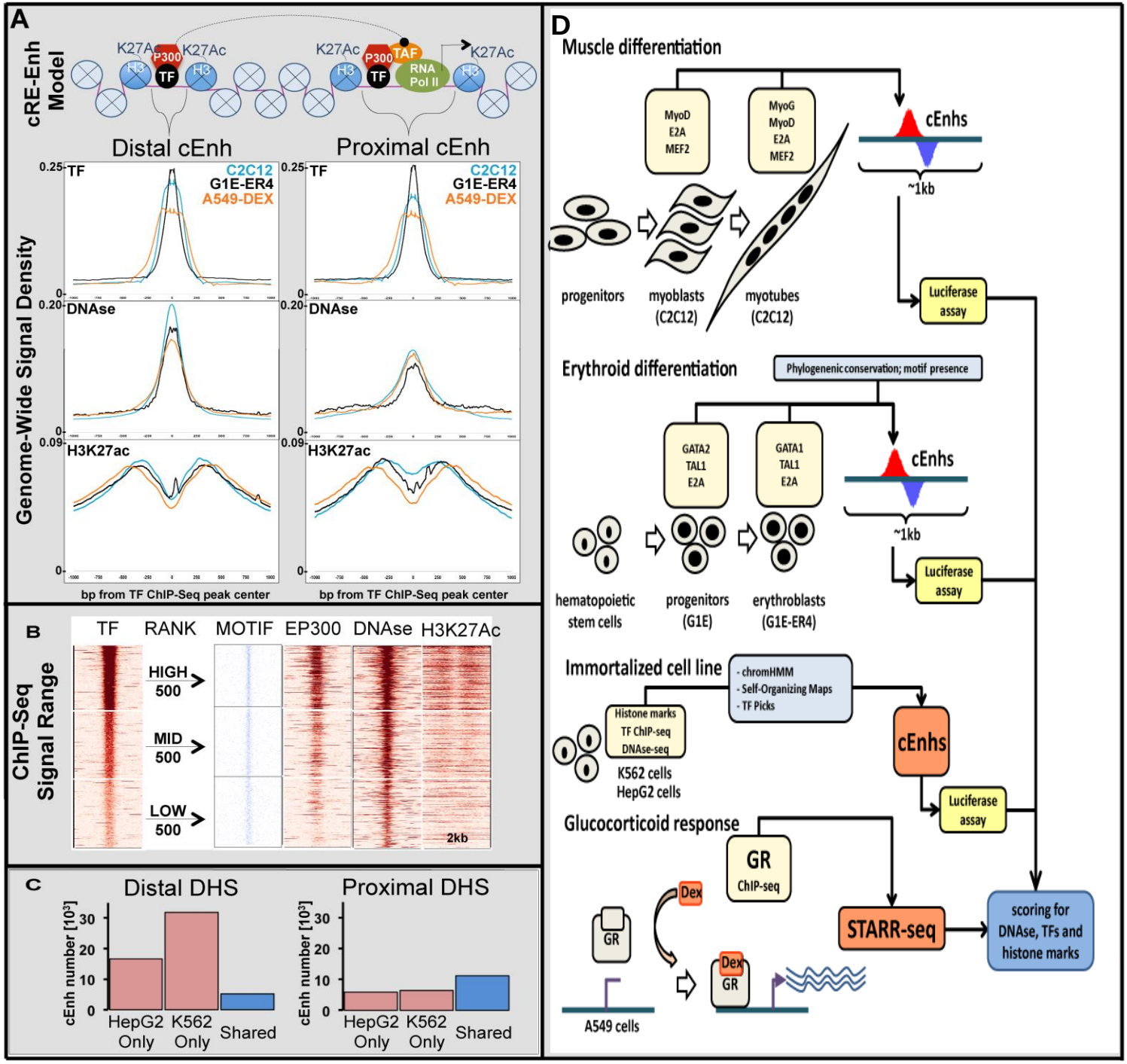

A
cRE-Enh Model
K27Ac
P300
TF
H3
TAF
RNA Pol II
Distal cEnh
Proximal cEnh
Genome-Wide Signal Density
C2C12
G1E-ER4
A549-DEX
TF
DNAse
H3K27ac
bp from TF ChIP-Seq peak center
B
ChIP-Seq Signal Range
TF
RANK
MOTIF
EP300
DNAse
H3K27Ac
HIGH 500
MID 500
LOW 500
2kb
C
Distal DHS
Proximal DHS
cEnh number [10³]
HepG2 Only
K562 Only
Shared
D
Muscle differentiation
MyoD E2A MEF2
MyoG MyoD E2A MEF2
cEnhs
~1kb
progenitors
myoblasts (C2C12)
myotubes (C2C12)
Luciferase assay
Erythroid differentiation
Phylogenenic conservation; motif presence
GATA2 TAL1 E2A
GATA1 TAL1 E2A
hematopoietic stem cells
progenitors (G1E)
erythroblasts (G1E-ER4)
Immortalized cell line
- chromHMM
- Self-Organizing Maps
- TF Picks
Histone marks TF ChIP-seq DNAse-seq
K562 cells
HepG2 cells
Glucocorticoid response
GR ChIP-seq
Dex
GR
STARR-seq
scoring for DNAse, TFs and histone marks
A549 cells

Fig1: Genome-wide Biochemical Signatures of Distal and Promoter-proximal candidate enhancers (cEnhs).

A1. Global biochemical profiles support a coherent model for cEnhs (A).

A2. Across six diverse cell types and states, the transcription factor (TF) occupancy, DNAse hypersensitivity (DHS) and H3K27ac global summed signatures are strikingly similar at both distal and proximal cEnhs. (A)

B1. Global TF measurements detect significant sites (IDR>0.05) that vary in signal intensity over 2+logs. The expected CAGSTG motif is found in 86.2% of the 32,278 myogenin IDR occupied regions.

C1. The vast majority of distal cEnhs use cell type differential locations in the genome, while the majority of TSS-proximal cEnhs are shared among cell types. (B)

D1) We selected candidate enhancers based on TF occupancy from a single cell state; while chromatin mark and DHS integrative methods sampled many datasets over a single (chromHMM) and multiple (SOM) tissues in a TF agnostic way to attempt to isolate candidate Enhancers.

We also selected as set of constrained motifs (Erythroid GATA1) without regard to the biochemistry.

A set of occupancy negative control elements that were characterized in the literature as T-Cell and axonal enhancers most of which contain a core “muscle” class e-box motif (dark blue – Figure 2) but were negative for occupancy of TF and DNAse were selected.

We tested the candidate enhancer elements only in the cells that match the biochemistry used for selection.

Figure 1 LEGEND:

Candidate enhancer regulatory element model.

Distal and proximal candidate enhancer elements shown. Non H3K27Ac modified histones are represented by the light blue core histones. P300 (dark grey) is recruited either directly by a TF, or through cofactors to acetylate adjacent histones (dark blue core), which are displaced, leaving a large DNAse hypersensitive (purple DNA) region where the TF is occupying the DNA. Both modified and unmodified histones largely protect the DNA from DNAse (dark blue DNA).

Distal candidate enhancers appear to be distinctly TF driven. Histones appear displaced around both promoter proximal and distal enhancers.

DNAse-seq and ChIP-seq experiments against H3K27ac and the TF myogenin in differentiated (myocyte) C2C12 cells[teal]; in Gata1 restored G1E-ER4 erythroid cells [black] and in GR activated (dex induced) A549 cells [orange] were analyzed. The 1kb radius region around the ChIP-Seq peak of statistically significant sites (IDR=0.05) were mapped and genome wide bp by bp density over the window is shown for the TF; DNAse and H3K27Ac measurements for TSS proximal (XXXbp radius) and distal candidate elements in the genome.

Distal candidate enhancers appear to be distinctly tissue specific elements.

Statistically significant DHS (IDR=0.05) from HepG2 (light grey) and K562 (dark grey) were analyzed and split into distal and TSS proximal (XXXbp radius). The DHS sites were then overlapped and compared. In total 53704 distal DHS sites were called of which only 5205 (9.6%) were shared (black). 23492 TSS Proximal DHS sites were found, of which 11,181 (47.6%) were shared between the two cell types.

ChIP-Seq signal range varies over XX logs.

DNAse-seq and ChIP-seq experiments against the TF MyoD, H3K27ac, p300 in undifferentiated (myoblast) C2C12 cells were analyzed. The 1kb radius region around the ChIP-Seq peak of statistically significant sites (IDR=0.05) were mapped and split into multiple subgroups depending on signal rank of the MyoD ChIP-Seq (top 500, mid 500, low 500). Any motif for MyoD (RRCAGSTG) within the same window were mapped and the signal for DNAse, H3K27Ac and p300 measurements for those regions are shown.

Candidate enhancer selection and enhancer assay.

Candidate enhancer elements were selected based on TF ChIP-Seq occupancy measurements in myelogenous (K562 – multiple TFs), Muscle (C2C12 – Myogenin/Myod) and Erythroid (G1E-ER4 – GATA1/Tal1). A set of conserved GATA1 motif containing regions were selected from the mouse genome for enhancer activity without regards to ChIP-Seq measurements prior to testing. Finally machine learning based candidate enhancer elements were selected in myogenenous (K562) cells based on chromatin segmentation (chromHMM) prediction and in liver (HepG2) based on DNAse/Histone SOM clusters predicted to be specifically active in that cell line. All candidate enhancers were tested in the corresponding cell lines, except the Erythroid mouse elements, which were tested in human K562 cells for activity.

Outline of experimental design, cRE selection and testing.

The elements selected for study, with the exception of negative control elements, were selected from regions of open or accessible chromatin as measured by ChIP-Seq experiments. A) emove this; now moved to gure 1. B) Elements in immortalized cell lines were picked based on machine learning algorithms. K562 were used to test chromHMM while HepG2 tested SOM based picks. C) Elements were selected from the dierentiation of muscle based on the major regulators of dierentiation (MyoD and MyoG). D) Erythroid candidate elements were selected based on the co-occupancy of GATA1/TAL1 (dierentiation factors). E) GR ChIP elements were cloned into a Starr-Seq libeary and assayed for activity.

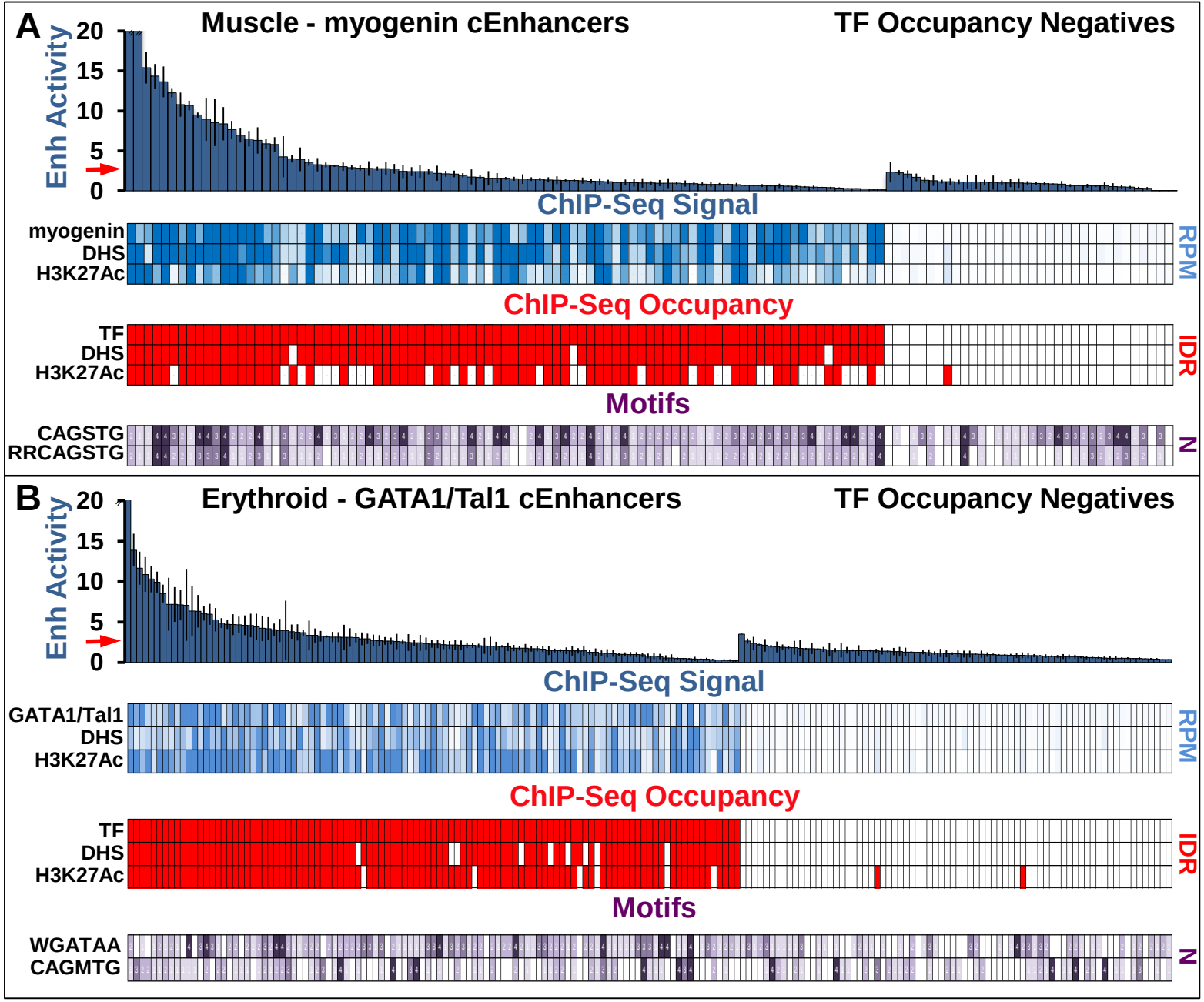

A
Muscle - myogenin cEnhancers
TF Occupancy Negatives
Enh Activity
20
15
10
5
0
ChIP-Seq Signal
myogenin
DHS
H3K27Ac
RPM
ChIP-Seq Occupancy
TF
DHS
H3K27Ac
IDR
Motifs
CAGSTG
RRCAGSTG
N
B
Erythroid - GATA1/Tal1 cEnhancers
TF Occupancy Negatives
Enh Activity
20
15
10
5
0
ChIP-Seq Signal
GATA1/Tal1
DHS
H3K27Ac
RPM
ChIP-Seq Occupancy
TF
DHS
H3K27Ac
IDR
Motifs
WGATAA
CAGMTG
N

Fig2: Enhancer activity and biochemical signal in tested Enhancers for (A) Muscle C2C12s and (B) Erythroid G1E-Er4 cell lines.

