Biochemical signatures and enhancer activity of ENCODE candidate regulatory elements

Gilberto DeSalvo[1], Georgi K. Marinov[6], Christopher Partridge[2], Christopher M. Vockley[4,7], Nergiz Dogan[3], Ricardo Ramirez[8,9], Timothy E. Reddy[4,5], Ali Mortazavi[8,9], Ross C. Hardison[3], Richard M. Myers[2], and Barbara J. Wold[1]

[1]*Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, USA.*
[2]*HudsonAlpha Institute for Biotechnology, 601 Genome Way, Huntsville, AL 35806, USA*
[3]*Dept. of Biochemistry and Molecular Biology, Penn State University, 304 Wartik Laboratory, University Park, PA 16802, USA*
[4]*Center for Genomic & Computational Biology, Duke University, Durham, NC 27708, USA*
[5]*Department of Biostatistics & Bioinformatics, Duke University, Durham, NC 27708, USA*
[6]*Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305*
[7]*Department of Cell Biology, Duke University, Durham, NC 27708, USA*
[8]*Department of Developmental and Cell Biology, University of California Irvine, Irvine, CA 92697-2300, USA*
[9]*Center for Complex Biological Systems, University of California Irvine, Irvine, CA 92697-2280, USA*

Abstract

An aspiration for functional genomics is to identify all *cis*-acting regulatory elements in the genome, and to define their activities. Genome-wide transcription factor occupancy and chromatin state profiles are widely used to identify candidate regulatory elements (cREs) such as enhancers, promoters and insulators, based on their characteristic biochemical signatures. Multiple large-scale efforts, in particular the ENCODE and NIH Roadmap Epigenomic Mapping consortia, have generated such datasets from a wide variety of cell lines, tissues and developmental stages in both human and mouse, serving as the foundation for cataloging the cRE repertoire on a genome-wide scale. However, accurately predicting the actual regulatory activity of cREs based on functional genomic maps alone has turned out to be difficult, and it is not yet clear to what extent active regulatory elements *in vivo* can be identified from biochemical signatures. To address these issues, we carried out large-scale tests for transcriptional enhancer activity for hundreds of candidate enhancer elements (cEnhs) from five human and mouse cell types, including both immortalized cell lines and models for cell differentiation during developmental transitions. Our cEnh collections were selected using biochemical signature criteria ranging from simple individual transcription factor (TF) occupancy to integrative machine learning models over multiple functional genomic datasets, and were designed to include full-length cREs. We supplement these tests with a massively parallel reporter assay (MPRA) characterization of cEnhs in one of the cell lines we studied. We find that irrespective of the selection criteria used, ∼50% of cEnh elements showing enhancer activity. Examination of the functional predictivity of biochemical signatures reveals that the majority of cEnhs genome-wide are likely to exhibit modest regulatory activity. In the same time, most active enhancers in the genome are characterized by relatively modest biochemical signature strength, due to the much lower abundance of cEnhs characterized by very strong biochemical signals. Finally, we discuss our results in the context of current models of the regulatory effect of enhancers on their cognate genes. We expect our findings to help guide future efforts towards cataloging the functional repertoires of mammalian genomes.

Introduction

The complete and accurate understanding of the relationship between the human genome and its corresponding phenotypes requires the comprehensive characterization of its compendium of functional elements. The results of the many genome-wide epigenomic and transcriptomic studies carried out over the last decade reveal a remarkable picture, in which non-coding regulatory elements constitute the bulk of such functional regions in the genome[65,66], with the ex-

pression of each gene (protein coding or non-coding) being controlled by the integrated output of multiple proximal and distal enhancer, insulator and silencing elements.

The genome-wide mapping and characterization of noncoding regulatory elements is thus a major goal of the field, and features prominently among the objectives of the **ENC**yclopedia **O**f **D**NA **E**lements (ENCODE) consortium[39]. However, achieving it, although greatly aided by the advent of high-throughput sequencing and epigenomic tools, is still not a simple task.

The biochemical activity associated with the function of regulatory elements results in certain biochemical signatures that can be captured by epigenomic assays. For example, active promoters in eukaryotes are classically associated with the trimethylation of lysine 4 on histone 3 (H3K4me3)[35], as well as other biochemical signatures, such as DNAse hypersensitivity[38]. Active enhancer elements have been proposed to exhibit their own biochemical signature, featuring DNAse hypersensitivity, the H3K27ac and H3K4me1 histone marks, and occupancy by the p300 acetyltransferase[38,61,62], as well as by sequence-specific transcription factors (Figure 1A).

These biochemical signatures enable the compilation of lists of candidate functional elements (cREs), but they do not on their own allow the conclusive identification of any given element as functional. While functional regulatory elements exhibit characteristic biochemical signatures, the reverse (that the presence of a biochemical signature necessarily means function) cannot be inferred straightforwardly[36]. Such inferences are further complicated by the observation that biochemical signatures are not binary but instead exist on a continuum between strong outstanding features, on one hand, and what is probably biochemical noise, on the other. For example, it is far from clear that all transcription factor binding sites that can be reproducibly identified using ChIP-seq and related techniques are in fact functional REs[37]. Therefore, individual cREs in the lists compiled by efforts such as the ENCODE and mouseENCODE consortia[65,66] have to be subsequently directly tested and functionally characterized in detail.

The ultimate functional characterization of cREs will involve a combination of loss-of-function assays and direct assays for activity. The former have been until recently technically challenging, but are becoming more commonplace with the advent of large-scale CRISPR/Cas9-mediated mutagenesis techniques[32,33]. Nevertheless, most work in the field has been based on testing cREs for regulatory activity using an exogenous plasmid construct combining a cRE, a promoter and a reporter gene. Classically, such testing has been done by cloning individual cREs into plasmids (or other vectors) and then assaying the expression of the reporter gene (luciferase activity in cell lines or staining for LacZ activity in embryos[34,37]). Numerous developmental enhancers have been characterized following that approach[34,47,48], starting from lists of cREs complied based on comparisons of evolutionary conservation and/or biochemical signatures derived from ChIP-seq datasets. However, such studies have often focused only on the most outstanding biochemical signatures[34], thus obtaining very high success rates that are likely to be nonrepresentative with respect to the genome-wide population of cREs.

High-throughput sequencing has enabled the development of assays that go beyond the testing of individual cREs, one by one; instead, very large numbers of sequences are analyzed in parallel, with the readout being based on sequencing DNA tags associated with the cRE or of the cRE itself. These are usually referred to as MPRAs[63], and in the last few years a number of variations of the principle have been successfully applied to a multitude of biological problems and systems[49–51,53,54,59,60], including the question of testing cREs for activity within the context of the ENCODE Consortium[55,56,58]. However, several issues complicate the interpretations of MPRA experiments.

First, the nature of MPRA designs is such that the elements tested are very short, in the 80–250bp neighborhood. This is significantly shorter than tens of thousands of blocks of conserved noncoding sequence that can be identified in the human genome by comparative genomic analysis (Supplementary Figure 1). Thus to what extent complete REs are assayed and what the corresponding false negative rate is have always been an obvious concerns regarding MPRAs.

Second, given that it is very difficult to control the number of constructs going into each individual cell in transfection experiments, and that an MPRA features large numbers of different cREs being tested in the same time, there is significant potential for crosstalk between active and inactive cREs that end up in the same cell, resulting in numerous false positives.

While the latter concern can be alleviated to an extent through the use of genome-integrated MPRA constructs[52,57], the short length of constructs tested remains a significant issue, and one that might be behind the low positive rates reported by MPRAs in the past[41]. There is therefore a major gap in the field that needs to be filled by testing a large number of individual constructs of large size (500-1000 bp) using a traditional luciferase assay, and examining the performance of biochemical signature predictions based on the resulting data.

To this end, as part of the ENCODE Project Consortium's efforts towards functional validation of cREs, we tested the regulatory activity of hundreds of candidate enhancer elements (cEnhs) using constructs of such lengths (Supplementary Figure 2) in several diverse mammalian cell lines, including both mouse and human systems. These cEnhs were selected from a wide range of biochemical signature strengths (Supplementary Figure 3), using both TF-centric selection criteria (identifying cEnhs based on ChIP-seq data for individual TFs) and machine learning "TF-agnostic" approaches (designed to find combinatorial signatures of enhancer elements from multiple epigenomic maps of histone modifications and DNAse hypersensitivity) for defining cREs.

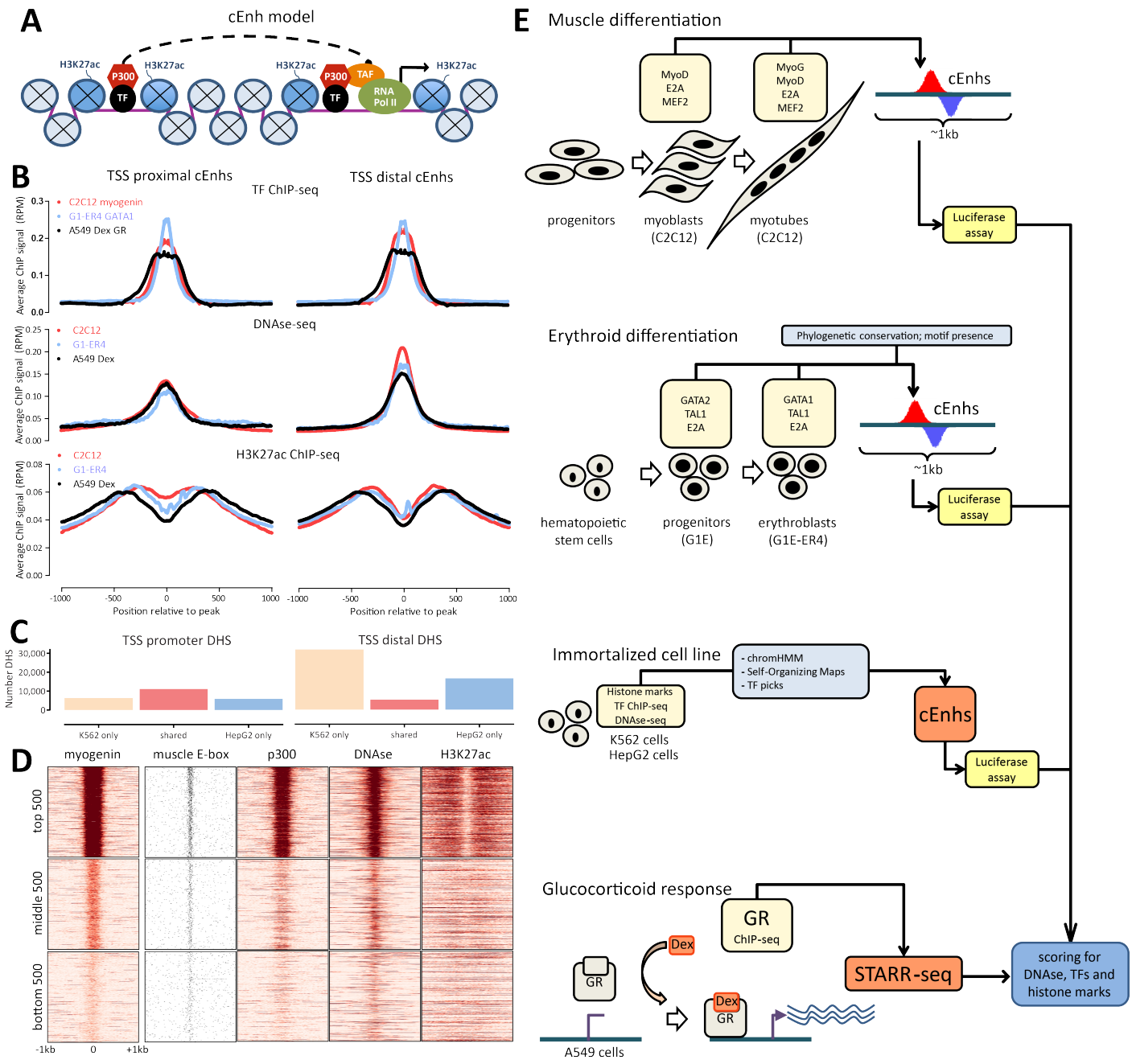


Figure 1: Biochemical signatures and functional testing of candidate enhancer elements (cEnhs) in mammalian genomes. (A) Biochemical signatures of cEnhs and promoters. Active enhancers are characterized by DNAse hypersensitivity due to nucleosome depletion, by p300 occupancy and by H3K27ac, as well H3K4me1 (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 GR upon dexamethasone stimulation of human A549 cells), DNAse hypersensitivity and H3K27ac; (C) 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; (D) The distribution of biochemical

(legend continued on next page)

We find that in general ∼50% of both TF-selected and TF-agnostic cEnhs showed significant enhancer activity in transfection assays, observing similar proportions across all cell lines and conditions examined. We observe that the presence of TF recognition motifs in cEnhs displays no correlation with enhancer activity. Our results indicate that DNAse and H3K27ac are generally more predictive of enhancer activity than TF occupancy alone, and that simple biochemical signatures such as the combination of chromatin accessibility and H3K27ac are specifically predictive of enhancer activity compared to regions of the genome that lack them. However, the presence of these signatures is not a strong indicator of enhancer activity as nearly half of cEnhs exhibiting them are not significantly active in transfection assays.

We also observe a positive correlation between biochemical signal strength and enhancer activity in transfection assays, with the highest fraction of cEnhs exhibiting significant enhancer activity being found among the most strongly occupied cEnhs. However, first, even among that latter group a large fraction of cEnhs displays no discernible enhancer activity, and second, because the distribution of biochemical signal strength is highly skewed towards weaker ChIP-seq and DNAse-seq peaks, the bulk of active enhancers in the genome are expected to be found among the larger population of cEnhs with modest biochemical signatures. Of note, enhancer activity as measured by transfection assays also exhibits a skewed pattern, with a smaller number of very highly active enhancers and a larger number of weaker ones.

Finally, we corroborated these findings by applying the STARR-seq MPRA to assay the activity of thousands of genomic regions occupied by the glucocorticoid receptor (GR) in stimulated A549 cells.

We expect that our findings will help guide efforts towards the comprehensive cataloging of functional elements in the human genome, and we discuss the implications of our findings in the context of models of gene regulation mediated by the action of distal enhancers.

Results

Large-scale functional activity testing of full-length cREs

One of the major goals of the ENCODE Project is to identify all functional regulatory elements controlling gene expression in the genome, to which end it has carried out mapping of DNAse hypersensitivity, dozens of histone modifications, numerous sequence-specific transcription factors, and RNA transcripts in hundreds of cell types, in both human and mouse[22,39,65,66]. As active regulatory regions typically exhibit certain biochemical signatures, these datasets provide a compendium of candidate regulatory elements. Such regions can be defined using highly multidimensional inputs incorporating multiple biochemical measurements, but because considerable redundancy exists between individual biochemical marks, cREs can also be identified following simpler rules. For example, active enhancers and promoters are typically occupied by sequence-specific transcription factors, marked by H3K27ac and exhibit DNA hypersensitivity (Figure 1B), and cEnhs can therefore be identified using TF occupancy, by the overlap of DNAse hypersensitive sites (DHS) and regions marked by H3K27ac, or by all three. As of the writing of this manuscript, based on DNAse-seq maps and histone marks profiles, more than a million and nearly half a million cREs in humans and mouse, respectively, have been identified by the ENCODE consortium[22].

However, there need not be a strict one-to-one relationship between these biochemical signatures and functional REs[36], as their presence alone does not necessarily imply that the cRE plays an active regulatory role. In addition, biochemical marks on themselves do not provide direct understanding of how exactly functional REs are specified or exercise their function. An additional highly useful criterion for assessing functionality is evolutionary conservation of cREs between the genomes of distant species, as functional elements are usually subject to selective constraint at the sequence level. However, the absence of conservation on its own does not imply nonfunctionality, as recently evolved lineage-specific REs do not appear as conserved in comparative genomic analyses. Therefore cREs identified

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 motif. (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 in K562 and HepG2 cells. 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.

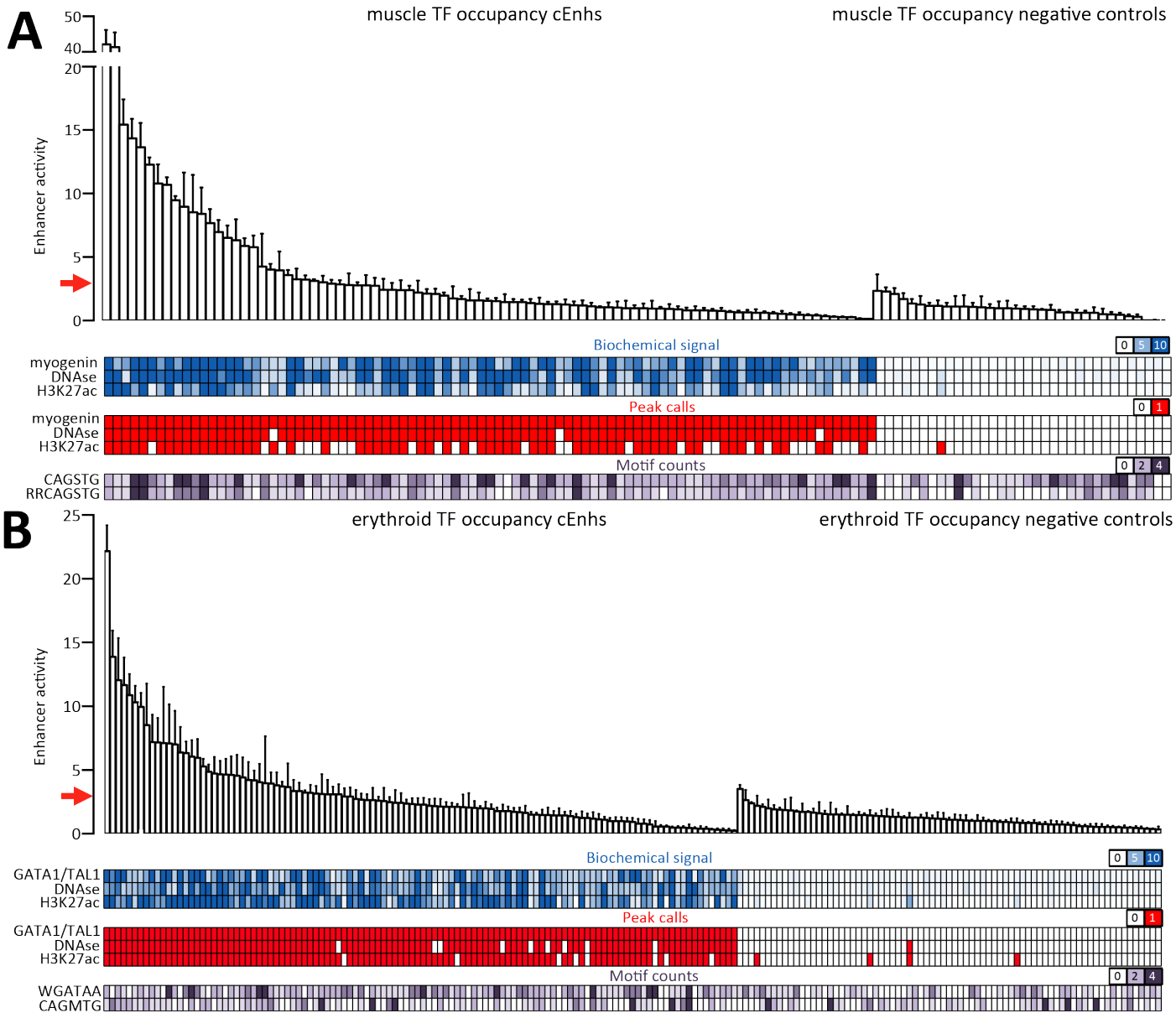


Figure 2: Functional testing of cEnh regulatory activity in mammalian cells. (A) Functional assay testing of cEnh regulatory activity in the context of muscle differentiation. Shown is luciferase assay fold activity in differentiated C2C12 myocytes across technical replicates ($n = 4$). The red arrow corresponds to the mean fold activity threshold above which elements are considered active. In addition, for each cEnh element, 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 element, 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.

using functional genomics tools have to be directly tested for function and subsequently dissected in detail if they are to be comprehensively understood.

In recent years, multiple high-throughput approaches for measuring regulatory activity have been devised, relying on a sequencing readout of the effect of a given cRE or custom-designed DNA sequence on the expression of a reporter gene[49–51,53–56,58–60]. On their own MPRAs are very powerful, however, they suffer from several shortcoming when it comes to testing cREs, in particular the short length (typically 80 to 250 bp) of the segments of DNA tested by most of them, which is likely to be significantly shorter than the size of functional regulatory elements in mammalian genomes.

To assess how prominent this issue might be, we examined the distribution of the lengths of conserved noncoding segments (i.e. excluding sequences overlapping with or in the vicinity of annotated exons) in the human genome (Supplementary Figure 1), and found that tens to hundreds of thousands (depending on the definition) of such blocks fall outside the range of testing of MPRAs. It is thus possible that MPRAs using short sequences exhibit substantial numbers of false negatives as they cannot assay the activity of complete, full-length REs. An additional concern with MPRAs is the possibility of cross-talk between different REs when multiple episomal constructs end up being transfected in the same cell, resulting in false positives. Alternative approaches towards testing the functionality of cREs are therefore needed.

