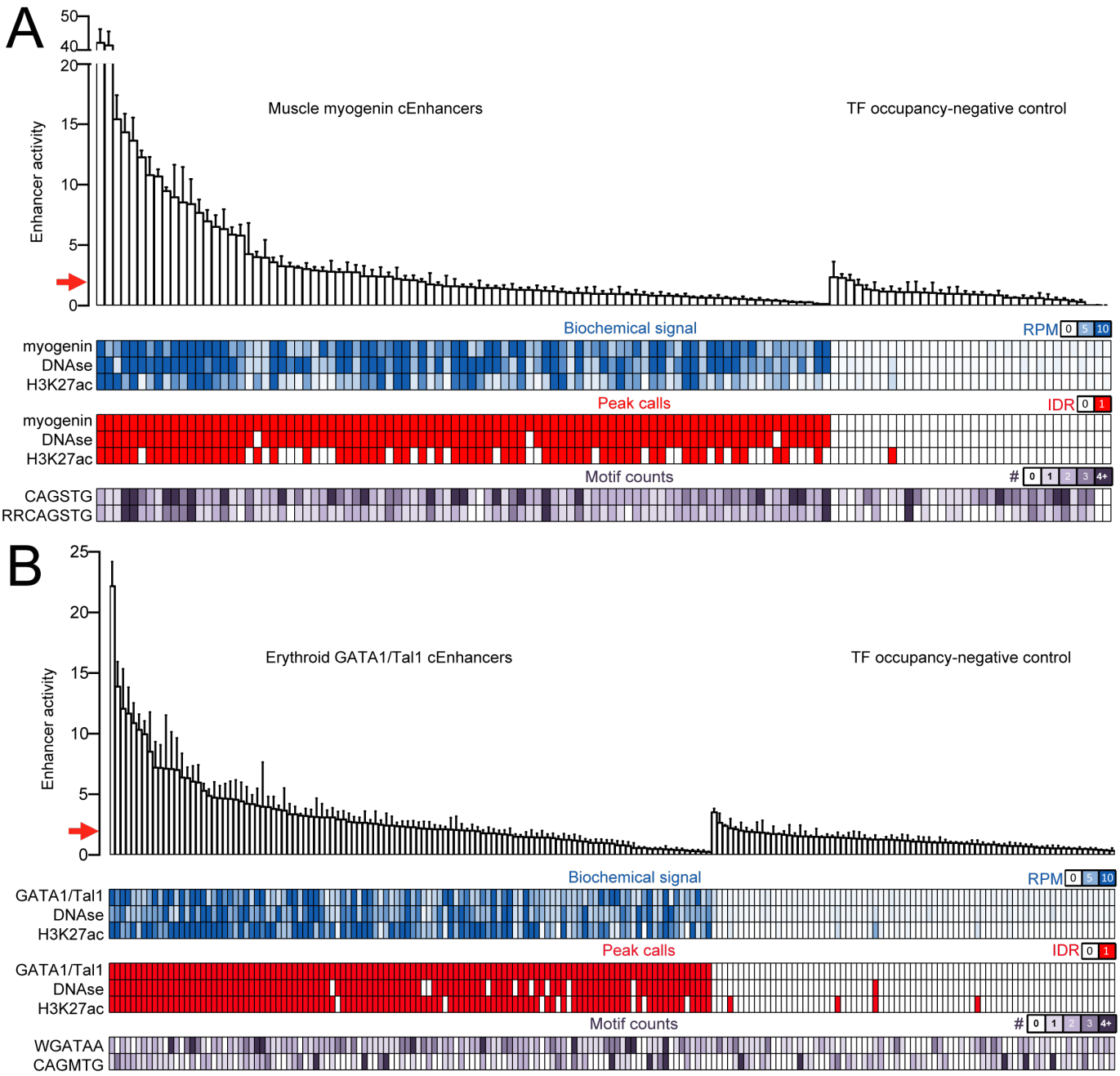


Figure 2: Functional testing of cEnh regulatory activity in mammalian cells. (A) Functional assay testing of cEnh regulatory activity in the context of muscle differentiation. Shown is luciferase assay fold activity in differentiated C2C12 myocytes across technical replicates ($n = 4$). The red arrow corresponds to the mean fold activity threshold above which elements are considered active. In addition, for each cEnh DNAse hypersensitivity, H3K27ac status, and myogenin occupancy are shown, both as RPM (Read Per Million) signal intensity values and as binary peak calls, as well as the number of myogenin motif (RRCAGSTG, derived from myogenin ChIP-seq data) occurrences. Tested cEnhs are sorted by mean fold activity. (B) Functional assay testing of cEnh regulatory activity in the context of erythropoesis. Shown is luciferase assay fold activity in K562 cells across biological ($n \in [1:9]$) and technical replicates ($n = 4$ for each biological replicate). The red arrow corresponds to the mean fold activity threshold above which elements are considered active. In addition, for each cEnh DNAse hypersensitivity, H3K27ac status, and GATA1/TAL1 occupancy are shown, both as RPM (Read Per Million) signal intensity values and as binary peak calls, as well as the number of TAL1 (CAGMTG) and GATA1 (WGATAA) motif occurrences. Tested cEnhs are sorted by mean fold activity.

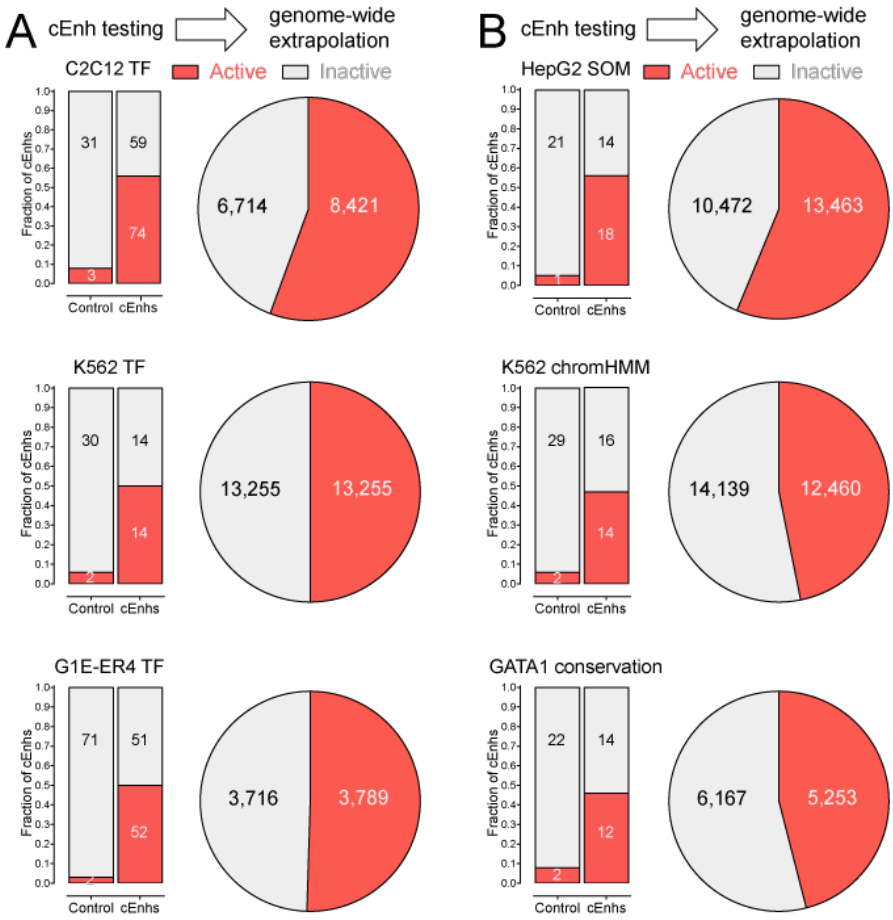


Figure 3: Summary of cEnh activity predictions by different selection criteria. (A) TF occupancy-centered selections. Tested eEnhs selected on the basis of TF occupancy in the context of mouse muscle differentiation and erythropoesis and in human K562 cells were further subselected with the additional requirement of exhibiting DNAse hypersensitivity and the H3K27ac histone mark. The fraction of active constructs in negative controls and cEnhs are shown on the left. The expected number of active cEnhs genome-wide is extrapolated on the left based on the number of $TF^+/DNAse^+/H3K27ac^+$ regions in the genome; (B) TF-occupancy agnostic selections. Tested eEnhs selected using Self-Organizing Maps in HepG2 cells, chromHMM in K562 cells, and evolutionary conservation of GATA1 motifs in G1E cells were further subselected with the additional requirement of exhibiting DNAse hypersensitivity and the H3K27ac histone mark. The fraction of active constructs in negative controls and cEnhs are shown on the left. The expected number of active cEnhs genome-wide is extrapolated on the left based on the number of $DNAse^+/H3K27ac^+$ (for HepG2 SOM and K562 chromHMM selections) $DNAse^+/H3K27ac^+$ regions with a conserved GATA1 motif (for GATA1 conservation selections) in the genome.

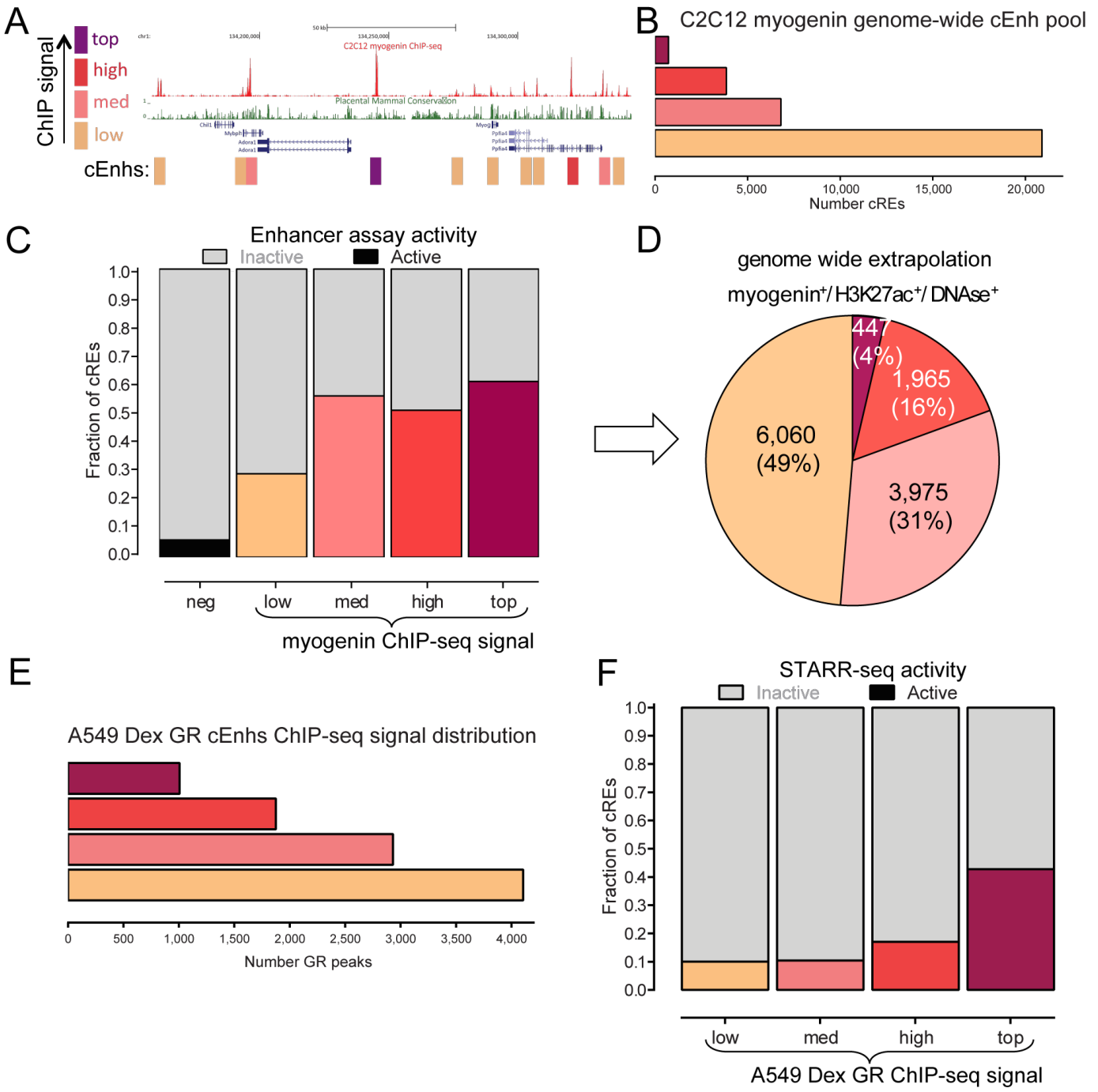


Figure 4: Enrichment of active cEnhs in different classes of cEnhs defined by the strength of their biochemical signatures. (A) cEnhs (rectangle boxes) belonging to different signal classes (based on ChIP-seq data for myogenin in C2C12 myocytes; "top": RPM ≥ 10; "high": RPM$\in [5, 10]$; "medium": RPM$\in [2.5, 5]$; "low" RPM≤ 2.5) in the neighborhood of the mouse *Myog* gene; (B) Genome-wide distribution of cEnhs in different signal classes based on ChIP-seq data for myogenin in C2C12 myocytes; (C) Fraction of active enhancers in different cEnh signal classes (based on ChIP-seq data for myogenin in C2C12 myocytes; "top": $n = 57$; "high": $n = 43$; "medium": $n = 39$; "low" $n = 38$) as well as in negative controls (with not myogenin occupancy; $n = 34$). Only cEnhs positive for myogenin, DNAse and H3K27ac were included; (D) Extrapolated number of active enhancers in C2C12 belonging to each signal strength class

(legend continued on next page)

based on the genome-wide number of myogenin$^+$/DNAse$^+$/H3K27ac$^+$ regions. (E) Genome-wide distribution of cEnhs in different signal classes based on the set of GR ChIP-STARR-seq cEnhs in A549 cells ("top": A549 Dex GR ChIP-seq RPM ≥ 10; "high": RPM$\in [5, 10]$; "medium": RPM$\in [2.5, 5]$; "low" RPM≤ 2.5). Only GR ChIP-seq regions significantly represented within STARR-seq libraries (i.e. with sufficiently many reads to score as active if they were in fact active) are shown for each signal class. (F) Fraction of cEnhs exhibiting significant activity in the GR ChIP-STARR-Seq assay in stimulated A549 cells for each signal strength class.