Approximately 50% of tested TF occupied candidates score as enhancers; there is poor correlation between the signal strength of the biochemical measurement and activity.

Most candidate enhancers are called as co-occupied but the signal intensity for individual biochemical measurements generally track poorly among each other even for active enhancers. This is likely because X-Seq signals are non-linear for physical occupancy of DNA.

The vast majority of muscle H3K27Ac negative cEnhancers are EP300 occupied indicating that an enhancer can function without requiring the presence of an established H3K27Ac mark in a plasmid system.

The enhancer activity measured in our assays is however specific for sites where biochemical occupancy is present. This is true for both randomly selected occupancy negative regions and literature validated enhancers of T-Cell and neuronal origin which are devoid of occupancy in muscle.

YYYadd motif disussion hereYYY

The vast majority of these selected occupancy negative candidates contain the same motifs found in the occupied counterparts. The number of motifs present within the tested region has no correlation with activity on the assay.

Figure 2 legend:

Functional assay testing of cEnh regulatory activity in C2C12 cells.

Fold activity in myocytes across technical replicates (n = 4) is shown. Candidate REs were required to be DNAse hypersensitive and then sorted 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, p300, MyoD and myogenin occupancy are shown for each cRE, both as RPM scores (highest signal in dark blue to low signal in yellow), and as binary (IDR=0.05) calls (red coloring indicates occupancy). These marks were used to extrapolate genome-wide activity for similarly biochemical marked sites (figure 3).

Functional assay testing of cEnh regulatory activity in Erythroid cells.

Functional assay testing of the regulatory activity of erythroid cEnhancers. Fold activity in K562 cells across biological replicates (n 2 [1; 9]) 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 (p=0.05). In addition, DNAse hypersensitivity, H3K27ac status, GATA1/Tal1 occupancy are shown for each cRE, both as binary (IDR=0.05) calls (red coloring indicates occupancy), and as RPM scores (heatmap with blue indicating high signal).

Fig3: Enhancer assay results and genome-wide predictions

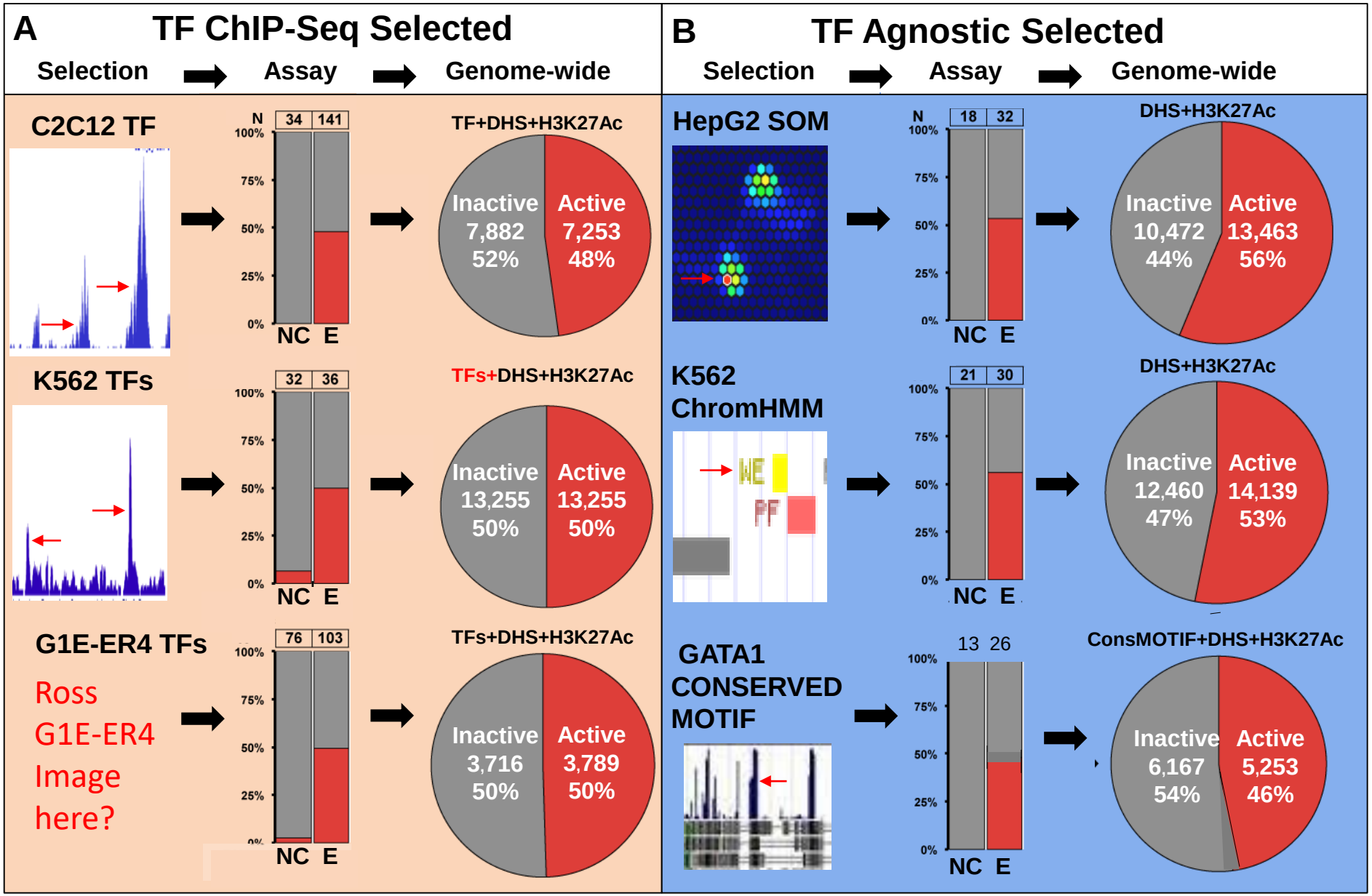


Fig3: Enhancer assay results and genome-wide predictions

~50% enhancers are found in TF occupied landscape across the entire spectrum of occupancy when annotated for DNAse and H3K27Ac measurements in Muscle, K562 and Erythroid cells.

~54% of DNAse/H3K27Ac candidate Enhancers selected by machine learning altorithms are functional as enhancers.

Overall the proportion of active enhancers is strikingly similar across different selection methods.

Conserved nucleotide elements containing the key TF motif but selected without any occupancy measures also return ~50% enhancer elements in the subset that are marked for DNAseHS and H3K27Ac; whereas again no function is found on the assay where there is no measurable occupancy.

We found no significant difference in the proportion of predicted active enhancers across different tissues, and extrapolate to similar predicted enhancer numbers across all tissues where the datasets are comparable.

FIGURE 3 LEGEND

Genome-wide predicted enhancer activity in TF occupied cEnhancers.

Functional assay testing of Enhancer regulatory activity was performed in C2C12 cells (Muscle); K562 (Myelogenous) and G1E-ER4 (Erythroid) for TF selected cEnhancers. Elements were subselected for stringent IDR (0.05) TF occupancy and presence of H3K27Ac+DHS. The number of elements scored for both a negative control set (NC) and experimental group (E) are represented above each plot individually. The proportion of elements active in the Experimental group is then used to extrapolate the predicted activity in similarly marked regions genome-wide. The negative control group was selected from genomic regions that show no TF occupancy or DHS; and in the cases of C2C12 and Erythroid required the presence of the TF occupancy motif.

Genome-wide predicted enhancer activity in TF agnostic cEnhancers.

Functional assay testing of Enhancer regulatory activity was performed in K562 cells (Myelogenous); HepG2 (Liver) and G1E-ER4 (Erythroid) for TF agnostic cEnhancers. Only elements that showed the presence of stringent (0.05) IDR H3K27Ac+DHS peak calls were selected. The proportion of elements active in the Experimental group is then used to extrapolate the predicted activity in similarly marked regions genome-wide. The negative control group was selected from genomic regions that show no DHS or H3K27Ac; for the Erythroid GATA1 conserved motif control group we required the presence of conserved TF occupancy motif with no biochemical marks.

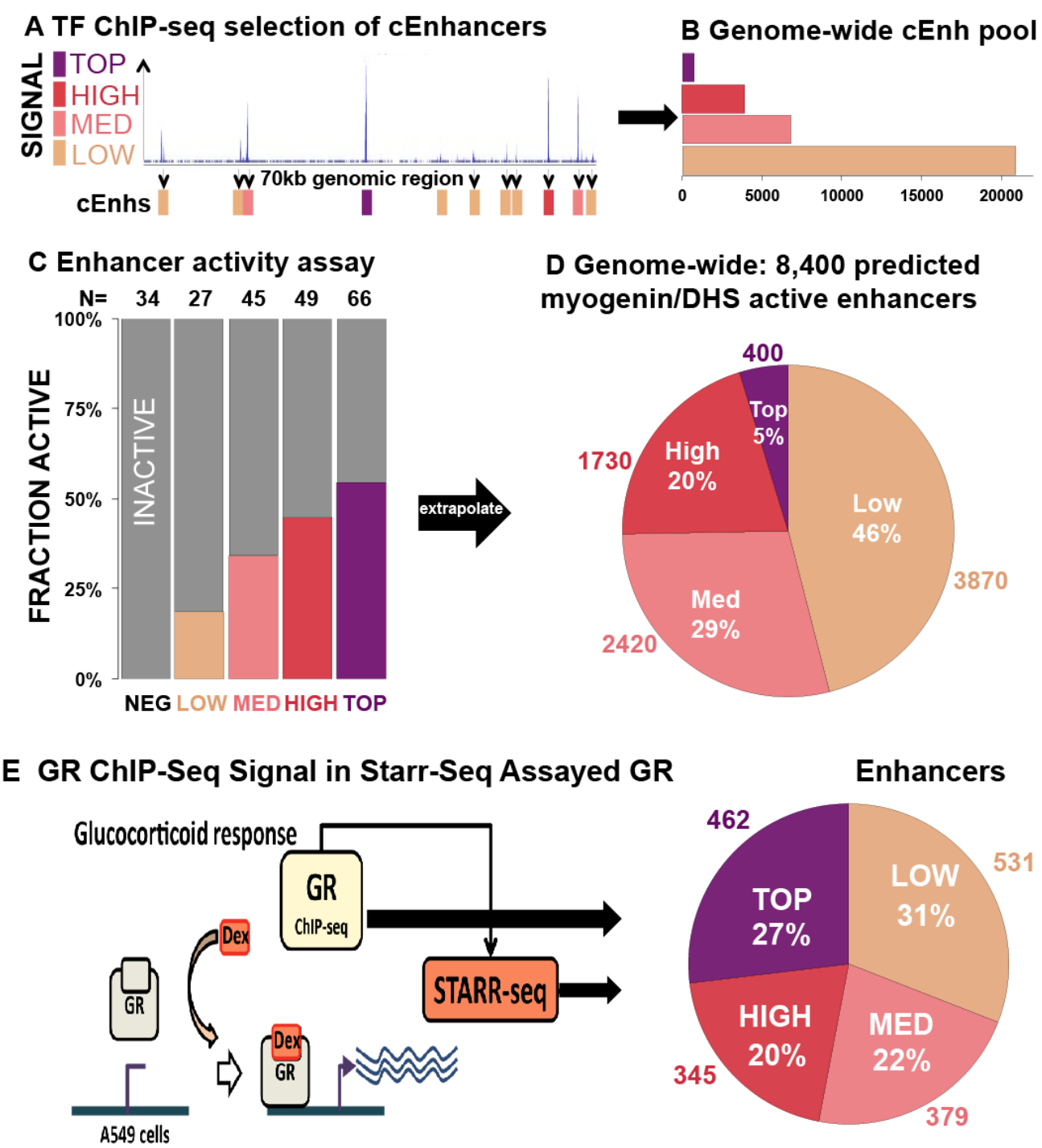

A TF ChIP-seq selection of cEnhancers
SIGNAL
TOP
HIGH
MED
LOW
70kb genomic region
cEnhs
B Genome-wide cEnh pool
0
5000
10000
15000
20000
C Enhancer activity assay
N= 34 27 45 49 66
100%
75%
50%
25%
0%
FRACTION ACTIVE
INACTIVE
NEG LOW MED HIGH TOP
extrapolate
D Genome-wide: 8,400 predicted myogenin/DHS active enhancers
400
Top 5%
1730
High 20%
Low 46%
3870
Med 29%
2420
E GR ChIP-Seq Signal in Starr-Seq Assayed GR Enhancers
Glucocorticoid response
GR ChIP-seq
STARR-seq
Dex
GR
Dex
GR
A549 cells
462
TOP 27%
531
LOW 31%
HIGH 20%
345
MED 22%
379