To address these issues, we carried out large-scale testing of cEnhs identified on the basis of biochemical signatures in a variety of mammalian systems using luciferase activity assays, which allow for much larger segments of DNA to be tested for enhancer activity. We aimed at comparing the predictivity of multiple approaches for identifying functional cEnhs, and at incorporating in our analysis a diversity of biological systems (Figure 1E), including immortalized cell lines (the human K562 and HepG2 cell lines), model systems for major developmental transitions (myogenesis and erythropoesis), and model systems for cellular response to exogenous signaling stimuli (activation of the glucocorticoid receptor in the prostate cancer cell line A549).

Immortalized cell lines were targeted as they have been extensively studied by the ENCODE consortium[39,66] and they are the source of a significant portion of ENCODE data, while differentiation and external signaling stimulation represent the two main types of dynamic transition of cellular states associated with regulatory alterations of chromatin states: the slower and typically irreversible differentiation of one cell type into another, and the much faster and reversible cellular response to signaling molecules.

The process of myogenesis transforms undifferentiated precursor myoblast cells into differentiated myocyte muscle cells, and is primarily regulated by four key bHLH TFs known as Myogenic Regulatory Factors (MRFs) along with numerous cofactors, such as MEF2, E2A, HEB, Pbx1 and others[16–19]. The epigenomic and occupancy landscape of these and other factors involved in the process is illustrated for reference in Supplementary Figure 5. These factors were profiled in the mouse C2C12 cell line[12], which has been for decades the main model system for studying myogenesis, a wealth of functional genomic data has been generated for it, and it was thus naturally also the focus of our study. The key specification MRF and the one expressed at high levels in myoblasts is MyoD, while myogenin is the most important differentiation TF and its expression is induced upon the onset of the process; the other two MRFs are Myf5 and Myf6. MyoD and myogenin occupy a highly overlappings set of sites (Supplementary Figure 5), the majority of which contain the classical muscle E-box sequence motif CAGSTG often in the extended RRCAGSTG form[20,21]. While the majority of ChIP-seq peaks contain this motif, these peaks include only a tiny fraction of the occurrences of the motif in the genome, underscoring the highly selective nature of in vivo occupancy by MRF TFs.

Lineage commitment during the process of erythropoesis is accomplished through the so called GATA switch[10]. The onset of terminal differentiation is marked by the replacement of the GATA2 transcription factor at thousands of occupied sites by GATA1, which then regulates the expression of genes involved in red blood cell development and functions, with SCL/TAL1 being an important cofactor of GATA1, often forming closely spaced heterodimers with it[7–9,11]. GATA1 binds to a WGATAA consensus recognition motif, while TAL1 is a bHLH factor targeting a CAGMTG E-box[11]. Mouse G1E cells have served as the key model system for erythropoesis for many years[6]. These cells are derived from in vitro differentiated mouse embryonic stem cells in which the endogenous GATA1 gene has been knocked out; a subclone of them, termed G1E-ER4, expresses constitutively a GATA1-ER fusion that can be specifically activated by estradiol exposure, allowing for differentiation to be triggered rapidly and in a controlled manner[4,5]. The epigenomic and occupancy landscape of the key TFs involved is illustrated for reference in Supplementary Figure 9, and it served as the basis for our cRE selections.

The transcriptional response to glucocorticoids, on the other hand, is characterized by a more rapid kinetics of gene expression activation, and by its general reversibility. We used the response of A549 cells to activation of the GR transcription factor by the cortisol analog dexamethasone (Dex) as a model system for our study. Upon activation by Dex GR rapidly associates with thousands of sites along the genome, both directly through its cognate motif and indirectly through association with cofactors such as AP-1[1–3], leading to changes in the expression of hundreds of genes. The epigenomic landscape of A549 cells during GR activation is illustrated for reference in Supplementary Figure 16.

While the full catalog of REs in the genome includes promoters, insulators, enhancers, silencers, and others. for the purposes of this study we focused on candidate transcriptional enhancers. A major reason for this choice is that cEnhs constitute the bulk of cREs distinguishing different cell types from each other, in contrast to, for example, active promoters, a major fraction of which is shared (Figure 1C, Supplementary Figures 4, 6, 10 and 12).

We also aimed to represent the full spectrum of cEnh biochemical signatures (Figure 1D and Supplementary Figure 3), as multiple studies have shown that the landscape of transcription factor occupancy, DNAse hypersensitivity and histone modification maps includes many more weaker sites than very strong peaks[27,36].

We applied several different strategies for compiling lists of cEnhs to be tested, broadly divided into two categories: TF-centric, based on TF ChIP-seq datasets, and TF-agnostic, based primarily on chromatin state signatures and evolutionary conservation.

In the context of muscle differentiation, we selected cEnh regions based on myogenin ChIP-seq data in differentiated C2C12 myocytes. We randomly selected a set of regions ($n = 89$) spanning the full range of myogenin occupancy levels, of which 88 contain a CAGSTG and 84 contain the extended RRCASGTG E-box. We also selected additional cEnh elements (XX n = ?? XX) associated with genes playing a well characterized role in the muscle development (XXX list genes XXX); the inclusion of these cEnhs allowed us to test whether cEnhs nearby known functionally relevant genes exhibit a higher levels of functionality that the genome-wide average. We also selected a group of negative controls ($n = 23$) out of a set of well characterized enhancer elements active in T cells and in neurons and not being occupied by myogenin in C2C12 cells; 21 of them contained E-box motifs. These negative controls were selected in order to test the specificity of biochemical signatures for prediction of functional activity in a given cell type. A second set of negative control elements ($n = 11$), 6 of them containing E-box motifs, were selected from regions without biochemical marks but located near genes highly expressed in C2C12 cells and near ChIP-seq positive regions. These were selected in order to assess the baseline functional activity of biochemically neutral regions.

TF-centric erythroid cEnhs ($n = 114$) and negative controls ($n = 74$) were selected on the basis of ChIP data for GATA1 and TAL1 in G1E-ER4 cells (see the Methods section for further details). Of note, even though they are not significantly occupied by GATA1 or TAL1, 45 of the negative control elements contain a WGATAA motif and 40 contain a CAGMTG E-box. Functional testing of cEnhs identified in G1E cells was then carried out in human K562 cells, which represent a similar developmental state but have the advantage of being much more easily transfectable.

XXX K562 TF selection XXX

TF-agnostic enhancer selection was based on evolutionary conservation and on the integration of measurements of multiple histone modifications and open chromatin.

A set of evolutionarily conserved erythropoetic cEnhs ($n = 46$) were selected by requiring both strong overall sequence conservation as assessed by multiple genome alignments of a collection of mammalian genomes and the conserved presence of WGATAA motifs.

Multiple computational approaches for integrating high-dimensional collections of functional genomic datasets into a small set of chromatin states have been devised over the last few years and applied to the problem in ENCODE cell lines, including the Hidden Markov Model-based Segway[23] and chromHMM[24], as well as Self-Organizing Maps[25] (SOM). We selected cEnhs in K562 cells based on Segway and chromHMM chromatin state assignments and the presence of DNAse hypersensitivity and H3K27ac ($n = 30$), with elements lacking both marks used as negative controls ($n = 21$). We also selected cEnhs based on SOMs trained on DNAse-seq and histone mark ChIP-seq data over multiple ENCODE cell types; these cEnh elements were picked so that they were specifically in an open chromatin state and marked by histone modifications associated with enhancer activity in HepG2 cells ($n = 32$). Elements lacking both marks ($n = 18$) were used as negative controls.

The collection of cEnh elements that we tested includes primarily long fragments, between 500 and 1000 bp (Supplementary Figure 2), thus most likely encompassing complete functional regulatory elements provided that such are indeed present within the cEnh region.

We tested these elements for functional activity using luciferase reporter assays XXX some details but not too many on how exactly this was done, refer to Methods for the rest XXX.

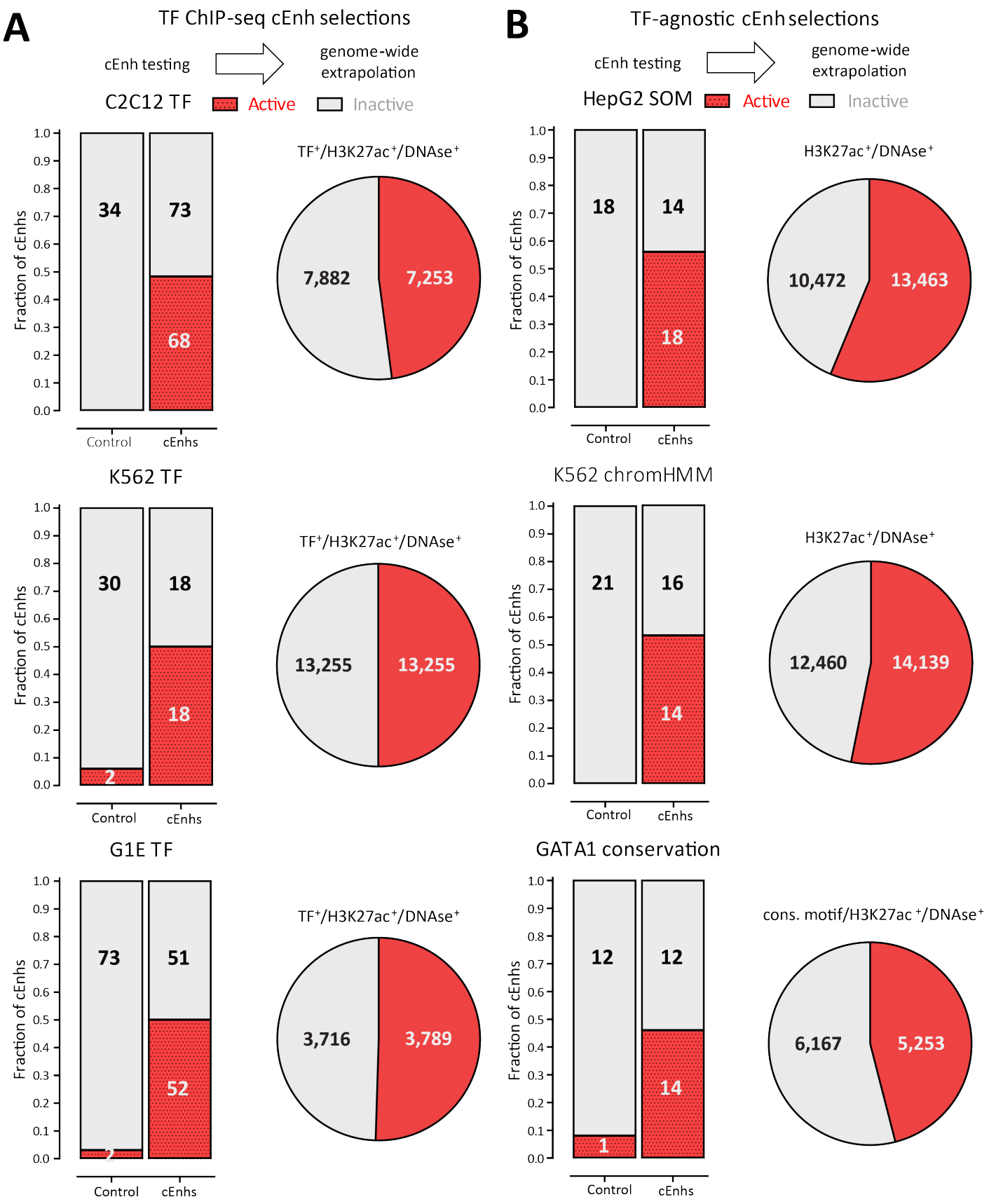
In addition to luciferase assays, we also incorporate ChIP-STARR-seq data from untreated and Dex-stimulated A549 cells[41]. The ChIP-STARR-seq libraries were generated by cloning GR ChIP-seq DNA fragments into a STARR-seq vector, then transfected into A549 cells, and RNA was sequenced from untreated and Dex-treated cells.

We subsequently compared measured activity with cEnh predictions based on various biochemical signatures to examine their predictivity of functional cEnhs.

Regulatory activity of TF-centric and TF-agnostic cEnh selections

Figure 2 shows the measured enhancer activities for TF-centric cEnh selections and negative controls in myogenesis (Figure 2A, and Supplementary Figure 7) and erythropoesis (Figure 2B). We found XXX out of XXX, or XX% of muscle cEnhs to pass the threshold of activity in contrast to XXX out of XXX, or XXX% of muscle negative controls. Similarly, XXX out of XXX, or XX% erythropoetic cEnhs and XXX out of XXX, or XX% of negative controls, were found to be active.

In K562 and HepG2 cells, we found XXX out of XXX, or XX%, and XXX out of XXX, or XX% of TF-centric cEnhs


A
TF ChIP-seq cEnh selections
cEnh testing
genome-wide extrapolation
C2C12 TF
Active
Inactive
Fraction of cEnhs
34
73
68
Control
cEnhs
TF+/H3K27ac+/DNAse+
7,882
7,253
K562 TF
30
18
2
18
13,255
13,255
G1E TF
73
51
2
52
3,716
3,789
B
TF-agnostic cEnh selections
HepG2 SOM
18
14
18
H3K27ac+/DNAse+
10,472
13,463
K562 chromHMM
21
16
14
12,460
14,139
GATA1 conservation
12
12
1
14
cons. motif/H3K27ac+/DNAse+
6,167
5,253

to be active, respectively, in contrast to XXX out of XXX (XX%) and XXX out of XXX (XX%) in the corresponding sets of negative controls (Supplementary Figure 13).

We then examined the subsets of cEnhs in each system bearing the simultaneous biochemical signature of TF ChIP-seq occupancy, H3K27ac demarcation and DNAse hypersensitivity, which would represent the most likely to be functional cEnhs. We observed 48.2% active such cEnhs in myogenic cEnhs, 50% among erythroid ones, and 49.5% in the set of cEnhs selected from human immortalized cell lines (Figure 3A).

The same analysis for TF-agnostic cEnh selections showed that XXX out of XXX (XX%) of SOM picks were active in HepG2 cells, XXX out of XXX (XX%) of chromHMM/Segway picks were active in K562 cells SUPP FIGURE SIMILAR TO FIG.2 FOR TF-agnostic picks, and XXX out of XXX (XX%) of GATA1 conservation selections exhibited significant activity SUPP FIGURE SIMILAR TO FIG.2 FOR TF-agnostic picks. Restricting our analysis to cEnhs positive for H3K27Ac and DNAse hypersensitivity, we observed 56% active cEnhs among HepG2 SOM picks, 46.6% among K562 chromHMM/Segway picks, and 46.1% for GATA1 conservation selections. We note that HepG2 SOM selections were biased towards more strongly H3K27ac/DNAse positive regions (Supplementary Figure 3), which might explain the higher levels of observed activity within that set.

Overall we find similar levels of activity in both the TF-centric and TF-agnostic sets of cEnhs predictions, around 50%, XX calculate proportions p-values once numbers are final XXX. Using the total number of cEnhs in each cell type and the observed proportions of cEnhs active in functional assays, we estimate that there are XXX active myogenin$^+$/H3K27ac$^+$/DNAse$^+$ enhancers in C2C12 cells, XXX GATA1$^+$/H3K27ac$^+$/DNAse$^+$ ones in erythroid cells, XXX TF$^+$/H3K27ac$^+$/DNAse$^+$ ones in K562 cells. Self Organizing Maps combined with functional testing predict XXX active enhancers in K562 cells while chromHMM/Segway predict XXX ones. The combination of GATA1 conservation and the H3K27ac/DNAse biochemical signature predicts XXX active enhancers in erythroid cells.

We also note that in all functional tests we carried out using luciferase assays, we find a skewed distribution of activity, similar what is observed for biochemical signal strength in ChIP-seq and other functional genomic experiments. A small number of cEnhs appear to be highly active, while the majority of even active cEnhs exhibits only modest activity (Figure 2, Supplementary Figure 7 and Supplementary Figure 13).

The bulk of enhancers in a given cell type are marked by modest biochemical signatures

We next examined the relationship between biochemical signature strength and functional activity. Figure 4A shows the myogenin occupancy landscape in the neighborhood of the mouse *Myog* and *Mybph* genes (both of these are well known muscle genes). A range of biochemical signal strengths is observed in cEnhs around these loci, from low to high, and this is typical for all biochemical marks. Figure 4B shows the genome-wide distribution of biochemical strength for cEnhs defined by myogenin occupancy in C2C12 cells; the bulk of cEnhs are characterized by low-level occupancy by myogenin.

In order to asses the relationship between occupancy strength and functional activity, we split the cEnhs we tested in C2C12s into four bins ("low", "medium", "high", and "top") according to the level of myogenin ChIP-seq signal observed (Figure 4C). We find that the fraction of active cEnhs increases steadily with the strength of myogenin signal, with only ∼20% of low-myogenin cEnhs exhibiting significant activity in contrast to ∼55% of the most strongly occupied ones. These observations superficially imply a close relationship between biochemical signal strength and functional activity, however, we note that even in the latter group a large fraction (nearly half) of cEnhs is inactive when directly functionally tested.

We then asked how many active enhancers genome-wide are likely to be found among each portion of the signal strength distribution. Figure 4D shows the extrapolated numbers of active enhancers in each bin of myogenin occupancy strength. While the strongest myogenin sites are mostly likely to be functionally active, the much greater numbers of weaker sites mean that the 75% of active en-

Figure 3 *(preceding page)*: 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 fractions 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 fractions 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) or DNAse$^+$/H3K27ac$^+$ regions with a conserved GATA1 motif (for GATA1 conservation selections) in the genome.

hancers in muscle cells are expected to be found among the sites belonging to the "low" and "medium" bins. Similar observations apply to other biochemical marks (Supplementary Figure 15).

We corroborated these finding by examining our A549 GR ChIP-STARR-seq data. Comparing STARR-seq reads to their input DNA libraries, and only including cEnhs with sufficiently deep representation in sequencing libraries (see the Methods section for details), we identified ∼15% of GR cEngs to be significantly active in Dex-stimulated A549 cells and ∼10% to be active in untreated cells (Supplementary Figure 15A). This fraction is considerably lower than what is observed for cEnhs with luciferase assays, an observation that is explained by a combination of the generally lower sensitivity of MPRAs and the shorter fragments being represented in STARR-seq libraries, which likely do not capture complete regulatory elements (although we note that we also carried out an activity analysis at the level of individual DNA fragments and did not observe longer DNA fragments to be preferentially active compared to shorter ones; Supplementary Figure 15C). We also note that, similar to luciferase functional assays, the majority of active enhancers in ChIP-STARR-seq datasets exhibit moderate levels of activity, with only a small minority of very highly active functional enhancers (Supplementary Figure 18).

The distribution of GR ChIP-seq signal strength in cEnhs tested by ChIP-STARR-seq is not as skewed in favor of low-occupancy sites as it is in other contexts (Figure 4E), which is due to the fact that representation in the ChIP-STARR-seq input libraries is biased towards stronger sites and that we excluded cEnhs with insufficient number of reads to evaluate their activity. Nevertheless, we do see more weaker sites than stronger ones, and we also observe much higher levels of activity (∼40%) in the "top" bin (Figure 4F) than within the "low" and "medium" ones (∼10%).

Thus we can conclude that even though the most visible biochemical signatures are most likely to correspond to active regulatory elements, in most biological contexts most functional enhancers in fact reside among the population of cEnhs characterized by only modest biochemical signatures, underscoring the complexity of the task of identifying active cEnhs from biochemical measurements alone.

Biochemical signature strength is not strongly correlated with enhancer strength

Finally, we examined the quantitative correlation between biochemical signal and functional activity. Figure 5 shows the distribution of active and inactive muscle (Figure 5A) and erythroid (Figure 5B) cEnhs relative to the spectrum of DNAse and H3K27ac signal genome-wide. While highly active enhancers are more often found among the most strongly H3K27ac$^+$/DNAse$^+$ regions in muscle cells, overall there is only modest correlation between biochemical marking and functional activity, and it is even less apparent in the erythroid context. We calculated the correlation between TF occupancy and biochemical marks on one hand, and functional activity on the other and found only a small positive quantitative correlation (Pearson $r^2 \leq 0.10$; Spearman rank correlation $r \leq 0.40$) between each of these signatures and enhancer activity (Figure 5C), observations that also hold in the other contexts we examined (Supplementary Figures 8A-B, 11A-B, 14A-B, and 17B).

We also evaluated the predictivity of biochemical signatures using a receiver operating characteristic curve analysis (Supplementary Figures 8C-F, 11C-D, and 14C-F). With the exception of erythroid cEnhs, where the combination of GATA1 and TAL1 was most predictive of functional activity, we find that DNAse and H3K27ac are most often the best predictors of functional enhancers. Their predictivity, however, is not incredibly strong, with AUROC values only exceeding 0.8 in K562 and HepG2 cells.

Overall, the combination of H3K27ac and DNAse hypersensitivity appears to be as reliable a predictor of functional activity as any other biochemical signature, however, even it is in no way absolutely predictive of function, with only approximately half of H3K27ac$^+$/DNAse$^+$ actually exhibiting significant enhancer activity.

Discussion

In this study, we have provided a comprehensive examination of cEnh activity in multiple mammalian systems and its relationship to biochemical signatures commonly used to select cEnh elements. Across cell types and methods for cEnh selection, approximately 50% of cEnhs simultaneously exhibiting significant H3K27ac marking and DNase hypersensitivity appear to function as active enhancers. We also demonstrated that enhancer assays activity is specific to genomic regions that are distinguished by characteristic biochemical signatures. By studying cEnhs sampled across the full spectrum of ChIP occupancy for multiple transcription factors, we have demonstrated that the most strongly biochemically marked cEnhs are highly enriched for functionality.

