Enhancer activity characterization of ENCODE biochemical signatures

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 and define the activities of the components regulating transcription. Genome-wide DNAse hypersentitivity, histone modifications and transcription factor occupancy measurements from the ENCODE and NIH Roadmap consortiae have been used to identify candidate regulatory elements (cREs) and profile them as candidate enhancers, promoters and insulators based on their characteristic biochemical signatures, now viewable on the SCREEN portal. However, accurately predicting the regulatory activity of cREs from the biochemical 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 gain initial insights into the predictivity of the biochemical signatures characteristic of enhancers we carried out tests for transcriptional enhancer activity for hundreds of candidate elements from five human or mouse cell types, including immortalized cells and cell lines modeling the early developmental differentiation to muscle or red blood cells. Our cEnh collections were selected using biochemical signature criteria ranging from individual transcription factor (TF) occupancy to tissue differential integrative machine learning models. We supplement these tests with a massively parallel reporter assay (MPRA) characterization of GR occupied elements. We find that irrespective of the selection criteria used, ∼50% of cEnhs exhibit enhancer activity with most being characterized by relatively modest biochemical signatures. In the same time, examination of the relative enhancer activities measured reveals that the majority of cEnhs genome-wide are likely to exhibit modest regulatory activity on their own. 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 repertoire 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[? ?], with the expression of each gene (protein coding or non-coding) being controlled by the integrated input of multiple proximal and distal enhancer, insulator and silencing elements.

The genome-wide mapping and characterization of non-coding 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[?] . The mapping and characterizing has been greatly aided by the advent of high-throughput sequencing and epigenomic tools. The biochemical annotations of known regulatory elements resulted in biochemical signatures for different types of cREs. For example, active promoters in eukaryotes are classically associated with the trimethylation of lysine 4 on histone 3 (H3K4me3)[?] , as well as other biochemical signatures, such as DNAse hypersensitivity[?] . Enhancer elements have shown to exhibit their own biochemical signature, featuring DNAse hypersensitivity, the H3K27ac and H3K4me1 post translational modifications on nearby histones (histone marks), and occupancy by the p300 acetyltransferase[? ? ?] orchestrated by sequence-specific transcription factors (Figure 1A).

These biochemical signatures enable the compilation and characterization of lists of candidate regulatory elements (cREs), but they do not on their own allow for the conclusive identification of any given element as having an active biological function. While active regulatory elements exhibit characteristic biochemical signatures, the reverse (that the presence of a biochemical signature necessarily means function) cannot be inferred straightforwardly[?] . 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 of a characterized trans-activator, which can be reproducibly identified using ChIP-seq are in fact active enhancers[?] . Therefore, individual cREs in the lists compiled by efforts such as the ENCODE and mouseENCODE consortia[? ?] have to be subsequently tested and functionally characterized in detail.

The ultimate functional characterization of candidate enhancers 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[? ?] .

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 for enhancer activity 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[? ?]).

Numerous developmental enhancers have been characterized following that approach[? ? ?] , starting from lists of candidate Enhancers (cEnhs) complied based on comparisons of evolutionary conservation and later from biochemical signatures derived from ChIP-seq datasets. However, such studies have often focused only on the most outstanding biochemical signatures[?] , thus obtaining very high success rates that is unlikely to be representative of the genome-wide populations.

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[?] , 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[? ? ? ? ? ? ?] , including the question of testing cREs for activity within the context of the ENCODE Consortium[? ? ?] . 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 (Figure 1). Thus to what extent complete REs are assayed and what the corresponding false negative rate is have always been an obvious concern 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, likely resulting in numerous false positives or negatives.

The concern of cross-talk can be alleviated through the use of genome-integrated MPRA constructs[? ?] , however the viral delivery vectors have a definite preference for integrating in dense regions and therefore likely impart the bias on activity of the region where they are integrated.[?] . The short length of constructs tested remains a significant issue, and one that might be behind the low positive rates reported by MPRAs[?] . There is therefore a major gap in the field that needs to be filled by testing a large number of individual constructs with larger sequence context (500-1000 bp) using a traditional luciferase assay, and using the resulting data to examine the performance of biochemical signature predictions.

To this end, as part of the ENCODE Project Consortium efforts towards functional validation of cREs, we tested the regulatory activity of hundreds of candidate enhancer elements (cEnh) using constructs of such lengths (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 (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.