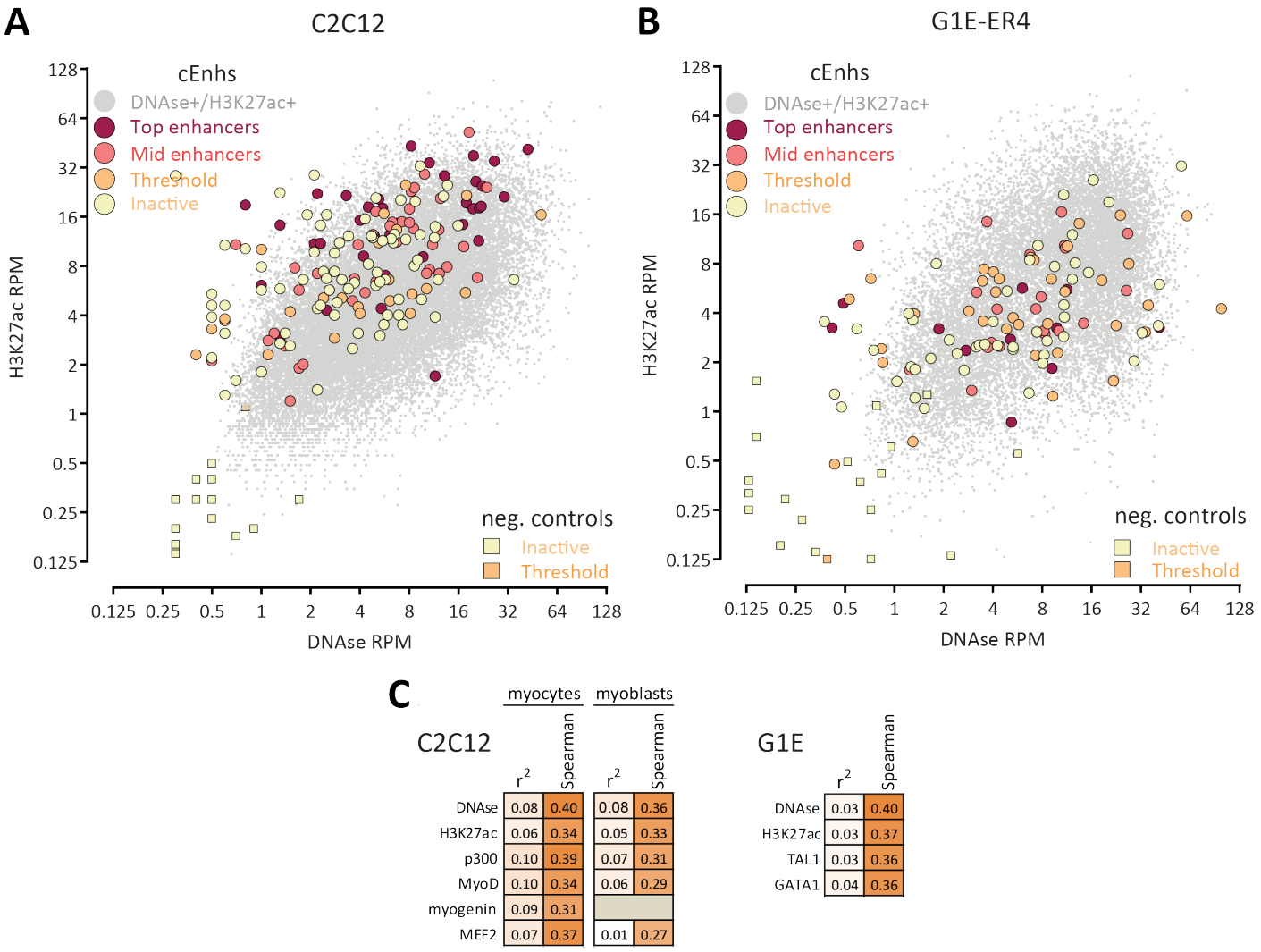
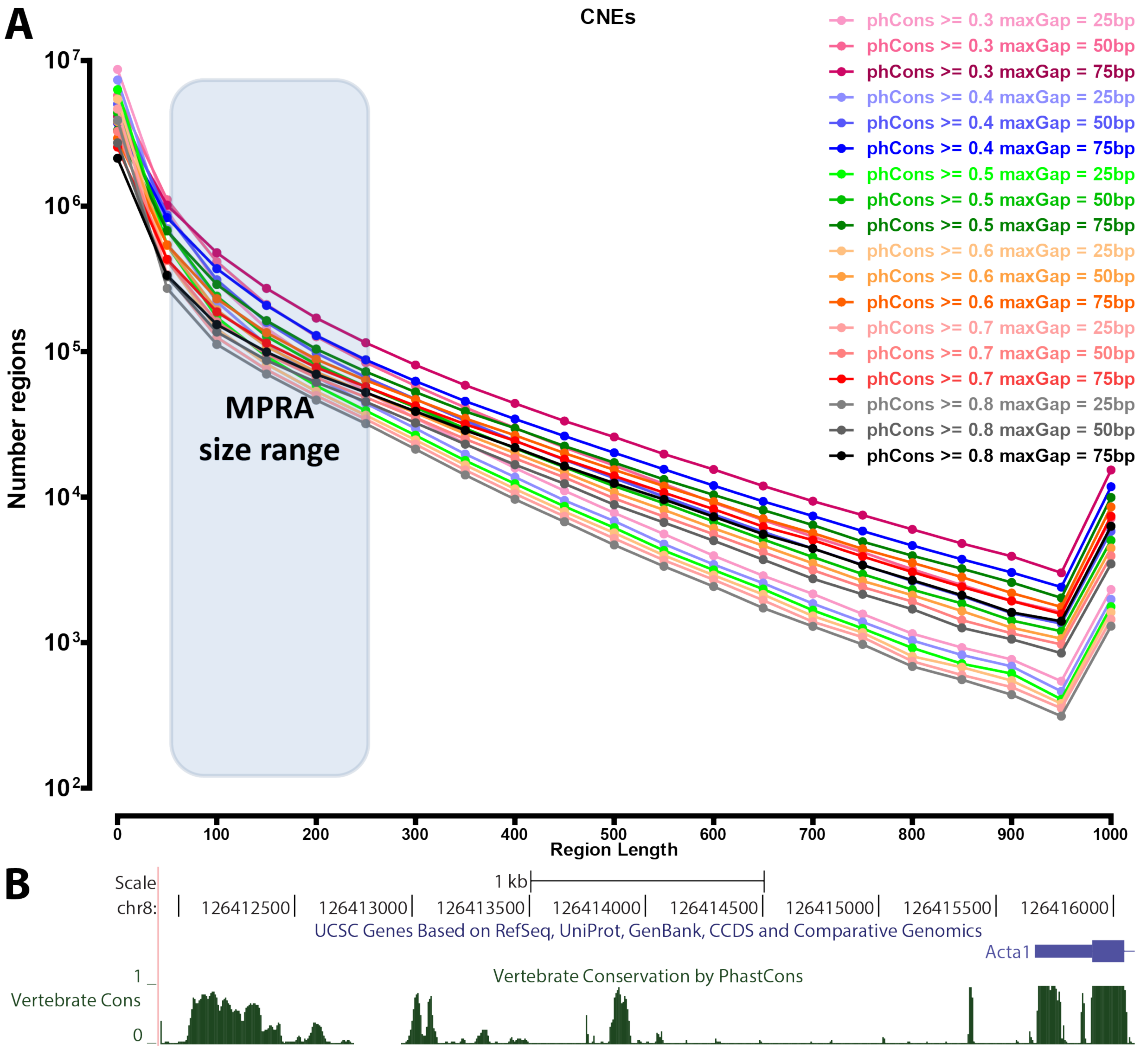
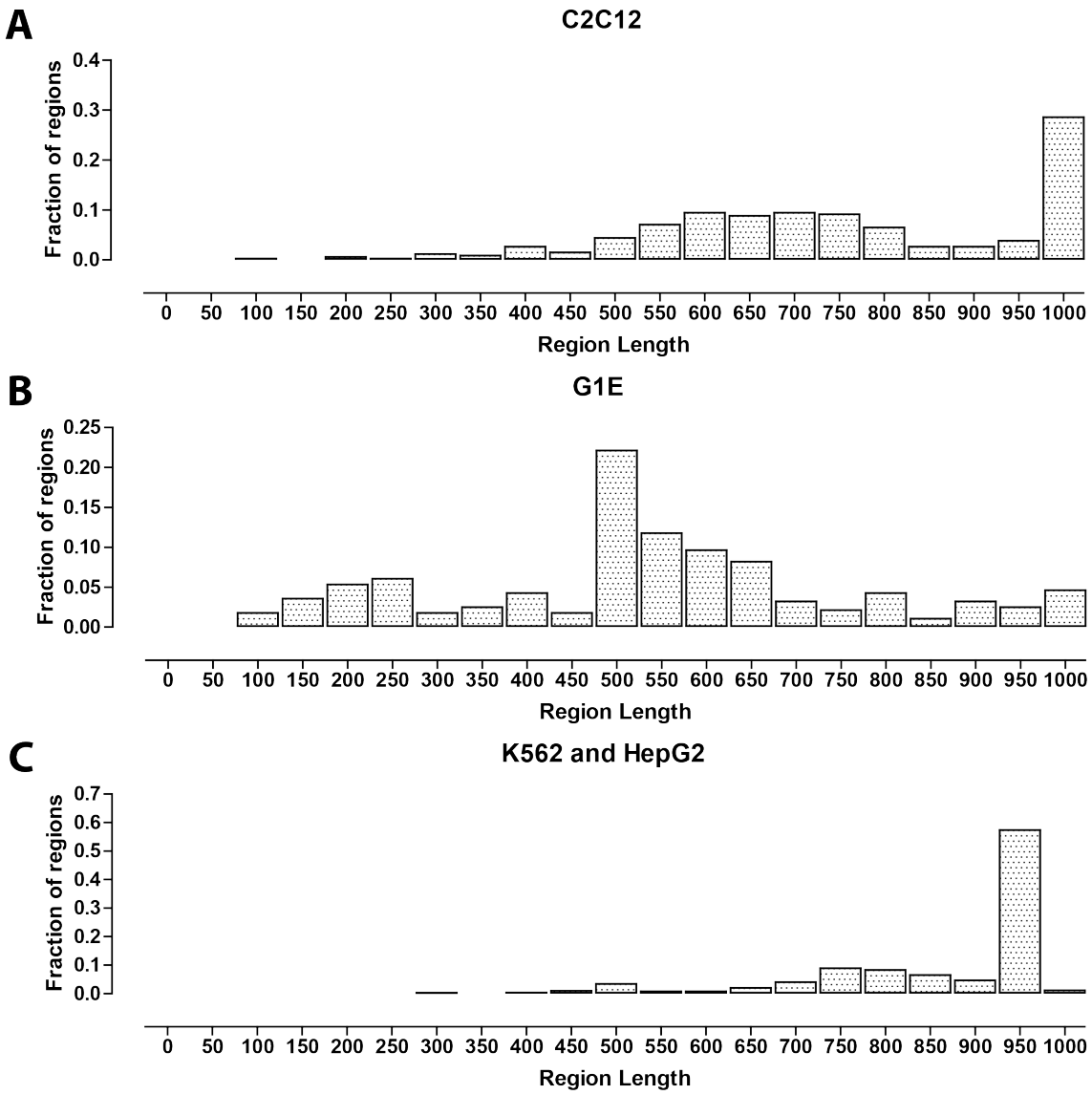


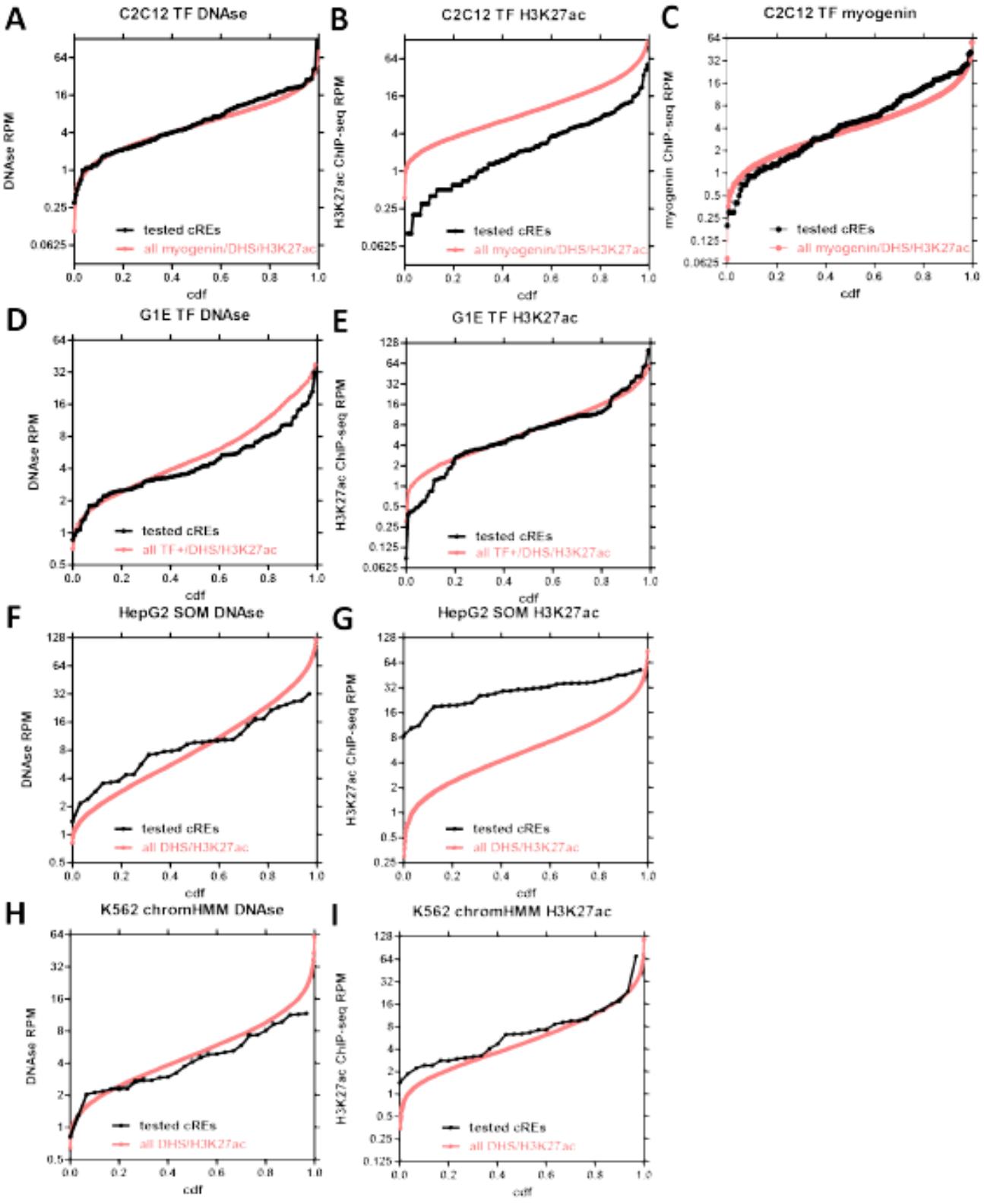
Figure 5: Absence of general strong correlation between biochemical signal strength and enhancer activity of cEnhs. (A) Distribution of tested cEnhs relative to the genome-wide DNAse and H3K27Ac signal distribution in C2C12 myocytes. Shown are DNAse and H3K27ac RPM values for all DNAse$^+$/H3K27ac$^+$ regions as well as for cEns tested for activity in C2C12 myocytes (outlined circles) and for occupancy negative control (outlined squares), with tested cEnhs separated into four classes based on their measured enhancer activity, from dark red (most active) to yellow (inactive). (B) Distribution of tested cEnhs relative to the genome-wide DNAse and H3K27Ac signal distribution in G1E-ER4 cells. Shown are DNAse and H3K27ac RPM values for all DNAse$^+$/H3K27ac$^+$ regions as well as for cEns tested for activity (outlined circles) and for occupancy negative control (outlined squares), with tested cEnhs separated into four classes based on their measured enhancer activity, from dark red (most active) to yellow (inactive). (C) Correlation between biochemical signals and measured enhancer activity in C2C12 and G1E cells. See also Supplementary Figures 4, 8, and 12 for more details.