However, first, active functional enhancers are also present throughout the whole biochemical signal spectrum, and because of the very large number of the latter, the bulk of active enhancers in any given cell type in fact resides in the population of cEnhs with modest biochemical signatures, and second, we do not observe a particularly strong correlation between the magnitude of enhancer activity in functional assays and strength of biochemical marks as measured using functional genomic assays.

These findings are in contrast to earlier studies, which reported over 80-90% activity for cEnhs defined using, for example p300 ChIP-seq[47]. This is most likely due to the fact the these studies only focused on elements selected among the most strongly enriched and likely to be functional cEnhs rather than the full spectrum of ChIP-seq signal.

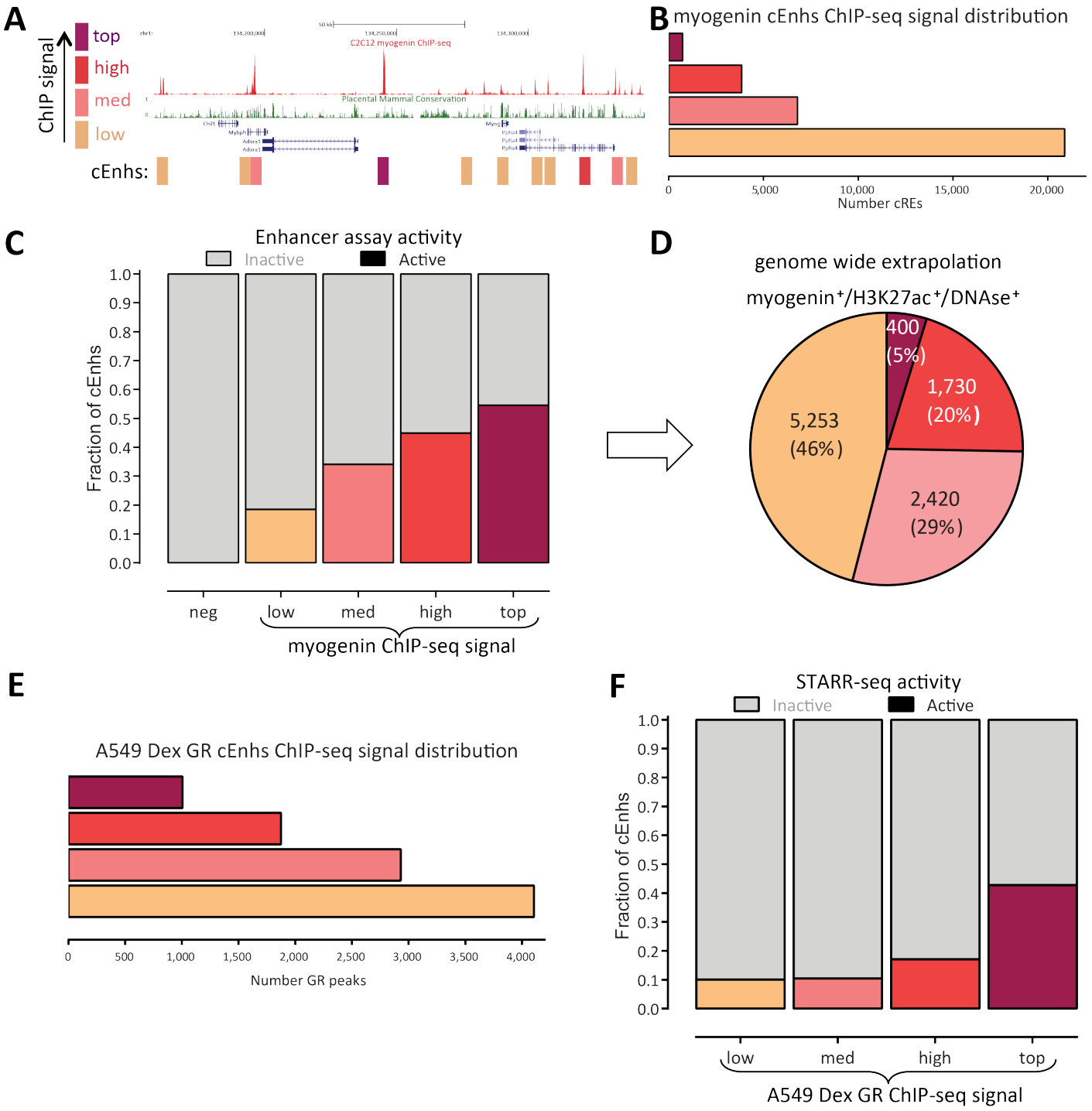


Figure 4: Enrichment of active cEnhs in different classes of cEnhs defined by the strength of their biochemical signatures. (A) cEnhs (rectangle boxes) belonging to different signal classes (based on ChIP-seq data for myogenin in C2C12 myocytes; "top": RPM ≥ 10; "high": RPM$\in [5, 10]$; "medium": RPM$\in [2.5, 5]$; "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": $n = 66$; "high": $n = 49$; "medium": $n = 45$; "low" $n = 27$) as well as in negative controls (with no myogenin occupancy; $n = 34$). Only cEnhs positive for myogenin, DNAse and H3K27ac were included; (D) Extrapolated numbers of active enhancers in C2C12 belonging to each signal strength class

(legend continued on next page)

We find a smaller fraction (15-25%) of active cEnhs using a high-throughput ChIP-STARR-seq MPRA, but similar qualitative patterns across the spectrum of biochemical signatures defining cEnhs. The reasons for the lower activity rates returned by MPRAs are manifold, and include (but are likely not limited to) the fact that the DNA fragments used as input to the MPRA are shorter than the length of fully functional regions, and that ChIP-STARR-seq libraries do not provide deep and complex representation of the original pools of ChIP-seq fragments, leaving many modestly active enhancers with insufficiently many reads to cross the thresholds of statistical significance; both of these factors are expected to lead to high false negative rates.

Promoter-enhancer specificity [67]

integration in genome [52,53,57]

Methods

Except where otherwise stated, all analyses were performed using custom-written python scripts. The v19 and vM4 versions of the GENCODE [29] annotations for human and mouse, respectively, were used for all analyses.

Cell culture

C2C12 cells

C2C12 myoblasts were maintained and seeded for transfection in 20% FBS supplemented DMEM medium. Upon reaching >80% confluency, the cells were were differentiated using 2% horse serum and 1 μM insulin in DMEM medium.

G1E cells

G1E cells were grown according to previously published protocols [13,15,64].

K562, HepG2 and A549 cells

K562; HepG2 and A549 cells were grown according to the approved ENCODE cell culture protocols publicly available through the ENCODE portal (`https://www.encodeproject.org/`).

Functional assays

Cloning and DNA purification

The specifics of each selection set, the promoters used for each cell line, and other details are publicly available through the ENCODE portal (`https://www.encodeproject.org/`).

Functional assay testing of cEnhs in C2C12 cells

Candidate REs and negative control regions were either PCR-amplified from female BALB/C purified mouse genomic DNA (Switchgear Genomics) or synthesized de novo (Genscript). The resulting DNA was cloned into a reporter vector 5′ of a custom TK promoter (SwitchGear Genomics) driving a high-turnover sequence-optimized luciferase reporter gene. Plasmids were purified using Miniprep kits (Qiagen) and standardized to 30 ng/μL using the Qubit® dsDNA HS (High Sensitivity) Assay Kit.

For the purpose of testing elements in the myoblast state, undifferentiated C2C12 cells were seeded in 96-well delta surface plates (NUNC) in quadruplicates 12 hours before transfection at a concentration of 2500 cells/well. For the purpose of testing elements in the myocyte state, undifferentiated C2C12 cells were seeded at a density of 3500 cells/well. Transfections were carried out with 50 ng of DNA per construct in each replicate using Lipofectamine LTX, after a 5 minute incubation with a 1:16 dilution with the PLUS reagent (Thermo Fisher). Myoblast plates were lysed using a Steady-Glo® kit, and luminescence was measured on a plate luminometer 24 hours post-transfection. Myocyte plates had their media exchanged with differentiation 12-16 hours post transfection and measured following the same procedure 24 hours later.

Aside from the plate reading step, the entirety of the transfection process was automated and carried out on a Tecan Freedom EVO 200 robot.

Functional assay testing of cEnhs in K562 and HepG2 cells

The set of K562 and HepG2 cEnh regions was PCR-amplified and cloned 5′ of the promoter of enhancer assay plasmids containing luciferase and renilla reporter genes; cloning was performed by SwitchGear Genomics. Each construct was quantified (using Qubit) and standardized to 30ng/μL before use in transfection assays. XXX details of transfection XXX. Chris Partridge please review this

Functional assay testing of erythropoetic cEnhs

G1E candidate enhancer regions were tested in K562 cells according to protocols publicly available through the ENCODE portal `https://www.encodeproject.org/`.

based on the genome-wide numbers 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.

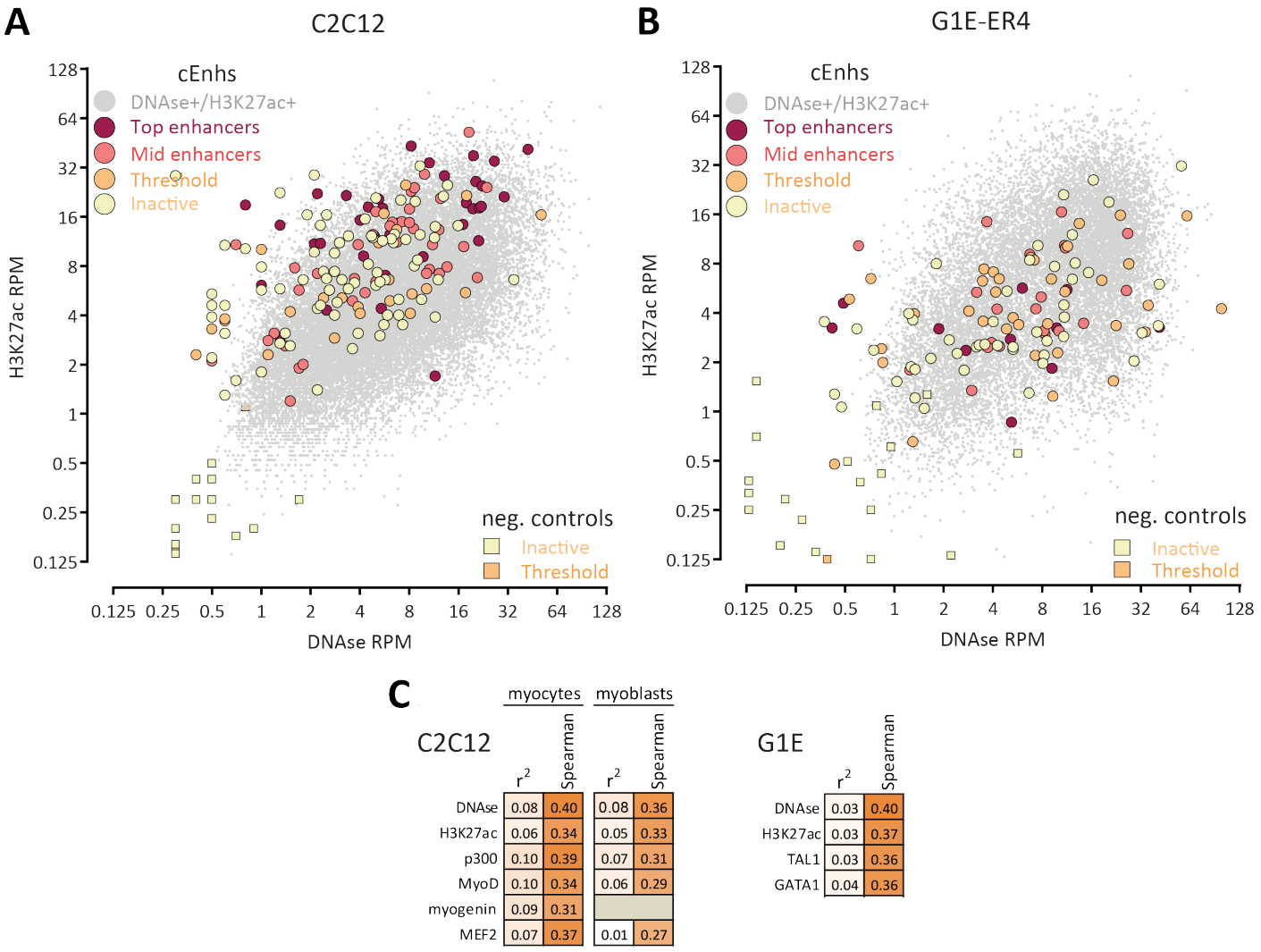


Figure 5: Absence of general strong correlation between biochemical signal strength and enhancer activity of cEnhs. (A) Distribution of tested cEnhs relative to the genome-wide DNAse and H3K27Ac signal distribution in C2C12 myocytes. Shown are DNAse and H3K27ac RPM values for all DNAse$^+$/H3K27ac$^+$ 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 cEnhs tested for activity (outlined circles) and for occupancy-negative controls (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 8, 11, and 14 for more details.

Functional Assay Data processing

For each cEnh or negative control measurement, the ratio between its value and the corresponding basal promoter vector (relative assay activity) was calculated. Active cREs were discriminated from inactive using a z-score analysis, comparing the population of test element technical replicate values to the set of negative controls. XXX this could be stated more explicitly with the actual formulas XX.

ChIP-seq experiments

Chromatin immunoprecipitation in A549 cells was performed as previously described (Reddy et al. 2009) using 2 $\times$ 10^7 A549 cells per replicate. Cells were sonicated using a Bioruptor XL (Diagenode) on the high setting until the resulting chromatin was fragmented to a median fragment size of 250 nt as assayed by agarose gel electrophoresis. GR ChIP was performed using 5 μg of a rabbit polyclonal α-GR

antibody (Santa Cruz Biotechnology sc-1003), and 200 μl of magnetic sheep anti-rabbit beads (Life Technologies M-280). H3K27ac ChIP was performed using XXX Ab source XXX. After reversal of formaldehyde crosslinks at 65 °C overnight, DNA was purified using MinElute DNA purification columns (QIAGEN). Illumina sequencing libraries were then generated using the Apollo 324 liquid handling platform according to manufacturer's specifications (Wafergen).

ChIP-seq in C2C12 cells was performed using chromatin from 2×10^7 nuclei, which was fragmented using a Misonix probe tip sonicator and subjected to immunoprecipitation using a robotic ChIP pipeline described before. The resulting purified DNA was then converted into sequencing libraries and sequenced on an HiSeq 2500 (Illumina) as described previously[42]. The following antibodies were used: α-myogenin (Santa Cruz Biotechnology SC-12732, lot K2311), α-MyoD (Santa Cruz Biotechnology SC-32758, lot J3115), α-MEF2 (Santa Cruz Biotechnology SC-17785, lot H1913), α-p300 (Santa Cruz Biotechnology SC-585, lot H3115), α-E2A (Santa Cruz Biotechnology SC-349X, lot B1207), α-H2B (Santa Cruz Biotechnology SC-357, F2305), and α-H3K27ac (Active Motif 39133, lot 34849).

In addition, publicly available[31] Pbx1 ChIP-seq and Control datasets were downloaded from GEO accession GSE76010.

For G1E, K562 and HepG2 cells, previously publicly available[39,65,66] ChIP-seq datasets were downloaded from the ENCODE portal `https://www.encodeproject.org/`.

DNAse-seq experiments

In C2C12 cells, DNAse-seq was carried out as follows: XXXXX DETAILS XXX

In A549 cells, DNAse-seq was carried out as follows: XXXXX DETAILS XXX.

For G1E, K562 and HepG2 cells, previously publicly available[39,65,66] DNAse-seq datasets were downloaded from the ENCODE portal `https://www.encodeproject.org/`.

STARR-seq experiments

The STARR-seq experiments previously published by Vockley et al.[41] were used in this study.

Genomic coordinate conversion

The regions to be tested using functional assays were designed based on the `mm8` and `mm9` verions of the mouse genome and the `hg19` version of the human genomes. Conversion of the original coordinates to `mm10` and `hg20` coordinates was performed using the `liftOver` tool from the UCSC Genome Browser Utilities[30].

Conservation analysis

Sequence conservation analysis were carried out using the `phastCons60way` and `phastCons100way` conservation tracks, which were downloaded from the UCSC Genome Browser[30].

ChIP-seq data processing and analysis

ChIP-seq reads were trimmed down to 36 bp in length and mapped against the `hg20` (for human samples; the male or female version depending on the sex of the cell line the sample originated from) and `mm10` (for mouse samples) using Bowtie[40] (version 1.0.1) with the following settings: `-v 2 -k 2 -m 1 --best --strata`. DNAse-seq reads were processed similarly except that they were trimmed down to 20bp for A549 samples and 36bp for C2C12 cells (due to differences in the experimental protocol used to generate the data).

Peak calling was carried out as follows. For DNAse and H3K27ac datasets, MACS2[43] (version 2.1.0) was run on individual replicates and on pseudoreplicates (generated by randomly splitting the pooled set of reads for both replicates into two) with relaxed settings (`--to-large -p 1e-1`). The top 100,000 peaks from each replicate or pseudoreplicate (ranked by q-value) were then used as input into IDR[44]. The number of peaks above a given IDR threshold called as reproducible between true replicates (N_t) and between pseudoreplicates (N_p) were recorded. Peak calling was then carried out on the pooled set of reads and the top $max(N_t, N_p)$ peaks were chosen as the final set of reproducible peaks. For point-source[45] datasets (transcription factors), peak calling was carried out following the same procedure but using SPP[46] (version 1.10.1), using the top 300,000 peaks as input to IDR.

The pooled sets of reads were also used to calculate RPM (reads per million) enrichment values over elements tested in functional assays.

STARR-seq data processing and analysis

STARR-seq and STARR-seq control/input reads (2×25mers) were mapped as paired ends to the `hg20` version of the human genome using Bowtie with the same settings as described above. Post-IDR peaks obtained from GR ChIP-seq were used as the list of candidate cEnhs to be scored using the STARR-seq data. For each STARR-seq and STARR-seq control/input replicate, raw fragment counts were obtained from every GR ChIP-seq peak; in addition, the rest of the genome (i.e. the regions that fall between the post-IDR GR ChIP-seq peaks) was split into bins of at most 50 kb length, and read counts were calculated for all such regions. The fragment counts for GR ChIP-seq peaks and for the intervening regions were combined together and used as input to DESeq2[26] for estimating differentially represented regions between STARR-seq and control/input libraries (at FDR-adjusted $p \leq 0.05$). The lowest average fragment counts value which was scored as significantly significant by DESeq2 was identified for each comparison, and all GR ChIP-seq regions with average fragment counts lower than this value were excluded

from subsequent analysis, as such regions were not sufficiently represented in the available sequencing data to be reliably scored as active or inactive. We also carried out a fragment-level analysis, in which read counts were calculated for each individual sequencing fragment (defined as the pair of positions $\{i, j\}$, where i and j are respectively the $5'$ and $3'$ ends of the first and the second sequencing reads in a pair), using the same DESeq2 framework.

Acknowledgments

Library generation and high-throughput sequencing for C2C12 ChIP-seq samples was performed by Igor Antoshechkin at the Millard and Muriel Jacobs Genetics and Genomics Laboratory. The authors would also like to thank Diane Trout and Henry Amrhein for technical assistance with maintaining the computational infrastructure used to carry out this study.

This material is based upon work supported by the National Science Foundation under Grant No. CNS-0521433.