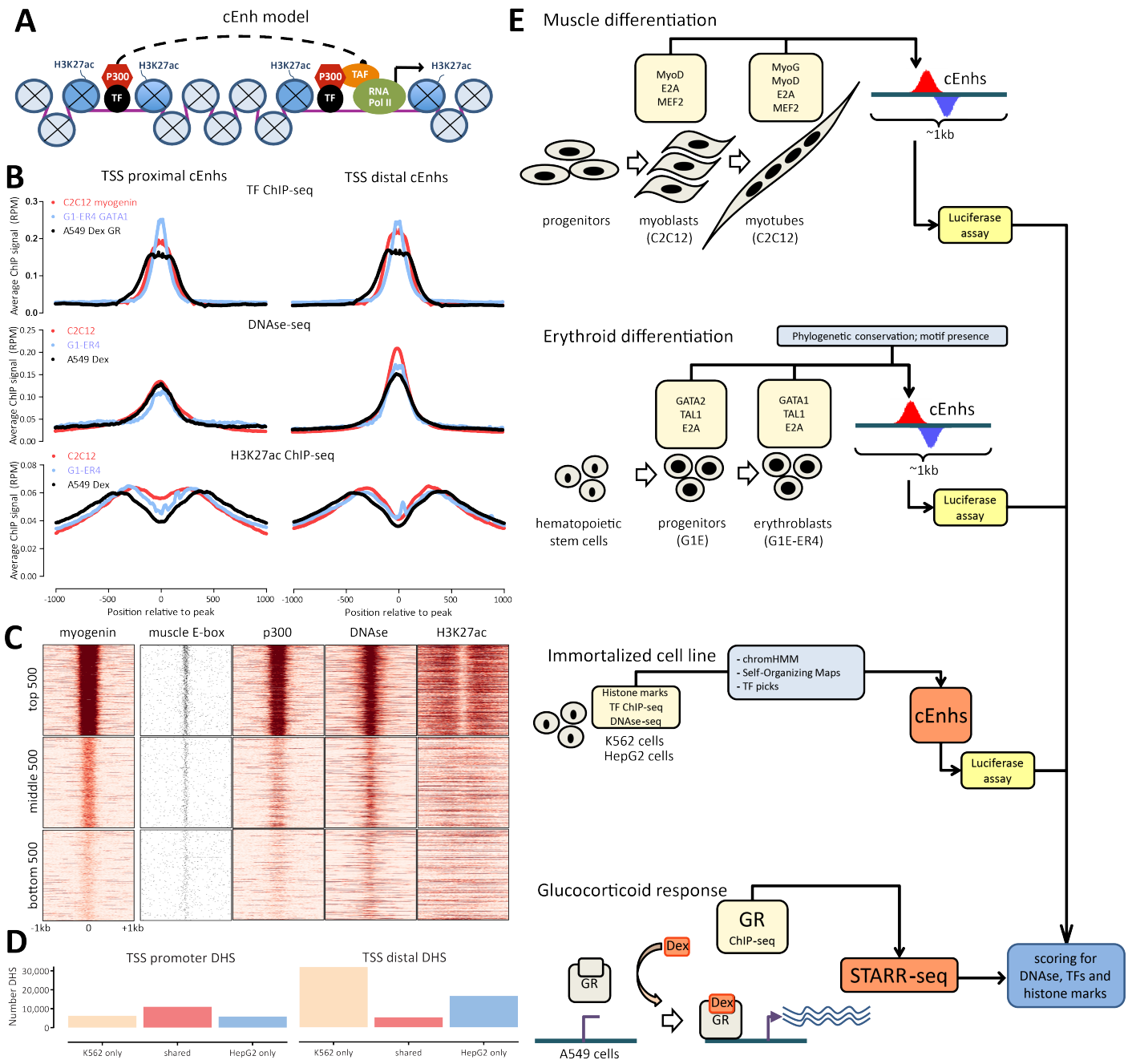


Figure 1: Biochemical signatures and functional testing of candidate enhancer elements (cEnhs) in mammalian genomes. (A) Biochemical signatures of cEnhs and promoters. Active enhancers are characterized by DNAse hypersensitivity due to nucleosome depletion, by p300 occupancy and by H3K27ac, as well H3K4me1 (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 upon Dexamethasone stimulation of human A549 cells), DNAse hypersensitivity and H3K27ac; (C) The distribution of biochemical signal strength varies over a large continuum. Shown are the signal distribution for myogenin, p300, DNAse-seq, and H3K27ac relative to the summits of the top 500, middle 500 and bottom 500 reproducible myogenin ChIP-seq sites (total $n = 32,278$) in differentiated C2C12 muscle cells, as well as the distribution of the cognate myogenin TF binding