Fig4: genome-wide enhancers predicted by strength of ChIP-Seq signal

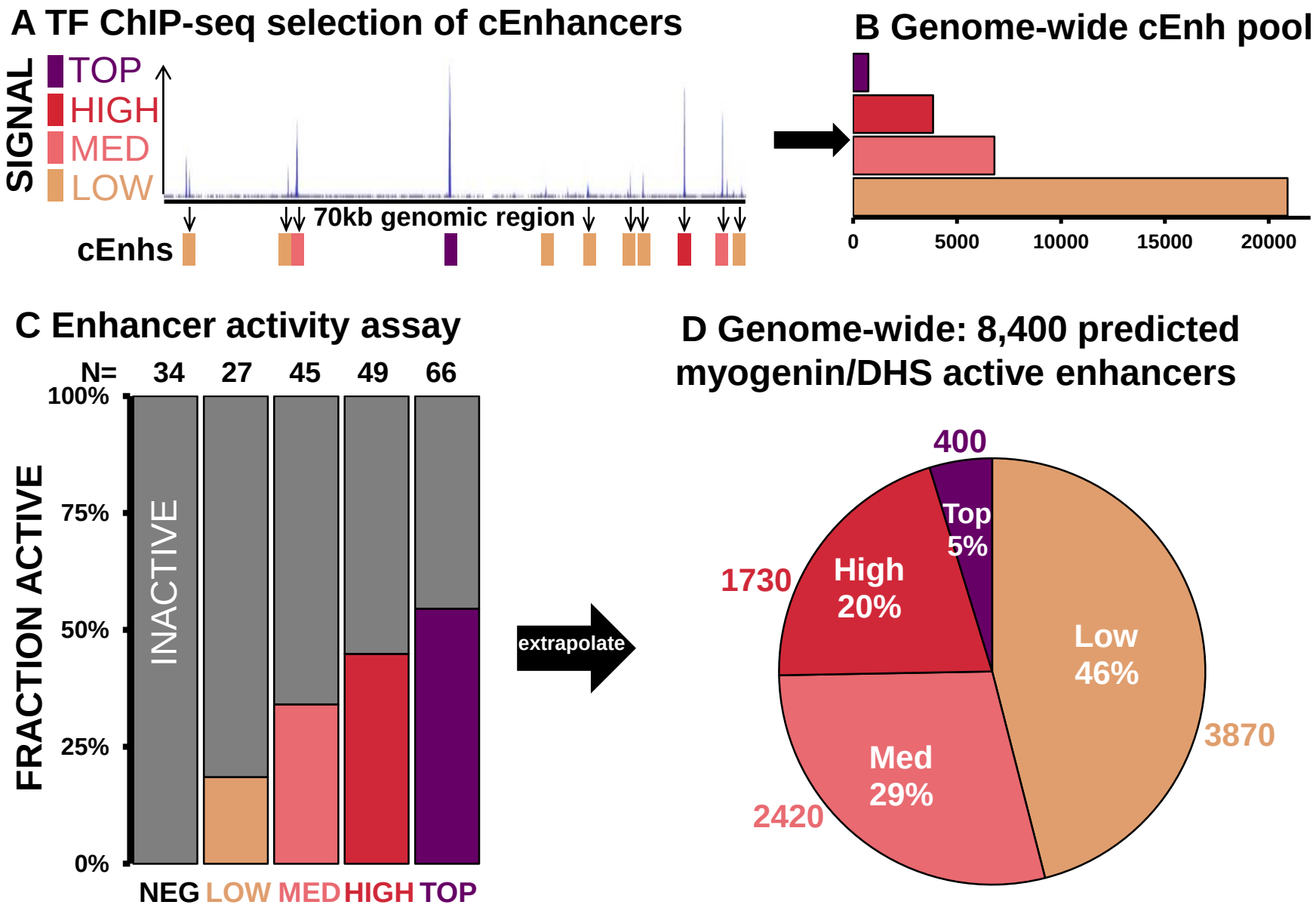
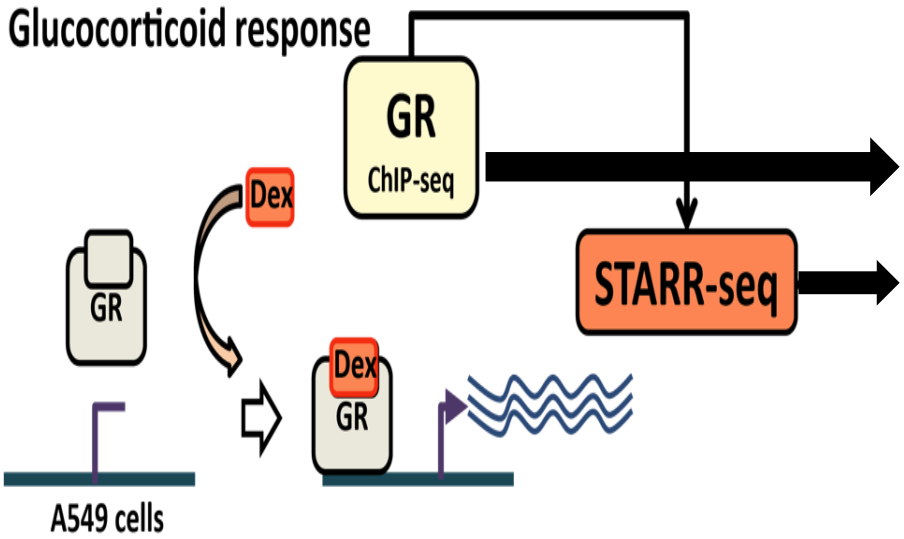
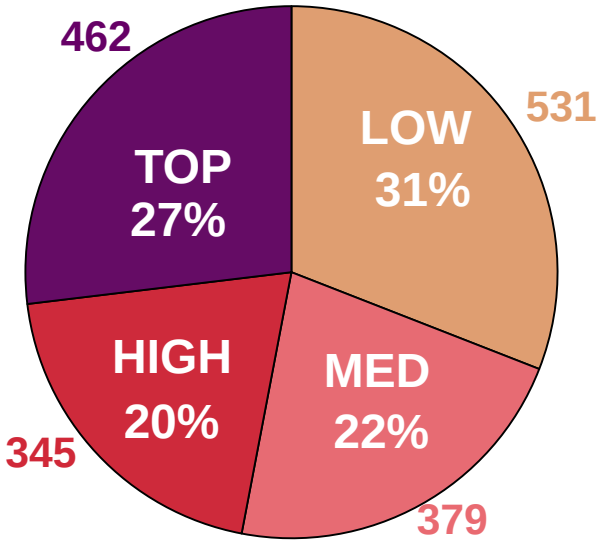


Fig 4

E GR ChIP-Seq Signal in Starr-Seq Assayed GR Enhancers





TF occupied candidate Enhancers in Muscle were picked from the full spectrum of myogenin signal. (A)

Small signals represent the majority of candidate enhancers in the genome. (B)

Enhancer assay shows high biochemical signals relatively enriched for function but function present throughout. (C)

Low-medium biochemical signals demark the majority of enhancers predicted in the genome. (D)

A similar contribution is found from the same elements annotated for H3K27Ac and DNAse signal; where small biochemical signals contribute the majority of enhancers. (SupFig6)

Modern Starr-Seq confirms that Low-Medium GR occupancy signals contribute the majority of enhancers detected in the genome. (E)

Figure 4 legend:

TF selected candidate enhancers derived from different biochemical signals.

Representative genomic region, showing sites that show a relative expected signal and derived candidate enhancers for top, high, medium, low signals that are statistically significant IDR>0.05 in a myogenin ChIP-Seq experiment.

Small signals represent the majority of candidate enhancers in the genome. Summed number of candidate enhancers by signal class in the genome.

Enhancer assay shows high biochemical signals relatively enriched for function. Candidate enhancers were tested from each signal class (N provided in figure), and a set of negative control elements that tests a set of similarly sized regions that contain the occupancy motif (CAGSTG) but show no biochemical occupancy signal at their relative location.

Low-medium biochemical signals demark the majority of enhancers in the genome. Predicted number of active enhancers from each signal class.

Fig5: Relationship of enhancer activity to measured biochemical marks in tested cEnhs

Relationship of Enhancer activity with measured DNAse and H3K27Ac signal in C2C12 (A) and G1E-ER4 (B) Cells

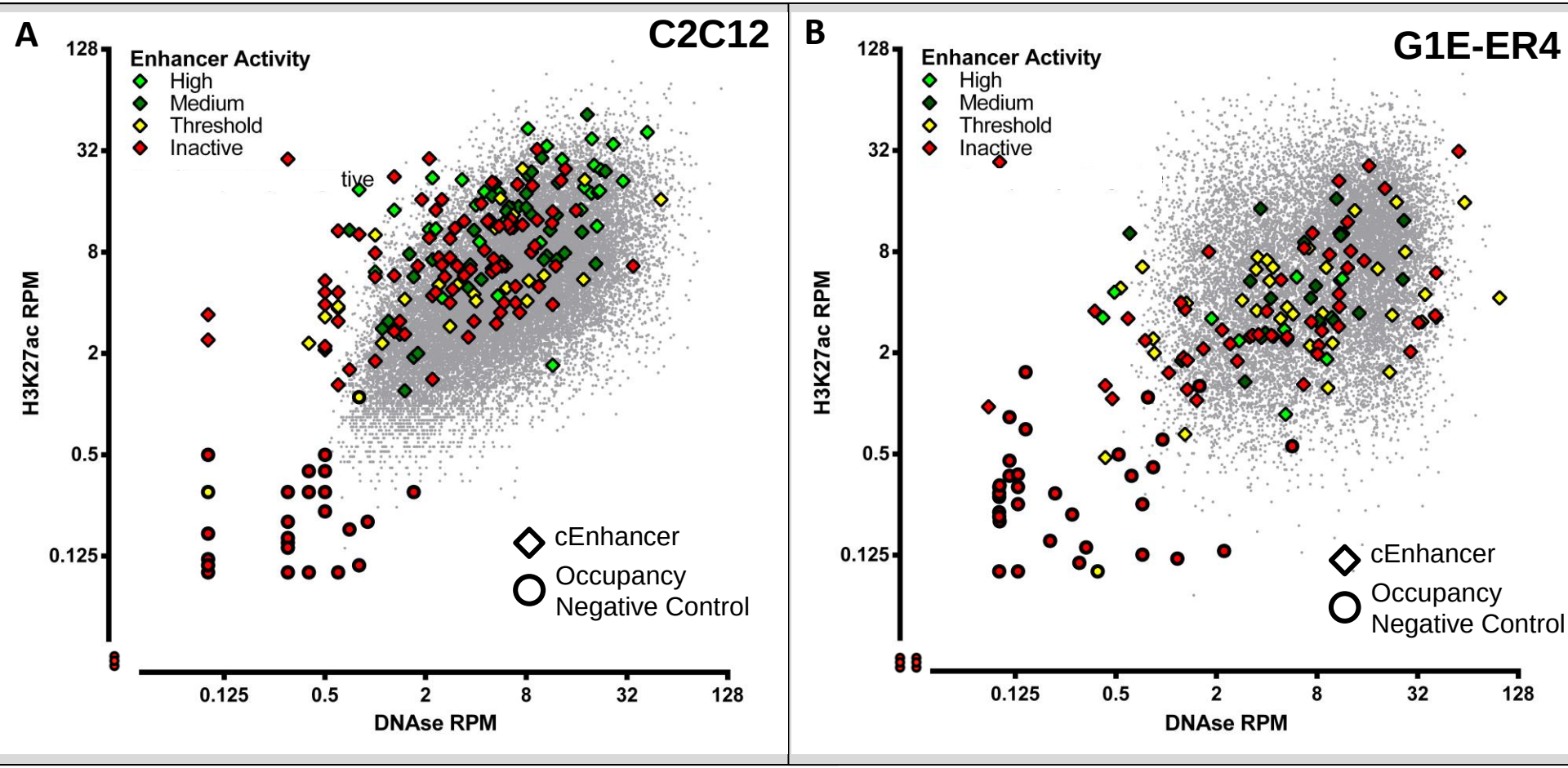


Fig5: Relationship of enhancer activity to measured biochemical marks in tested cEnhs

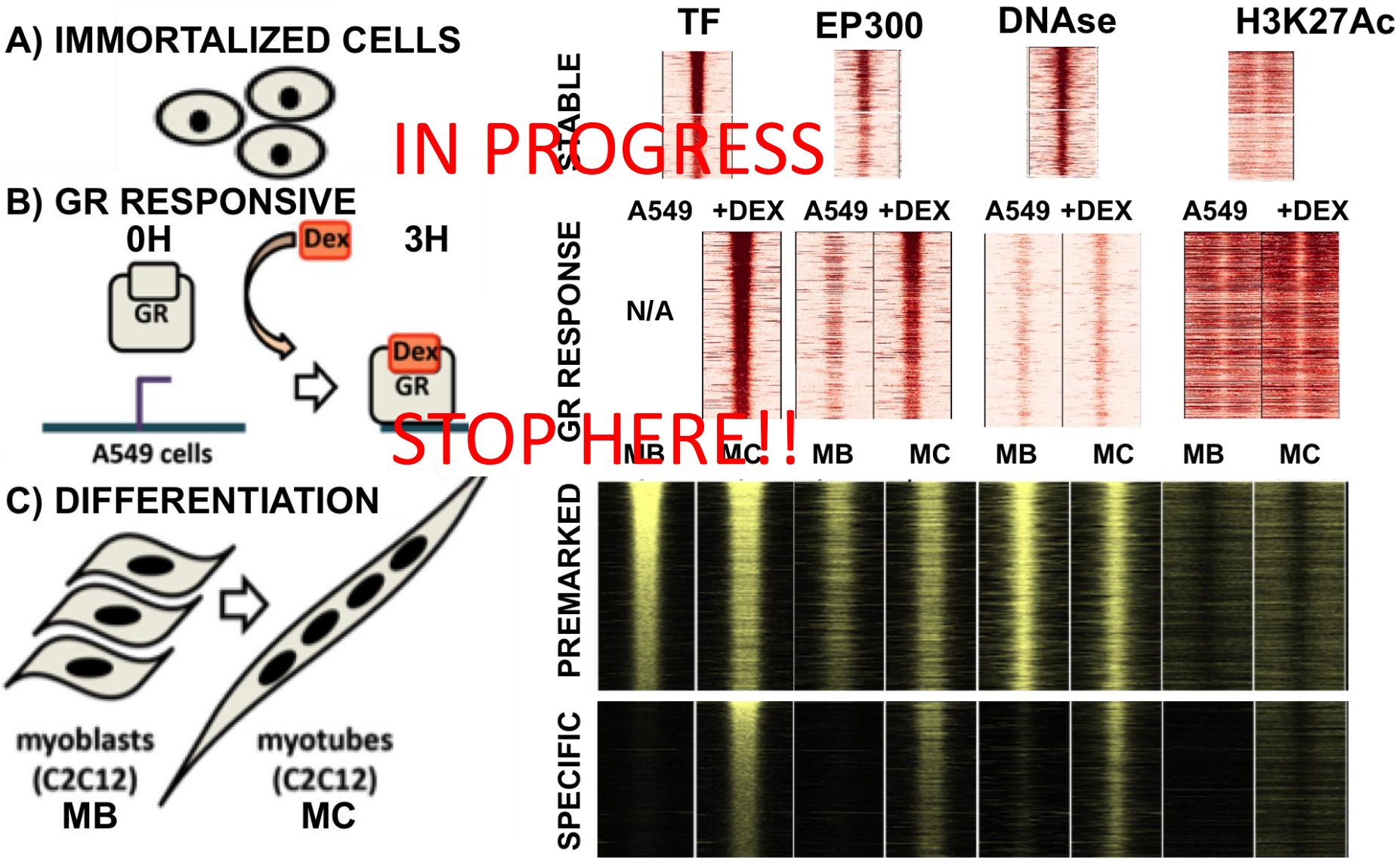
- Biochemical signal strength is unreliable in predicting enhancer functional output.
- Many weakly biochemically marked sites are part of the top enhancers as measured by the assay.
- Most of highly marked by DNAse are active elements; but the proportion of active enhancers drops off quickly as the biochemical measurement signal decreases.