References

1. Gertz J, Savic D, Varley KE, Partridge EC, Safi A, Jain P, Cooper GM, Reddy TE, Crawford GE, Myers RM. 2013. Distinct properties of cell-type-specific and shared transcription factor binding sites. *Mol Cell* **52**(1):25–36.
2. So AY, Chaivorapol C, Bolton EC, Li H, Yamamoto KR. 2007. Determinants of cell- and gene-specific transcriptional regulation by the glucocorticoid receptor. *PLoS Genet* **3**(6):e94.
3. Reddy TE, Pauli F, Sprouse RO, Neff NF, Newberry KM, Garabedian MJ, Myers RM. 2009. Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. *Genome Res* **19**(12):2163–2171.
4. Tsang AP, Visvader JE, Turner CA, Fujiwara Y, Yu C, Weiss MJ, Crossley M, Orkin SH. 1997. FOG, a multitype zinc finger protein, acts as a cofactor for transcription factor GATA-1 in erythroid and megakaryocytic differentiation. *Cell* **90**(1):109–119.
5. Rylski M, Welch JJ, Chen YY, Letting DL, Diehl JA, Chodosh LA, Blobel GA, Weiss MJ. 2003. GATA-1-mediated proliferation arrest during erythroid maturation. *Mol Cell Biol* **23**(14):5031–5042.
6. Weiss MJ, Yu C, Orkin SH. 1997. Erythroid-cell-specific properties of transcription factor GATA-1 revealed by phenotypic rescue of a gene-targeted cell line. *Mol Cell Biol* **17**(3):1642–1651.
7. Tripic T, Deng W, Cheng Y, Zhang Y, Vakoc CR, Gregory GD, Hardison RC, Blobel GA. 2009. SCL and associated proteins distinguish active from repressive GATA transcription factor complexes. *Blood* **113**(10):2191–2201.
8. Yu M, Riva L, Xie H, Schindler Y, Moran TB, Cheng Y, Yu D, Hardison R, Weiss MJ, Orkin SH, Bernstein BE, Fraenkel E, Cantor AB. 2009. Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. *Mol Cell* **36**(4):682–695.
9. Wu W, Morrissey CS, Keller CA, Mishra T, Pimkin M, Blobel GA, Weiss MJ, Hardison RC. 2014. Dynamic shifts in occupancy by TAL1 are guided by GATA factors and drive large-scale reprogramming of gene expression during hematopoiesis. *Genome Res* **24**(12):1945–1962.
10. Kaneko H, Shimizu R, Yamamoto M. 2010. GATA factor switching during erythroid differentiation. *Curr Opin Hematol* **17**(3):163–168.
11. Han GC, Vinayachandran V, Bataille AR, Park B, Chan-Salis KY, Keller CA, Long M, Mahony S, Hardison RC, Pugh BF. 2015. Genome-Wide Organization of GATA1 and TAL1 Determined at High Resolution. *Mol Cell Biol* **36**(1):157–172.
12. Yaffe D, Saxel O. 1977. Serial passaging and differentiation of myogenic cells isolated from dystrophic mouse muscle. *Nature* **270**(5639):725–727.
13. Wang H, Zhang Y, Cheng Y, Zhou Y, King DC, Taylor J, Chiaromonte F, Kasturi J, Petrykowska H, Gibb B, Dorman C, Miller W, Dore LC, Welch J, Weiss MJ, Hardison RC. 2006. Experimental validation of predicted mammalian erythroid *cis*-regulatory modules. *Genome Res* **16**(12):1480–1492.
14. Taylor J, Tyekucheva S, King DC, Hardison RC, Miller W, Chiaromonte F. 2006. ESPERR: learning strong and weak signals in genomic sequence alignments to identify functional elements. *Genome Res* **16**(12):1596–1604.
15. Cheng Y, King DC, Dore LC, Zhang X, Zhou Y, Zhang Y, Dorman C, Abebe D, Kumar SA, Chiaromonte F, Miller W, Green RD, Weiss MJ, Hardison RC. 2008. Transcriptional enhancement by GATA1-occupied DNA segments is strongly associated with evolutionary constraint on the binding site motif. *Genome Res* **18**(12):1896–1905.
16. Taylor SM, Jones PA. 1979. Multiple new phenotypes induced in 10T1/2 and 3T3 cells treated with 5-azacytidine. *Cell* **17**(4):771–779.
17. Davis RL, Weintraub H, Lassar AB. 1987. Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell* **51**(6):987–1000.
18. Hasty P, Bradley A, Morris JH, Edmondson DG, Venuti JM, Olson EN, Klein WH. 1993. Muscle deficiency and neonatal death in mice with a targeted mutation in the myogenin gene. *Nature* **364**(6437):501–506.
19. Wright WE, Sassoon DA, Lin VK. 1989. Myogenin, a factor regulating myogenesis, has a domain homolo-

gous to MyoD. *Cell* **56**(4):607–617.

20. Moncaut N, Rigby PW, Carvajal JJ. 2013. Dial M(RF) for myogenesis. *FEBS J* **280**(17):3980–3990.
21. Tapscott SJ. 2005. The circuitry of a master switch: Myod and the regulation of skeletal muscle gene transcription. *Development* **132**(12):2685–2695.
22. Moore J, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, Kawli T, Davis CA, Dobin A, Kaul R, Halow J, Van Nostrand EL, Freese P, Gorkin DU, He Y, Mackiewicz M, The ENCODE Consortium, Cherry MJ, Myers RM, Ren BGraveley BR, Stamatoyannopoulos JA, Gerstein MB, Pennacchio LA, Gingeras T, Snyder MP, Bernstein1 BE, Wold B, Hardison RC, Weng Z. *submitted*. ENCODE Phase III: Building an Encyclopaedia of candidate Regulatory Elements for Human and Mouse
23. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. 2012. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* **9**(5):473–476.
24. Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**(3):215–216.
25. Mortazavi A, Pepke S, Jansen C, Marinov GK, Ernst J, Kellis M, Hardison RC, Myers RM, Wold BJ. 2013. Integrating and mining the chromatin landscape of cell-type specificity using self-organizing maps. *Genome Res* **23**(12):2136–2148.
26. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**(12):550.
27. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, Chen Y, DeSalvo G, Epstein C, Fisher-Aylor KI, Euskirchen G, Gerstein M, Gertz J, Hartemink AJ, Hoffman MM, Iyer VR, Jung YL, Karmakar S, Kellis M, Kharchenko PV, Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, Park PJ, Pazin MJ, Perry MD, Raha D, Reddy TE, Rozowsky J, Shoresh N, Sidow A, Slattery M, Stamatoyannopoulos JA, Tolstorukov MY, White KP, Xi S, Farnham PJ, Lieb JD, Wold BJ, Snyder M. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22**(9):1813–1831.
28. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012b. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**(3):562–578.
29. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigó R, Hubbard TJ. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**(9):1760–1774.
30. Tyner C, Barber GP, Casper J, Clawson H, Diekhans M, Eisenhart C, Fischer CM, Gibson D, Gonzalez JN, Guruvadoo L, Haeussler M, Heitner S, Hinrichs AS, Karolchik D, Lee BT, Lee CM, Nejad P, Raney BJ, Rosenbloom KR, Speir ML, Villarreal C, Vivian J, Zweig AS, Haussler D, Kuhn RM, Kent WJ. 2017. The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res* **45**(D1):D626–D634.
31. Dell'Orso S, Wang AH, Shih HY, Saso K, Berghella L, Gutierrez-Cruz G, Ladurner AG, O'Shea JJ, Sartorelli V, Zare H. 2016. The Histone Variant MacroH2A1.2 Is Necessary for the Activation of Muscle Enhancers and Recruitment of the Transcription Factor Pbx1. *Cell Rep* **14**(5):1156–1168.
32. Fulco CP, Munschauer M, Anyoha R, Munson G, Grossman SR, Perez EM, Kane M, Cleary B, Lander ES, Engreitz JM. 2016. Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354**(6313):769–773.
33. Korkmaz G, Lopes R, Ugalde AP, Nevedomskaya E, Han R, Myacheva K, Zwart W, Elkon R, Agami R. 2016. Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat Biotechnol* **34**(2):192–198.
34. Visel A, Minovitsky S, Dubchak I, Pennacchio LA. 2007. VISTA Enhancer Browser–a database of tissue-specific human enhancers. *Nucleic Acids Res* **35**(Database issue):D88–92.
35. Vermeulen M, Timmers HT. 2010. Grasping trimethylation of histone H3 at lysine 4. *Epigenomics* **2**(3):395–406.
36. Kellis M, Hardison RC, Wold BJ, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, Dunham I, Elnitski L, Farnham PJ, Feingold EA, Gerstein M, Giddings MC, Gilbert DM, Gingeras TR, Green ED, Guigo R, Hubbard TJP, Kent WJ, Lieb JD, Myers RM, Pazin MJ, Ren B, Stamatoyannopoulos J, Weng Z, White KP, Members of the ENCODE Consortium. 2014. Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A* 111(17):6131–6138.
37. Fisher WW, Li JJ, Hammonds AS, Brown JB, Pfeiffer BD, Weiszmann R, MacArthur S, Thomas S, Stamatoyannopoulos JA, Eisen MB, Bickel PJ, Biggin MD, Celniker SE. 2012. DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in *Drosophila*. *Proc Natl Acad*

Sci U S A **109**(52):21330–21335.

38. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kutyavin T, Lajoie B, Lee BK, Lee K, London D, Lotakis D, Neph S, Neri F, Nguyen ED, Qu H, Reynolds AP, Roach V, Safi A, Sanchez ME, Sanyal A, Shafer A, Simon JM, Song L, Vong S, Weaver M, Yan Y, Zhang Z, Zhang Z, Lenhard B, Tewari M, Dorschner MO, Hansen RS, Navas PA, Stamatoyannopoulos G, Iyer VR, Lieb JD, Sunyaev SR, Akey JM, Sabo PJ, Kaul R, Furey TS, Dekker J, Crawford GE, Stamatoyannopoulos JA. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**(7414):75–82.

39. ENCODE Project Consortium. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9**(4):e1001046.

40. Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**(3):R25.

41. Vockley CM, D'Ippolito AM, McDowell IC, Majoros WH, Safi A, Song L, Crawford GE, Reddy TE. 2016. Direct GR Binding Sites Potentiate Clusters of TF Binding across the Human Genome. *Cell* **166**(5):1269–1281.

42. Gasper WC, Marinov GK, Pauli-Behn F, Scott MT, Newberry K, DeSalvo G, Ou S, Myers RM, Vielmetter J, Wold BJ. 2014. Fully automated high-throughput chromatin immunoprecipitation for ChIP-seq: identifying ChIP-quality p300 monoclonal antibodies. *Sci Rep* **4**:5152.

43. Feng J, Liu T, Qin B, Zhang Y, Liu XS. 2012. Identifying ChIP-seq enrichment using MACS. *Nat Protoc* **7**(9):1728–1740.

44. Li Q, Brown J, Huang H, Bickel P. 2011. Measuring reproducibility of high-throughput experiments. *Ann Appl Stat* **5**:1752–1779.

45. Pepke S, Wold B, Mortazavi A. 2009. Computation for ChIP-seq and RNA-seq studies. *Nat Methods* **6**(11 Suppl):S22–32.

46. Kharchenko PV, Tolstorukov MY, and Park PJ. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* **26:**1351–1359.

47. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Ren B, Rubin EM, Pennacchio LA. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**(7231):854–858.

48. May D, Blow MJ, Kaplan T, McCulley DJ, Jensen BC, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Afzal V, Simpson PC, Rubin EM, Black BL, Bristow J, Pennacchio LA, Visel A. 2011. Large-scale discovery of enhancers from human heart tissue. *Nat Genet* **44**(1):89–93.

49. Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**(6123):1074–1077.

50. Patwardhan RP, Lee C, Litvin O, Young DL, Pe'er D, Shendure J. 2009. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol* **27**(12):1173–1175.

51. Kinney JB, Murugan A, Callan CG Jr, Cox EC. 2010. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl Acad Sci U S A* **107**(20):9158–9163.

52. Inoue F, Kircher M, Martin B, Cooper GM, Witten DM, McManus MT, Ahituv N, Shendure J. 2016. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res* **27**(1):38–52.

53. Murtha M, Tokcaer-Keskin Z, Tang Z, Strino F, Chen X, Wang Y, Xi X, Basilico C, Brown S, Bonneau R, Kluger Y, Dailey L. 2014. FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells. *Nat Methods* **11**(5):559–565.

54. Kwasnieski JC1, Mogno I, Myers CA, Corbo JC, Cohen BA. 2012. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc Natl Acad Sci U S A* **109**(47):19498–19503.

55. Kwasnieski JC, Fiore C, Chaudhari HG, Cohen BA. 2014. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res* **24**(10):1595–1602.

56. Ernst J, Melnikov A, Zhang X, Wang L, Rogov P, Mikkelsen TS, Kellis M. 2016. Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat Biotechnol* **34**(11):1180–1190.

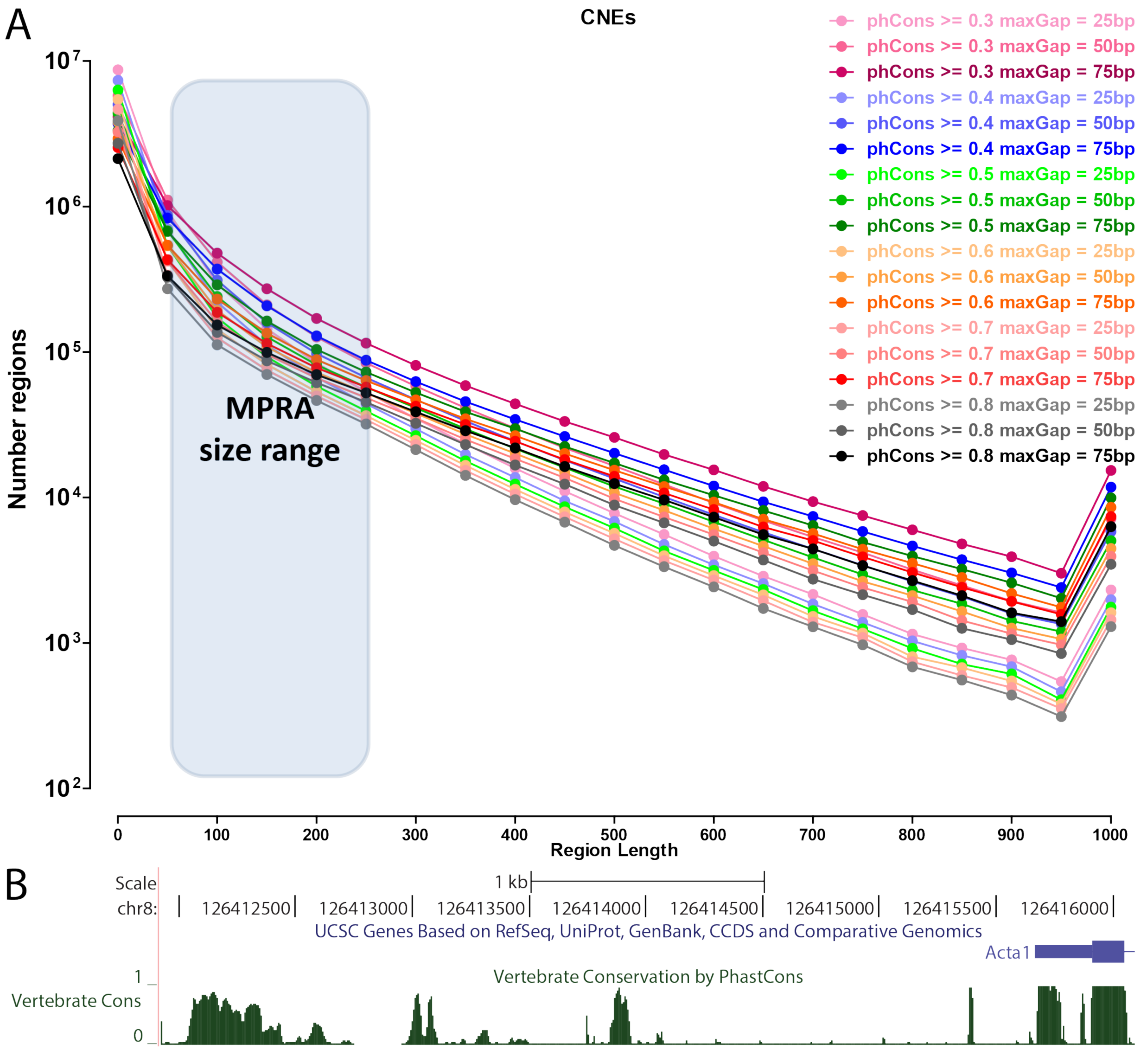
57. Maricque BB, Dougherty JD, Cohen BA. 2016. A genome-integrated massively parallel reporter assay reveals DNA sequence determinants of cis-regulatory activity in neural cells. *Nucleic Acids Res* **45**(4):e16.

58. Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, Alston J, Mikkelsen TS, Kellis M. 2013. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res* **23**(5):800–811.

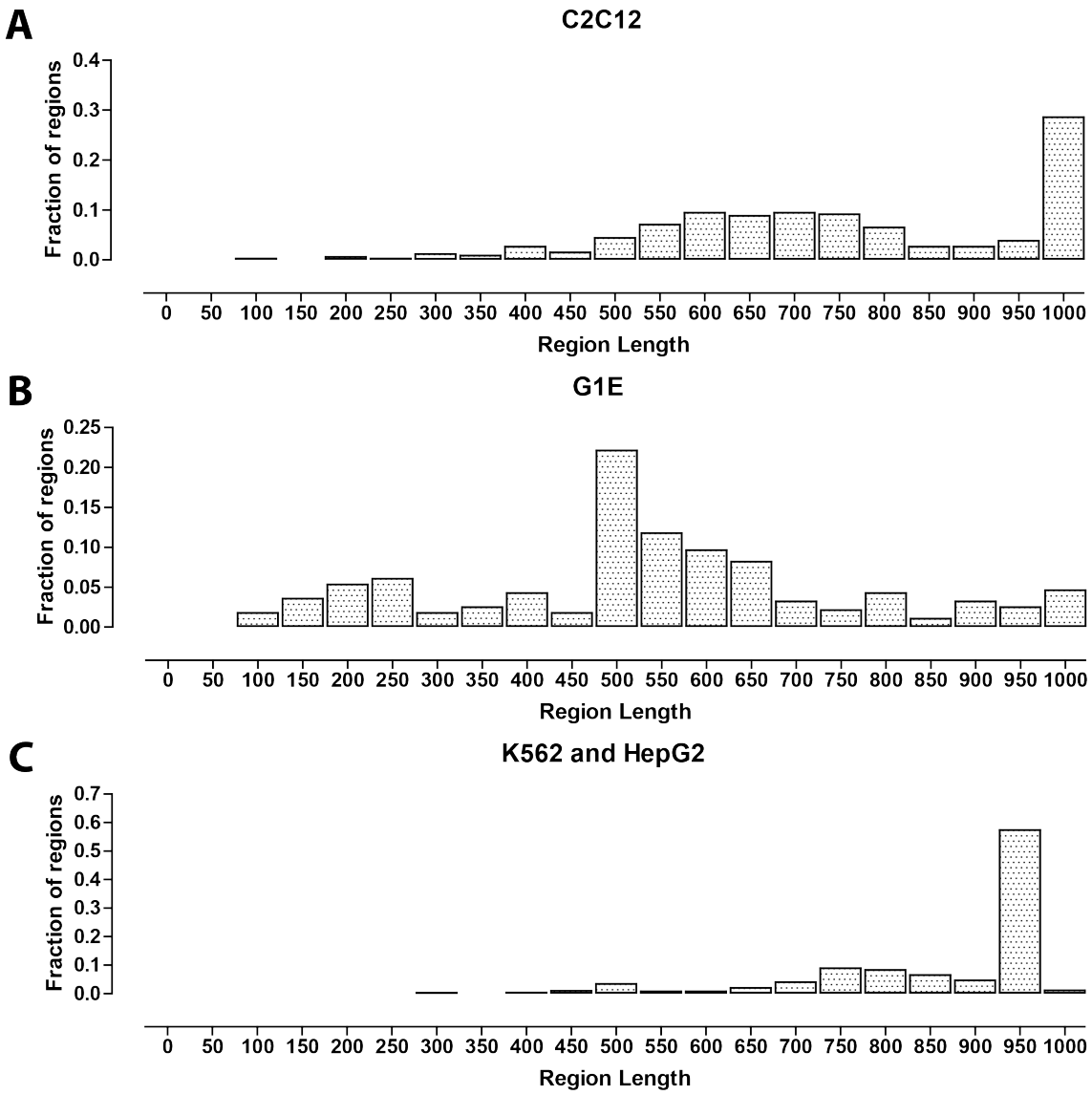
59. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG Jr, Kinney JB, Kellis M, Lander ES, Mikkelsen TS. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter as-

say. *Nat Biotechnol* **30**(3):271–277.