(legend continued on next page)

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 or number 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 enhancer testing of full size 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 the consortium 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[? ? ? ?] . Active enhancers and promoters are typically occupied by sequence-specific transcription factors, marked by H3K27ac and exhibit DNA hypersensitivity (Figure 1B), and cREs 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[?] . However, there need not be a strict one-to-one relationship between these biochemical signatures and functional REs[?] , as their presence alone does not necessarily imply that the cRE plays an active regulatory role. In addition, biochemical marks by themselves do not provide direct understanding of how exactly functional REs are specified or exercise their function.

While the full catalog of REs in the genome includes promoters, insulators, enhancers, silencers, and elements with other function; for the purposes of this study we focused primarily on distal candidate transcriptional enhancers (cEnhs). A major reason for this choice is that distal cEnhs constitute the bulk of cREs distinguishing different cell types from each other, in contrast to, for example, active promoters or TSS proximal enhancers, a major fraction of is biochemically occupied in multiple tissues making predictions for cell type specific active enhancers difficult. (Figure 1C).

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 within a biochemically occupied region on its own does not imply non-

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

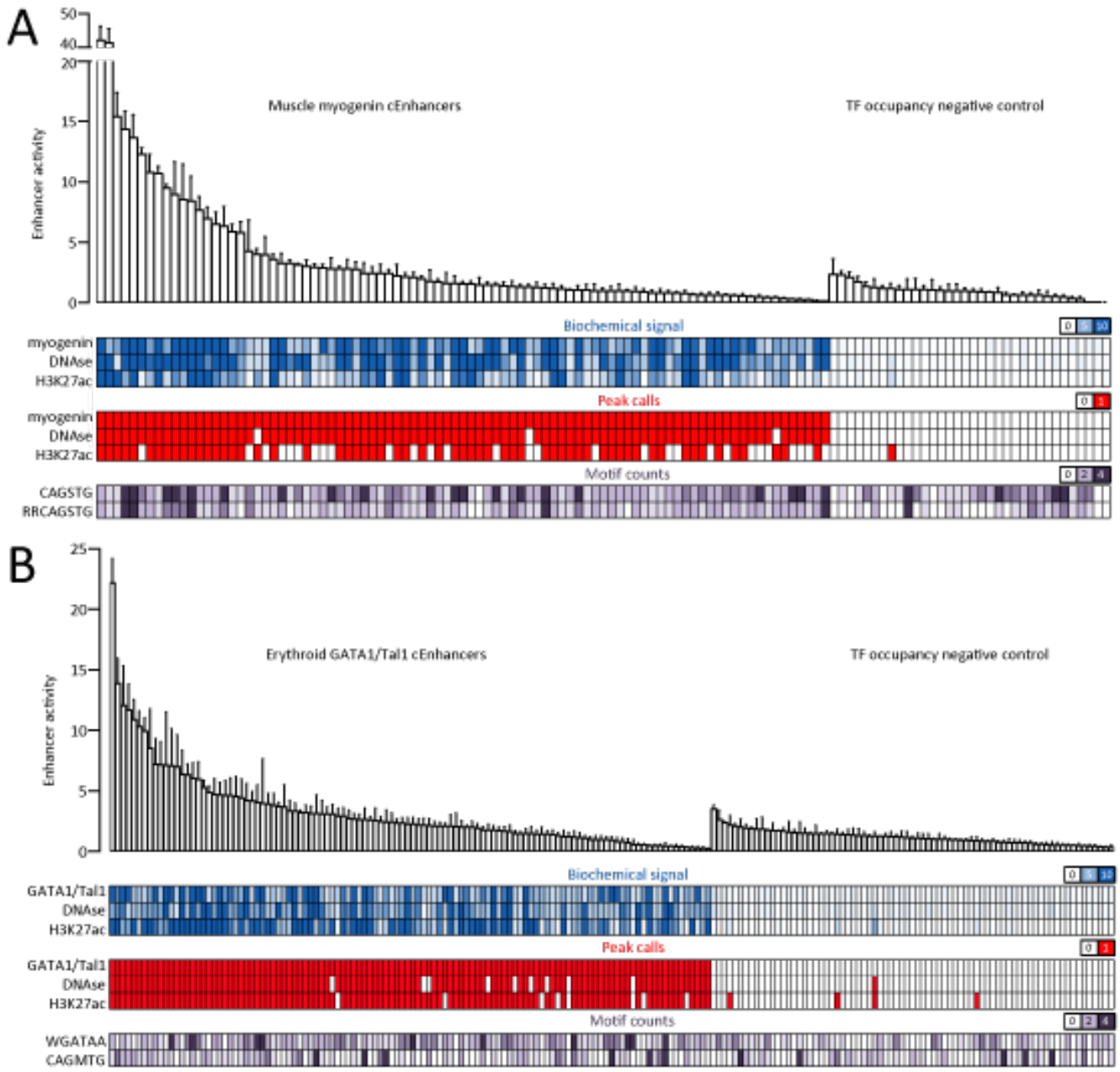


Figure 2: Functional testing of cEnh regulatory activity in mammalian cells. (A) Functional assay testing of cEnh regulatory activity in the context of muscle differentiation. Shown is luciferase assay fold activity in differentiated C2C12 myocytes across technical replicates ($n = 4$). The red arrow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

functionality, as demonstrated by recently evolved lineage-specific REs which eluded comparative genomic analyses. Therefore cREs identified 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[? ? ? ? ? ? ? ? ? ? ?] . 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, and accurately asses the functional predictivity of biochemical marks we carried out our tests of cEnhs in a variety of mammalian systems using transient luciferase reported enhancer activity assays, which allow for much larger segments of DNA to be tested for enhancer activity, between 500 and 1000 bp (Figure 2), thus most likely encompassing complete functional regulatory elements. 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[? ?] .

Enhancers driving the differentiation of muscle and red blood cells