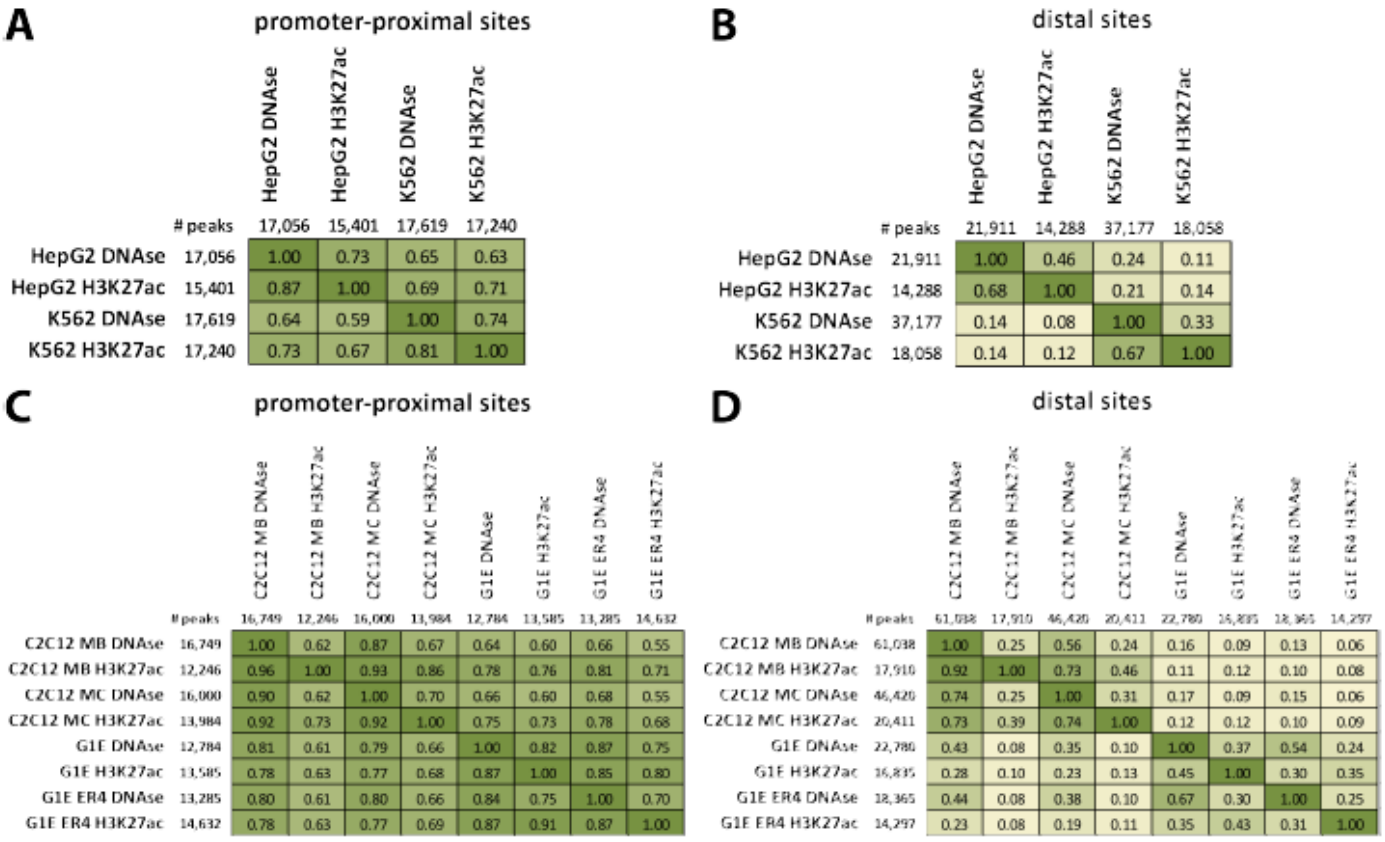
Supplementary Figure 1: The length of thousands of conserved noncoding elements in mammalian genomes greatly exceeds the size range of MPRA constructs. (A) The length distribution of conserved noncoding regions in the human genome. The `phastCons100way` conservation track for the `hg20` version of the human genome was downloaded from the UCSC Genome Browser. Blocks of conservation, in which all nucleotides have `phastCons` scores higher than the indicated minimum (`phCons`), were identified, and then merged into larger regions if the length of the gaps between them was smaller than the indicated `maxGap` parameter. The distribution of the lengths of the resulting sets of regions was plotted. This approach captures the properties of enhancer elements observed in the genome, which often consist of multiple blocks of highly conserved sequences separated by gaps of less conserved sequences, resulting in an enhancer element of up to a few hundred base pairs in length or more. (B) Such an example is shown for the *Acta1* gene in mouse.



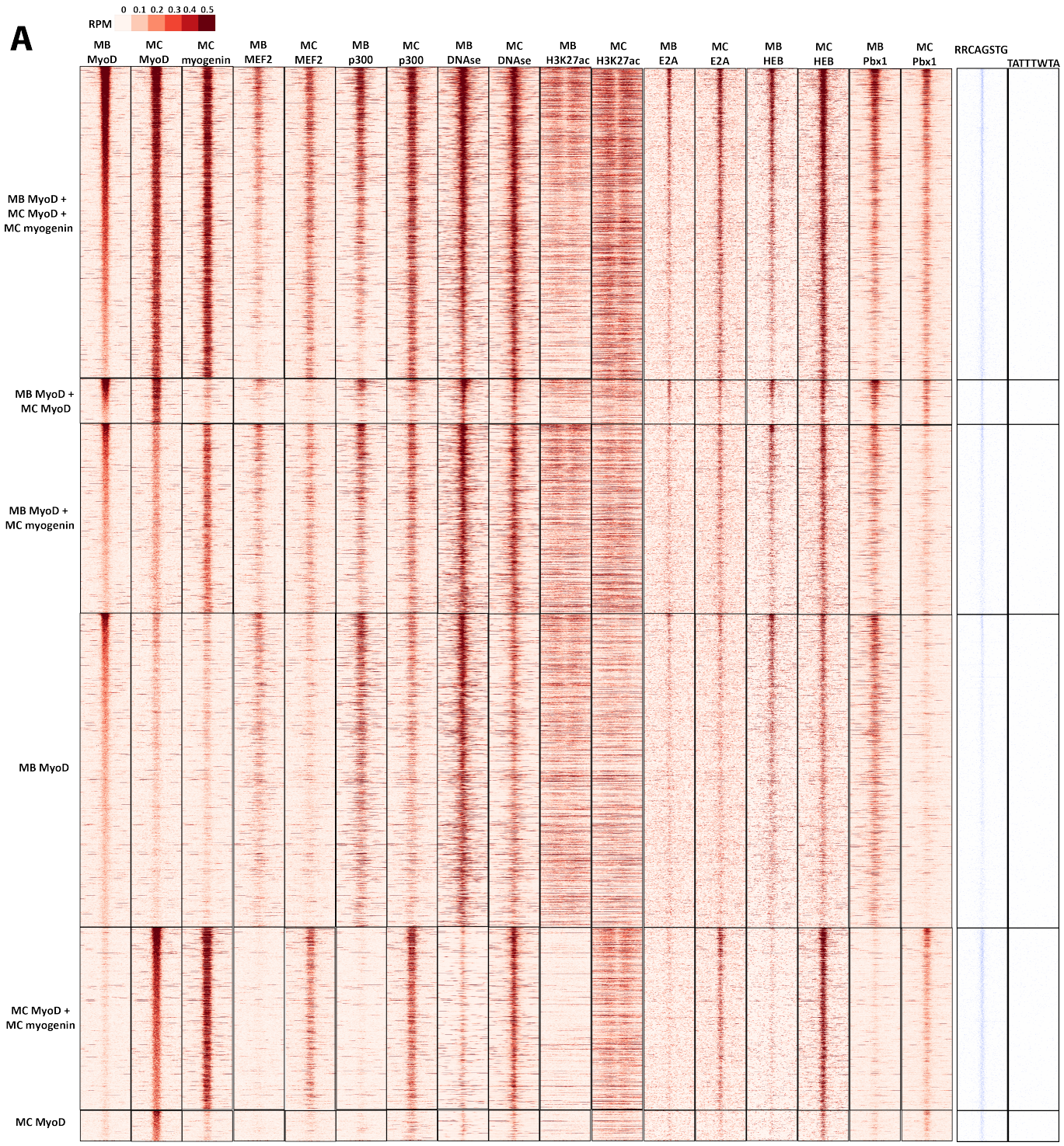
Supplementary Figure 2: Length distribution of functional assays constructs used to test cREs in this study. (A) Distribution of functional assay construct lengths tested in this study in C2C12 cells. (B) Distribution of functional assay construct lengths tested in this study in G1E cells. (C) Distribution of functional assay construct lengths tested in this study in K562 and HepG2 cells.