Fig5: Relationship of enhancer activity to measured biochemical marks in tested cEnhs

DNAse and H3K27Ac signal in co-occupied IDR 0.05 regions in C2C12 myocytes (A – grey circles) and G1E-ER4 cells (B – grey circles). cEnhancers tested are in the shape of black outlined romboid; while occupancy negative control elements as black outlined circles. The activity from the transfection assay is broken down in four classes: high active (bright green); medium active (green); threshold (yellow) and inactive elements (red)

- STOP HERE -> figure 6 is still in progress

Fig6: Relationship of H3K27Ac predictivity in dynamic processes.



There are key caveats to a simple model that requires the presence of DNAse and H3K27Ac (Fig-6A) to annotate candidate enhancers in the ENCODE encyclopedia:

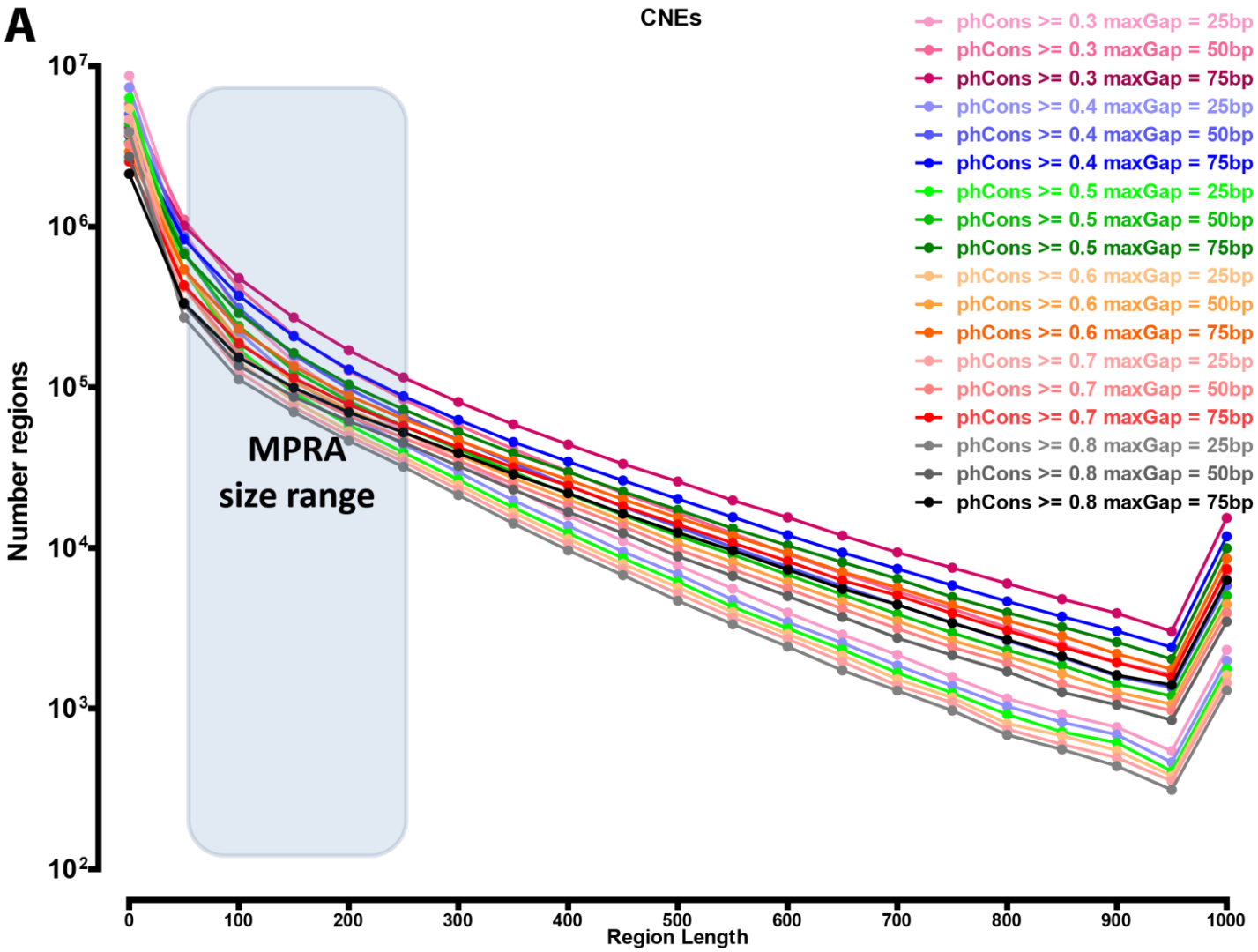
- The biochemical locations of GR response is largely predestined by DNAse/H3K27Ac (Fig-6B) at the vast majority of sites sites prior to GR occupancy; with EP300 being recruited upon induction. These sites would be impossible to discern using the measurements of the most modern ENCODE encyclopedia alone.

- During development of muscle there are thousands of TF occupied sites where EP300 is co-localized but H3K27 has not yet been acetylated.

  As demonstrated in the assay results presented in figure 2; many of these elements are likely functional enhancer elements that would also be excluded from an enhancer characterization in the ENCODE encyclopedia.

Supplementary Figures

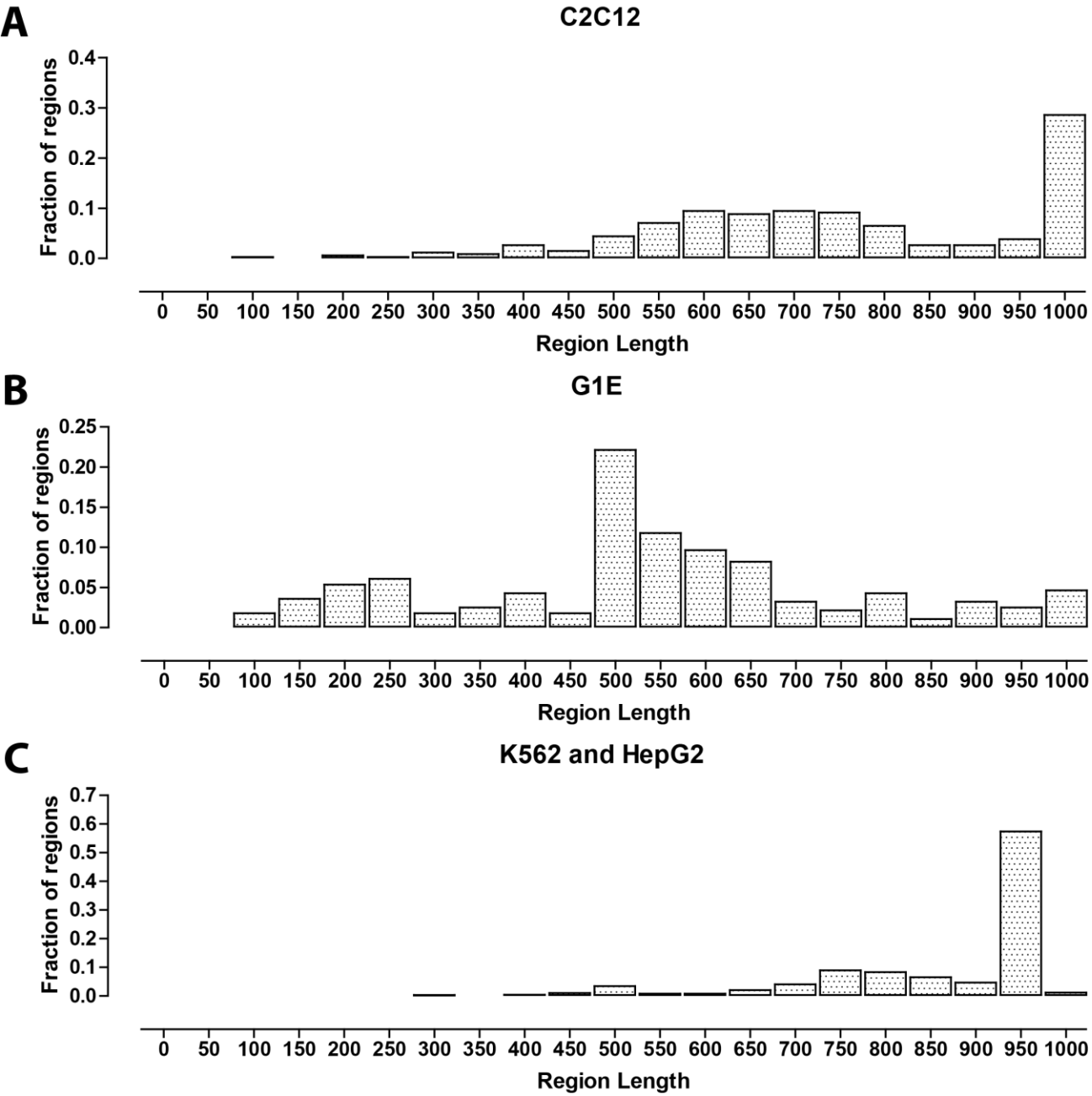
Supplementary figure 1



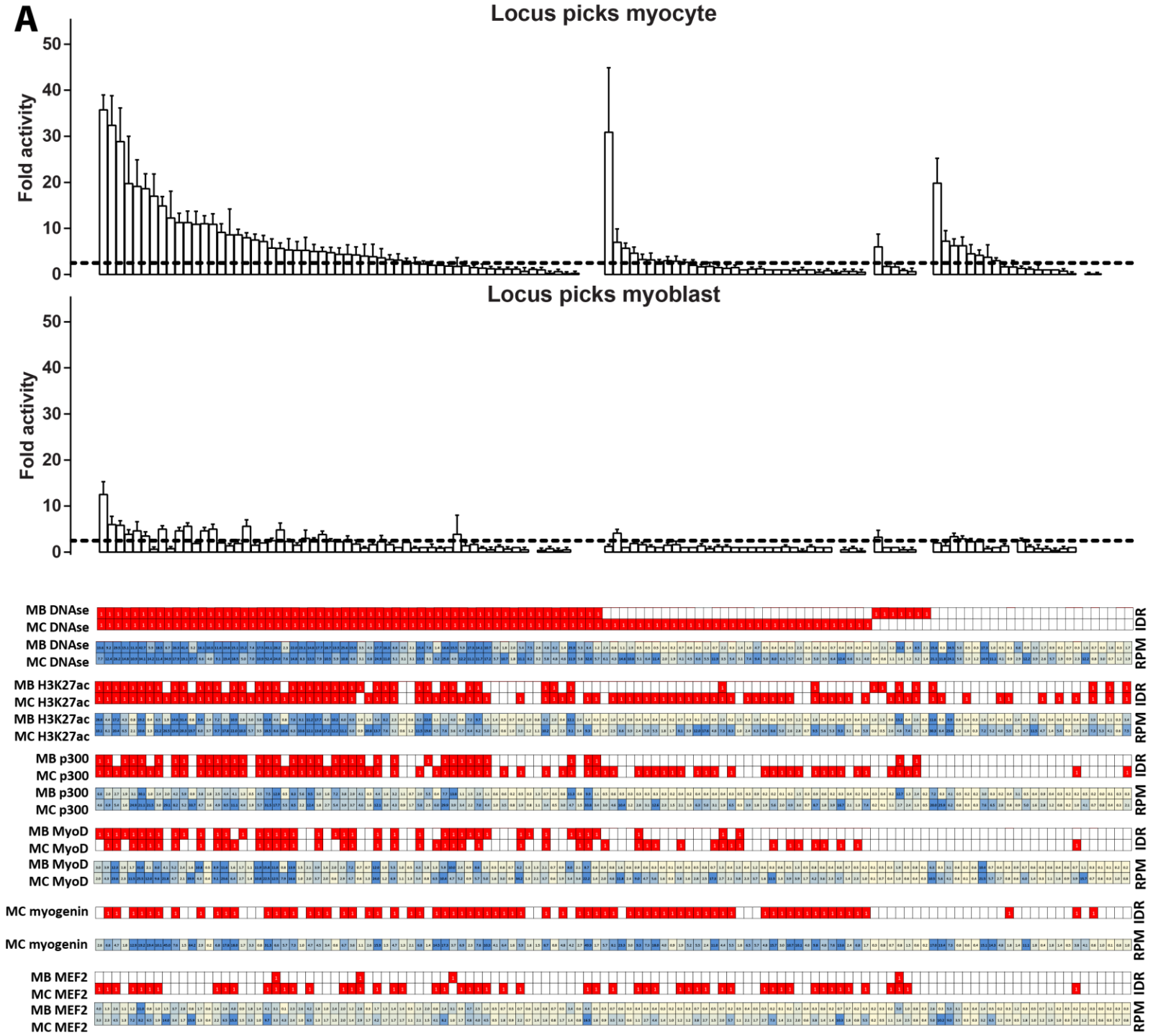
The structure and length of thousands of conserved noncoding elements in mammalian genomes greatly exceeds the size range of MPRA constructs

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 identied, 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. (D) Such an example is shown for the Acta1 gene in mouse.