The differentiation of a precursor into a specified cell type has long been used to characterize regulatory elements as it provides a key differential where biochemical measurements can be contrasted to a prior state. To this end we selected cEnhancers based on cell type specific TFs known to function as trans-activators of differentiation in their systems.

As an example, 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[? ? ? ?] . The occupancy landscape of these factors and other major regulators of the differentiation of muscle are provided for reference (Supplementary Figure 10). 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 4), the majority of which contain the classical muscle E-box sequence motif CAGSTG often in the extended RRCAGSTG form[? ?] . While the majority of ChIP-Seq regions contain this motif, they still cover less than a percent of the available motifs in the genome implying that occupancy is highly selective. The mouse C2C12 cell line[?] 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 the focus of our study.

In the context of muscle differentiation, we selected cEnh regions based on ChIP-seq data for the myogenin TF in differentiated C2C12 myocytes. We randomly selected 89 regions spanning the full range of myogenin occupancy levels. Of these 88 contain CAGSTG and 84 contain RRCASGTG E-boxes. We selected additional elements to fully sample cEnhancers affiliated with genes well studied in the development of muscle cells, in order to ask if affiliation with these genes imparted a functional bias even with similar underlying biochemistry. We also selected a group of 23 occupancy negative controls out of a set of characterized T-cell and neuronal enhancer regions 21 also containing E-boxes to ascertain if observed functionality with our assay is specific for regions biochemically marked in muscle cells. A second set of 11 negative control elements, 6 of which contained eboxes, were selected near expressed genes and ChIP-Seq positive regions in order to assay the baseline function of biochemically neutral regions of similar size to the cEnhancers.

The process of hematopoiesis has also been characterized over decades, looking at the steps that transform progenitor cells into red blood cells. The transcription factor GATA1 is a key regulator of this process, being responsible for the differentiation and proliferation of red blood cells (6), and when knocked out results in an anemic phenotype. TAL1, a BHLH protein, is required for multiple functions in hematopoiesis, including terminal differentiation of red blood cells (7) and its occupancy overlaps heavily with GATA1. The G1E-ER4 is a mouse origin GATA1-knockout cell line that can rescue the presence of GATA1 by estradiol exposure. The occupancy landscape of these factors and other major regulators of the differentiation of red blood cells are provided for reference (Supplementary Figure 10)