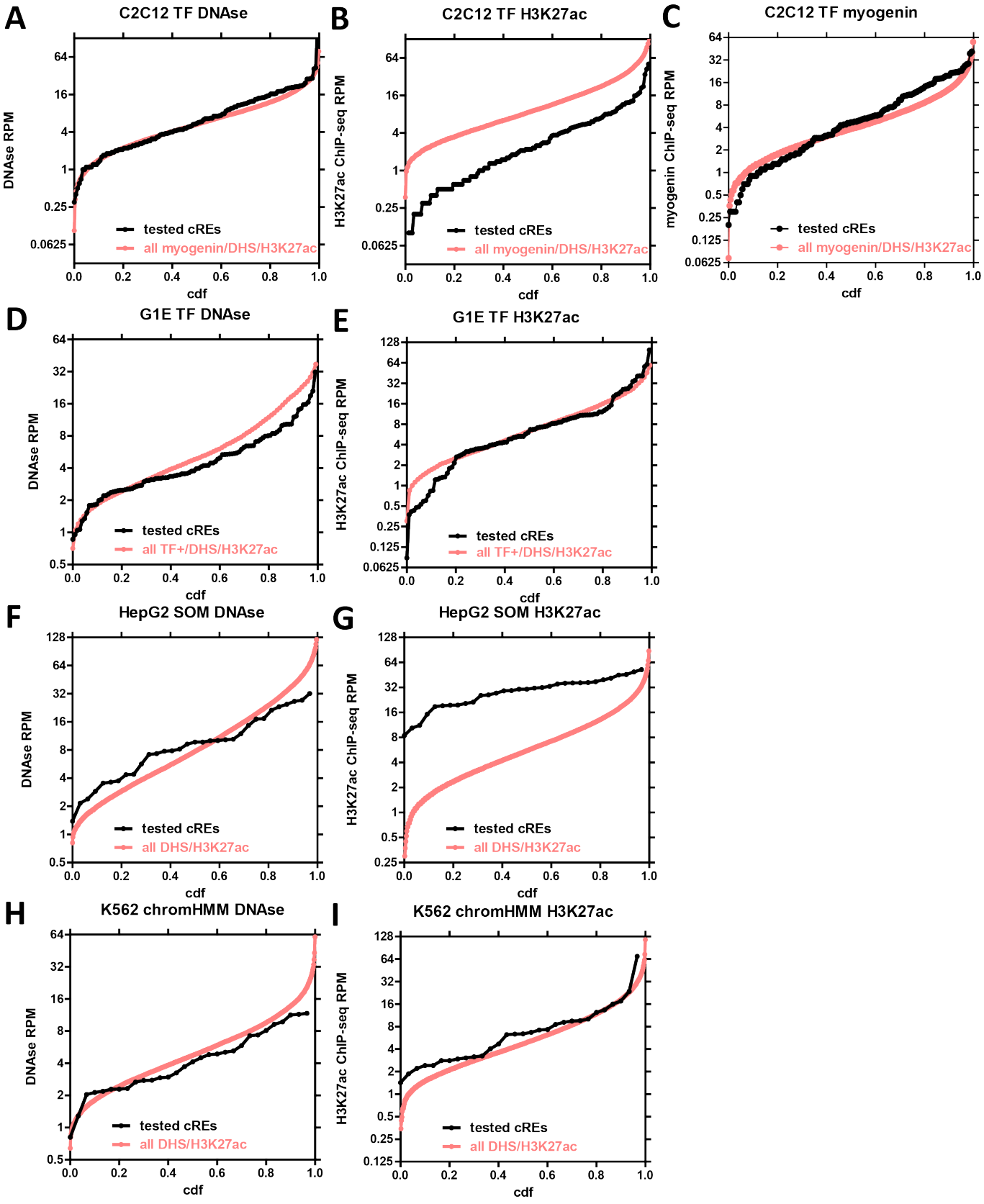
60. Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee SI, Cooper GM, Ahituv N, Pennacchio LA, Shendure J. 2012. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* **30**(3):265–270.
61. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA, Young RA, Jaenisch R. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**(50):21931–21936.
62. Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. 2011. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**(7333):279–283.
63. Inoue F, Ahituv N. 2015. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**(3):159–164.
64. Dogan N, Wu W, Morrissey CS, Chen KB, Stonestrom A, Long M, Keller CA, Cheng Y, Jain D, Visel A, Pennacchio LA, Weiss MJ, Blobel GA, Hardison RC. 2015. Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. *Epigenetics Chromatin* **8**:16.
65. Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, Pope BD, Shen Y, Pervouchine DD, Djebali S, Thurman RE, Kaul R, Rynes E, Kirilusha A, Marinov GK, Williams BA, Trout D, Amrhein H, Fisher-Aylor K, Antoshechkin I, DeSalvo G, See LH, Fastuca M, Drenkow J, Zaleski C, Dobin A, Prieto P, Lagarde J, Bussotti G, Tanzer A, Denas O, Li K, Bender MA, Zhang M, Byron R, Groudine MT, McCleary D, Pham L, Ye Z, Kuan S, Edsall L, Wu YC, Rasmussen MD, Bansal MS, Kellis M, Keller CA, Morrissey CS, Mishra T, Jain D, Dogan N, Harris RS, Cayting P, Kawli T, Boyle AP, Euskirchen G, Kundaje A, Lin S, Lin Y, Jansen C, Malladi VS, Cline MS, Erickson DT, Kirkup VM, Learned K, Sloan CA, Rosenbloom KR, Lacerda de Sousa B, Beal K, Pignatelli M, Flicek P, Lian J, Kahveci T, Lee D, Kent WJ, Ramalho Santos M, Herrero J, Notredame C, Johnson A, Vong S, Lee K, Bates D, Neri F, Diegel M, Canfield T, Sabo PJ, Wilken MS, Reh TA, Giste E, Shafer A, Kutyavin T, Haugen E, Dunn D, Reynolds AP, Neph S, Humbert R, Hansen RS, De Bruijn M, Selleri L, Rudensky A, Josefowicz S, Samstein R, Eichler EE, Orkin SH, Levasseur D, Papayannopoulou T, Chang KH, Skoultchi A, Gosh S, Disteche C, Treuting P, Wang Y, Weiss MJ, Blobel GA, Cao X, Zhong S, Wang T, Good PJ, Lowdon RF, Adams LB, Zhou XQ, Pazin MJ, Feingold EA, Wold B, Taylor J, Mortazavi A, Weissman SM, Stamatoyannopoulos JA, Snyder MP, Guigo R, Gingeras TR, Gilbert DM, Hardison RC, Beer MA, Ren B; Mouse ENCODE Consortium. 2014. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**(7527):355–364.
66. ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414):57–74.
67. Zabidi MA, Arnold CD, Schernhuber K, Pagani M, Rath M, Frank O, Stark A. 2015. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**(7540):556–559.



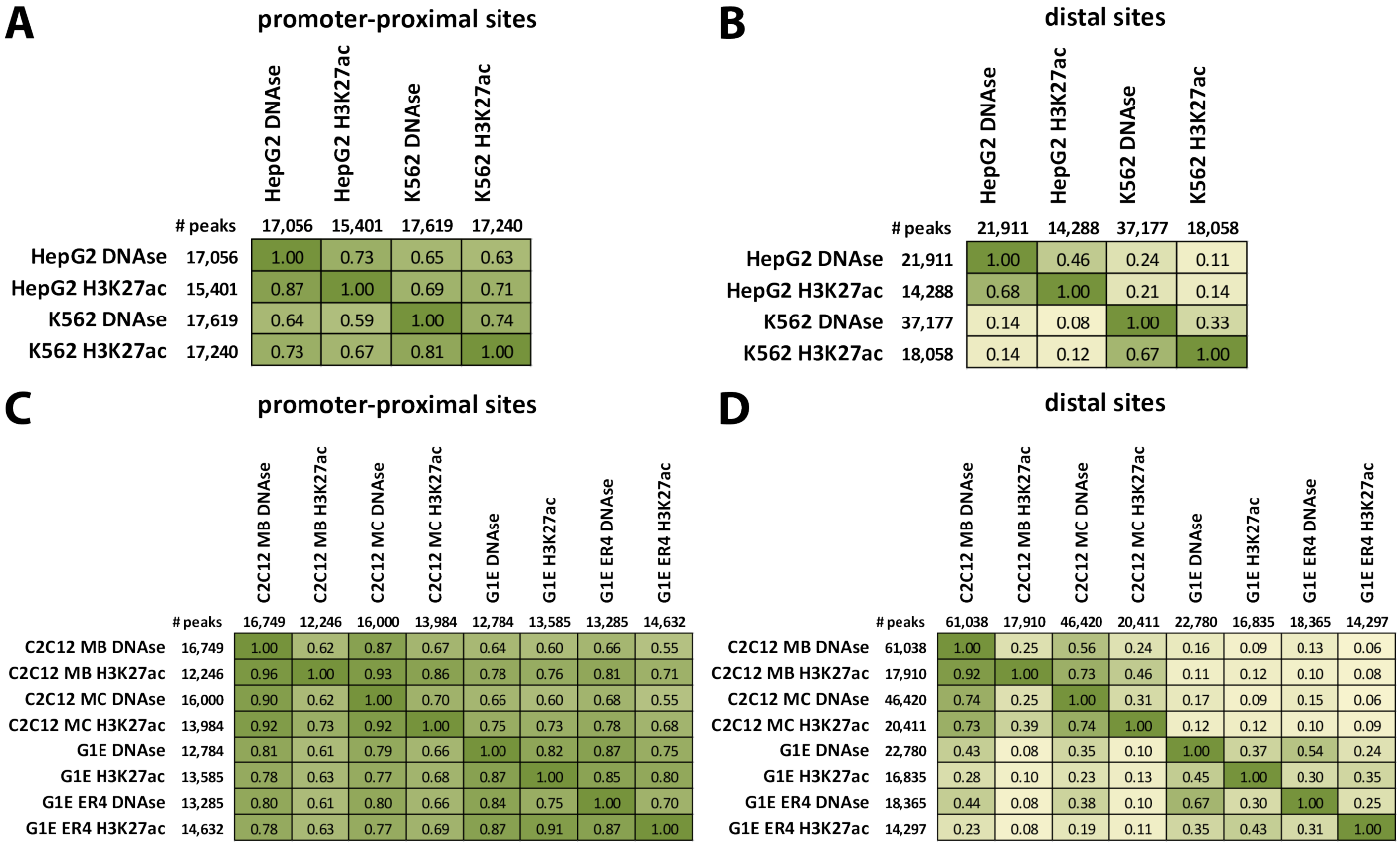
Supplementary Figure 1: The length of thousands of conserved noncoding elements in mammalian genomes greatly exceeds the size range of MPRA constructs. (A) The length distribution of conserved noncoding regions in the human genome. The `phastCons100way` conservation track for the `hg20` version of the 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.



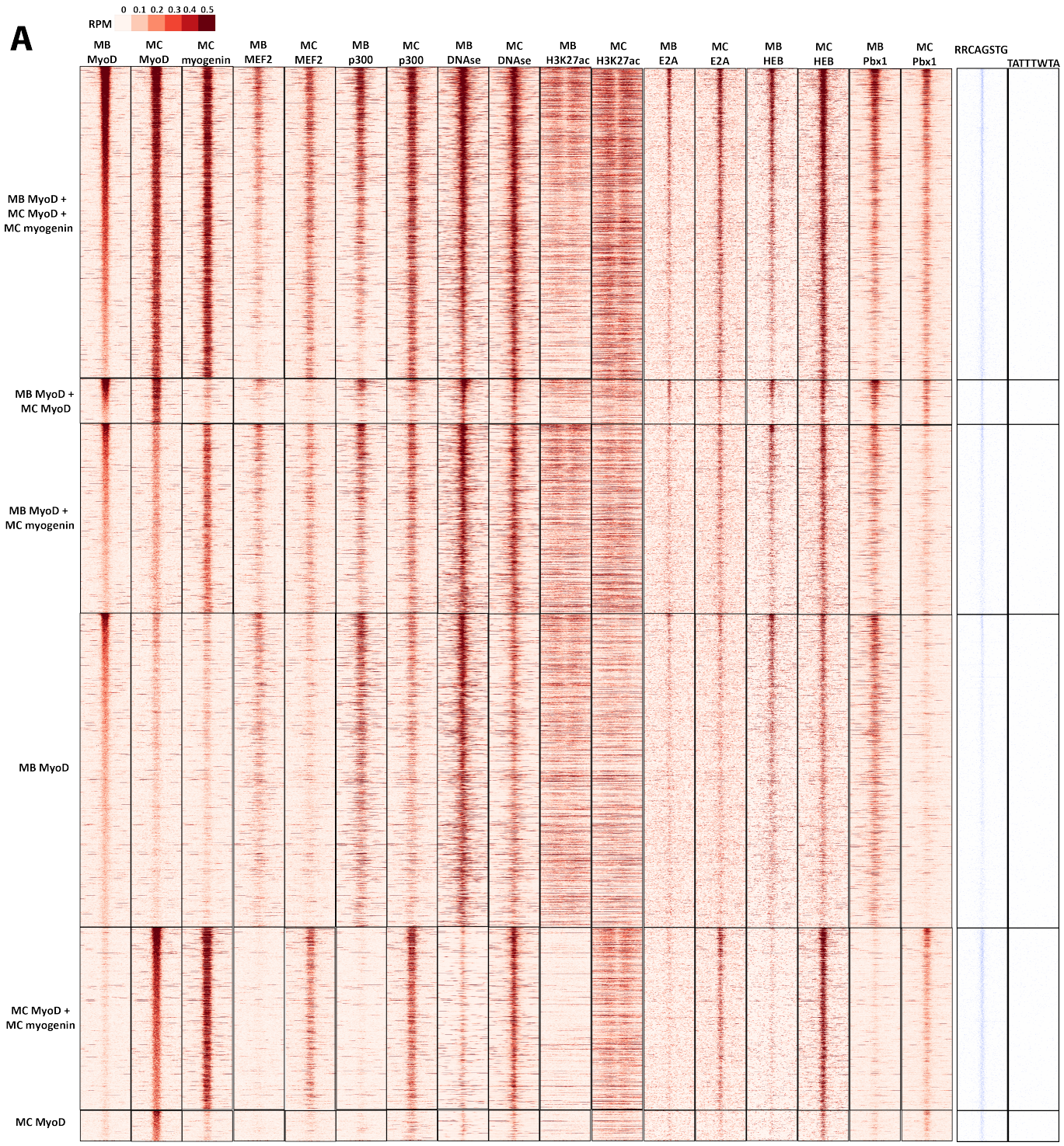
Supplementary Figure 2: Length distribution of functional assays constructs used to test cREs in this study. (A) Distribution of functional assay construct lengths tested in this study in C2C12 cells. (B) Distribution of functional assay construct lengths tested in this study in G1E cells. (C) Distribution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

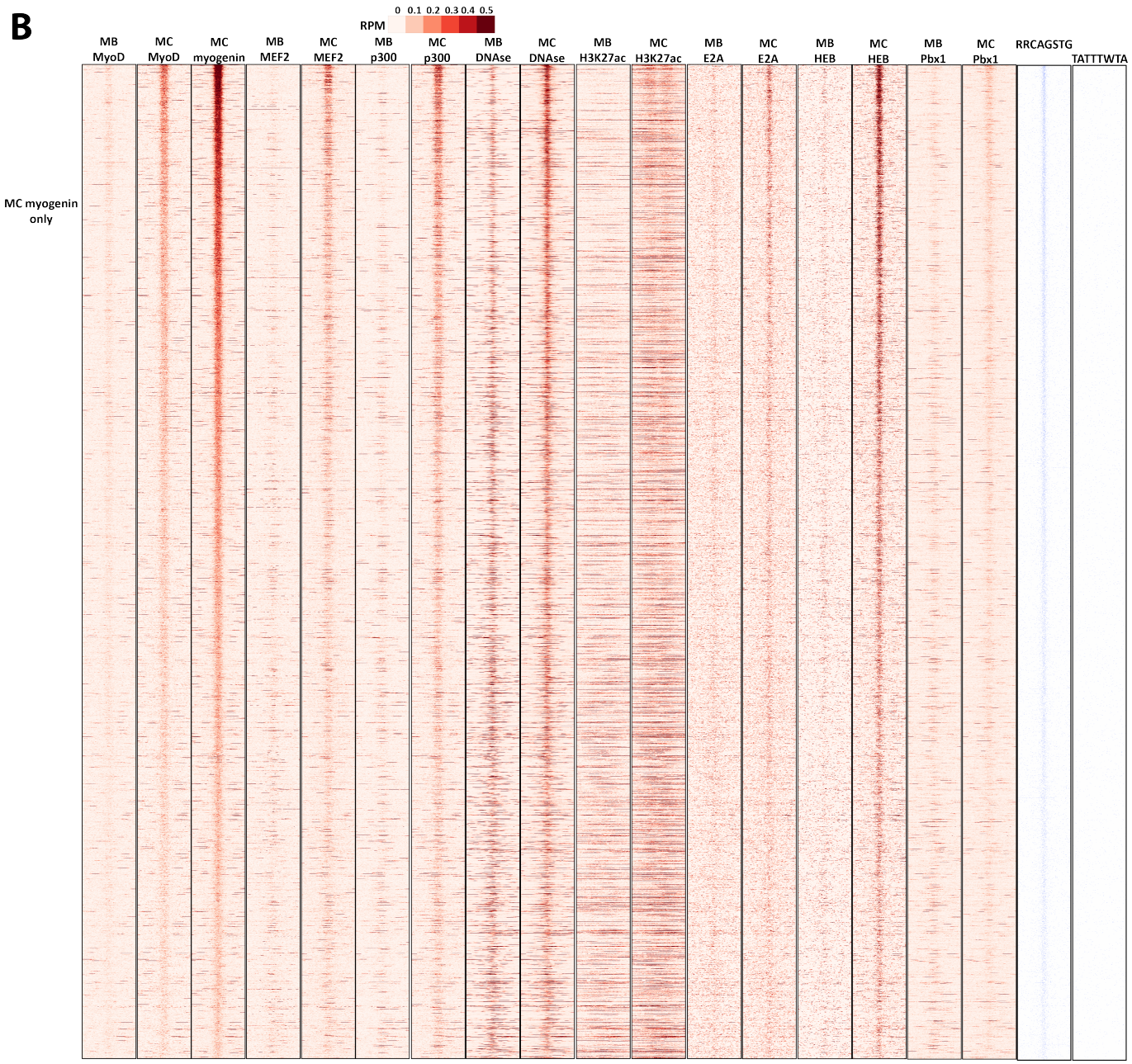


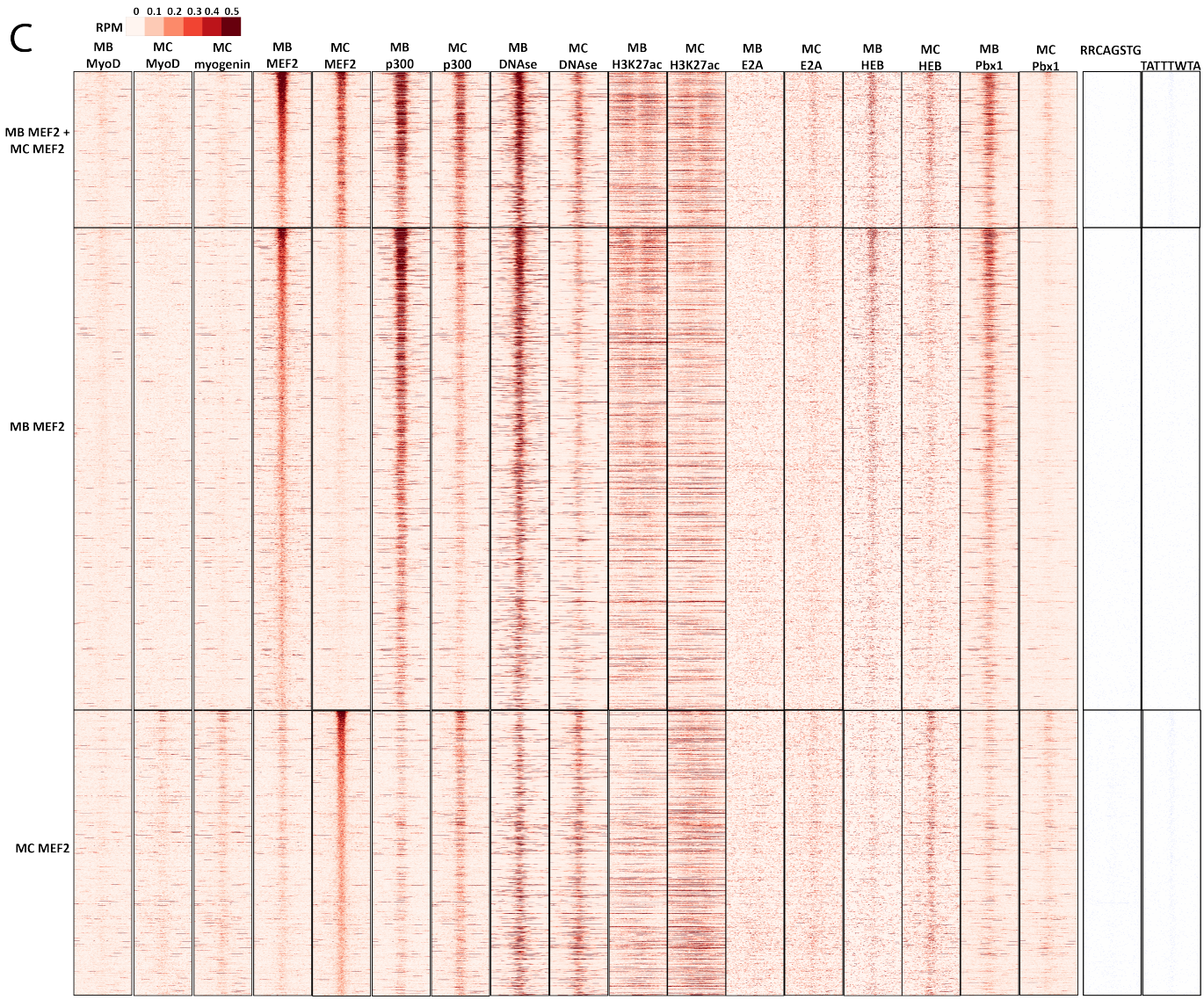
Supplementary Figure 3: Distribution of biochemical signal in tested cEnhs and genome-wide. Shown is the distribution of ChIP-seq or DNAse-seq RPM values for the set of cEnhs tested and for the genome-wide set of cEnh with similar biochemical signatures shown in Figure 3.