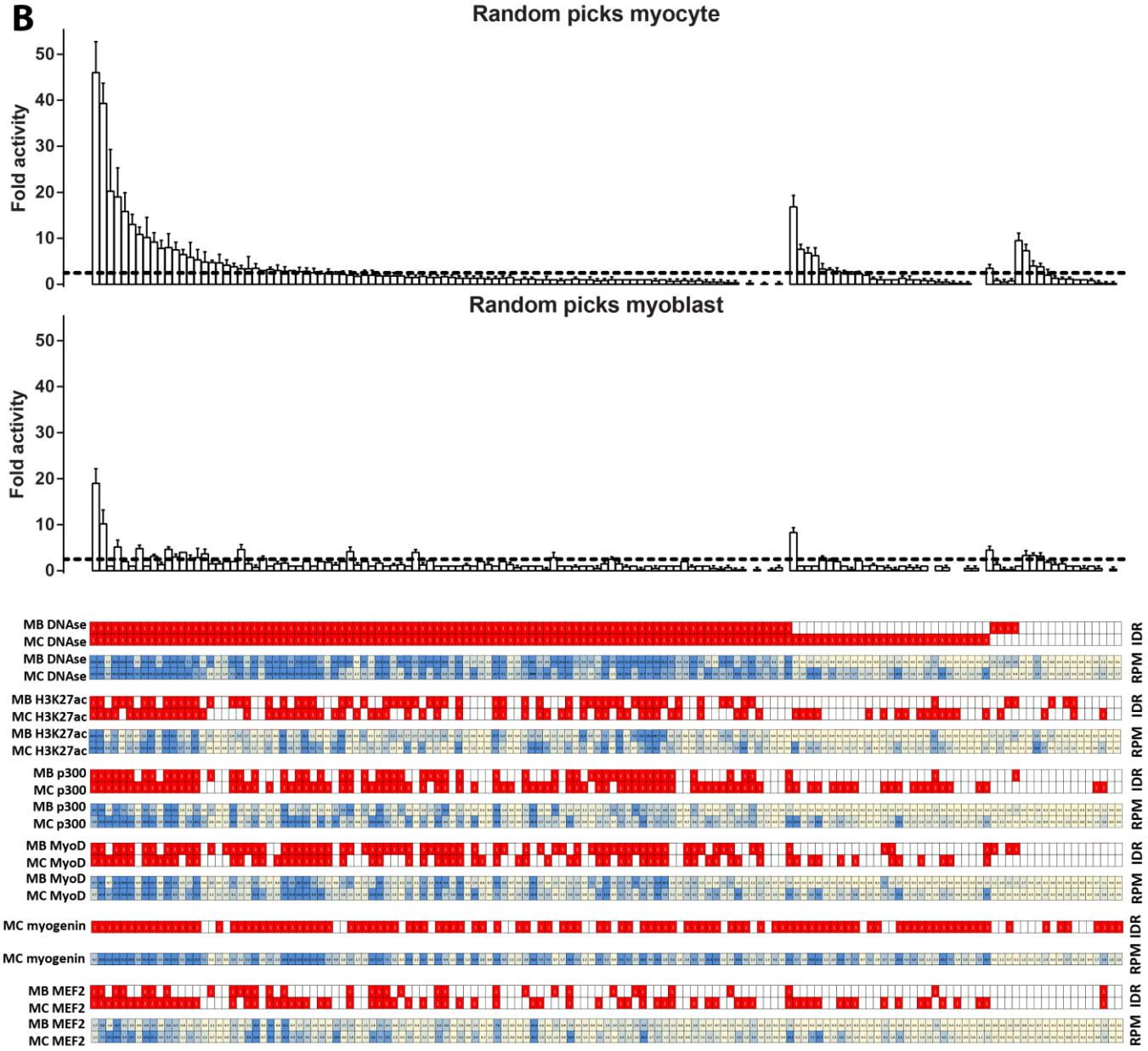
Supplementary figure 2



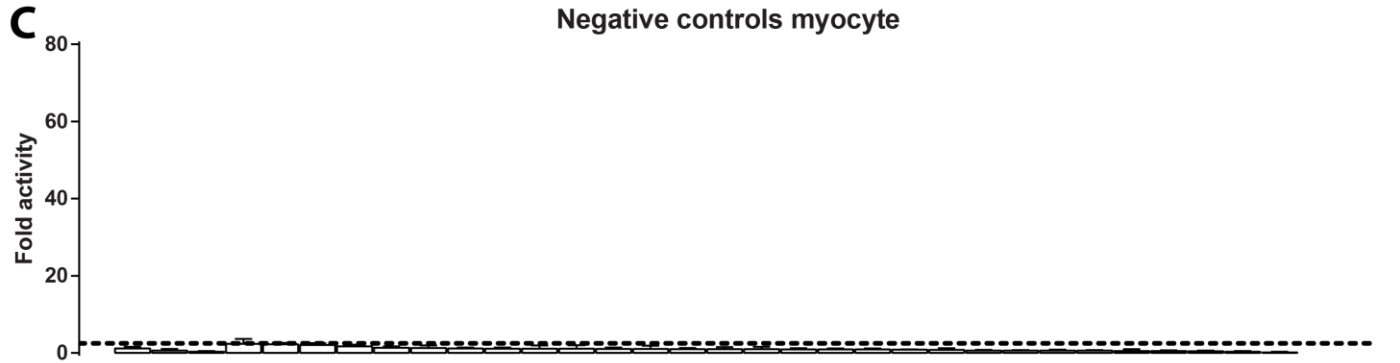
Length distribution of functional assays constructs used to test cREs in this study. (A) Distribution of functional assay construct lengths tested in this study in C2C12 cells. (B) Distribution of functional assay construct lengths tested in this study in G1E cells. (C) Distribution of functional assay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

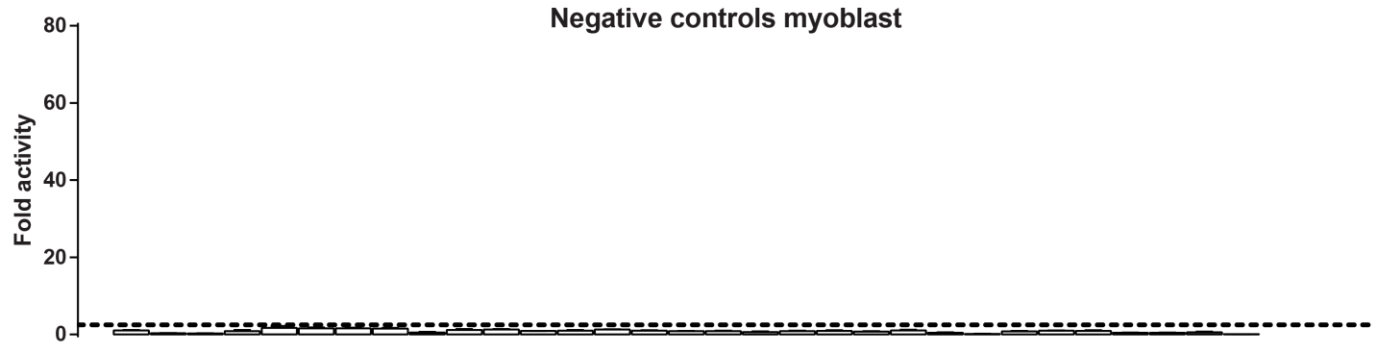

A
Locus picks myocyte
Fold activity
50
40
30
20
10
0
Locus picks myoblast
Fold activity
50
40
30
20
10
0
MB DNAse
MC DNAse
IDR
MB DNAse
MC DNAse
RPM
MB H3K27ac
MC H3K27ac
IDR
MB H3K27ac
MC H3K27ac
RPM
MB p300
MC p300
IDR
MB p300
MC p300
RPM
MB MyoD
MC MyoD
IDR
MB MyoD
MC MyoD
RPM
MC myogenin
IDR
MC myogenin
RPM
MB MEF2
MC MEF2
IDR
MB MEF2
MC MEF2
RPM

Supplementary figure 3A - Muscle



C

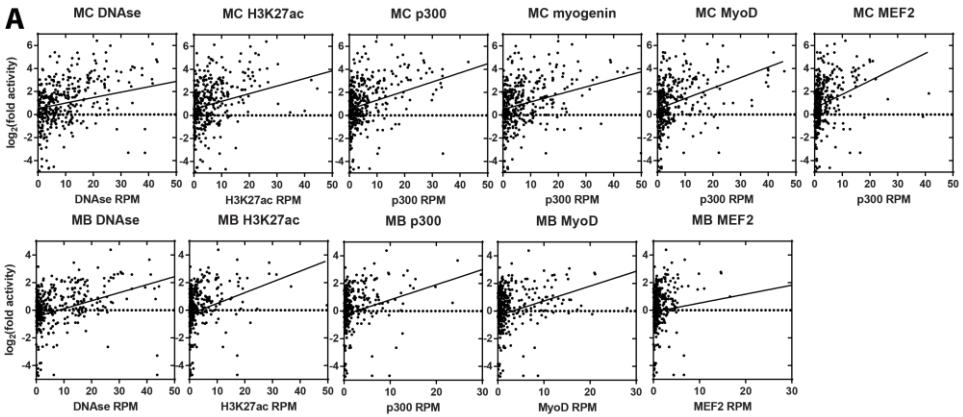




MB DNAse	1	1	1																													IDR
MC DNAse																																
MB DNAse	1.7	1.7	2.1	0.1	0.1	0.1	0.2	0.1	0.2	0.0	0.1	0.1	0.1	0.1	0.4	0.1	0.2	0.2	0.1	0.3	0.0	0.2	0.1	0.0	0.0	0.2	0.1	0.0	0.4	0.1	0.1	0.1	0.3	0.1	RPM
MC DNAse	2.7	0.8	0.9	0.1	0.3	0.8	0.6	0.1	0.3	0.1	0.2	0.1	0.1	0.5	0.5	0.3	0.4	0.5	0.3	0.4	0.0	0.5	0.3	0.0	0.1	0.3	0.1	0.1	1.7	0.2	0.4	0.3	0.5	0.4	
MB H3K27ac																																IDR
MC H3K27ac	1																															
MB H3K27ac	0.4	0.1	0.2	0.3	0.4	0.3	0.2	0.5	0.3	0.1	0.2	0.1	0.2	0.5	0.1	0.0	0.3	0.3	0.1	0.7	0.0	0.5	0.3	0.1	0.2	0.4	0.3	0.1	0.4	0.3	0.3	0.3	0.3	0.2	RPM
MC H3K27ac	4.7	0.1	0.2	0.3	0.3	1.1	0.1	0.5	0.1	0.1	0.3	0.1	0.1	0.2	0.1	0.1	0.1	0.3	0.2	0.4	0.0	0.4	0.1	0.1	0.0	0.2	0.1	0.1	0.3	0.0	0.3	0.1	0.5	0.4	
MB p300	1																															IDR
MC p300	1																															
MB p300	0.9	0.6	0.6	0.5	0.3	0.1	0.5	0.1	0.1	0.2	0.1	0.2	0.1	0.1	0.2	0.1	0.4	0.5	0.6	0.3	0.0	0.4	0.2	0.0	0.3	0.3	0.2	0.1	0.2	0.1	0.1	0.2	0.3	0.3	RPM
MC p300	2.7	0.2	0.2	0.4	0.1	0.3	0.3	0.1	0.2	0.1	0.1	0.1	0.2	0.2	0.1	0.1	0.2	0.8	0.4	0.8	0.0	0.5	0.2	0.1	0.1	0.2	0.1	0.1	0.5	0.3	0.4	0.3	0.3	0.3	
MB MyoD																																IDR
MC MyoD																																
MB MyoD	0.3	0.7	0.5	0.6	0.4	0.3	0.7	0.4	0.2	0.2	0.3	0.2	0.2	0.1	0.5	0.2	0.4	1.0	0.4	0.2	0.0	0.3	0.1	0.1	0.3	0.6	0.1	0.2	0.8	0.2	0.3	0.3	0.5	0.2	RPM
MC MyoD	0.4	0.3	0.3	0.5	0.2	0.3	0.6	0.3	0.1	0.2	0.0	0.3	0.2	0.2	0.2	0.3	0.5	0.2	0.3	0.5	0.0	0.5	0.1	0.1	0.1	0.4	0.2	0.2	0.6	0.1	0.3	0.6	0.6	0.2	
MC myogenin																																IDR
MC myogenin	0.4	0.3	0.4	0.3	0.5	0.3	0.4	0.3	0.1	0.2	0.2	0.1	0.2	0.1	0.2	0.2	0.2	0.3	0.4	0.4	0.0	0.7	0.2	0.0	0.1	0.7	0.1	0.0	0.6	0.1	0.6	0.3	0.7	0.6	RPM
MB MEF2																																IDR
MC MEF2																																
MB MEF2	0.6	0.3	0.6	0.5	0.7	0.2	0.3	0.3	0.2	0.1	0.2	0.3	0.0	0.2	0.4	0.0	0.5	0.3	0.5	0.5	0.0	0.3	0.3	0.0	0.2	0.6	0.1	0.1	0.6	0.2	0.3	0.1	0.3	0.3	RPM
MC MEF2	1.1	0.6	0.4	0.6	0.3	0.5	0.4	0.2	0.0	0.1	0.1	0.4	0.1	0.2	0.3	0.2	0.5	0.3	0.4	0.8	0.0	0.3	0.3	0.2	0.0	0.4	0.1	0.1	0.3	0.2	0.2	0.3	0.4	0.3	

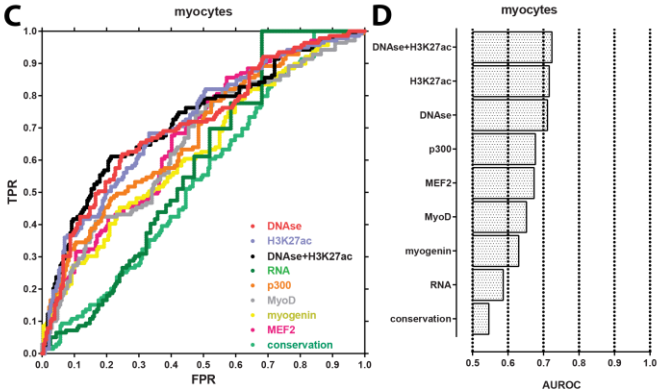
Functional assay testing of cRE regulatory activity in C2C12 cells. Fold activity in myocytes (top) and myoblasts (bottom) across biological replicates (n = 4) and technical replicates (n = 4 for each biological replicate) is shown. Candidate REs were sorted rst by their DNAse status and then by their 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, 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. (A) cREs selected for their physical proximity to loci known for their importance to muscle development (\locus picks"); (B) randomly selected from the genome-wide set of MyoD/myogenin-occupied regions; (C) negative controls.