In order to isolated Erythroid cEnhancers were selected

from GATA1 and Tal1 ChIP-Seq occupied regions of the mouse genome and tested in the easily transfectable human K562 cells. The 113 erythroid cREs were selected from both GATA1 and Tal1 occupancy regions of the mouse genome. GATA1 ChIP-chip data in the cell line G1E-ER4 was used to select GATA1 cREs. Elements were subjected to independent validation by ChIP-qPCR, with 53 validated comprising candidate Regulatory Elements. Genes induced by restoration of GATA1 are frequently association of TAL1-bound DNA segments. TAL1 ChIP-seq data was used to select 60 ChIP-seq positive regions that were tested for enhancer activity after transient transfection in K562 cells. Genomic elements that were not significantly occupied by GATA1 were selected as occupancy negative controls for the assay[? ?] , totaling 74 elements. Of these, GATA1 ChIP-chip that failed ChIP-qpcr validation comprised 63 of the negative control elements. Another set of 11 DNA segments were not called as peaks in the ChIP-chip analysis, these are labeled GHN for GATA1 hit negative. Although they are not significantly occupied, 45 of the negative control elements contain a GATA1 motif (WGATAA); while 40 contain a Tal1 occupancy motif (CAGMTG) providing a specificity control for biochemically occupied regions.

We provide a summary of the enhancer assay results from each of these two systems where we selected candidates based on TF occupancy in Figure 2; together with the biochemical signal measured and statistically significant peak calls (IDR0.05) for the key TF; DNAse-seq and H3K27ac ChIP-Seq. We also provide the number of DNA motifs present within each cEnhancer tested. We note that even though in both cases the occupancy is highly selective for a tiny fraction of the available motifs, only $\sim 50\%$ of candidate Enhancers are active in each system and that both the strength or presence of the biochemical signals or the number of motifs appears to be non-predictive of enhancer function. We found no measurable difference in the predictivity of activity of TF occupancy between elements randomly selected 2 in the genome and the locus affiliated cEnhancers presented in the supplemental figure **??**. remove from supp fig7 the negatives, random picks. keep only the section for the locus picks

Biochemical signature strength is not strongly correlated with enhancer strength

The signature combination of H3K27ac and DNAse hypersensitivity outside of promoter regions has been used by ENCODE to characterize cREs with the potential for enhancer function. The overlap of H3K27ac and DNAse hypersensitivity signal is significant but the two measurements of the key identifiers have no correlation to each other in either C2C12 and G1E-ER4. **??** Even though the TSS distal genome-wide TF occupancy and DNAse hypersensitivity have almost identical profiles /refFig1B the measured correlation between DNAse hypersensitivity signal and TF occupancy measurements is mediocre at best.

With such lack consistency of signal between even closely associated biochemical measurements it is no surprise that we found no correlation between the magnitude enhancer assay activity measured and the biochemical signal strength in 5 for either C2C12 or G1E-ER4 cells. We include detailed correlations of activity on the enhancer assay to all the major TFs and biochemical measurements in each system together with ROC predicticity curves in the supplementary figures **??**.

Generic trans-activator TFs are equally predictive of enhancer activity as their tissue specific counterparts

The K562 and HepG2 comprised the primary Immortalized cell lines selected by the ENCODE consortium for extensive annotation[? ?] and as such are the source of a significant portion of ENCODE data. Overall both these cell lines displayed a similar relationship between DNAse hypersensitivity and H3K27 acetylation to that found in mouse C2C12 cells with similar numbers of candidate enhancers found in both TSS proximal and distal regions. **??**

We aimed at comparing the predicitivy of cEnhancers selected from a cell-type specific trans-activator TF to that of a collection of overlapping occupancy by non cell-type specific positive acting factors mapped by ENCODE in K562. One might expect that by requiring the occupancy of multiple factors, one would improve the predictivity of active enhacer function.

To this end we used the vast throve of biochemical annotations in K562 cells to generate cEnhancers ($n = 36$) selected occupied by a large ($n = 4+$) number of known generic trans-activator TFs, to contrast the predicitivity of cell type specific TFs used to select cEnhs in our differentiation models. A set of regions occupied in HepG2 ($n = 32$), but void of occupancy in K562 were used as negative control elements.

The results of these enhancer assays based on TF-centric cEnhancers and occupancy negative control elements are summarized in **??**. We find no significant difference in the predictivity of cEnhancers selected by using cell-lineage specific contrasted with selection of cEnhancers based on generic trans-activator TFs.

Biochemical marks are similarly predictive of active enhancers in multiple cell types

Next, 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 **??**). As such, we applied several different strategies for compiling lists of cEnhs to be tested, broadly contrasting the TF-centric with TF-agnostic picks based primarily on chromatin state signatures and evolutionary conservation in order to compare their predictivity for active enhancers.