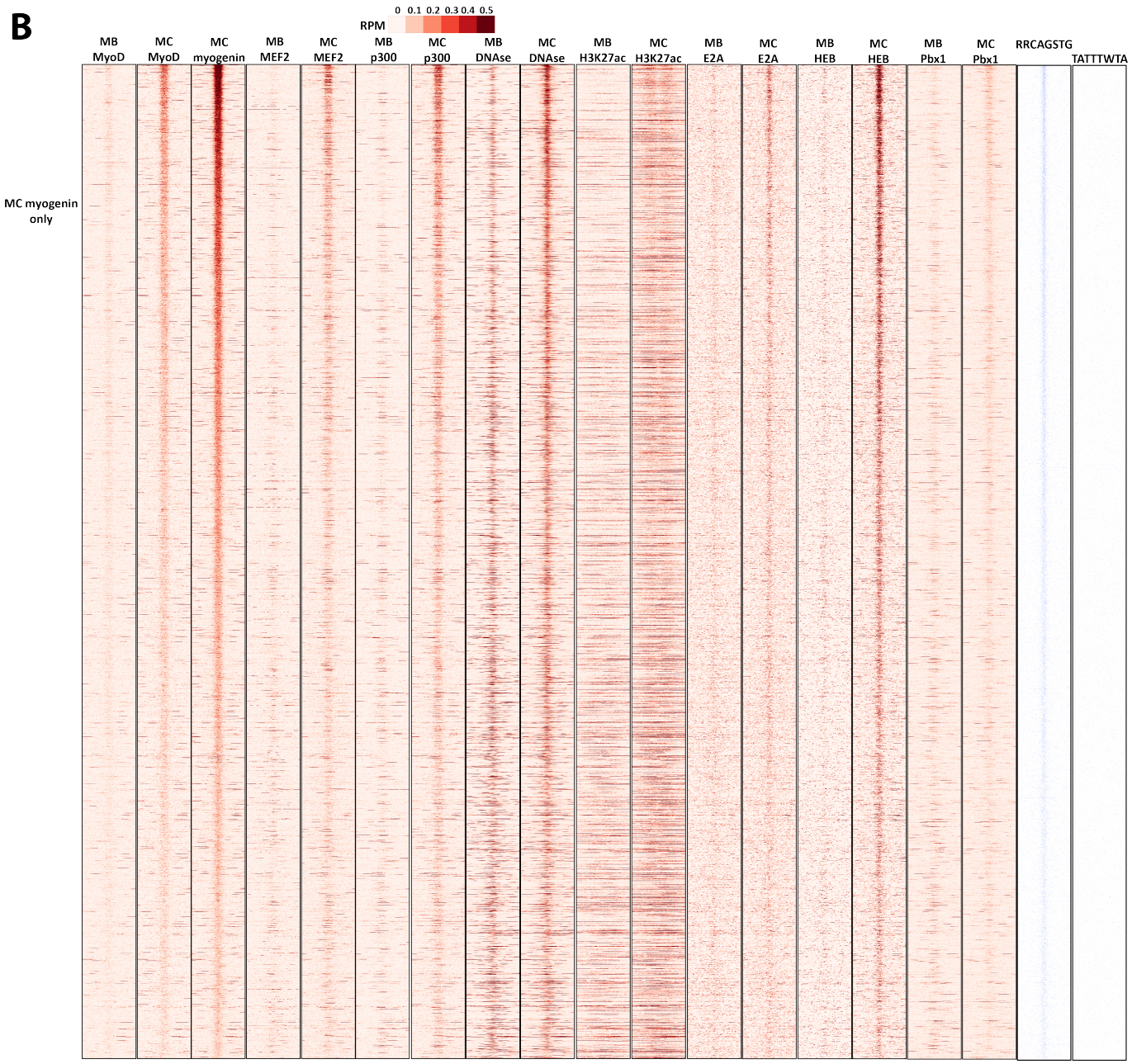


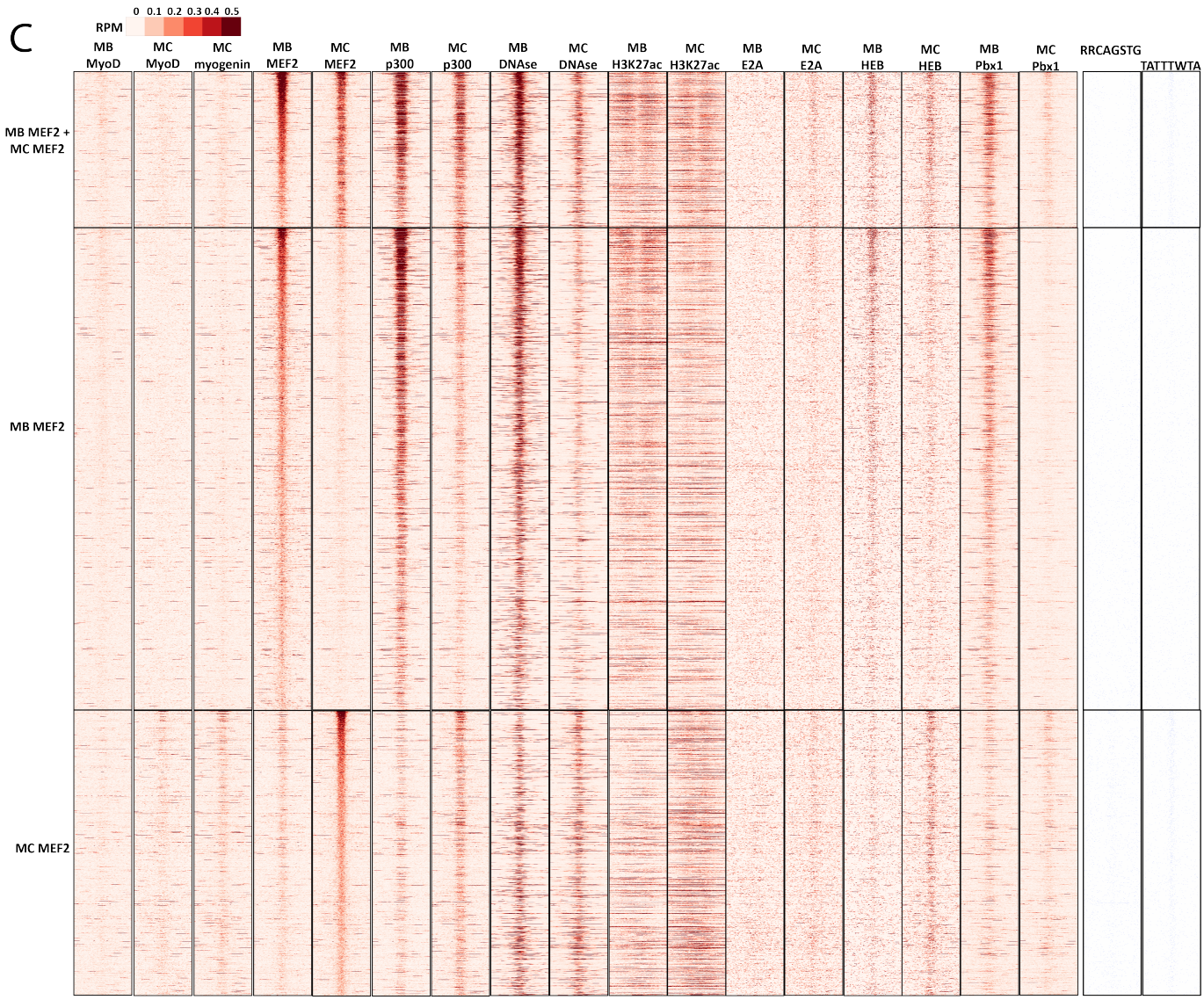
Supplementary Figure 3: Distribution of biochemical signal in tested cEnhs and genome-wide. Shown is the distribution of ChIP-seq or DNAse-seq RPM values for the set of cEnhs tested and for the genome-wide set of cEnh with similar biochemical signatures shown in Figure 3.



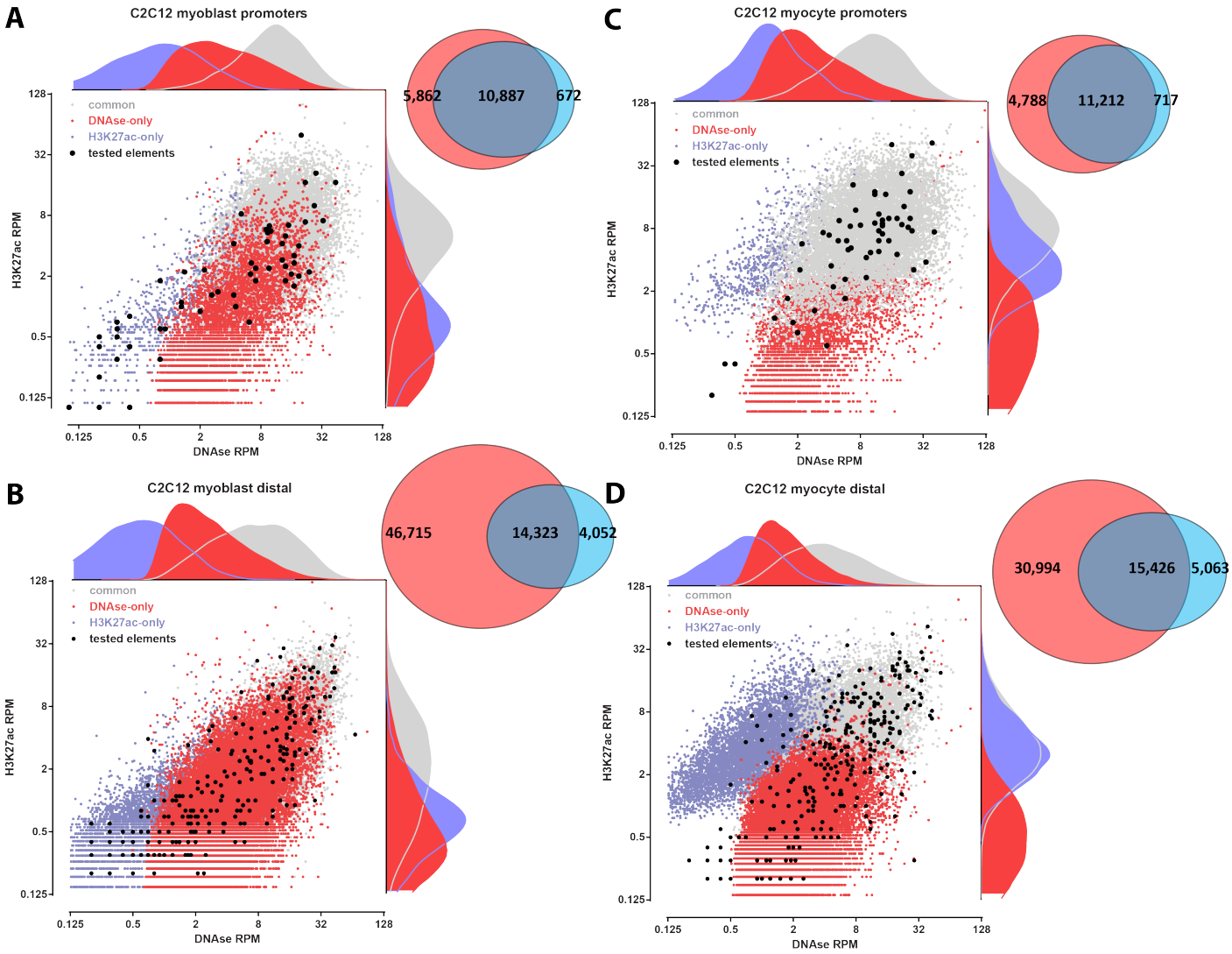
Supplementary Figure 4: Differential marking of proximal and distal cREs by DNAse and H3K27ac between different cell types and cell states. (A) Promoter-proximal (within ≤1 kb of an annotated TSS) sites in K562 and HepG2 cells; (A) Distal (≥1 kb from an annotated TSS) sites in K562 and HepG2 cells; (C) Promoter-proximal (within ≤1 kb of an annotated TSS) sites in differentiated and undifferentiated C2C12 and G1E cells; (D) Distal (≥1 kb from an annotated TSS) sites in differentiated and undifferentiated C2C12 and G1E cells. The overlap score (O_{xy}) shown in each cell (x, y) indicates the fraction of peaks in the dataset on the y-axis that are also found in the dataset on the x-axis, i.e. $O_{xy} = |X \cap Y|/|Y|$.


A
RPM
0 0.1 0.2 0.3 0.4 0.5
MB MyoD
MC MyoD
MC myogenin
MB MEF2
MC MEF2
MB p300
MC p300
MB DNAse
MC DNAse
MB H3K27ac
MC H3K27ac
MB E2A
MC E2A
MB HEB
MC HEB
MB Pbx1
MC Pbx1
RRCAGSTG
TATTTWTA
MB MyoD + MC MyoD + MC myogenin
MB MyoD + MC MyoD
MB MyoD + MC myogenin
MB MyoD
MC MyoD + MC myogenin
MC MyoD


B
RPM
0 0.1 0.2 0.3 0.4 0.5
MB MyoD
MC MyoD
MC myogenin
MB MEF2
MC MEF2
MB p300
MC p300
MB DNAse
MC DNAse
MB H3K27ac
MC H3K27ac
MB E2A
MC E2A
MB HEB
MC HEB
MB Pbx1
MC Pbx1
RRCAGSTG
TATTTWTA
MC myogenin only



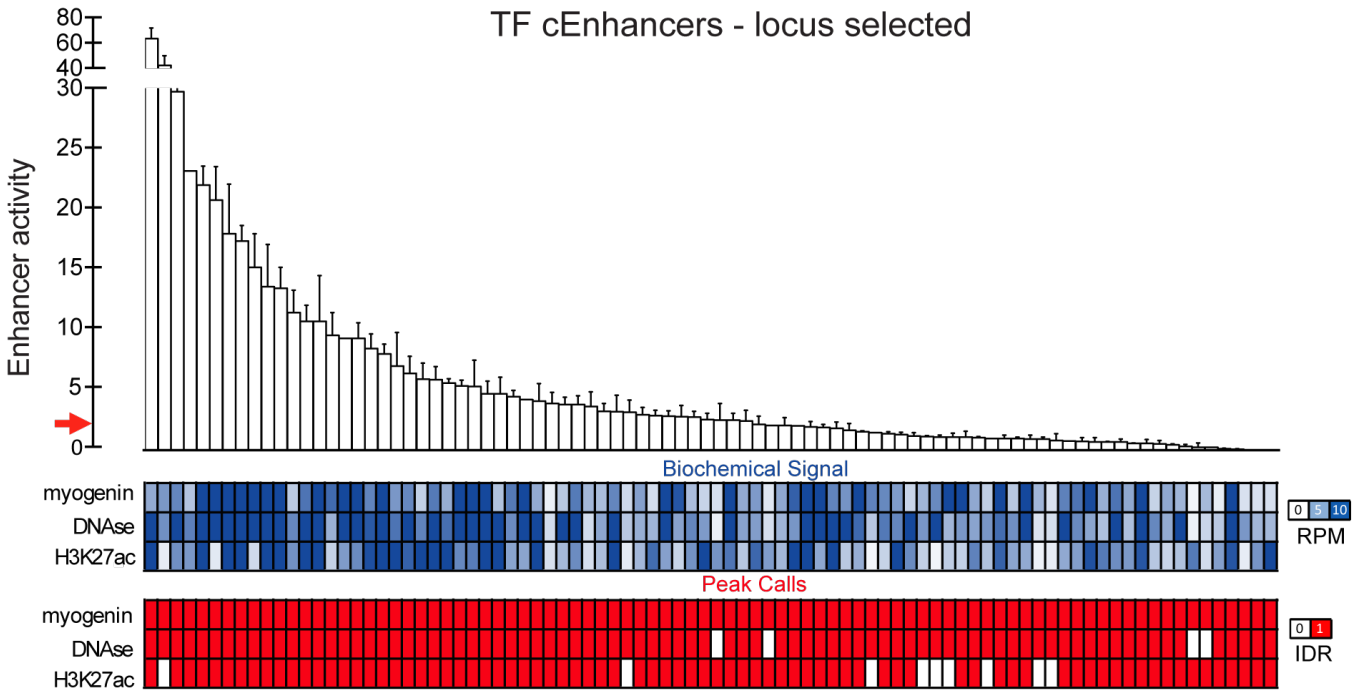
Supplementary Figure 5: Regulatory landscape of muscle differentiation. DNAse-seq and ChIP-seq experiments against H3K27ac, p300, the MRFs MyoD and myogenin, and cofactors (MEF2, E2A/TCF3, HEB/TCF12, and Pbx1) in undifferentiated (myoblast, or "MB") and differentiated (myocyte, or "MC") C2C12 cells were analyzed. Sites were split into multiple subgroups depending on regulatory factor occupancy (at IDR=0.05) – MyoD-positive (in either condition) sites (A), myogenin-only sites (B), and MEF2-only sites (C) – then sorted by MRF ChIP-seq signal (in the following order of priority: myoblast MyoD, myocyte MyoD, myocyte myogein, myoblast MEF2, myocyte MEF2); the signal in the 500bp-radius region around the ChIP-seq peak position is shown.