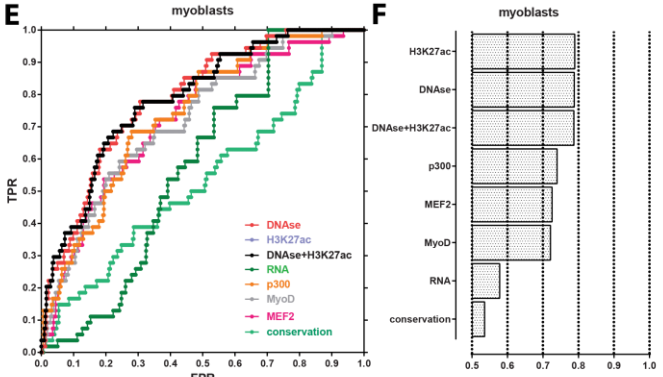
Supplementary figure 3B - Muscle



B

	myocytes		myoblasts	
	r^2	Spearman	r^2	Spearman
DNAse	0.09	0.44	0.09	0.37
H3K27ac	0.08	0.38	0.05	0.34
p300	0.11	0.42	0.05	0.31
MyoD	0.12	0.38	0.06	0.30
myogenin	0.11	0.36		
MEF2	0.09	0.41	0.01	0.29

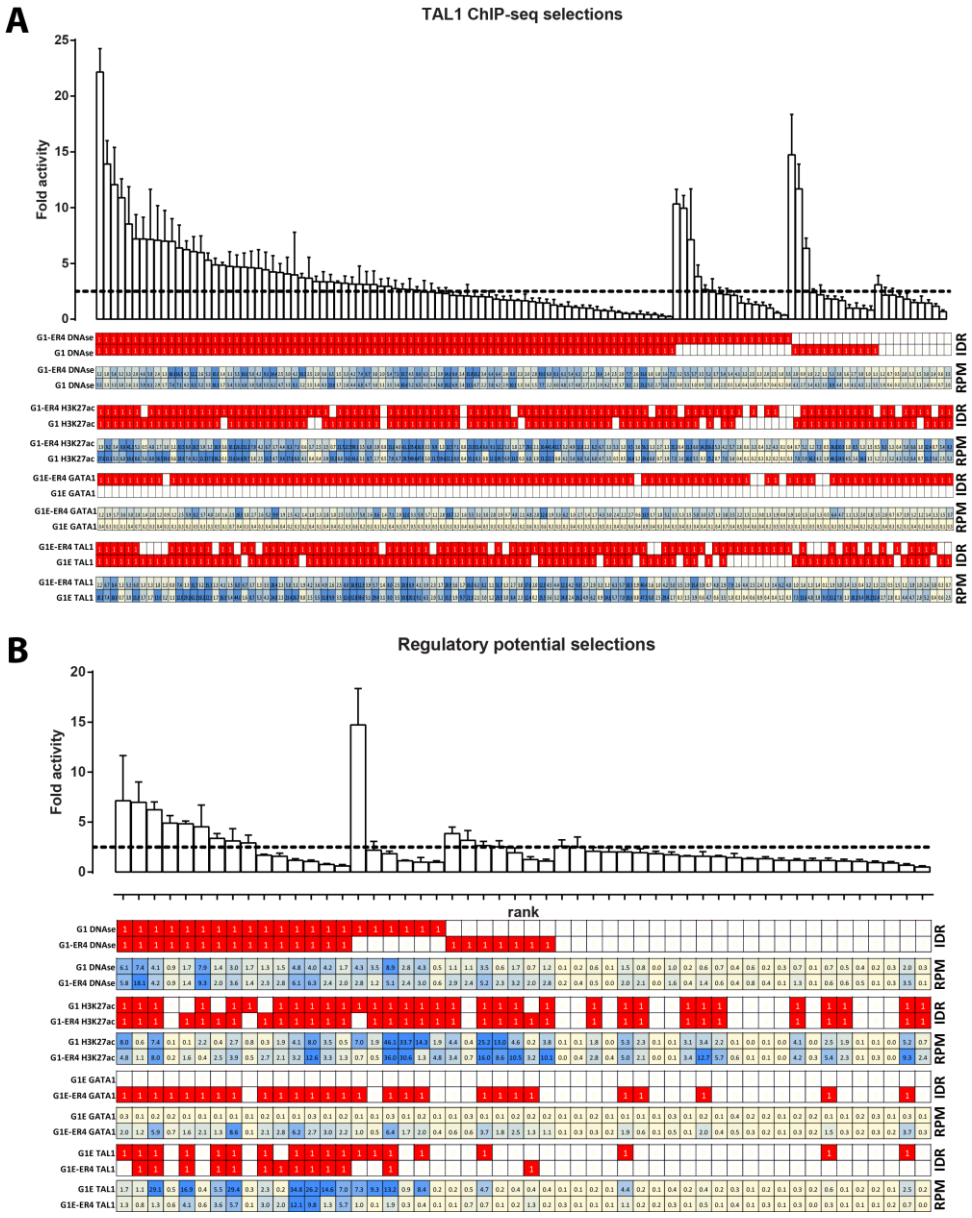




Correlation between regulatory activity and biochemical marks in C2C12 cells. (A and B) Correlation between fold activity and DNAse hypersensitivity, H3K27ac, p300, myogenin, MyoD and MEF2 occupancy in myoblasts and myocytes; (C) ROC curves showing biochemical mark predictivity of cRE fold activity in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

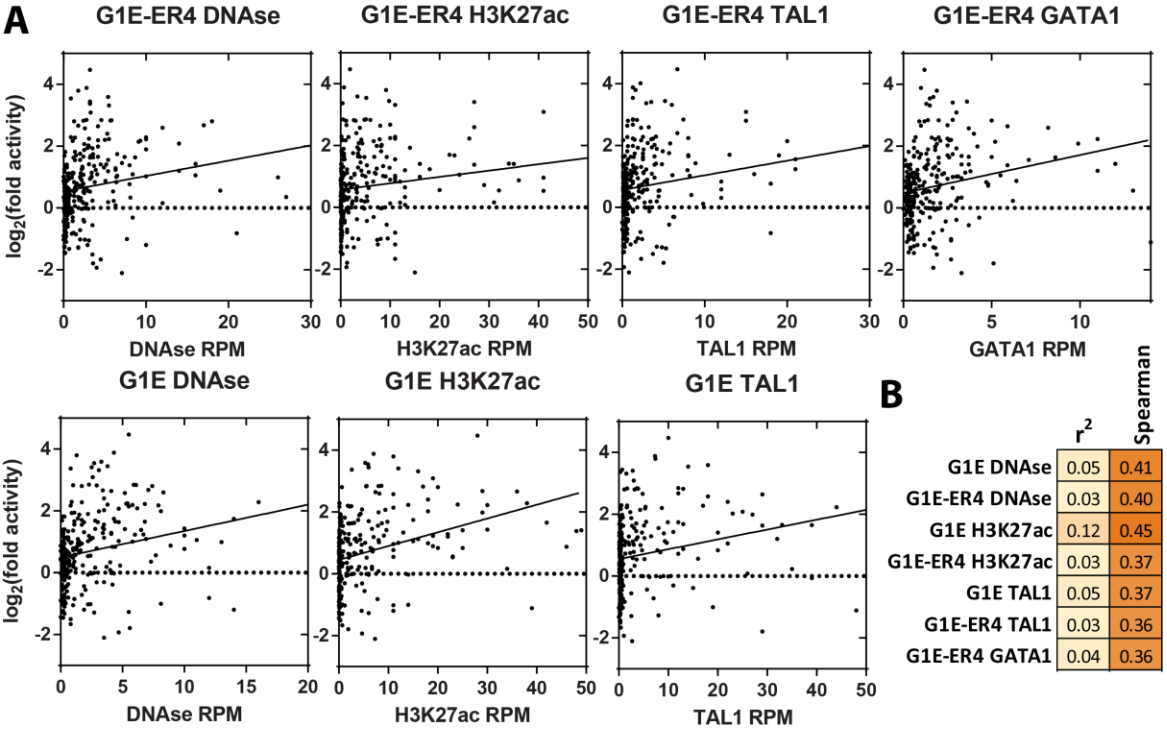
mark predictivity of cRE fold activity in myoblasts; (F) AUROC values for dierent biochemical marks in myoblasts.

Supplementary figure 4A - Erythroid



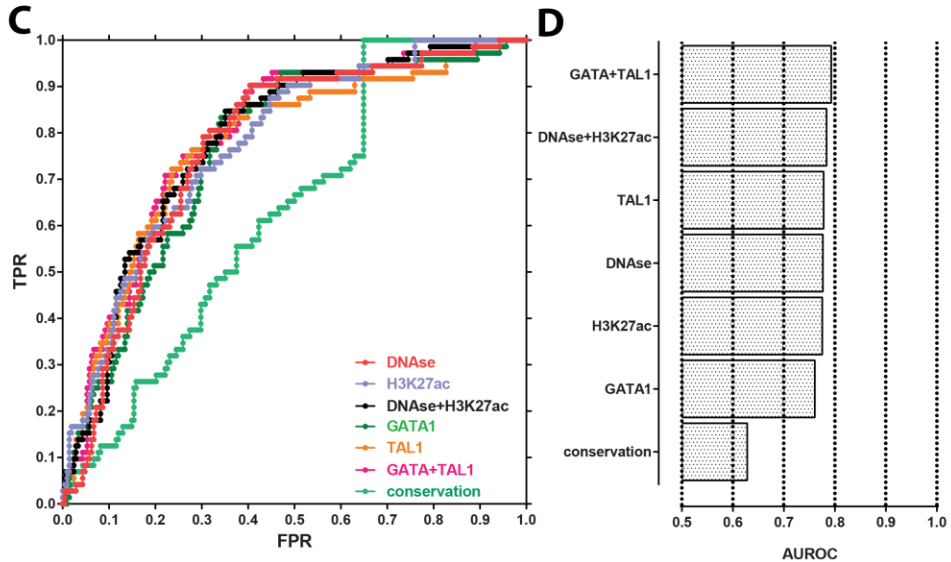
Functional assay testing of the regulatory activity of erythroid cREs. Fold activity in K562 cells across biological replicates (n 2 [1; 9]) and technical replicates (n = 4 for each biological replicate) is shown. Candidate REs were sorted rst by their DNAse status and then by their mean fold activity. 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 4B - Erythroid



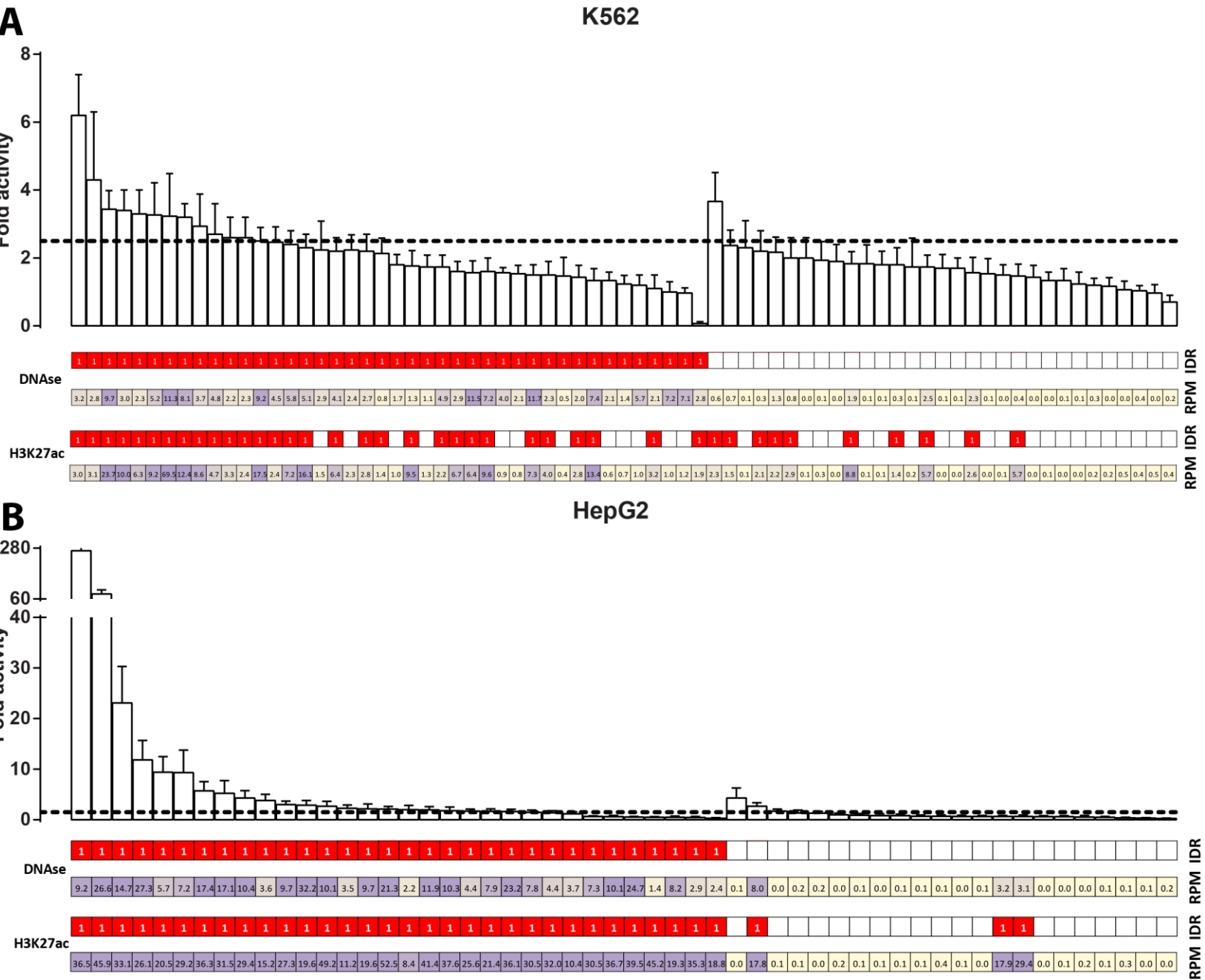
B

	r^2	Spearman
G1E DNAse	0.05	0.41
G1E-ER4 DNAse	0.03	0.40
G1E H3K27ac	0.12	0.45
G1E-ER4 H3K27ac	0.03	0.37
G1E TAL1	0.05	0.37
G1E-ER4 TAL1	0.03	0.36
G1E-ER4 GATA1	0.04	0.36