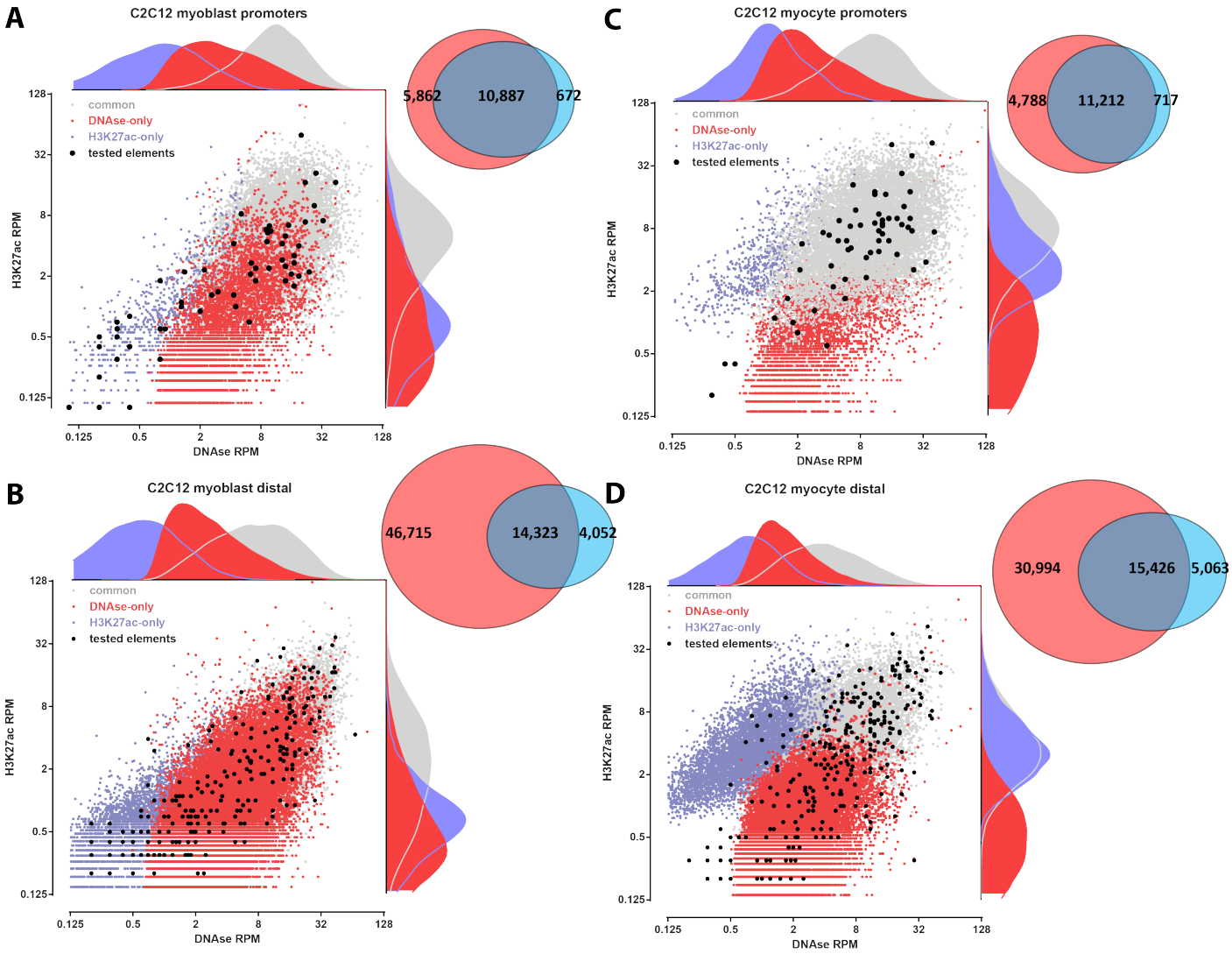
Supplementary Figure 4: Differential marking of proximal and distal cREs by DNAse and H3K27ac between different cell types and cell states. (A) Promoter-proximal (within $\leq$1 kb of an annotated TSS) sites in K562 and HepG2 cells; (A) Distal ($\geq$1 kb from an annotated TSS) sites in K562 and HepG2 cells; (C) Promoter-proximal (within $\leq$1 kb of an annotated TSS) sites in differentiated and undifferentiated C2C12 and G1E cells; (D) Distal ($\geq$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



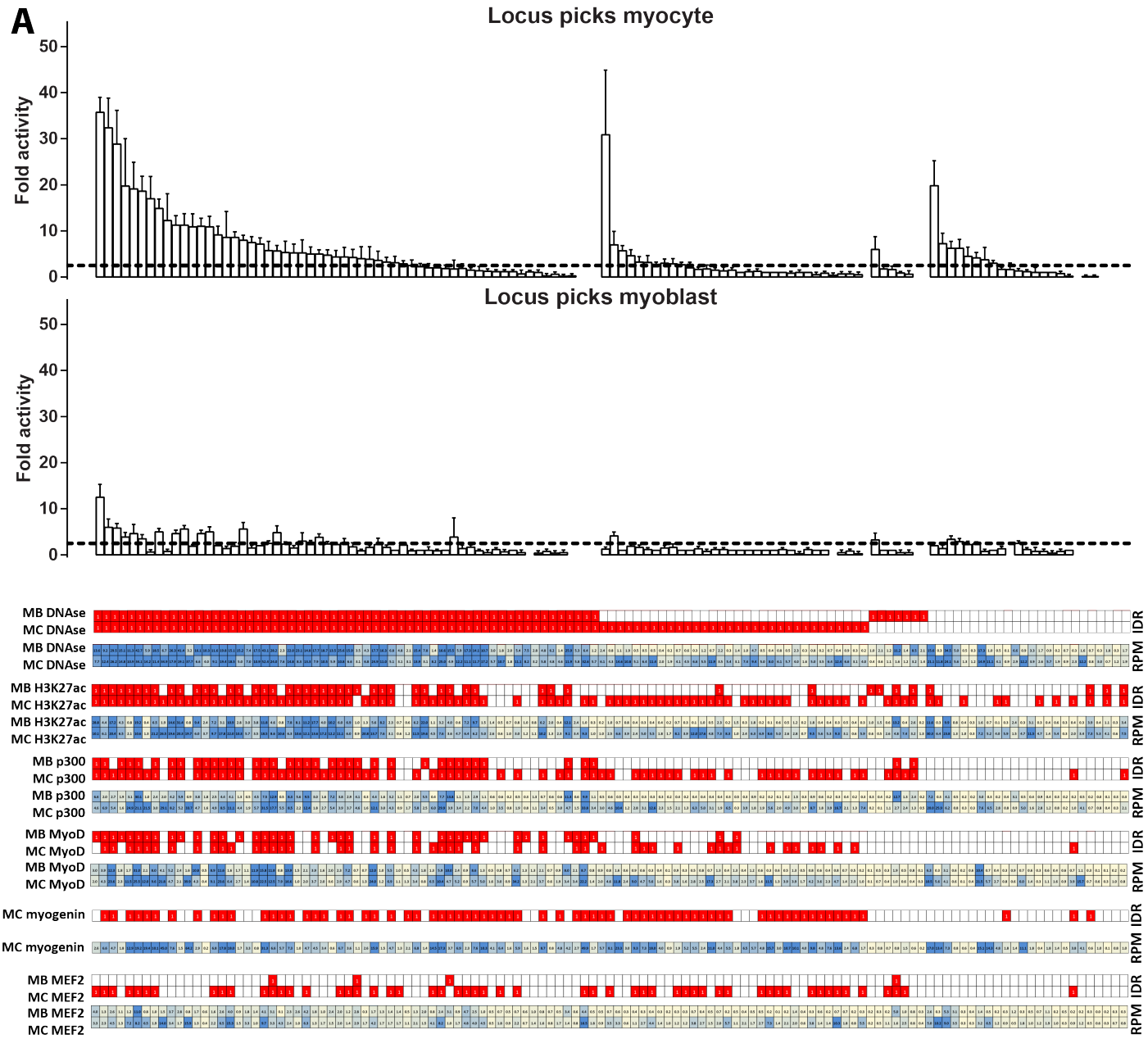
Supplementary Figure 5: Regulatory landscape of muscle differentiation. DNAse-seq and ChIP-seq experiments against H3K27ac, p300, the MRFs MyoD and myogenin, and cofactors (MEF2, E2A/TCF3, HEB/TCF12, and Pbx1) in undifferentiated (myoblast, or "MB") and differentiated (myocyte, or "MC") C2C12 cells were 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.

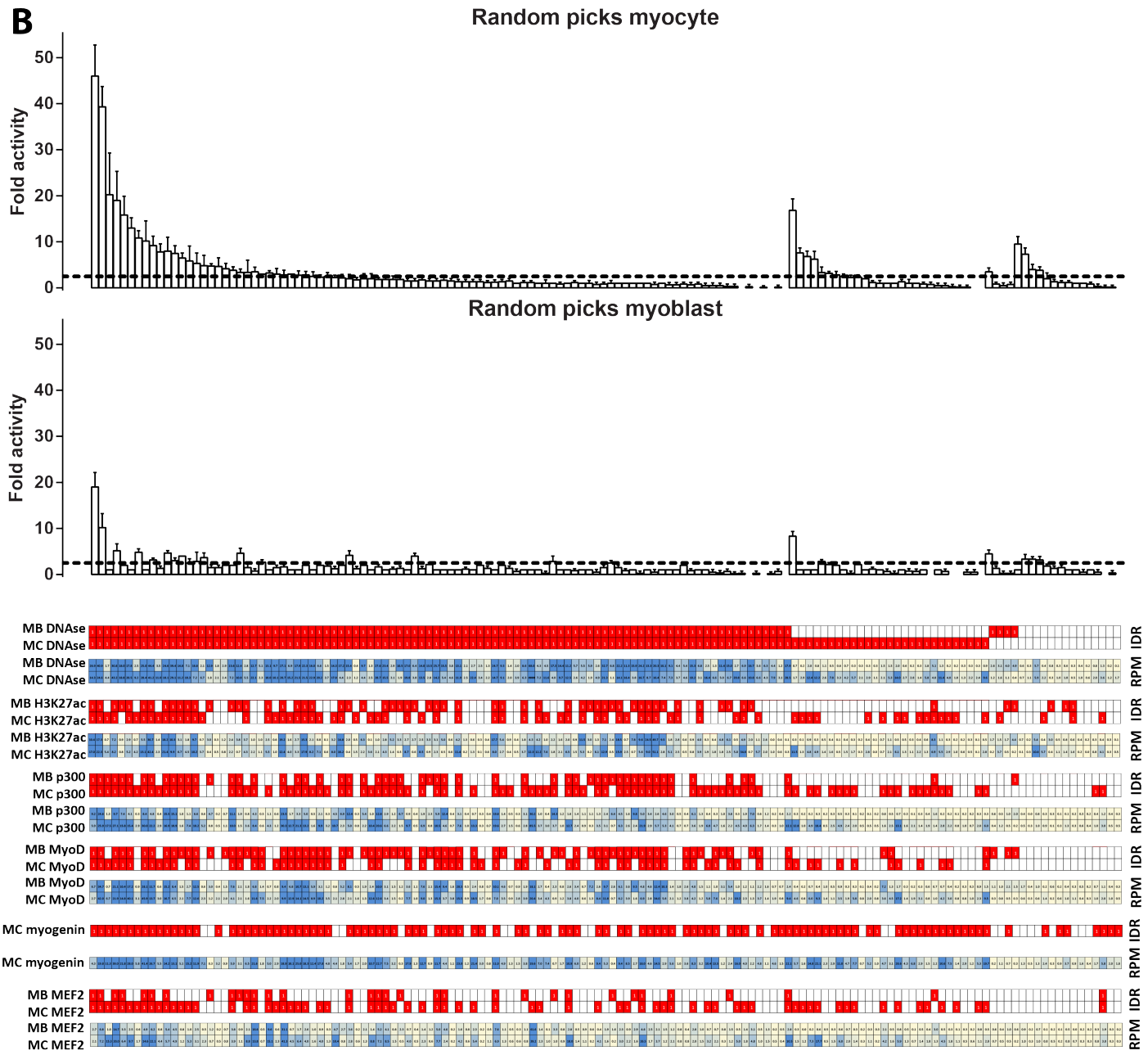


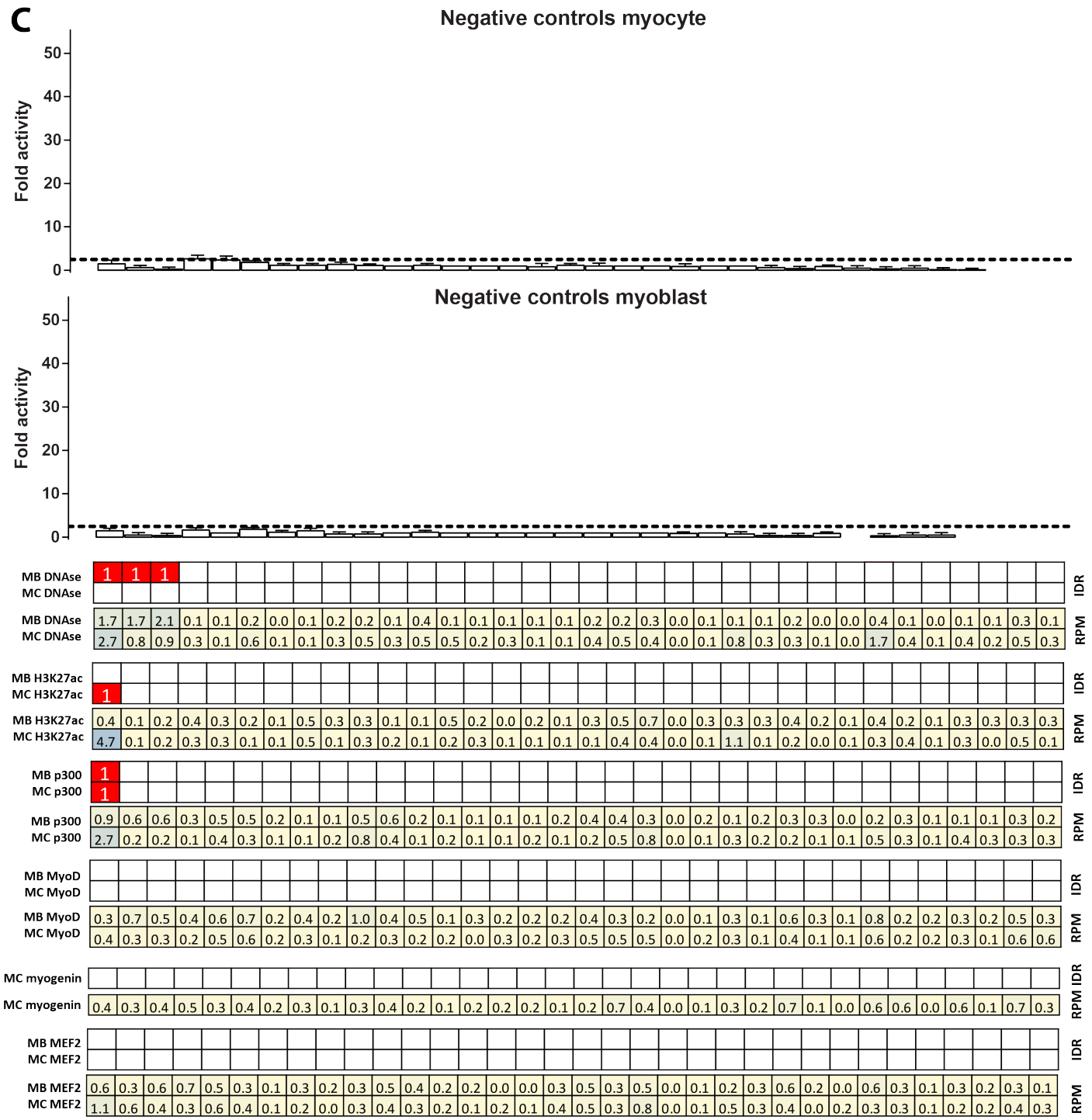
E

	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%

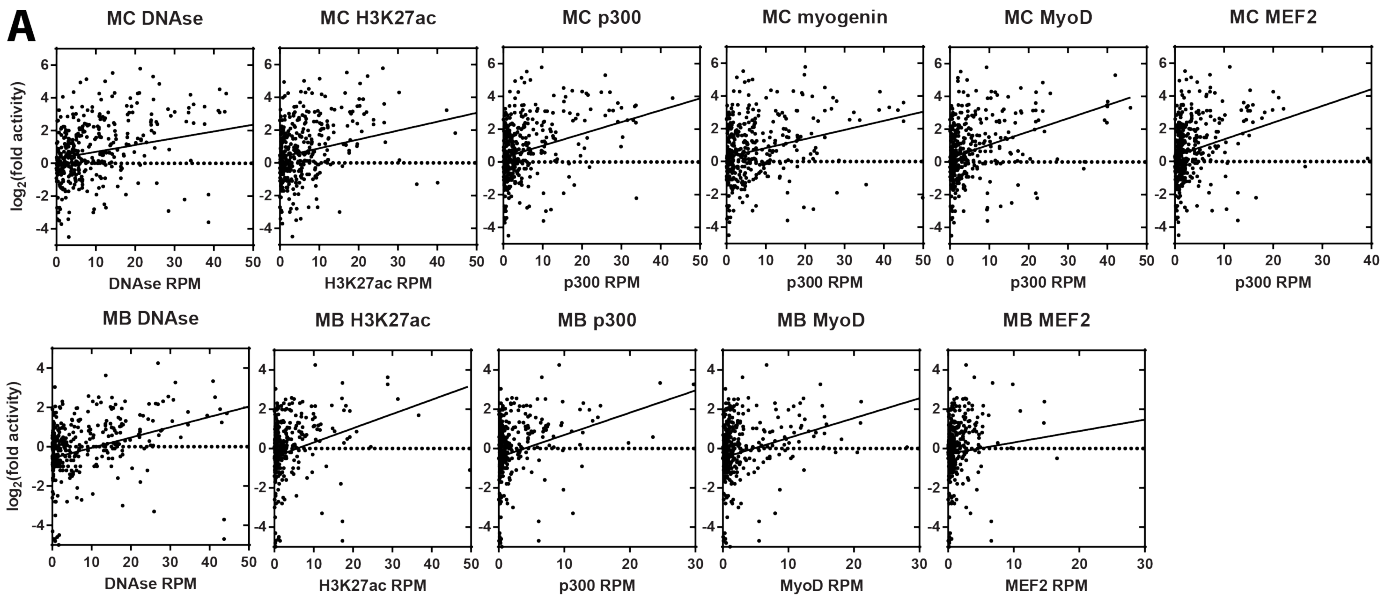
Supplementary Figure 6: Relationship between DNAse hypersensitivity and H3K27 acetylation during muscle differentiation. (A) Overlap between DNAse hypersensitive and H3K27ac-positive promoter-proximal regions in C2C12 myoblasts; (B) Overlap between DNAse hypersensitive and H3K27ac-positive promoter-proximal 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.


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
MB DNAse
MC DNAse
IDR
RPM
MB H3K27ac
MC H3K27ac
MB H3K27ac
MC H3K27ac
IDR
RPM
MB p300
MC p300
MB p300
MC p300
IDR
RPM
MB MyoD
MC MyoD
MB MyoD
MC MyoD
IDR
RPM
MC myogenin
MC myogenin
IDR
RPM
MB MEF2
MC MEF2
MB MEF2
MC MEF2
IDR
RPM


B
Random picks myocyte
Fold activity
50
40
30
20
10
0
Random 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

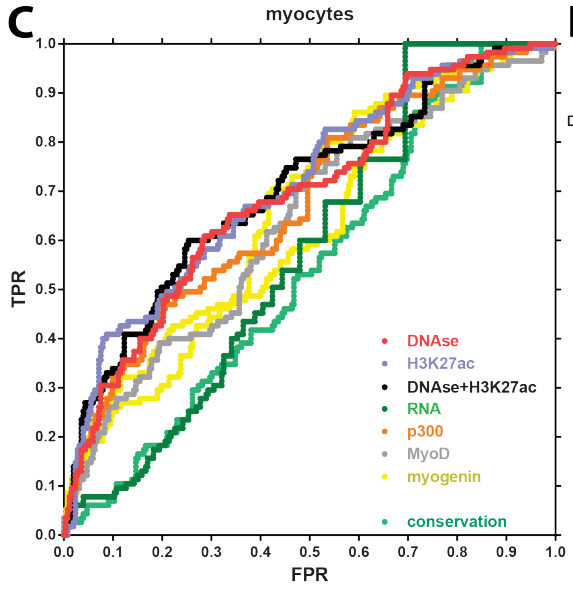


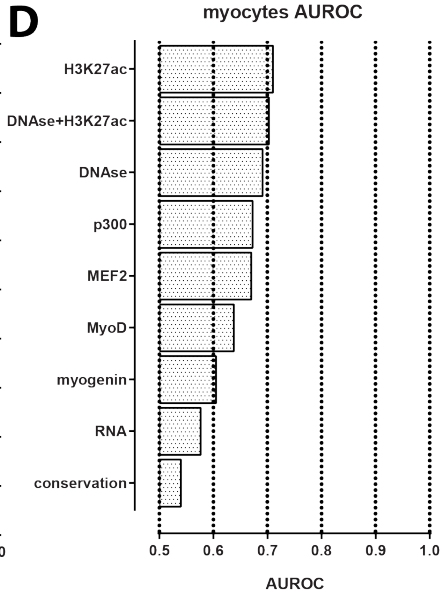
Supplementary Figure 7: Functional assay testing of cRE regulatory activity in C2C12 cells. Fold activity in myocytes (top) and myoblasts (bottom) across biological replicates ($n = 4$) and technical replicates ($n = 4$ for each biological replicate) is shown. Candidate REs were sorted first by their 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, 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.