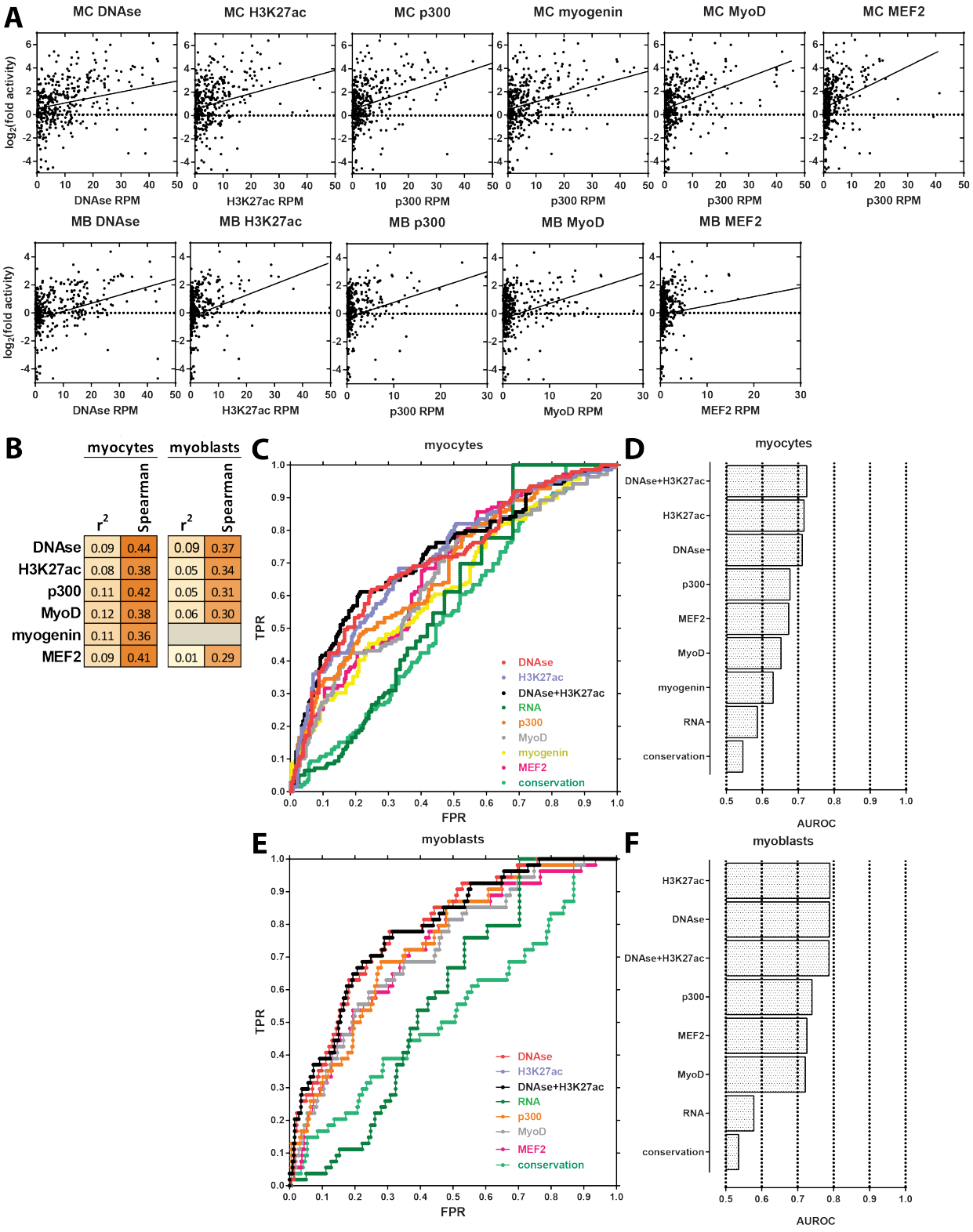
E

	promoter DNAse +			distal DNAse +		
	common	MB only	MC only	common	MB only	MC only
	14,789	1,960	1,697	36,163	24,875	13,791
MB H3K27ac +	66.98%	26.17%	12.91%	33.30%	16.92%	2.46%
MB H3K27ac -	33.02%	73.83%	87.09%	66.70%	83.08%	97.54%
MC H3K27ac +	72.91%	25.51%	35.65%	36.14%	11.25%	21.21%
MC H3K27ac -	27.09%	74.49%	64.35%	63.86%	88.75%	78.79%
	promoter H3K27ac +			distal H3K27ac +		
	common	MB only	MC only	common	MB only	MC only
	10,568	1,678	3,665	8,667	9,243	12,266
MB DNAse +	98.20%	92.25%	81.17%	97.17%	89.69%	58.24%
MB DNAse -	1.80%	7.75%	18.83%	2.83%	10.31%	41.76%
MC DNAse +	97.41%	83.37%	87.15%	90.24%	61.02%	67.46%
MC DNAse -	2.59%	16.63%	12.85%	9.76%	38.98%	32.54%

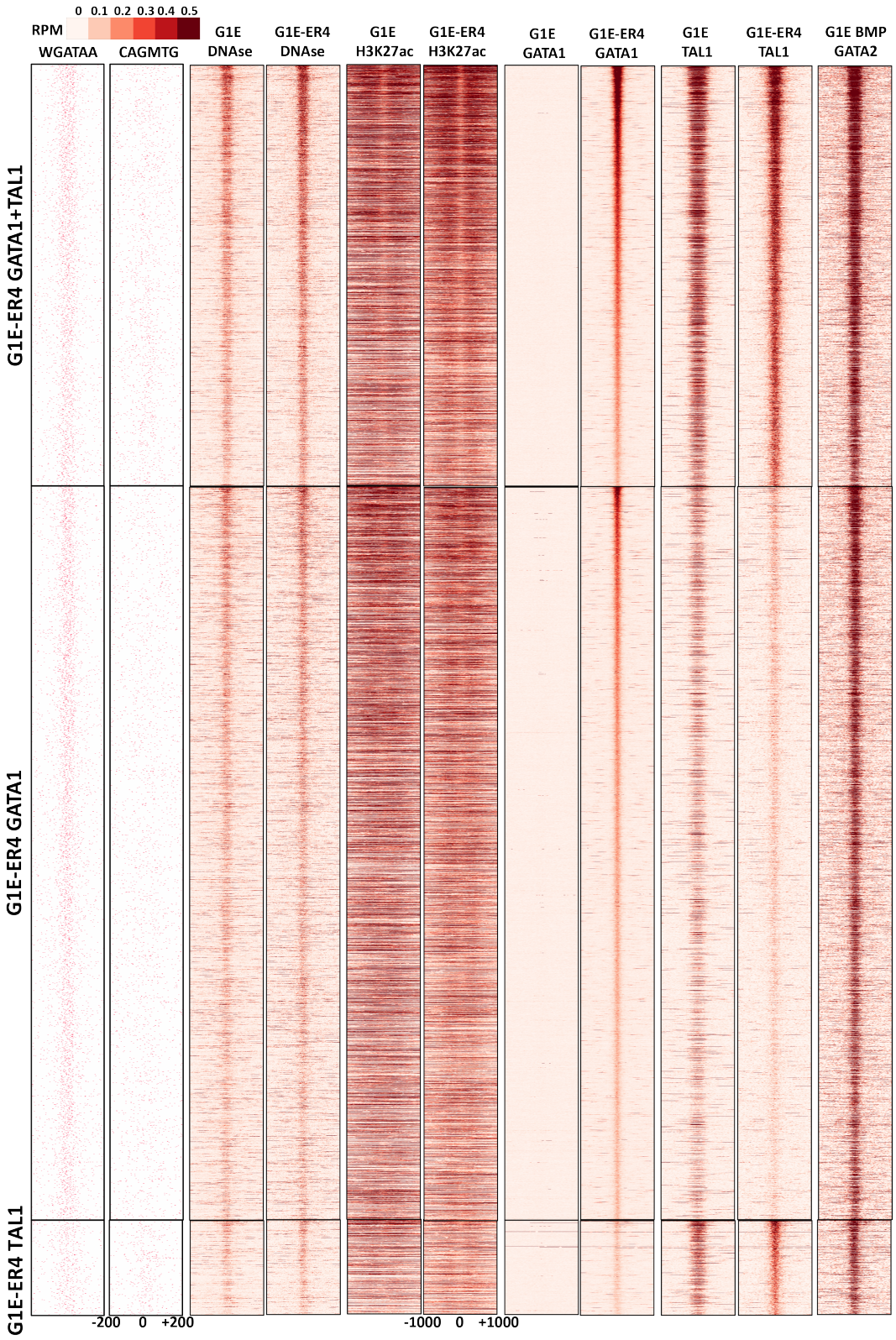
Supplementary Figure 6: Relationship between DNAse hypersensitivity and H3K27 acetylation during muscle differentiation. (A) Overlap between DNAse hypersensitive and H3K27ac-positive promoter-proximal regions in C2C12 myoblasts; (B) Overlap between DNAse hypersensitive and H3K27ac-positive promoter-proximal regions in C2C12 myocytes; (C) Overlap between DNAse hypersensitive and H3K27ac-positive distal regions in C2C12 myoblasts; (D) Overlap between DNAse hypersensitive and H3K27ac-positive distal regions in C2C12 myocytes; the kernel density of the ChIP-seq/DNAse-seq signal distribution for each class of sites is overlaid over the scatter plots, and the distribution of tested cREs is shown in black; (E) Dynamic changes in DNAse hypersensitivity and H3K27 acetylation upon differentiation for promoter-proximal and distal sites.



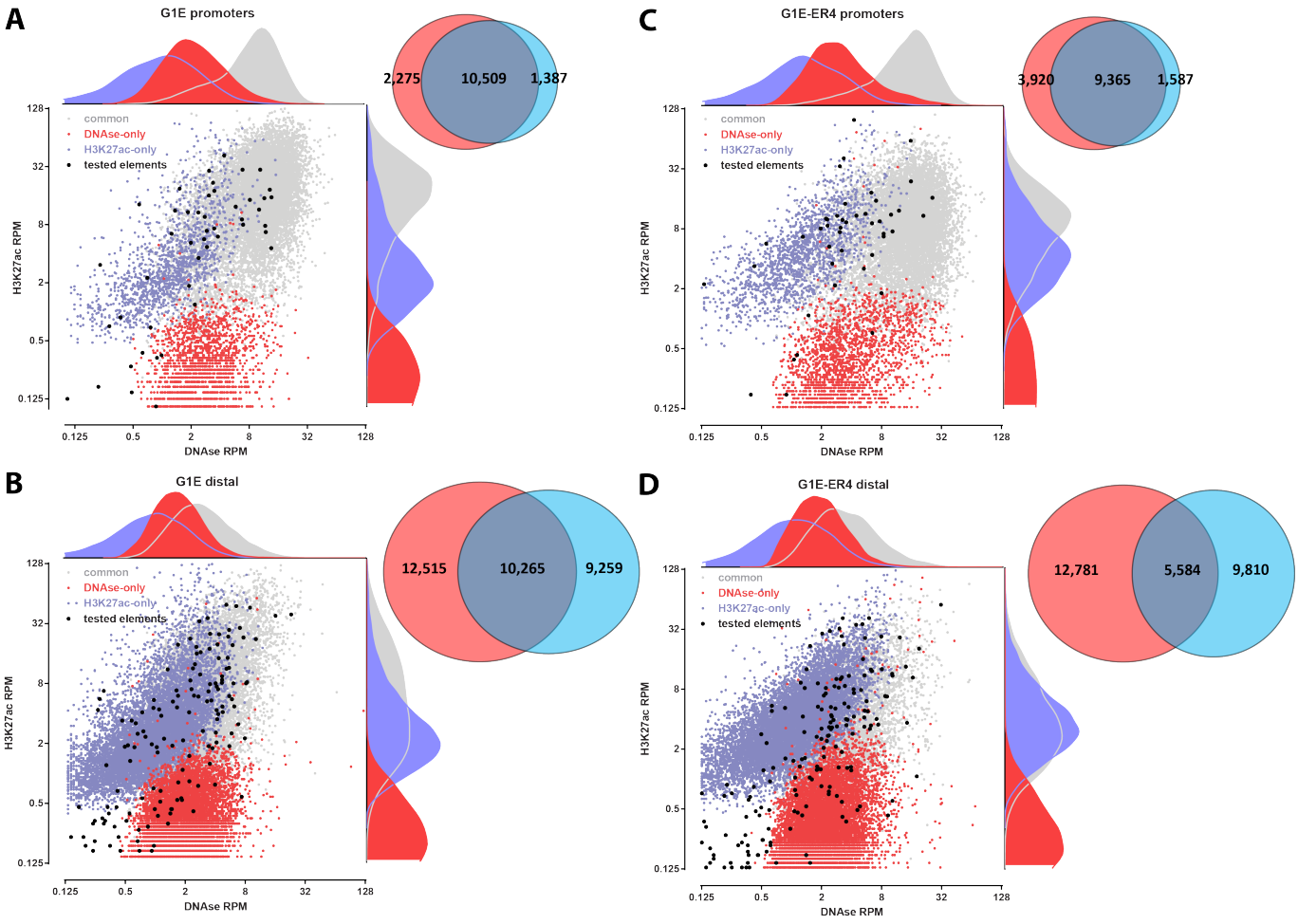
Supplementary Figure 7: Functional assay testing of cREs selected for their physical proximity to genes known for their importance to muscle development. Fold activity in myocytes across biological replicates ($n = 4$) and technical replicates ($n = 4$ for each biological replicate) is shown. Candidate REs were sorted by their mean fold activity. The red arrow corresponds to the mean fold activity threshold above which elements are considered active. In addition, DNAse hypersensitivity, H3K27ac status, p300, MyoD and myogenin occupancy are shown for each cRE, both as binary (IDR=0.05) calls (red coloring indicates occupancy), and as RPM scores (blue heatmap).



Supplementary Figure 8: Correlation between regulatory activity and biochemical marks in C2C12 cells. (A and B) Correlation between fold activity and DNAse hypersensitivity, H3K27ac, p300, myogenin, MyoD and MEF2 occupancy in myoblasts and myocytes; (C) ROC curves showing biochemical mark predictivity of cRE fold activity in myocytes; (D) AUROC (area under ROC curve) values for different biochemical marks in myocytes; (E) ROC curves showing biochemical mark predictivity of cRE fold activity in myoblasts; (F) AUROC values for different biochemical marks in myoblasts.



Supplementary Figure 9: Regulatory landscape of eryhtroid differentiation. DNAse-seq and ChIP-seq experiments against H3K27ac, GATA1, TAL1 and GATA2 G1E and G1E-ER4 were analyzed. Sites were split into subgroups depending on GATA1 and TAL1 occupancy (IDR=0.05), then sorted by ChIP-seq signal (in the following order of priority: G1E-ER4 GATA1, G1E-ER4 TAL1); the signal in the 500bp-radius region around the ChIP-seq peak position is shown.