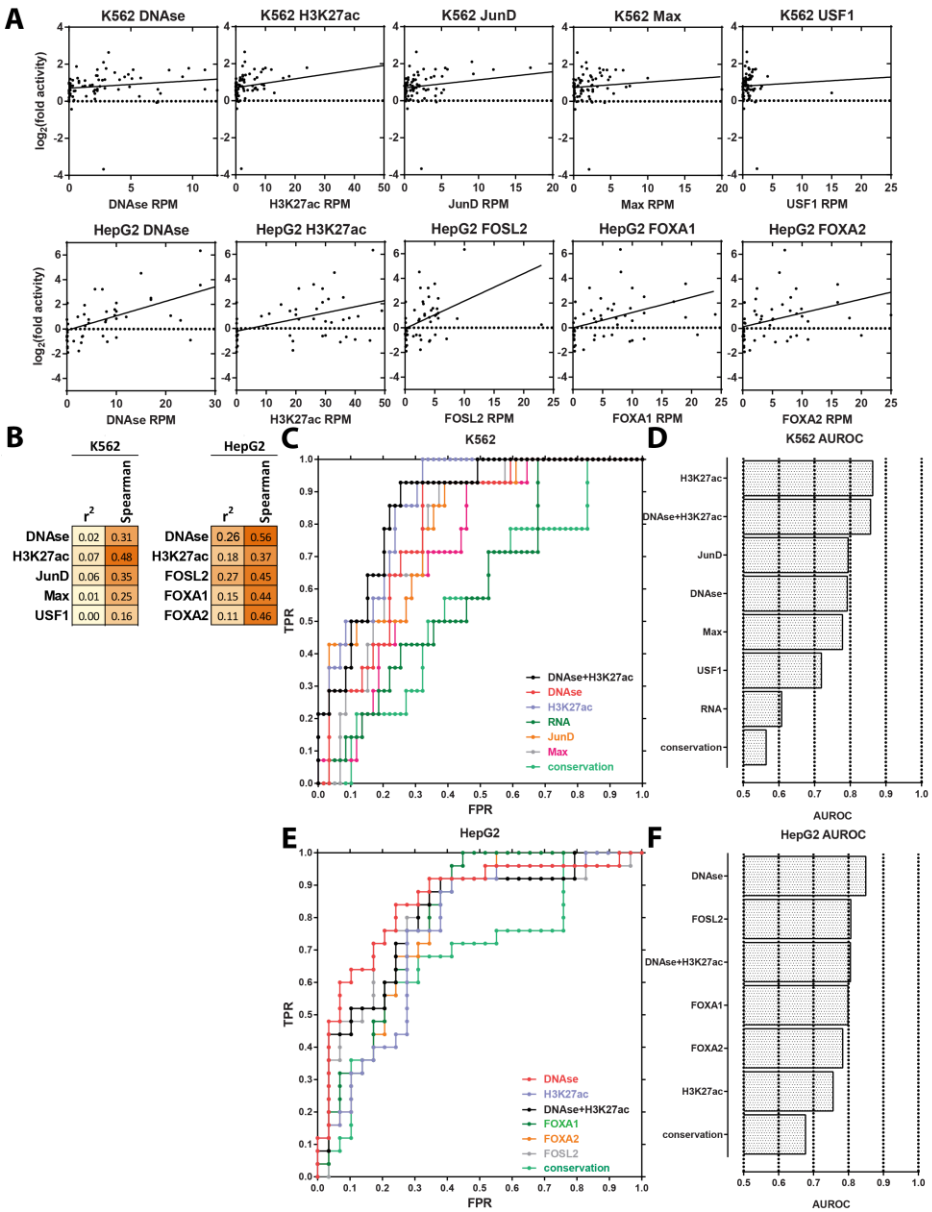
Correlation between regulatory activity and biochemical marks in erythroid cells. (A and B) Correlation between fold activity in K562 cells and DNAse hypersensitivity, H3K27ac, TAL1, and GATA1 occupancy in G1E and G1E-ER4 cells; (C) ROC curves showing biochemical mark predictivity of cRE fold activity;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

Supplementary figure 5A – K562/HepG2 Agnostic picks



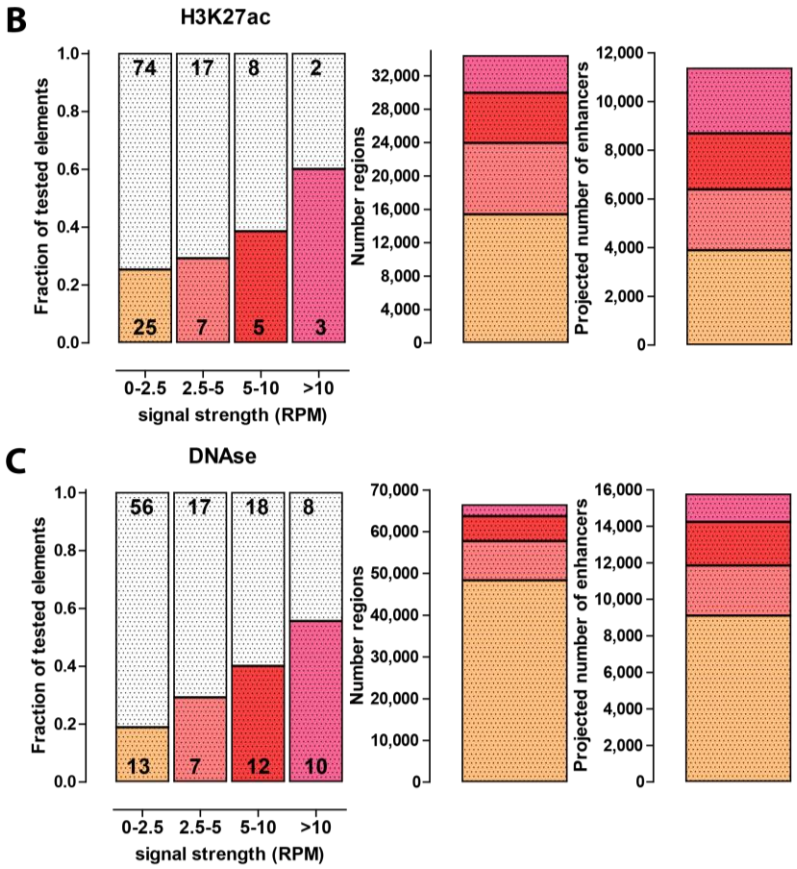
Functional assay testing of cRE regulatory activity in human immortalized cell lines. Fold activity across biological replicates (n = ???) and technical replicates (n = ??? for each biological replicate) is shown. Candidate REs were sorted rst by their DNAse status and then by their mean fold activity. 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 5B – K562/HepG2

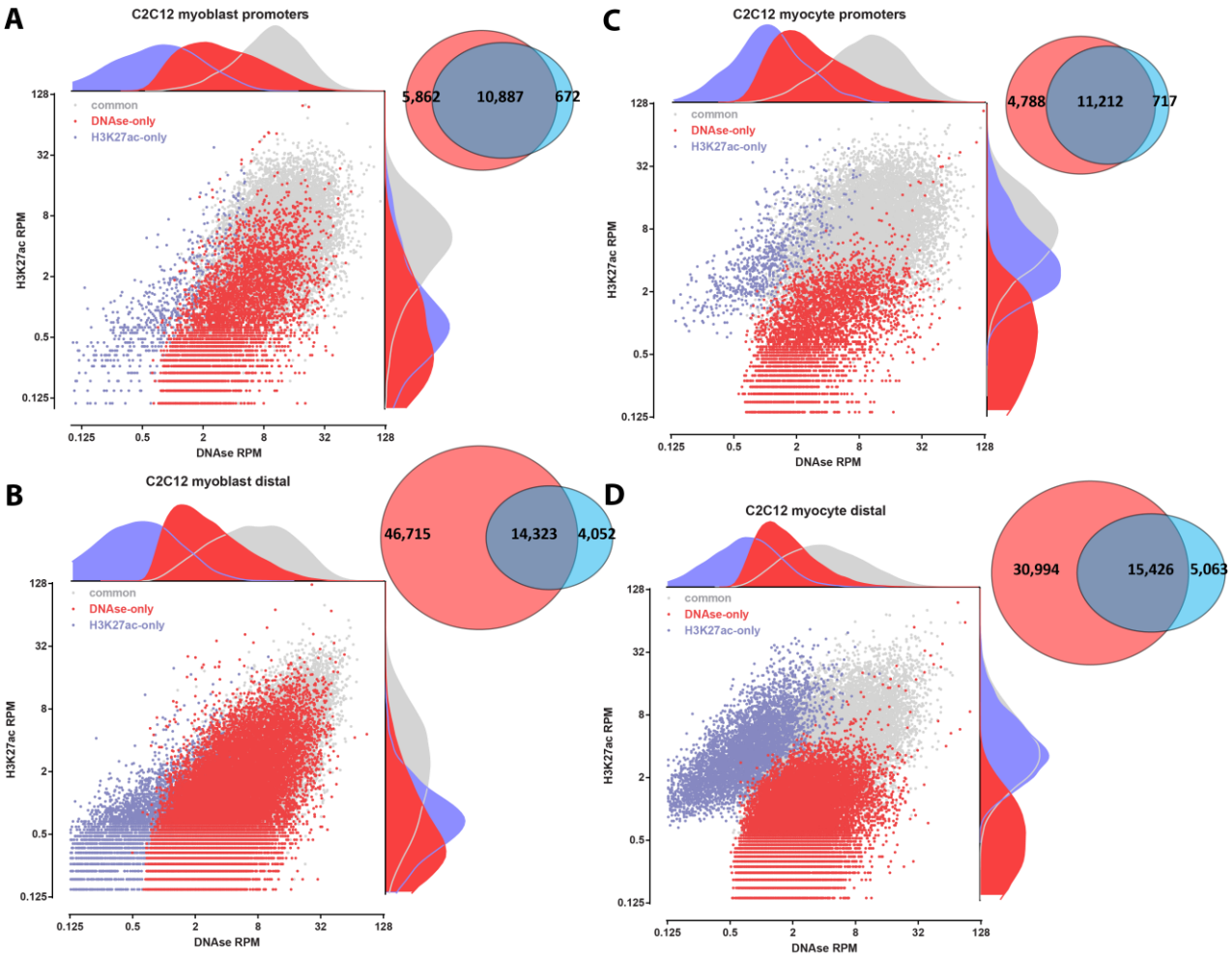


Correlation between regulatory activity and biochemical marks in human immortalized cell lines. (A and B) Correlation between fold activity in K562 cells and DNAse hypersensitivity, and transcription factor occupancy in K562 and HepG2 cells; (C) ROC curves showing biochemical mark predictivity of cRE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

Supplementary figure 6



Enhancer predictions in dierent signal classes for DNAseHS and H3K27Ac. (B)Fraction of elements active by H3K27Ac signal bin. Number of regions in the genome by signal strength bin. Number of projected enhancer ranked signal bin (low to high). (C)Fraction of elements active by DNAse signal bin. Number of regions in the genome by signal strength bin. Number of projected enhancer ranked signal bin (low to high).

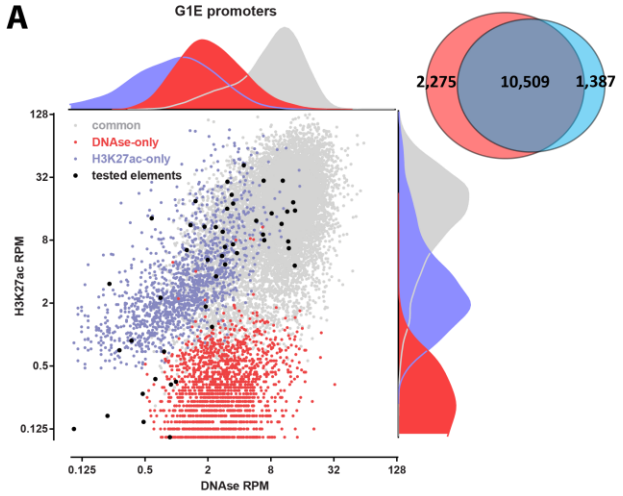


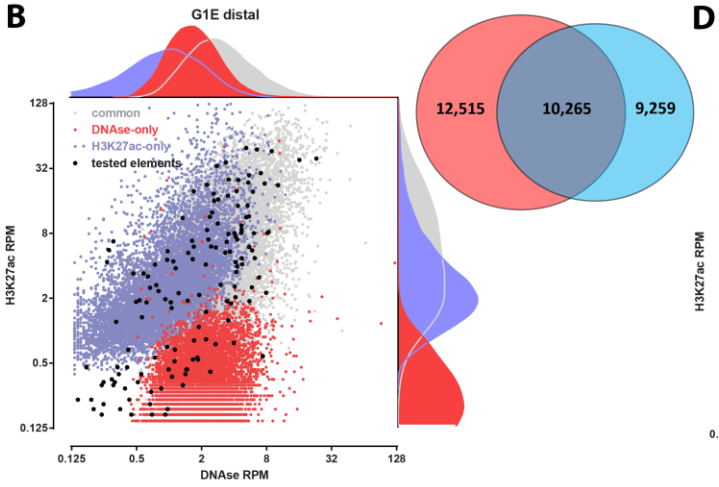
E

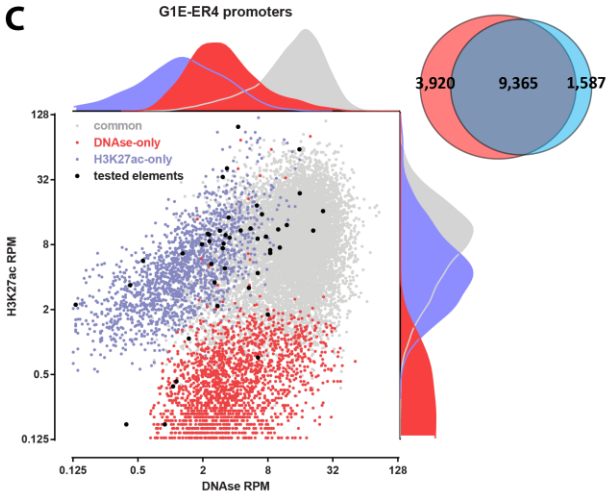
	promoter DNAse +			distal DNAse		
	common	MB only	MC only	common	MB only	MC only
	14,789	1,960	1,697	36,163	24,875	13,791
MB H3K27ac +	66.98%	26.17%	12.91%	33.30%	16.92%	2.46%
MB H3K27ac -	33.02%	73.83%	87.09%	66.70%	83.08%	97.54%
MC H3K27ac +	72.91%	25.51%	35.65%	36.14%	11.25%	21.21%
MC H3K27ac -	27.09%	74.49%	64.35%	63.86%	88.75%	78.79%

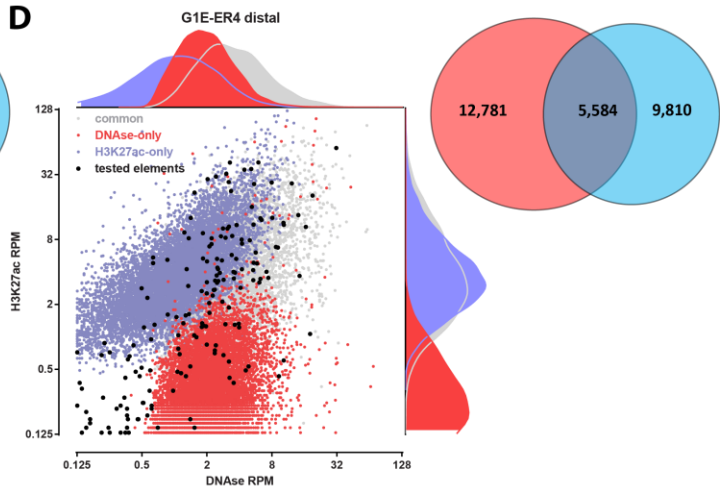
	promoter H3K27ac			distal H3K27ac		
	common	MB only	MC only	common	MB only	MC only
	10,568	1,678	3,665	8,667	9,243	12,266
MB DNAse +	98.20%	92.25%	81.17%	97.17%	89.69%	58.24%
MB DNAse -	1.80%	7.75%	18.83%	2.83%	10.31%	41.76%
MC DNAse +	97.41%	83.37%	87.15%	90.24%	61.02%	67.46%
MC DNAse -	2.59%	16.63%	12.85%	9.76%	38.98%	32.54%

Relationship between DNAse hypersensitivity and H3K27 acetylation during muscle dierentiation. (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 dierentiation for promoter-proximal and distal sites.