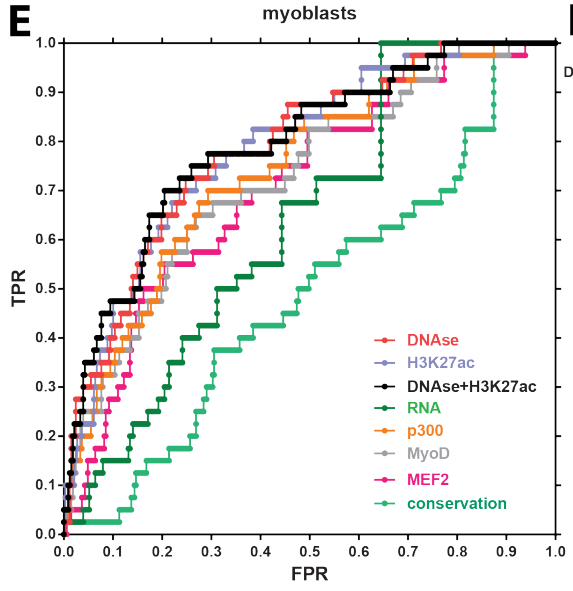


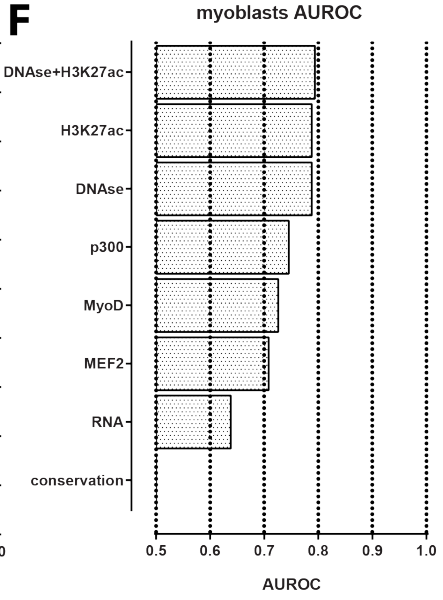
B

	myocytes		myoblasts	
	r^2	Spearman	r^2	Spearman
DNAse	0.08	0.40	0.08	0.36
H3K27ac	0.06	0.34	0.05	0.33
p300	0.10	0.39	0.07	0.31
MyoD	0.10	0.34	0.06	0.29
myogenin	0.09	0.31		
MEF2	0.07	0.37	0.01	0.27

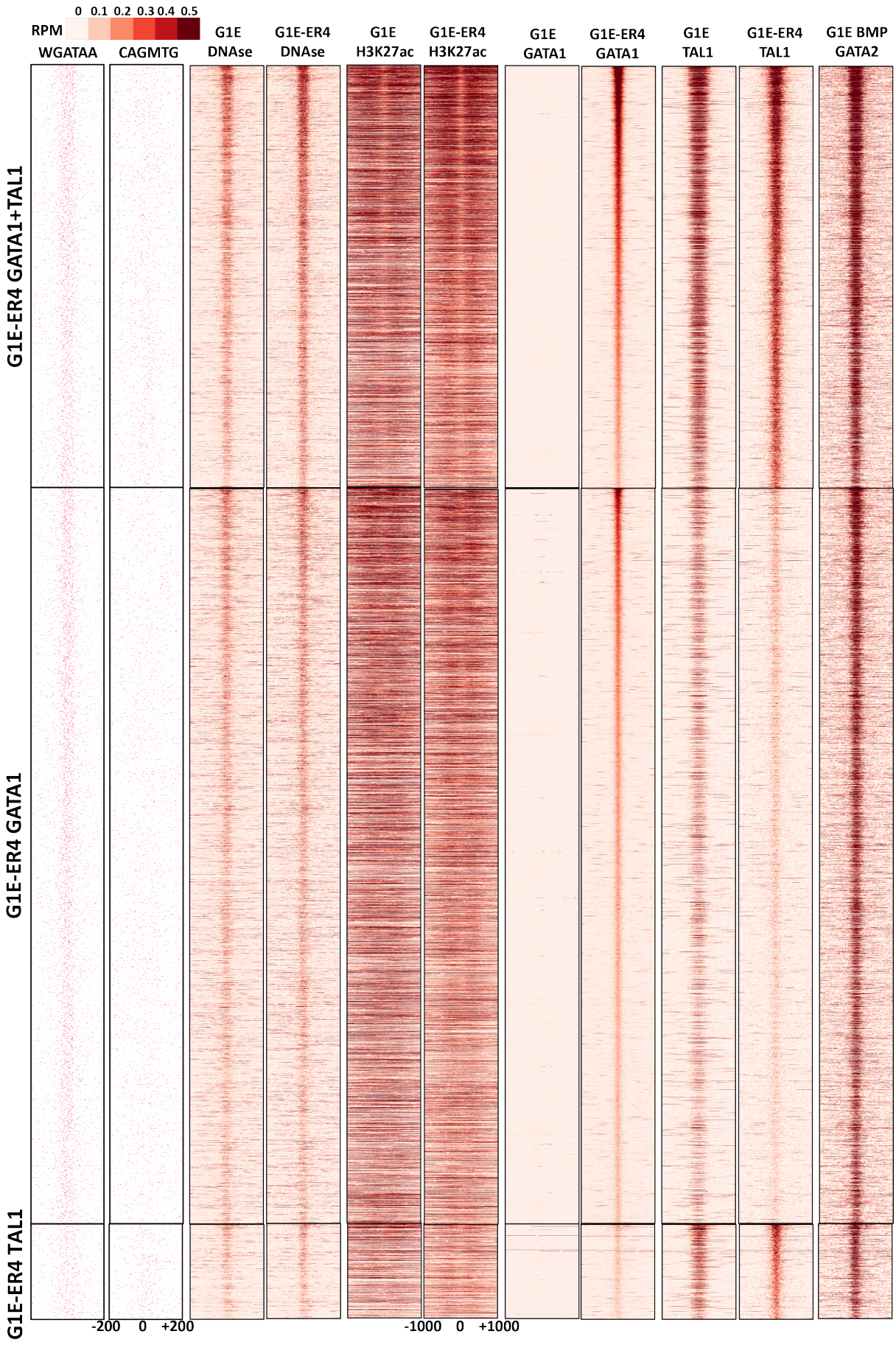




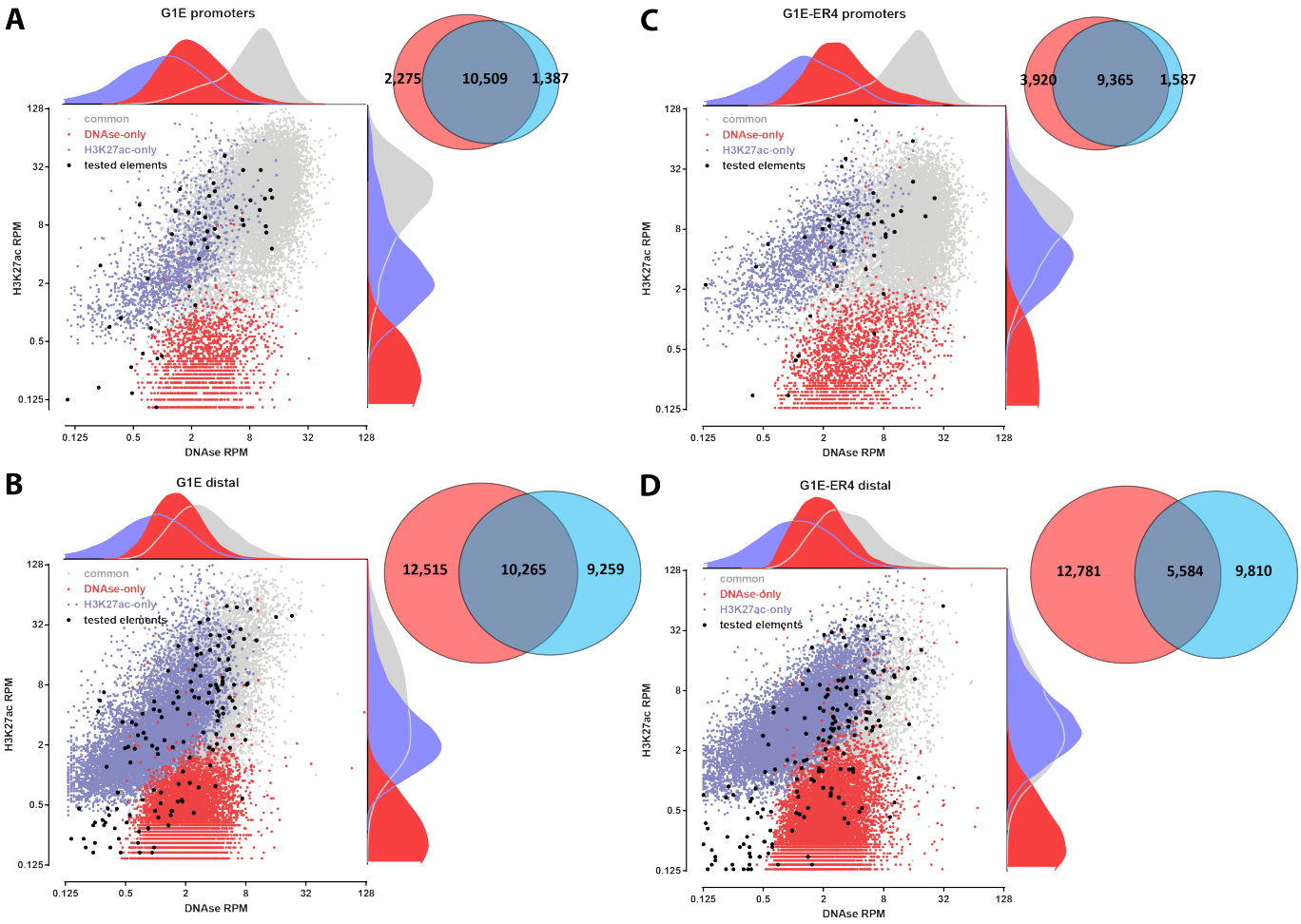




Supplementary Figure 8 *(preceding page)*: Correlation between regulatory activity and biochemical marks in C2C12 cells. (A and B) Correlation between fold activity and DNAse hypersensitivity, H3K27ac, p300, myogenin, MyoD and MEF2 occupancy in myoblasts and myocytes; (C) ROC curves showing biochemical 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.


RPM
0 0.1 0.2 0.3 0.4 0.5
WGATAA
CAGMTG
G1E DNAse
G1E-ER4 DNAse
G1E H3K27ac
G1E-ER4 H3K27ac
G1E GATA1
G1E-ER4 GATA1
G1E TAL1
G1E-ER4 TAL1
G1E BMP GATA2
G1E-ER4 GATA1+TAL1
G1E-ER4 GATA1
G1E-ER4 TAL1
-200 0 +200
-1000 0 +1000

Supplementary Figure 9 *(preceding page)*: 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.

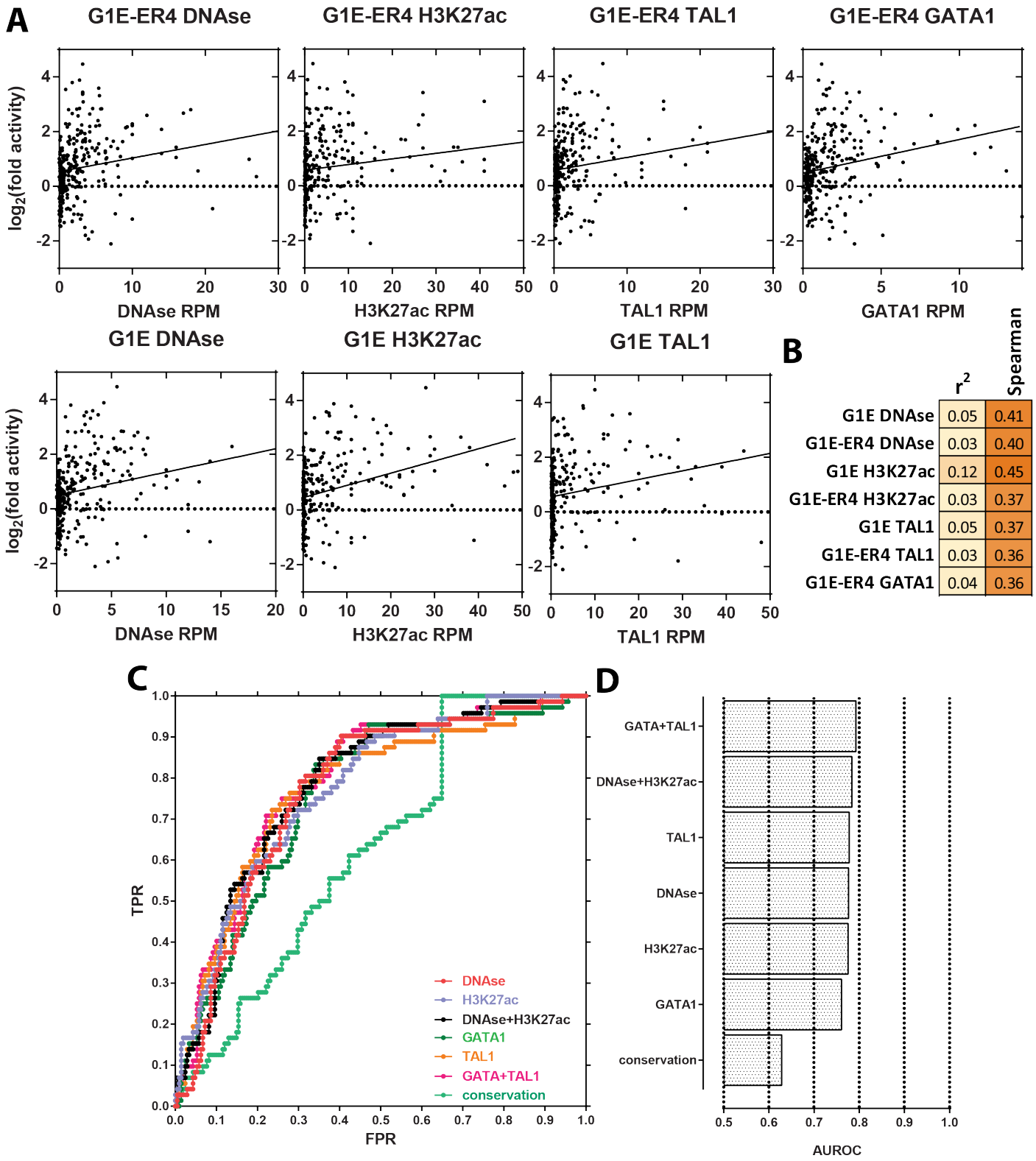


E

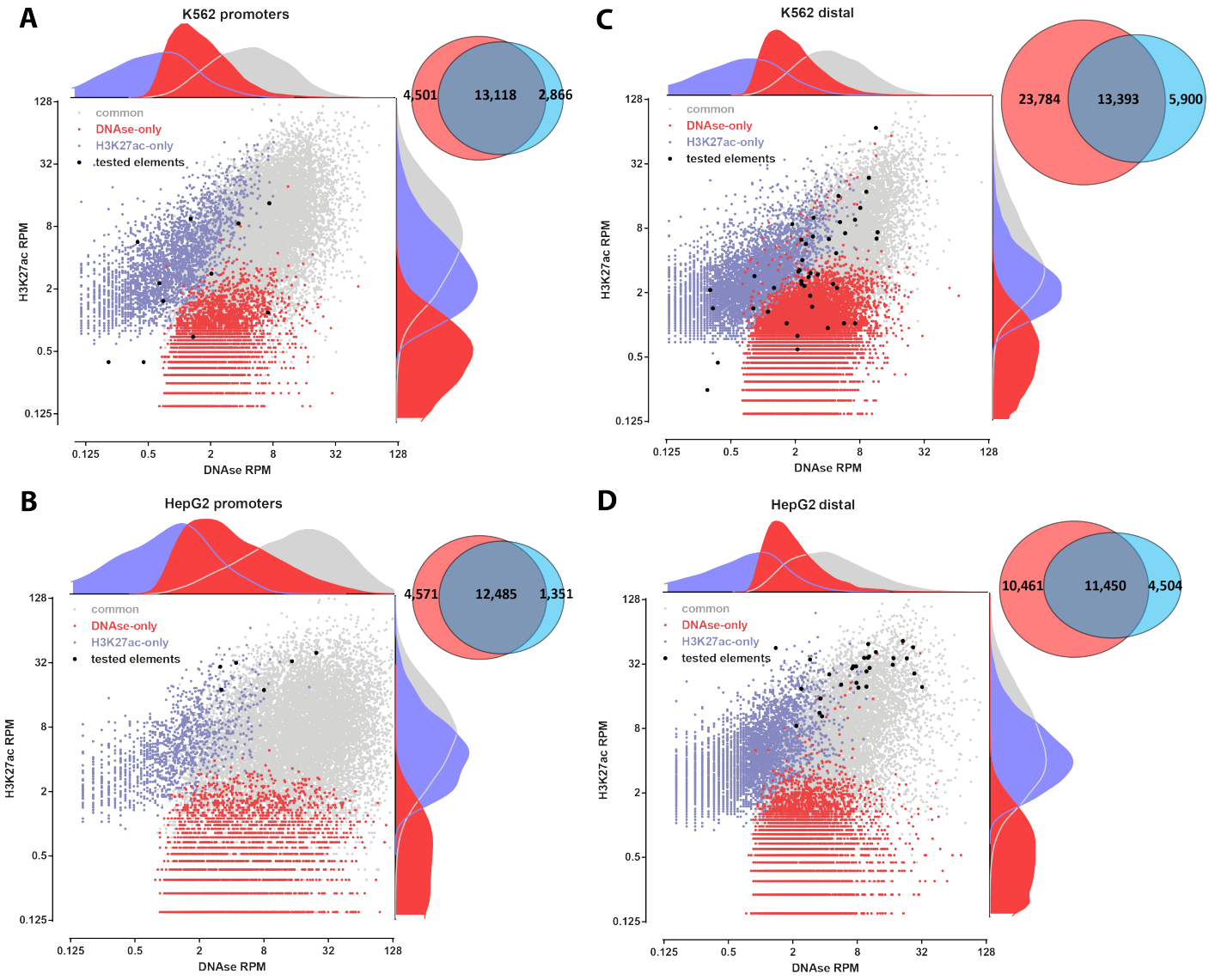
	promoter DNAse			distal DNAse		
	common	G1E only	G1E-ER4 only	common	G1E only	G1E-ER4 only
	11,217	1,567	2,040	12,288	10,492	5,953
G1E H3K27ac +	86.43%	51.95%	13.33%	44.17%	46.10%	20.02%
G1E H3K27ac -	13.57%	48.05%	86.67%	55.83%	53.90%	79.98%
G1E-ER4 H3K27ac +	80.73%	31.84%	13.63%	33.52%	23.42%	23.65%
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
	11,030	2,555	1,223	6,498	10,337	7,128
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%

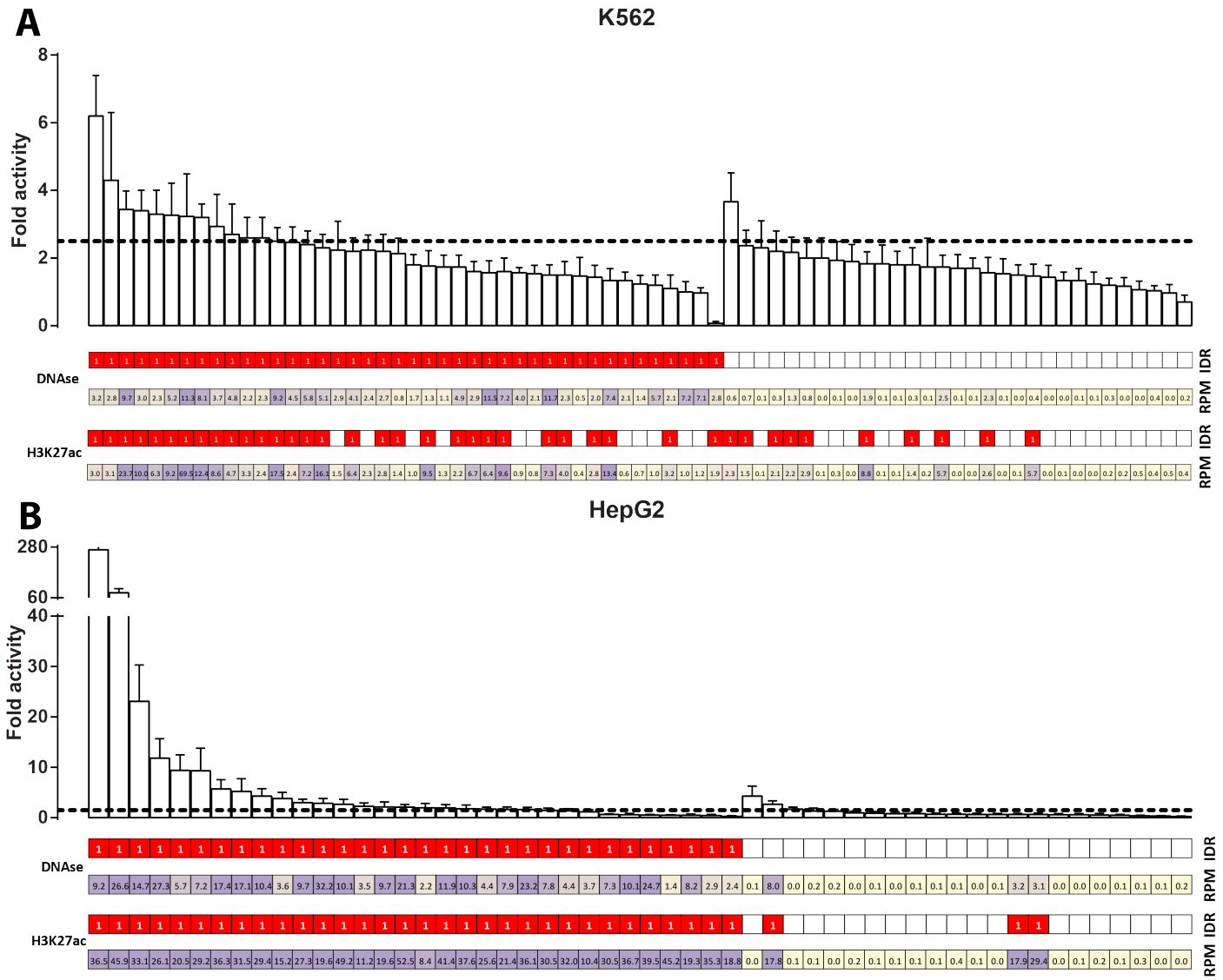
Supplementary Figure 10: Relationship between DNAse hypersensitivity and H3K27 acetylation during erythroid differentiation. (A) Overlap between DNAse hypersensitive and H3K27ac-positive promoter-proximal regions in G1E cells; (B) Overlap between DNAse hypersensitive and H3K27ac-positive promoter-proximal 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.