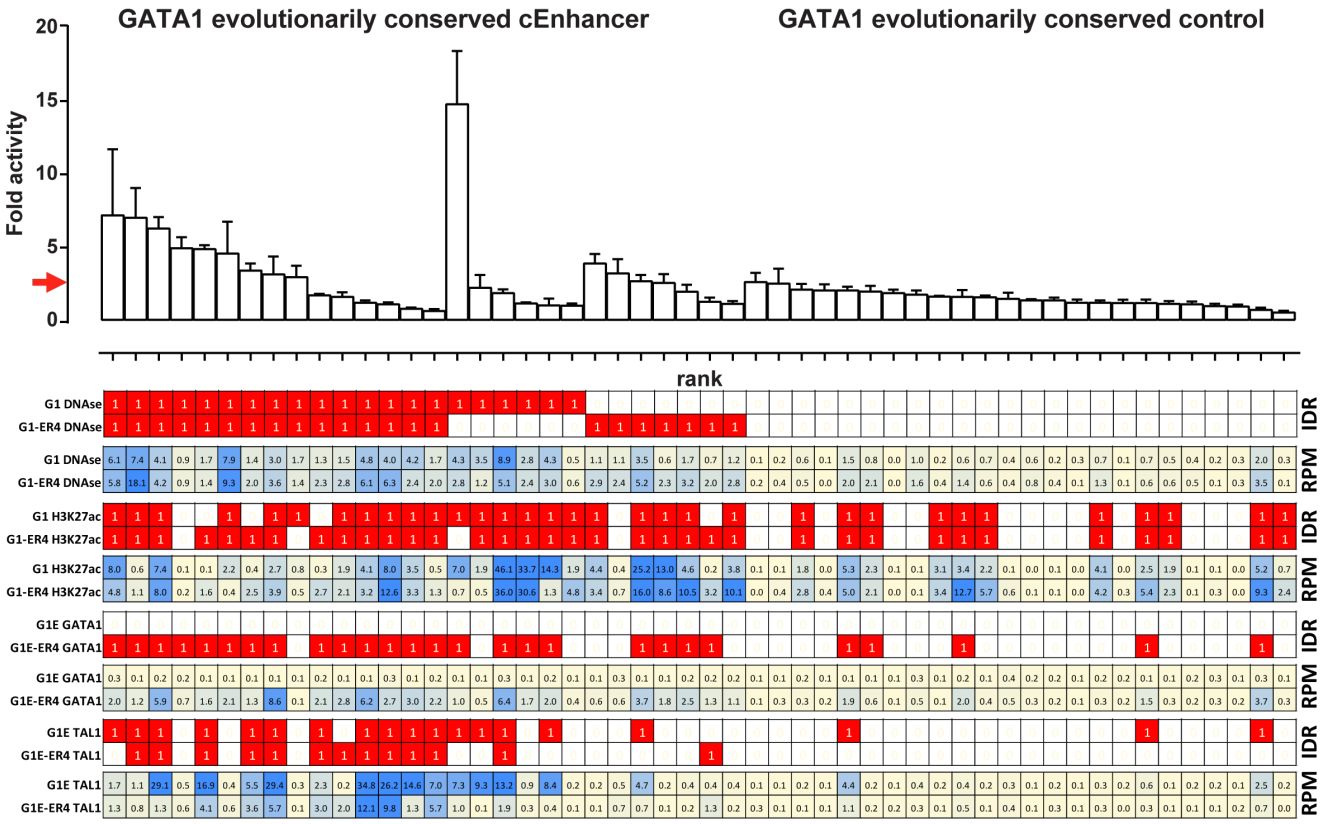


E

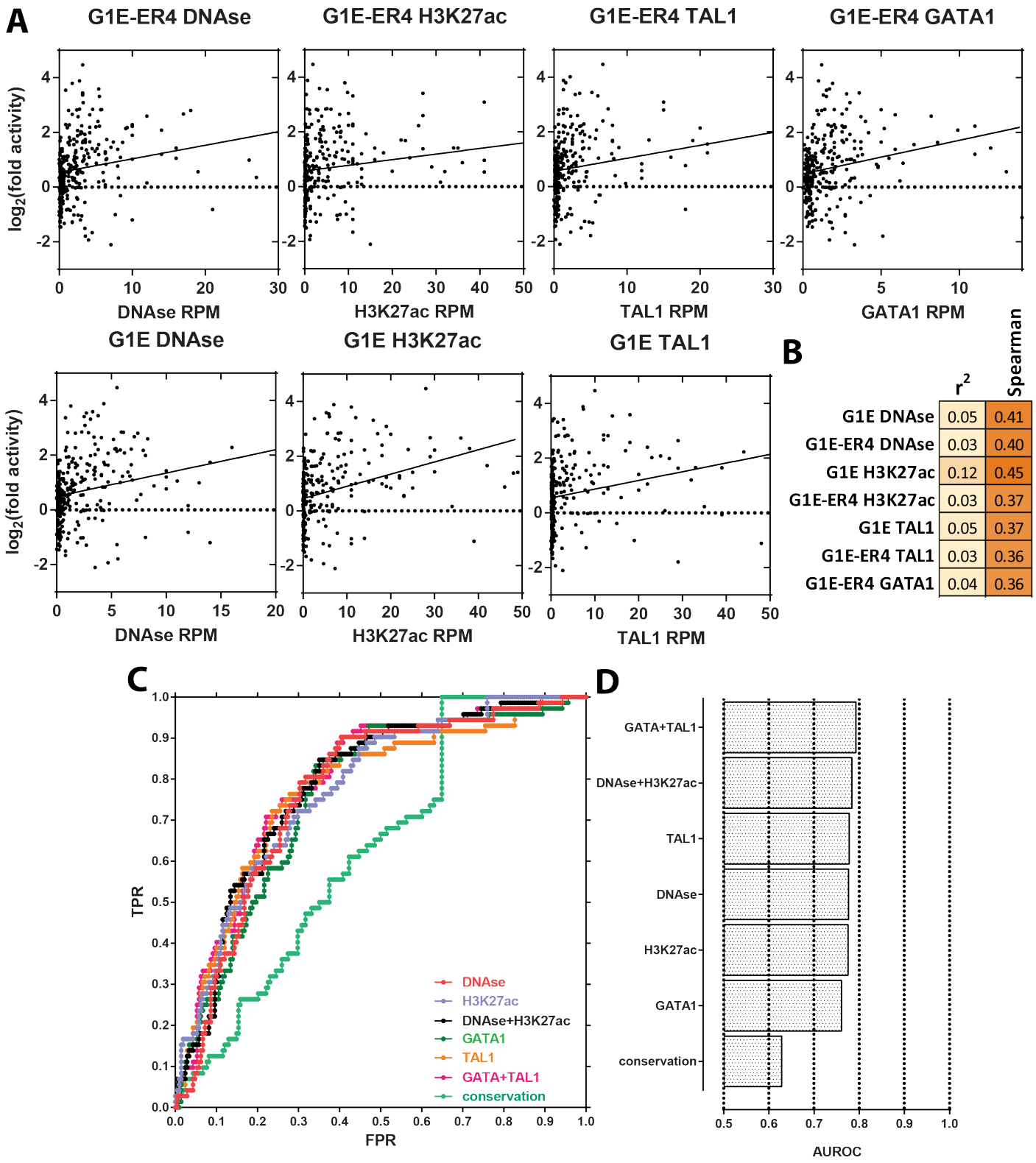
	promoter DNAse +			distal DNAse		
	common	G1E only	G1E-ER4 only	common	G1E only	G1E-ER4 only
	14,789	1,960	1,697	36,163	24,875	13,791
G1E H3K27ac +	86.43%	51.95%	13.33%	44.17%	46.10%	20.02%
G1E H3K27ac -	13.57%	48.05%	86.67%	55.83%	53.90%	79.98%
G1E-ER4 H3K27ac +	80.73%	31.84%	13.63%	33.52%	23.42%	23.65%
G1E-ER4 H3K27ac -	19.27%	68.16%	86.37%	66.48%	76.58%	76.35%

	promoter H3K27ac			distal H3K27ac		
	common	G1E only	G1E-ER4 only	common	G1E only	G1E-ER4 only
	10,568	1,678	3,665	8,667	9,243	12,266
G1E DNAse +	95.78%	63.91%	22.00%	59.51%	35.88%	9.68%
G1E DNAse -	4.22%	36.09%	78.00%	40.49%	64.12%	90.32%
G1E-ER4 DNAse +	94.91%	54.64%	26.98%	48.58%	18.93%	13.64%
G1E-ER4 DNAse -	5.09%	45.36%	73.02%	51.42%	81.07%	86.36%

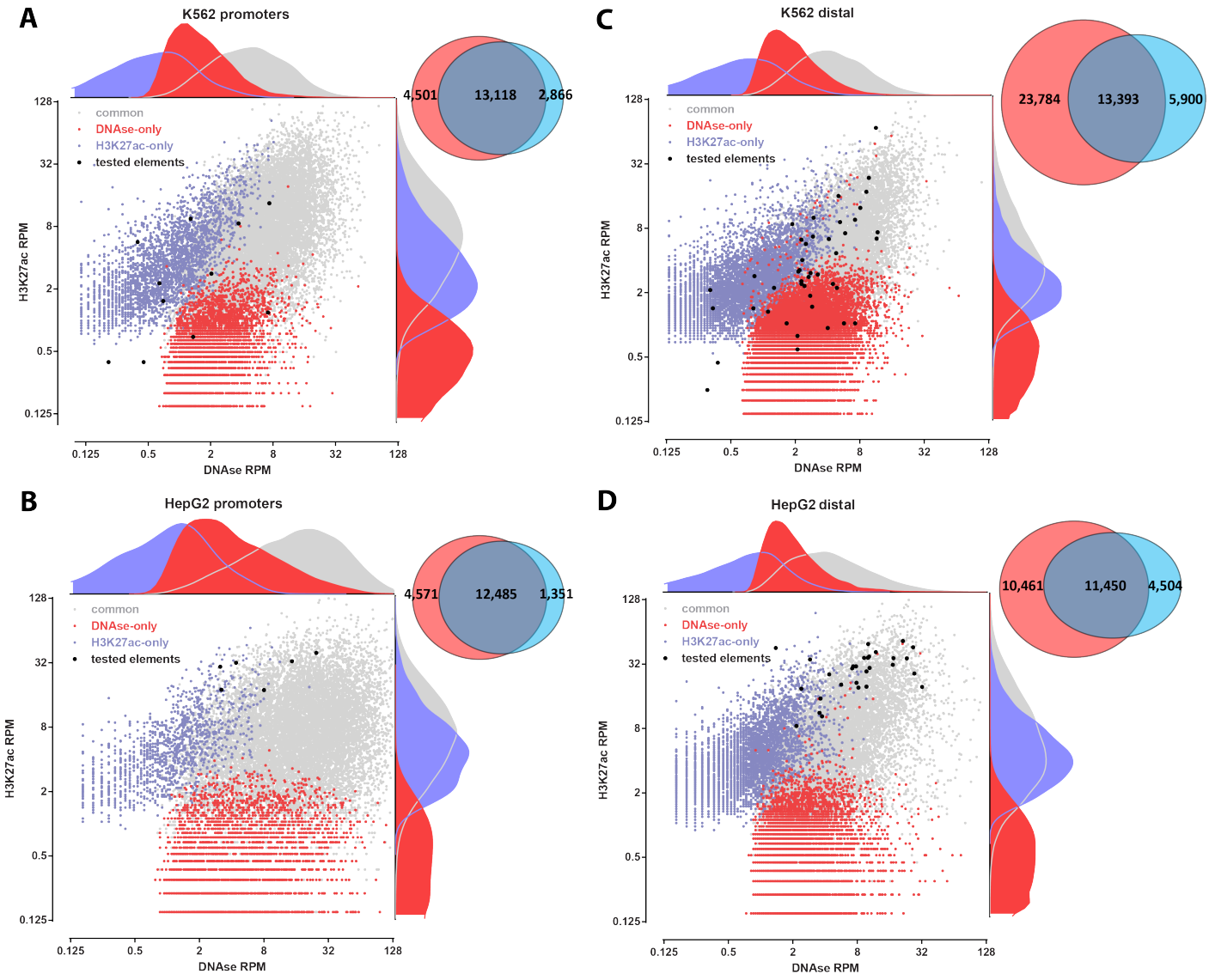
Supplementary Figure 10: Relationship between DNAse hypersensitivity and H3K27 acetylation during erythroid differentiation. (A) Overlap between DNAse hypersensitive and H3K27ac-positive promoter-proximal regions in G1E cells; (B) Overlap between DNAse hypersensitive and H3K27ac-positive promoter-proximal regions in G1E-ER4 cells; (C) Overlap between DNAse hypersensitive and H3K27ac-positive distal regions in G1E cells; (D) Overlap between DNAse hypersensitive and H3K27ac-positive distal regions in G1E-ER4 cells; the kernel density of the ChIP-seq/DNAse-seq signal distribution for each class of sites is overlaid over the scatter plots, and the distribution of tested cREs is shown in black; (E) Dynamic changes in DNAse hypersensitivity and H3K27 acetylation upon differentiation for promoter-proximal and distal sites.



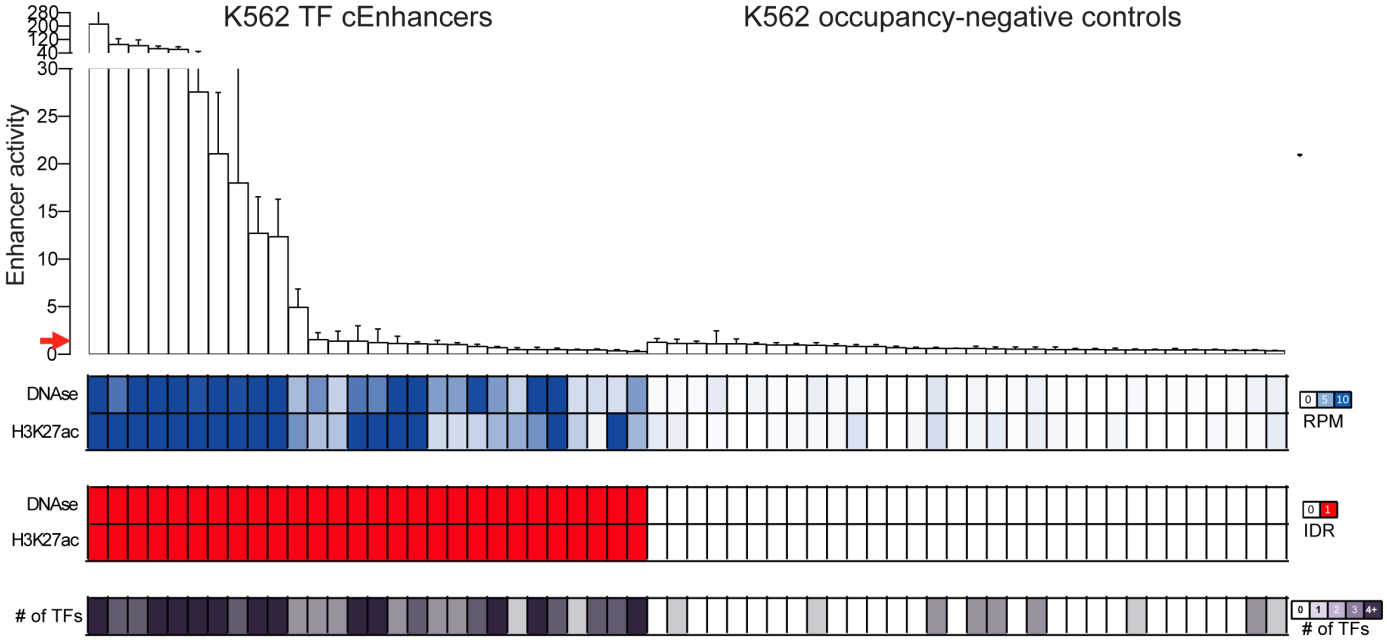
Supplementary Figure 11: Functional assay testing of the regulatory activity of erythroid cREs. Fold activity in K562 cells across biological replicates ($n \in [1, 9]$) and technical replicates ($n = 4$ for each biological replicate) is shown. Candidate REs were sorted first by their DNAse status and then by their mean fold activity. The horizontal dotted line corresponds to the mean fold activity threshold above which elements are considered active. In addition, DNAse hypersensitivity, H3K27ac status, GATA1, and TAL1 occupancy are shown for each cRE, both as binary (IDR=0.05) calls (red coloring indicates occupancy), and as RPM scores. (A) cREs randomly selected from the genome-wide set of GATA1/TAL1-occupied regions; (B) cREs selected among the set of highly evolutionarily constrained non-coding elements that contain a GATA1 motif ("regulatory potential selections").