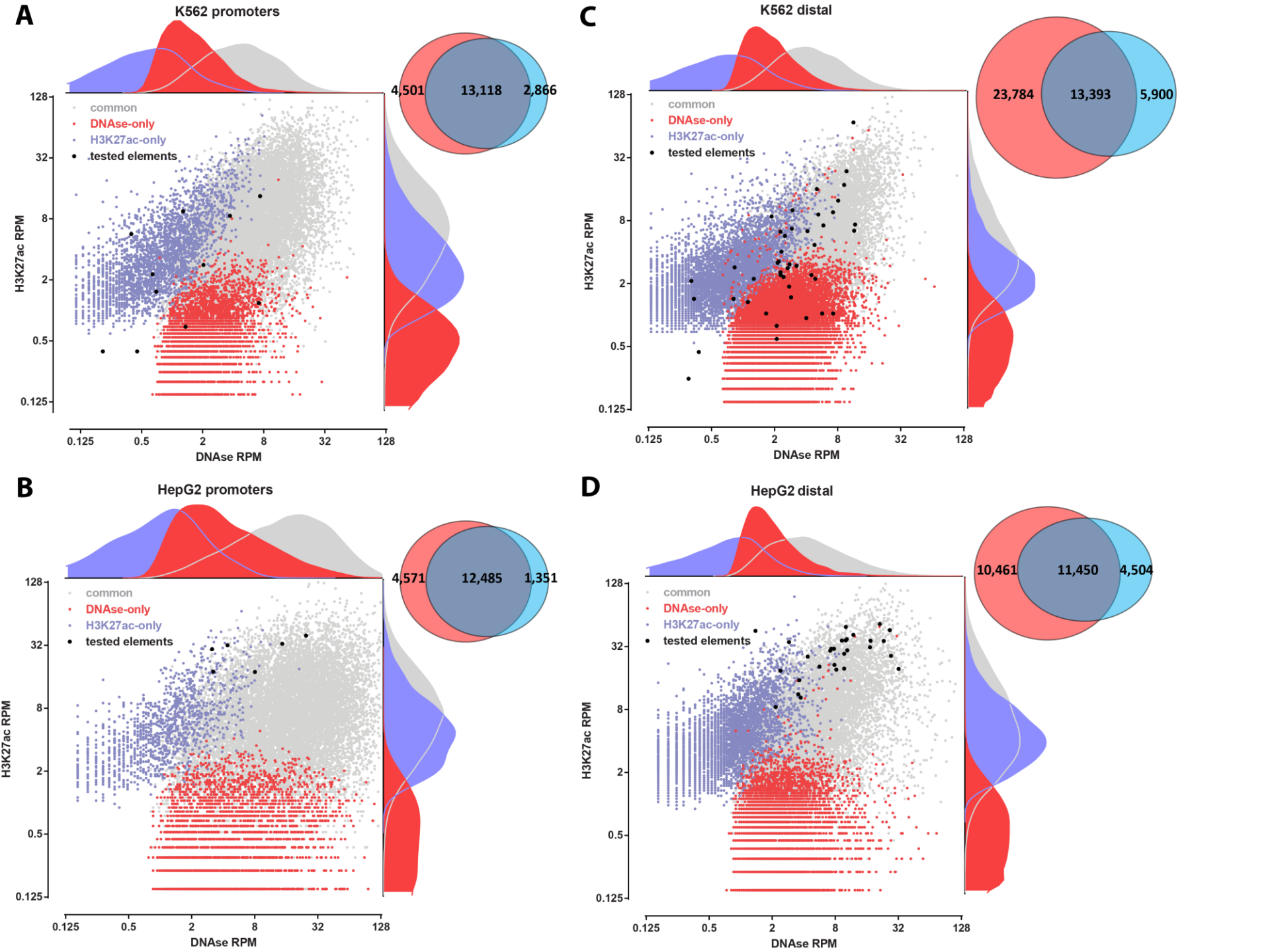


E

	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%

Relationship between DNAse hypersensitivity and H3K27 acetylation during erythroid dierentiation. (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 dierentiation for promoter-proximal and distal sites.


A
K562 promoters
common
DNAse-only
H3K27ac-only
tested elements
4,501
13,118
2,866
H3K27ac RPM
DNAse RPM
0.125
0.5
2
8
32
128
B
HepG2 promoters
4,571
12,485
1,351
C
K562 distal
23,784
13,393
5,900
D
HepG2 distal
10,461
11,450
4,504

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 hypersensi-

tive and H3K27ac-positive distal regions in HepG2 cells; the kernel density of the ChIP-seq/DNAse-seq signal distribution

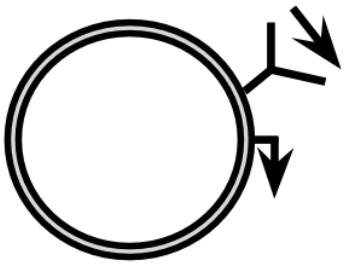
for each class of sites is overlaid over the scatter plots, and the distribution of tested cREs is shown in black.

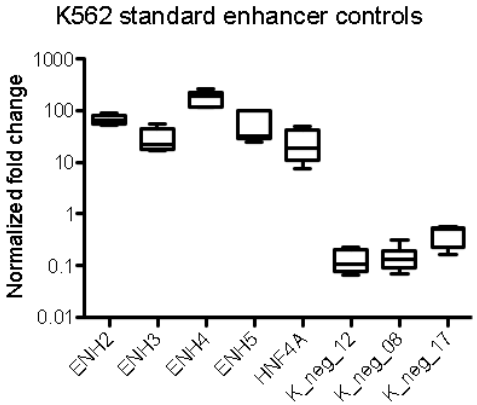
- QUARANTINED UNTIL ORI DATA IS LOOKED AT.

Supplementary figure 7

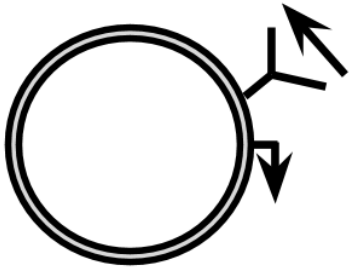
Testing control enhancer elements in multiple plasmid sites and in both orientations

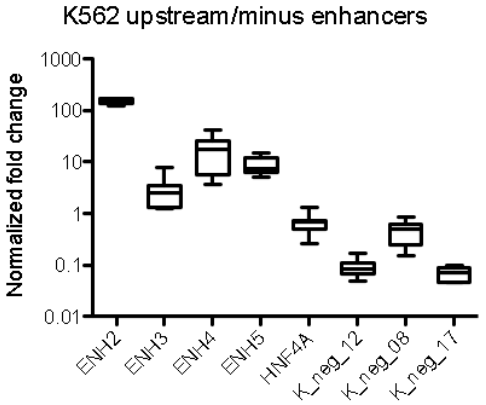
"Upstream", "plus"



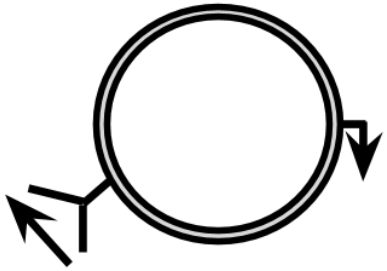


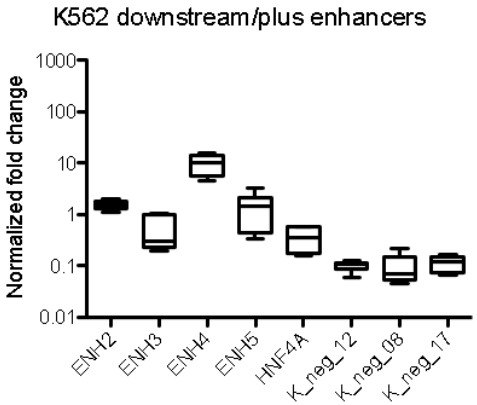
"Upstream", "minus"



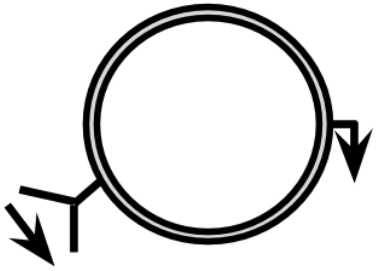


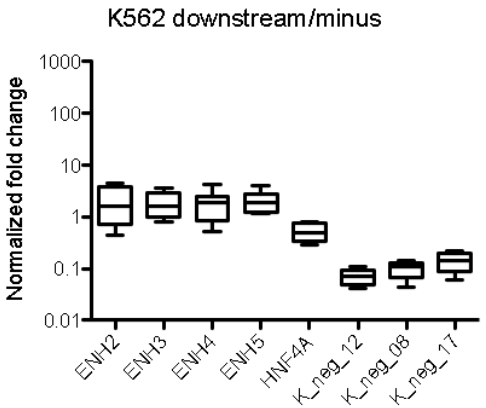
"Downstream", "plus"





"Downstream", "minus"



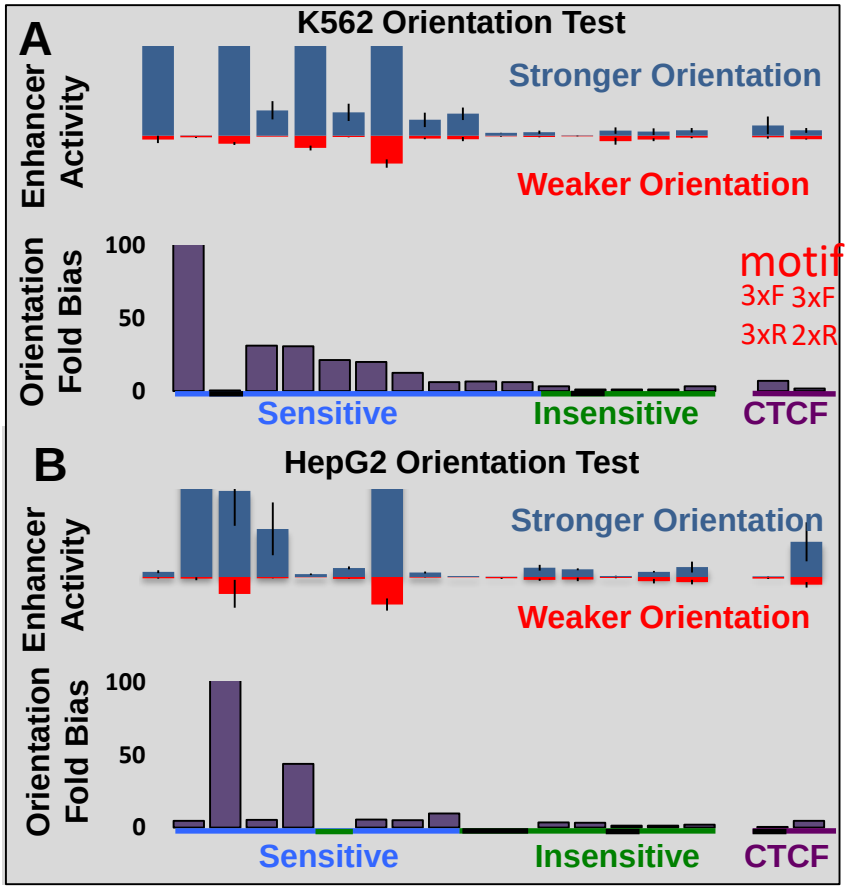


When tested 5’ both orientations function as enhancers.

Signal is lost at 3’ end in both orientations; Ori interference as cause?

Testing 5’ an issue? Modest set here says no;

Also picked mainly TSS distal (Gencode) cEnhancers



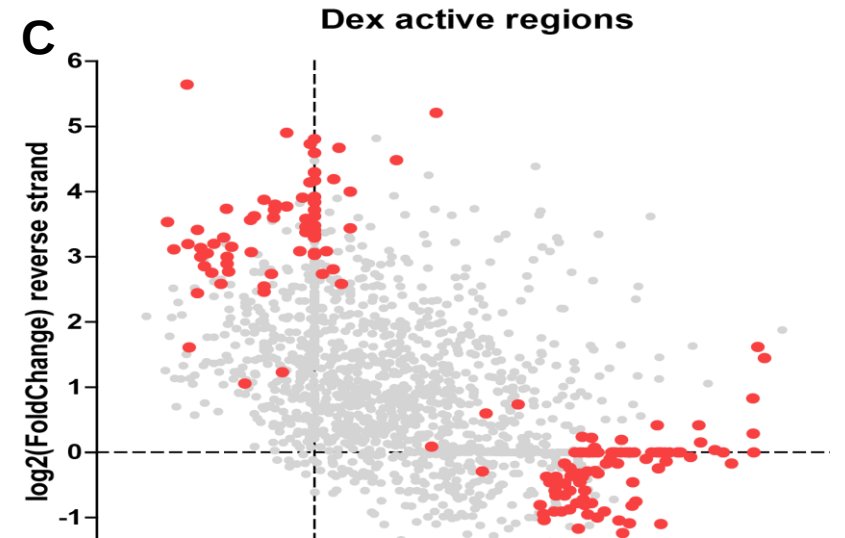


Figure 9

Recent experiments have found evidence of CTCF be orientation sensitive for its function in looping chron locations.

We tested a modest set of candidate enhancers in b directions to assert if the poor predictivity of the bio signal in the candidate regions could be explained.

We found only a modest increase in overall predictivity of enhancer function but that some eler filtered for annotated TSS and CAGE/rampage TSS ca to be strongly orientation biased for their enhancer

This orientation bias could not be tied to CTCF occup although we find a similar effect in TF and CTCF co-o enhancer regions.

Using Starr-Seq we can predict that ~10% of enhanc strongly orientation sensitive. (plot is not stripped fo TSS peaks)

Starr-Seq activity correlation to DNAse Signal
-> no better than muscle

Orientation test of enhancer elements in K562 cells.

Functional assay testing of the regulatory activity of both orientations for functional enhancer elements in K562 cells.

Fold activity in K562 cells across biological replicates (n = 2) and technical replicates (n = 6 for each biological replicate) is shown with the stronger fold activity orientation in blue and the weaker in red. Enhancer elements were selected; sorted first by whether they are orientation sensitive or insensitive; then sorted by the fold bias for the stronger orientation. CTCF occupied elements are marked by &; and do not represent a significant set of the tested elements.

YYY Legend for Starr-Seq part YYY