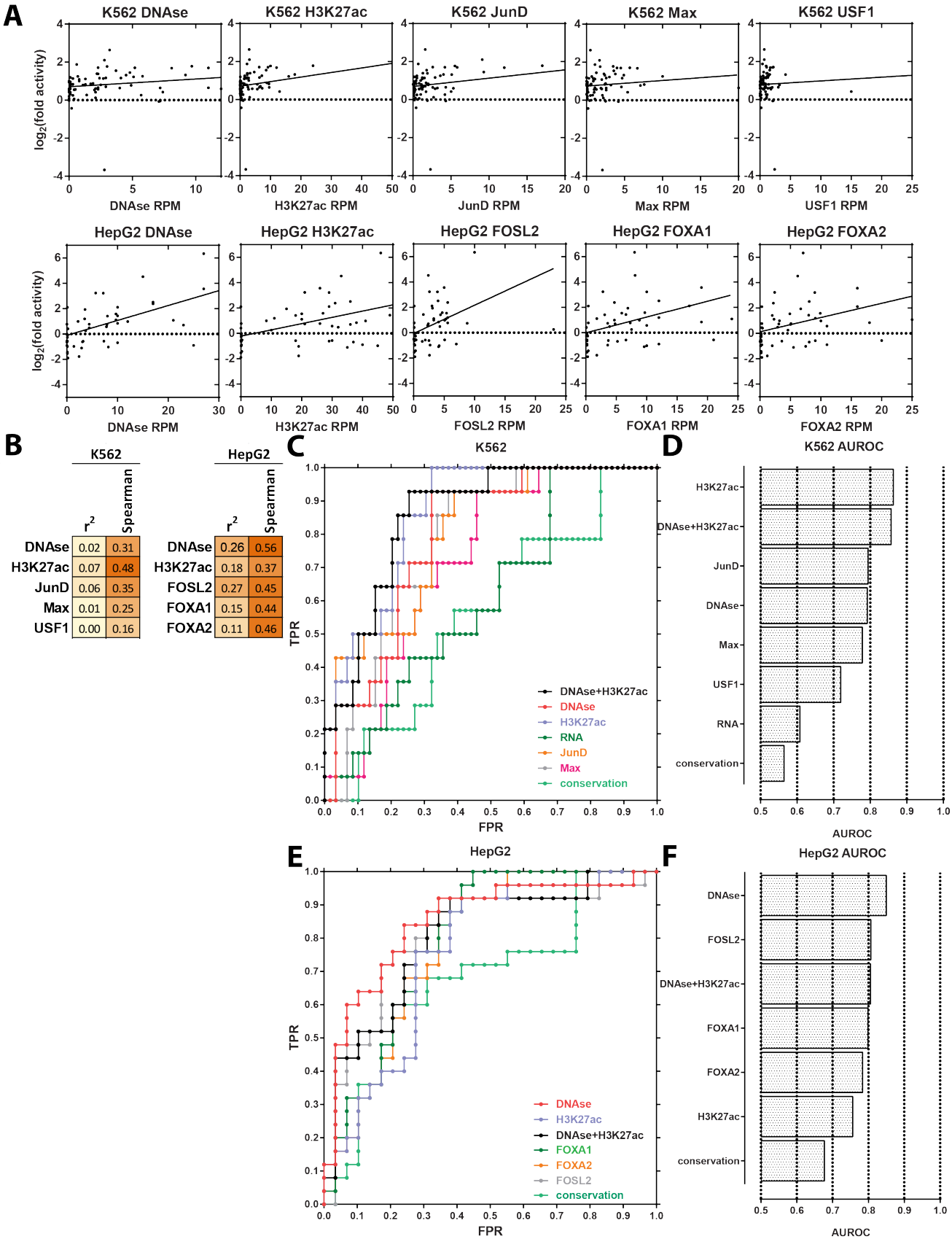
Supplementary Figure 11: Correlation between regulatory activity and biochemical marks in erythroid cells. (A and B) Correlation between fold activity in K562 cells and DNAse hypersensitivity, H3K27ac, TAL1, and GATA1 occupancy in G1E and G1E-ER4 cells; (C) ROC curves showing biochemical mark predictivity of cRE fold activity; (D) AUROC (area under ROC curve) values for different biochemical marks.



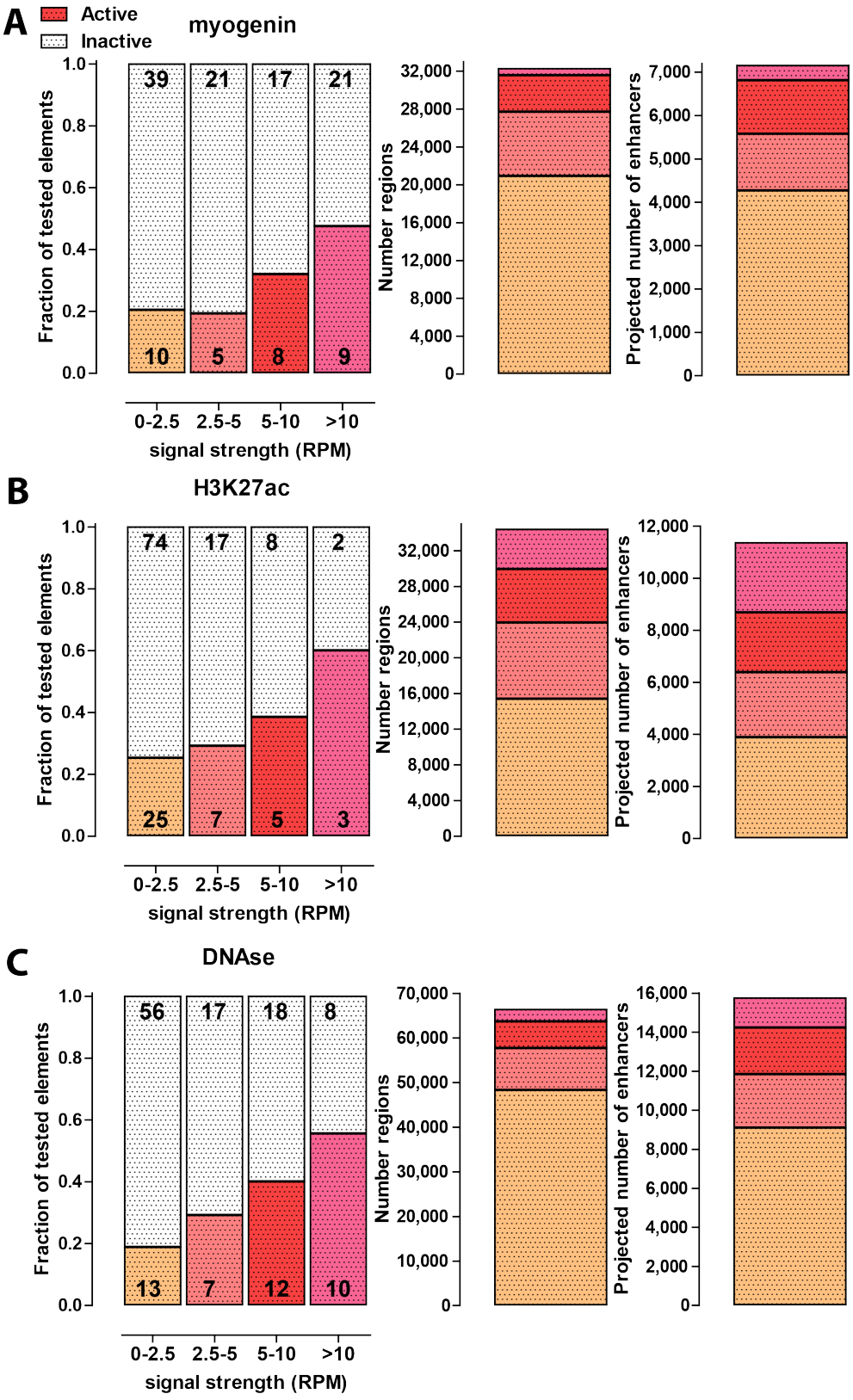
Supplementary Figure 12: Relationship between DNAse hypersensitivity and H3K27 acetylation in immortalized human cell lines. (A) Overlap between DNAse hypersensitive and H3K27ac-positive promoter-proximal regions in K562 cells; (B) Overlap between DNAse hypersensitive and H3K27ac-positive distal regions 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.



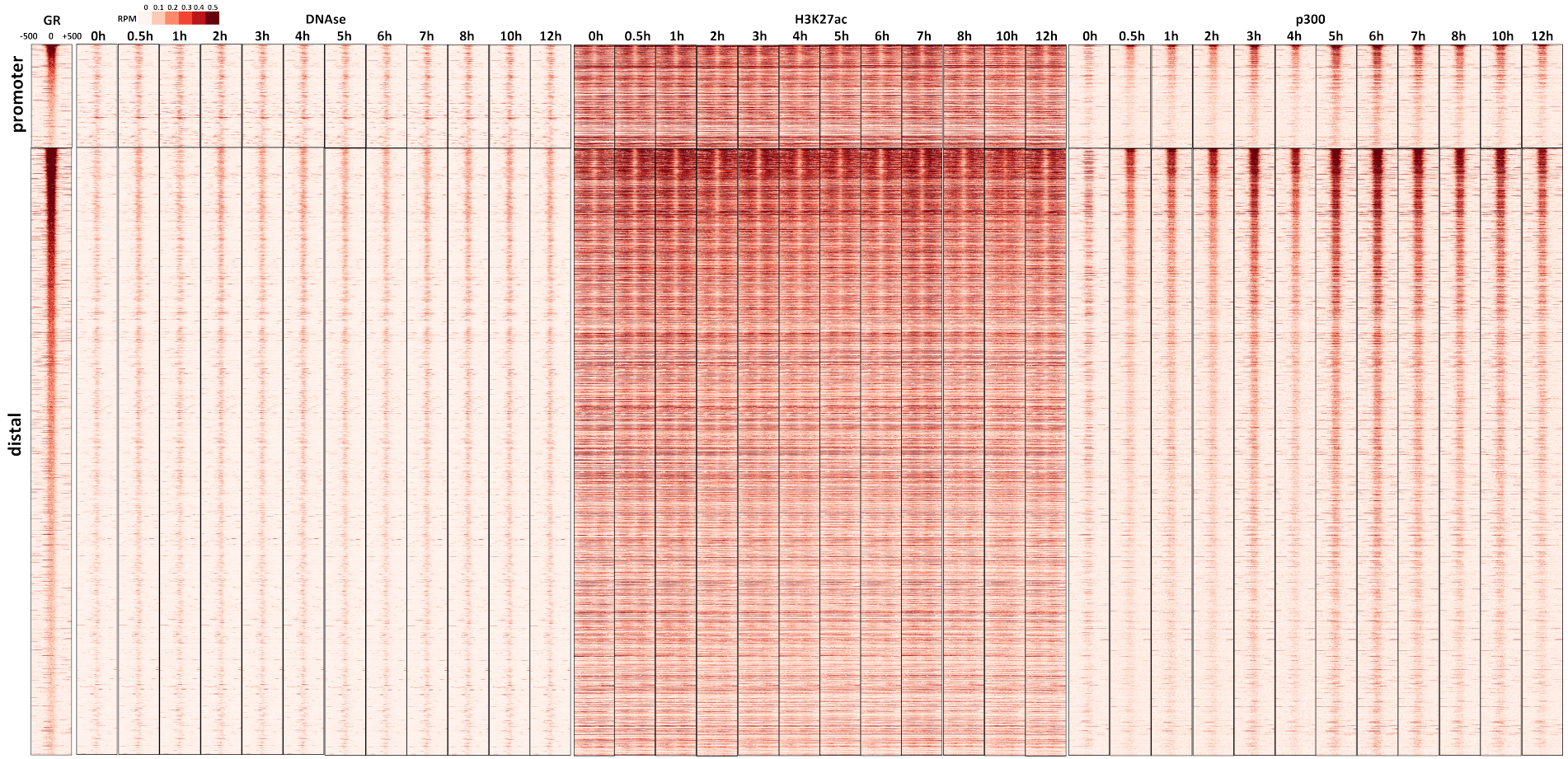
Supplementary Figure 13: Functional assay testing of cRE regulatory activity in human immortalized cell lines. Fold activity across biological replicates (n = ???) and technical replicates (n = ??? for each biological replicate) is shown. Candidate REs were sorted first by their DNAse status and then 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.


A
K562 DNAse
K562 H3K27ac
K562 JunD
K562 Max
K562 USF1
HepG2 DNAse
HepG2 H3K27ac
HepG2 FOSL2
HepG2 FOXA1
HepG2 FOXA2
log2(fold activity)
DNAse RPM
H3K27ac RPM
JunD RPM
Max RPM
USF1 RPM
FOSL2 RPM
FOXA1 RPM
FOXA2 RPM
B
K562
HepG2
r2
Spearman
DNAse 0.02 0.31
H3K27ac 0.07 0.48
JunD 0.06 0.35
Max 0.01 0.25
USF1 0.00 0.16
DNAse 0.26 0.56
H3K27ac 0.18 0.37
FOSL2 0.27 0.45
FOXA1 0.15 0.44
FOXA2 0.11 0.46
C
K562
TPR
FPR
DNAse+H3K27ac
DNAse
H3K27ac
RNA
JunD
Max
conservation
D
K562 AUROC
H3K27ac
DNAse+H3K27ac
JunD
DNAse
Max
USF1
RNA
conservation
AUROC
E
HepG2
TPR
FPR
DNAse
H3K27ac
DNAse+H3K27ac
FOXA1
FOXA2
FOSL2
conservation
F
HepG2 AUROC
DNAse
FOSL2
DNAse+H3K27ac
FOXA1
FOXA2
H3K27ac
conservation
AUROC

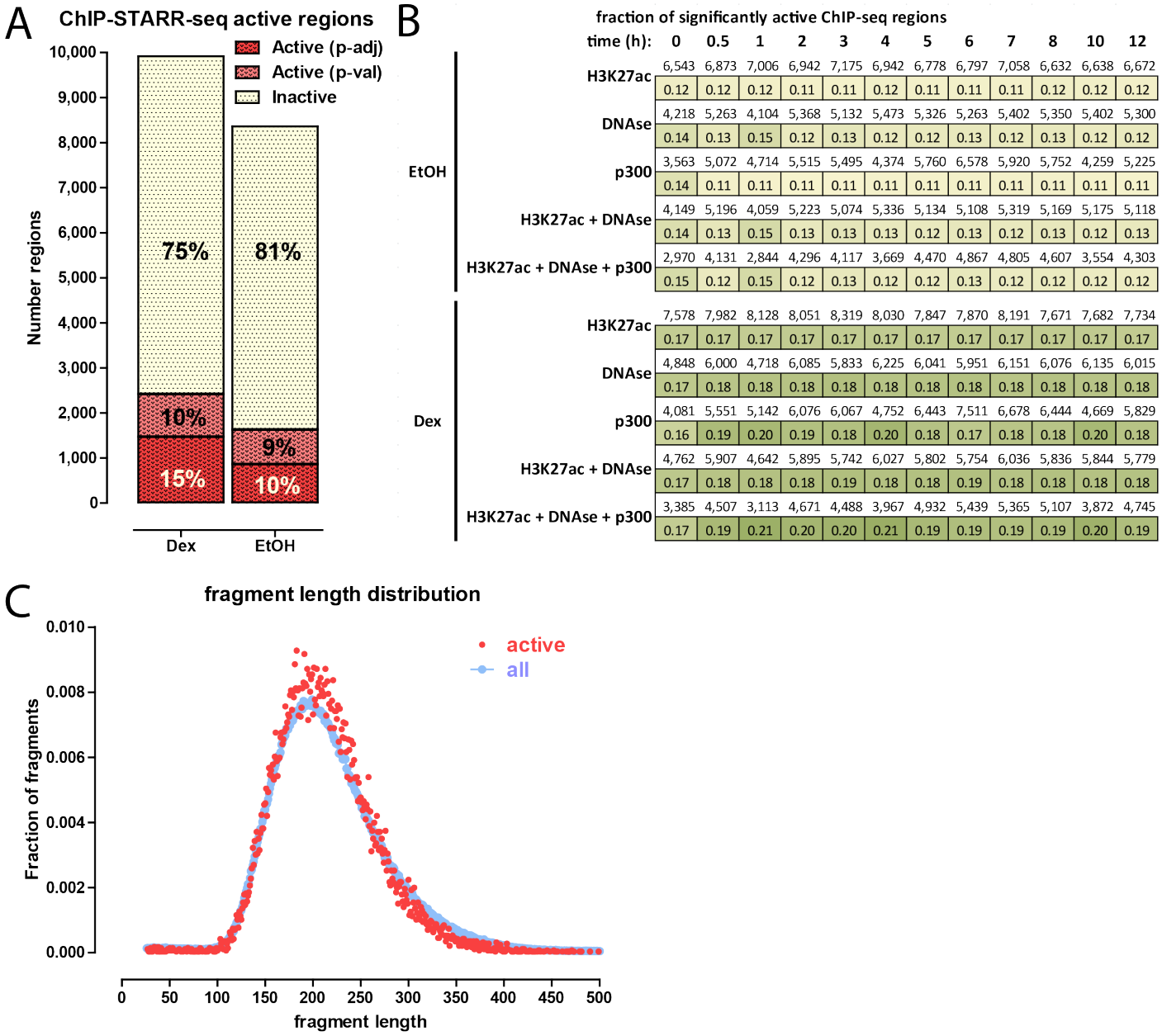
Supplementary Figure 14 *(preceding page)*: Correlation between regulatory activity and biochemical marks in human immortalized cell lines. (A and B) Correlation between fold activity in K562 cells and DNAse hypersensitivity, and transcription factor occupancy in K562 and HepG2 cells; (C) ROC curves 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.



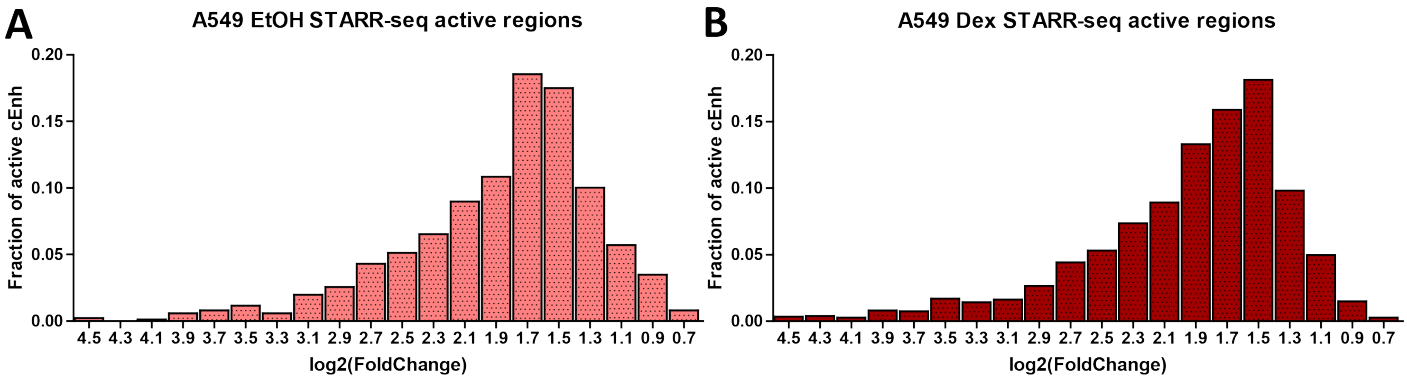
Supplementary Figure 15: Enrichment of active cEnhs in different classes of myogenic cEnhs defined by the strength of their biochemical signatures. (A) According to myogenin ChIP-seq signal strength; (B) According to H3K27ac ChIP-seq signal strength; (C) According to DNAse-seq signal strength. Shown on the right are the total number of cEnhs biochemically marked my myogenin, H3K27ac or DNAse, and the extrapolated number of active enhancer elements among them.



Supplementary Figure 16: Regulatory landscape of GR response in A549 cells. GR ChIP-seq peaks were split into promoter-proximal and promoter-distal sites, then ranked by ChIP-seq signal strength within each group. Shown is the distribution of p300 and H3K27ac ChIP-seq and DNAse-seq signal around each site.



Supplementary Figure 17: 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 18: 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.