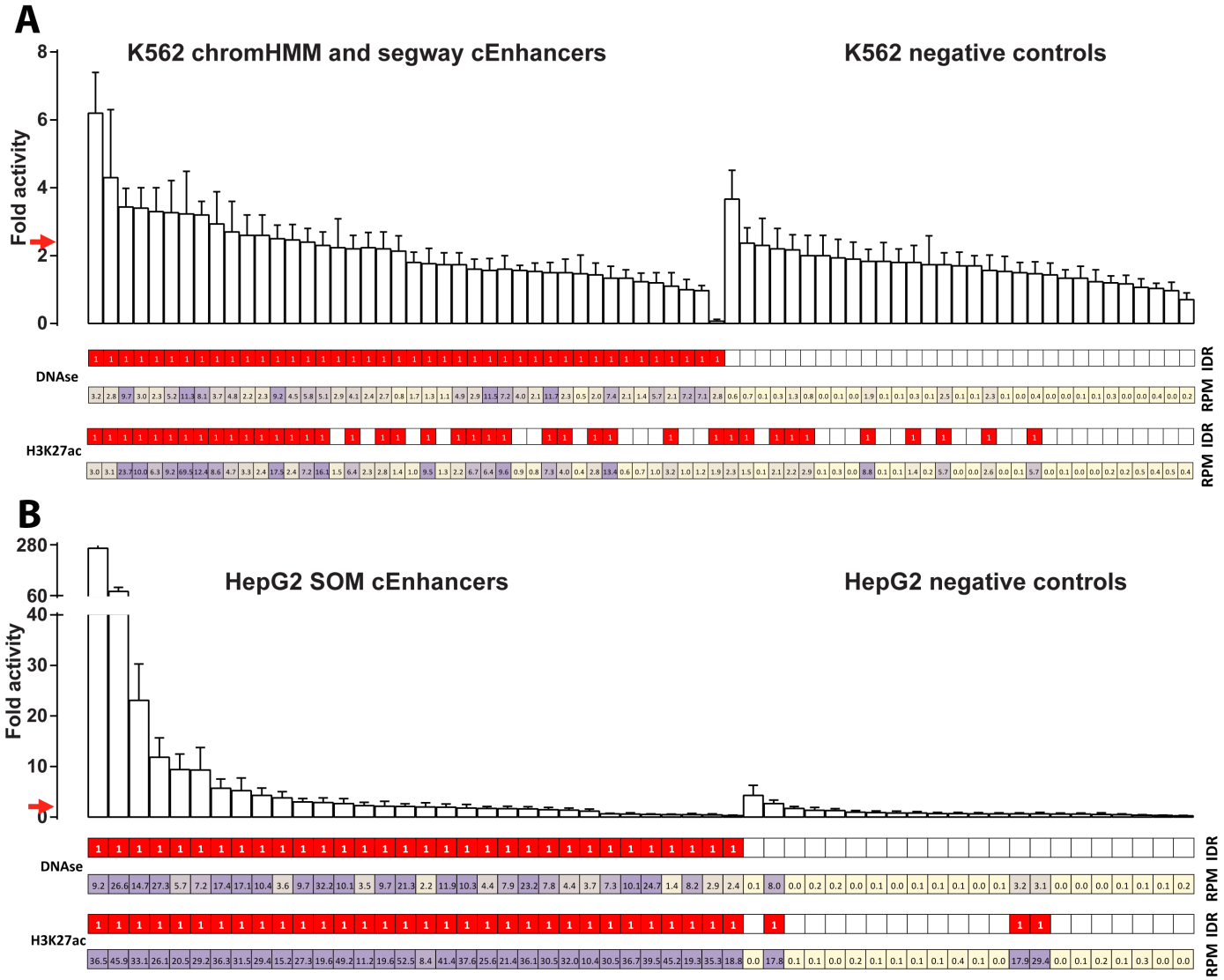
Supplementary Figure 12: Correlation between regulatory activity and biochemical marks in erythroid cells. (A and B) Correlation between fold activity in K562 cells and DNAse hypersensitivity, H3K27ac, TAL1, and GATA1 occupancy in G1E and G1E-ER4 cells; (C) ROC curves showing biochemical mark predictivity of cRE fold activity; (D) AUROC (area under ROC curve) values for different biochemical marks.



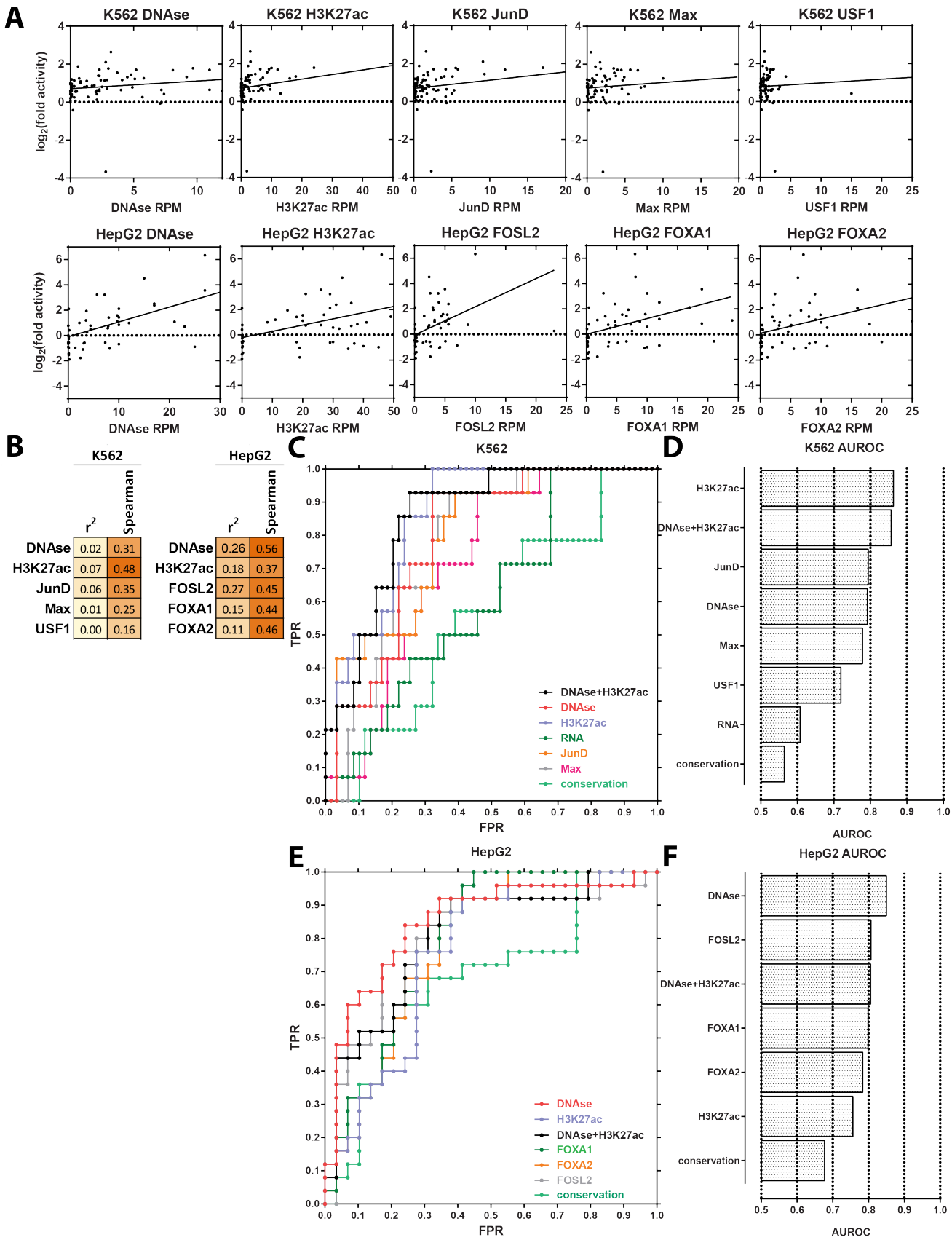
Supplementary Figure 13: Relationship between DNAse hypersensitivity and H3K27 acetylation in immortalized human cell lines. (A) Overlap between DNAse hypersensitive and H3K27ac-positive promoter-proximal regions in K562 cells; (B) Overlap between DNAse hypersensitive and H3K27ac-positive distal regions in K562 cells; (C) Overlap between DNAse hypersensitive and H3K27ac-positive promoter-proximal regions in HepG2 cells; (D) Overlap between DNAse hypersensitive and H3K27ac-positive distal regions in HepG2 cells; the kernel density of the ChIP-seq/DNAse-seq signal distribution for each class of sites is overlaid over the scatter plots, and the distribution of tested cREs is shown in black.



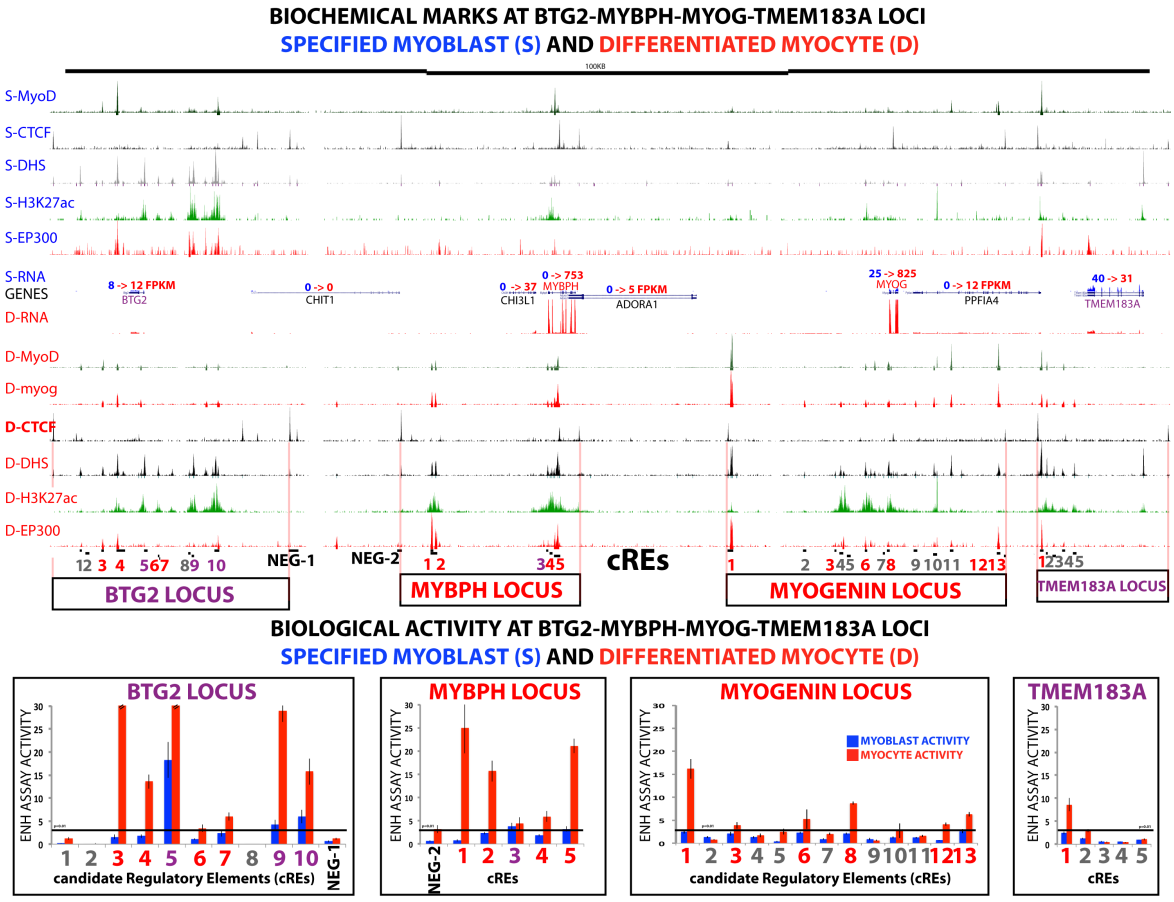
Supplementary Figure 14: Functional assay testing of TF selected cEnhs in human immortalized cell lines. Fold activity across biological replicates (n =?? ???) and technical replicates (n =?? ???) for each biological replicate) is shown. Candidate REs were sorted first by their DNAse status and then by their mean fold activity. The horizontal dotted line corresponds to the mean fold activity threshold above which elements are considered active. In addition, DNAse hypersensitivity and H3K27ac status are shown for each cRE, both as binary (IDR=0.05) calls (red coloring indicates occupancy), and as RPM scores. (A) cREs tested in K562 cells (B) cREs tested in HepG2 cells.



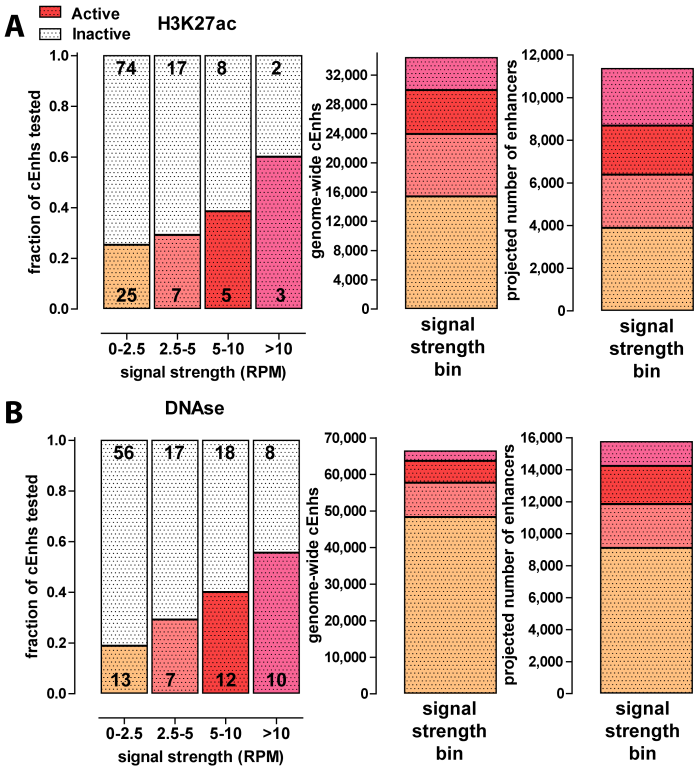
Supplementary Figure 15: Functional assay testing of machine learning selected cEnhs in human immortalized cell lines. Fold activity across biological replicates (n =? ???) and technical replicates (n =?? ??? for each biological replicate) is shown. Candidate REs were sorted first by their DNAse status and then by their mean fold activity. The horizontal dotted line corresponds to the mean fold activity threshold above which elements are considered active. In addition, DNAse hypersensitivity and H3K27ac status are shown for each cRE, both as binary (IDR=0.05) calls (red coloring indicates occupancy), and as RPM scores. (A) cREs tested in K562 cells (B) cREs tested in HepG2 cells.



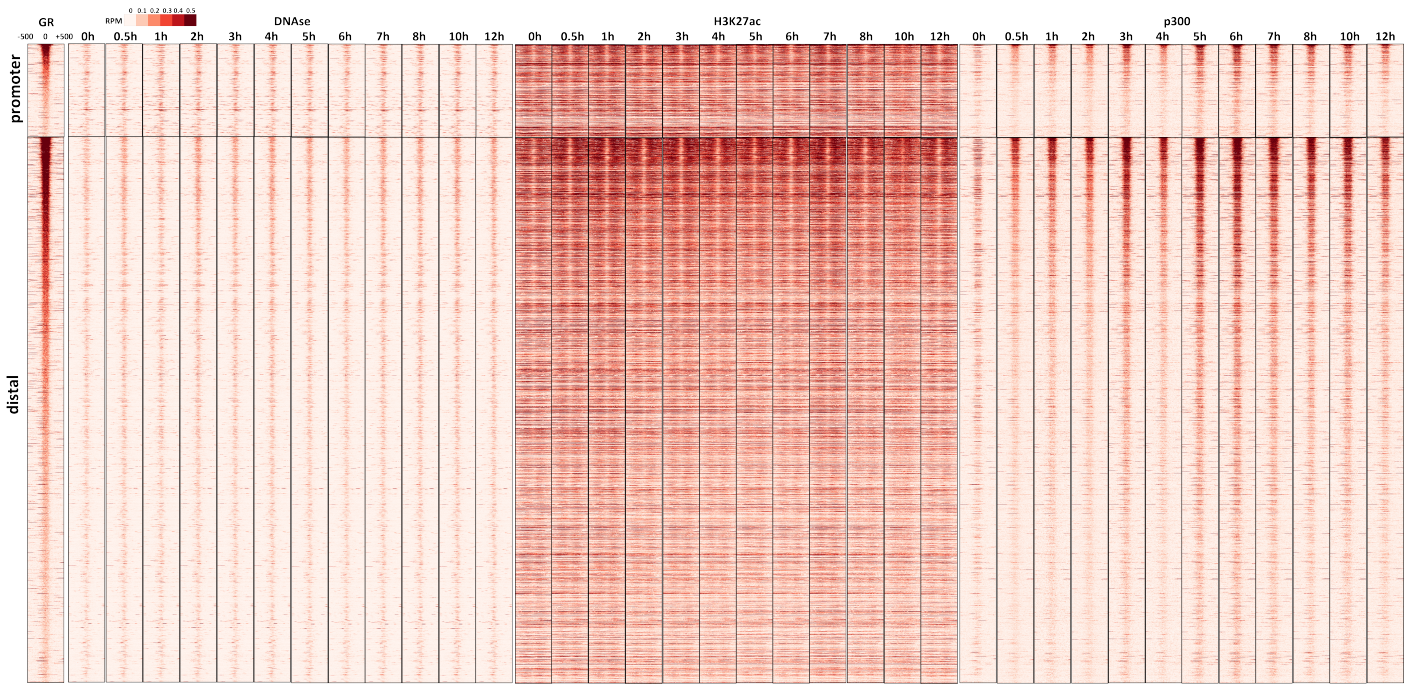
Supplementary Figure 16: Correlation between regulatory activity and biochemical marks in human immortalized cell lines. (A and B) Correlation between fold activity in K562 cells and DNAse hypersensitivity, and transcription factor occupancy in K562 and HepG2 cells; (C) ROC curves showing biochemical mark predictivity of cRE fold activity in K562 cells; (D) AUROC (area under ROC curve) values for different biochemical marks in K562 cells; (E) ROC curves showing biochemical mark predictivity of cRE fold activity in K562 cells; (F) AUROC (area under ROC curve) values for different biochemical marks in K562 cells.



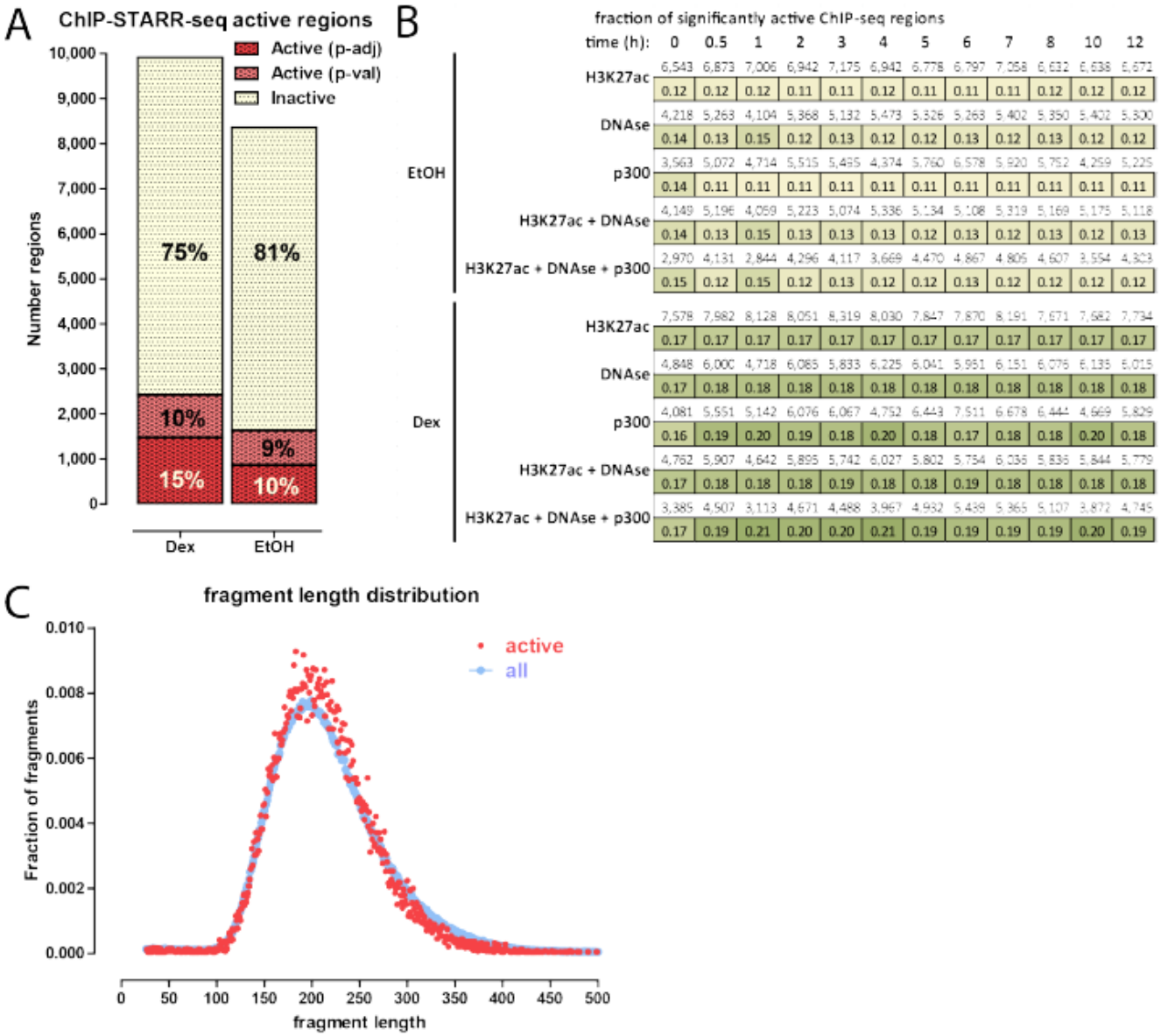
Supplementary Figure 17: CAPTION. (A) CAPTION GOES HERE



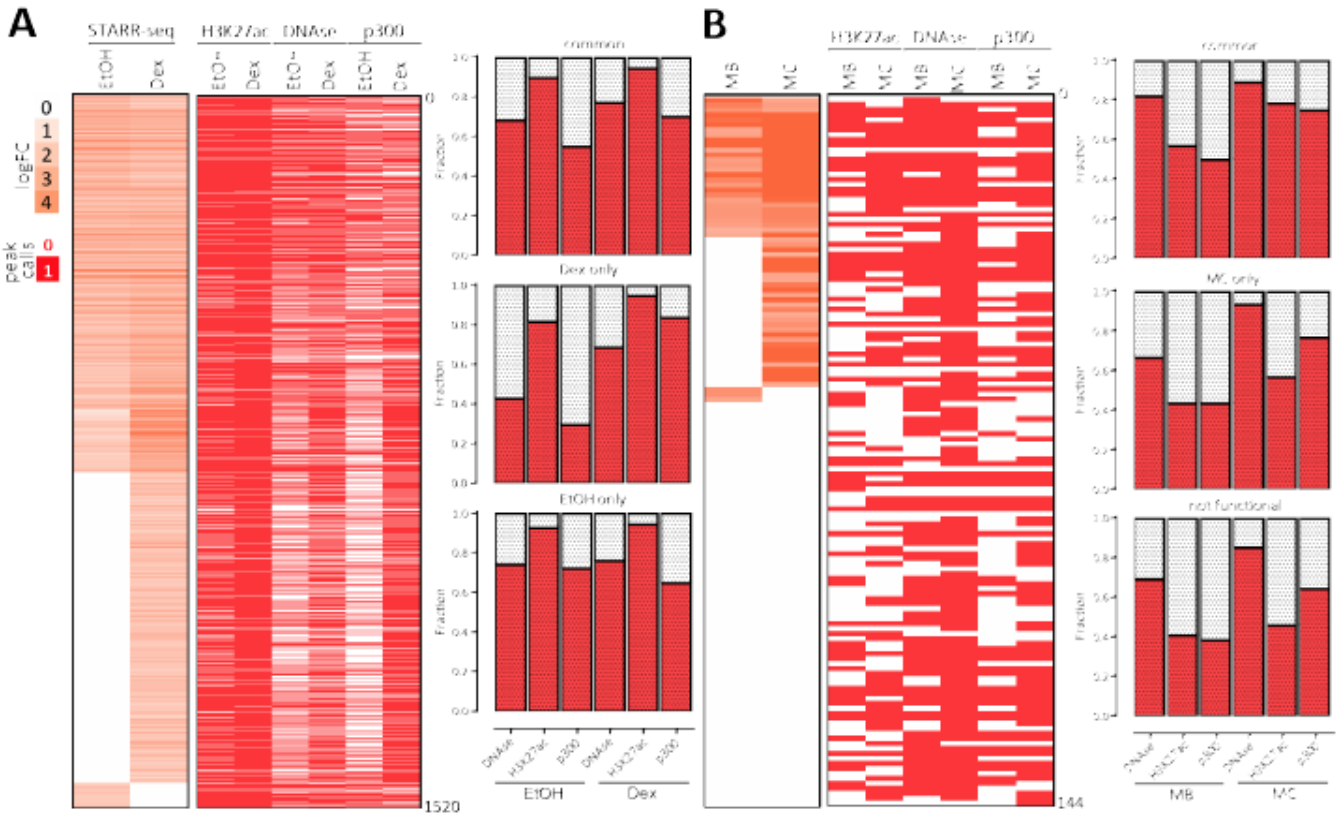
Supplementary Figure 18: Enrichment of active cEnhs in different classes of cEnhs defined by the strength of their biochemical signatures. (A) Fraction of active enhancers in different cEnh signal classes based on ChIP-seq data for H3K27ac in C2C12 myocytes. Genome-wide distribution and extrapolated number of active enhancers in C2C12 belonging to each signal strength class. (B) Fraction of active enhancers in different cEnh signal classes based on DNAse hypersensitivity data for H3K27ac in C2C12 myocytes. Genome-wide distribution and extrapolated number of active enhancers in C2C12 belonging to each signal strength class.



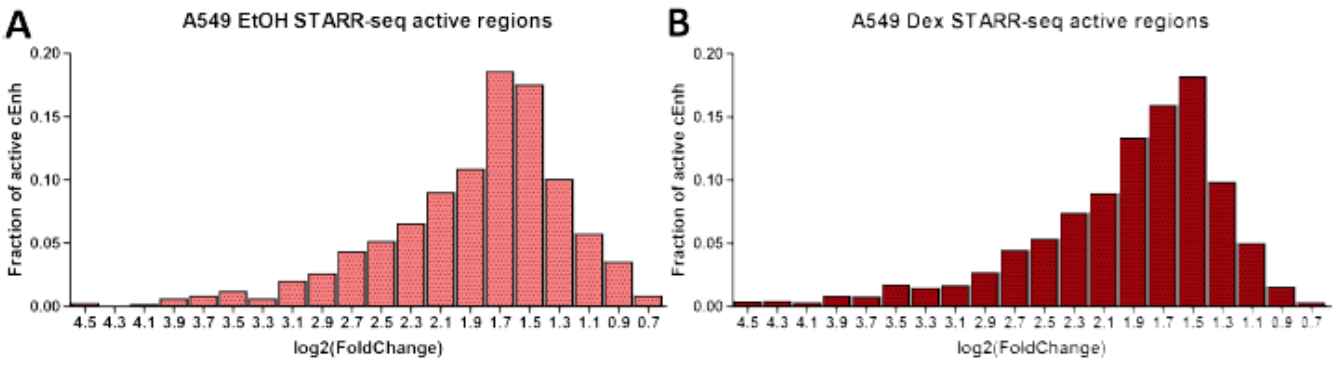
Supplementary Figure 19: Regulatory landscape of GR response in A549 cells. DNAse-seq and ChIP-seq experiments against H3K27ac across a 12 hour A549 dex response timecourse were analyzed within GR occupancy (IDR=0.05). Sites were split using GENCODE annotations into promoter proximal and distal subgroups then sorted by ChIP-seq signal for GR; the signal in the 500bp-radius region around the ChIP-seq peak position is shown.



Supplementary Figure 20: Testing of cEnhs for activity using ChIP-STARR-seq for GR in A549 cells with and without Dexamethasone stimulation . (A) Fraction of active cEnhs detected in each condition. Shown is the number of cEnhs that passed the minimum representation threshold (see the Methods section for more details) and were identified as active using DESeq2. (B) Fraction of significantly active (FDR-corrected p-value ≤ 0.05) biochemically marked individually on in combinations by H3K27ac, DNAse, p300. (C) Length distribution of active and inactive STARR-seq fragments as defined by DESeq2.



Supplementary Figure 21: Marking of common and cell state-specific active cEnhs by H32K7ac, DNAse and p300. (A) STARR-seq data in A549 cells with and without Dexamethasone treatment (epigenomic datasets from the 3 hour time point were used); (B) Luciferase assay data in differentiated and undifferentiated C2C12 cells.



Supplementary Figure 22: Distribution of STARR-seq activity in A549 cells. Shown is the distribution of log_2(FoldChange) values (defined by DESeq2) for STARR-seq experiments in resting EtOH-treated (A) and Dexamethasone-treated (B) A549 cells.