We relied on the multiple computational approaches for integrating high-dimensional collections of functional genomic datasets into a small set of chromatin states that

have been devised over the last few years and applied to the problem in ENCODE cell lines, including the Hidden Markov Model-based Segway[?] and chromHMM[?] , as well as Self-Organizing Maps[?] (SOMs). We selected cEnhs in K562 cells based on Segway and chromHMM chromatin state assignments and the presence of DNAse and H3K27ac ($n = 30$), with elements lacking both marks used as negative controls ($n = 21$). We also selected cEnhs based on SOMs trained DNAse and histone mark ChIP-seq data over multiple ENCODE cell types; these cEnhs 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 were also derived from other SOM regions that lacked both marks in HepG2 ($n = 18$) were used as negative controls. The individual enhancer assay results are summarized in the supplemental figure **??** while the correlations to the biochemical data, which are comparable to those found in C2C12 and G1E-ER4, are found in the supplemental figure **??**.

In the context of hematopoiesis, a TF-agnostic cEnhancer selection was based on evolutionary conservation and on the integration of measurements of multiple histone modifications and open chromatin. A total of 46 cREs were selected from regions conserved in alignments of multiple mammalian species that were characteristic of known regulatory regions. 6 additional DNA segments highly conserved outside of mammals that have a match to a GATA1 binding site motif were also included. These elements were annotated for DHS and H3K27Ac in G1E-ER4 cells; with 26 elements scored positive for both biochemical marks and were used as cEnhancers while the 13 elements lacked both of the modern marks were used to score as a negative control set.[? ?]

The results of these enhancer assays and occupancy negative control elements are summarized in **??**.

figure3 GATA1 conservation bar plot has an error; 14/12 seems reversed proportion?

Overall, we find no major difference in the predictivity of enhancer activity for either TF-centric or TF-agnostic selections across multiple tissues, and estimate that approximately half of the population of biochemically DNAse/H3K27ac co-marked cEnhancers are biologically active.

Similarly TF agnostic integrative methods; or cEnhs selected based on evolutionarily conservation all found no significantly improved predictivity of enhancer activity within biochemically marked regions. However, as expected the GATA1 motif containing evolutionarily conserved picked cEnhs scored blind to the biochemistry would have yielded a much reduced fraction (38 vs 53 percent) of predicted enhancers.

A detailed breakdown of the DNAse/H3K27ac signature combinations underlying all cEnhancers tested is provided in **??**.

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

The biochemical signals measured by a ChIP-seq experiment vary widely within statistically significant regions. **??** The genomic region surrounding the BTG2-MYBPH-myogenin genes **??** shows that the general range of ChIP-Seq signals is fairly well represented even near outstandingly transcribed genes. As expected from the biochemistry found, there was no significant difference in the predictivity or relative activity measured on the assay between the subpopulation of cEnhancers culled from a random selection of TF occupied sites compared to ones affiliated with outstanding genes.

Overall smaller biochemical signals make up the vast majority of the catalogue of myogenin occupancy based cEnhancers in the genome **??**. One key unresolved question is if biological activity is equally present within the spectrum of the biochemical measurement. To address this question we selected our cEnhancers to sample the spectrum of ChIP-Seq occupancy found in the genome and assayed them for enhancer activity. **??**

Overall we found that the vast majority of enhancers in a given cell type are expected to be marked by relatively myogenin signals **??**. This finding also held true when comparing the signal for DNAse hypersensitivity and H3K27ac contained within the same cEnhancer regions **??**

Finally, we sought to the contrast the contribution of enhancers found in differentiation with external signaling stimulation as 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 A549s are a human lung epithelial cell line which has been used to study the genomic response to the stimulation of the glucocorticoid receptor (GR), normally involved in the suppression of inflammatory responses. The biochemical landscape of GR reponse is provided in the supplemental figures **??**. Using an IP for GR against the DNA of A549 cells induced by exposure to dexamethasone; we created a library to be used in a self-reporting enhancer assay (STARR-seq) cite tim/vockley. Unlike our previous selections, which were heavily biased for distal regions, this provided an assay that tested both TSS proximal and distal cEnhancers in reporter vector that positioned the regions 3' relative to the contained promoter element.

We selected the genomic regions which were significantly represented them in the library and mapped the corresponding GR ChIP-seq signal **??**. While the assay biased for the overall presence of relatively strong biochemical signals the overall proportion of enhancers found active in each signal class remains remarkably similar **??**.

were any occupancy negative regions significantly represented in the library for Starr-Seq? If so we need to state this.

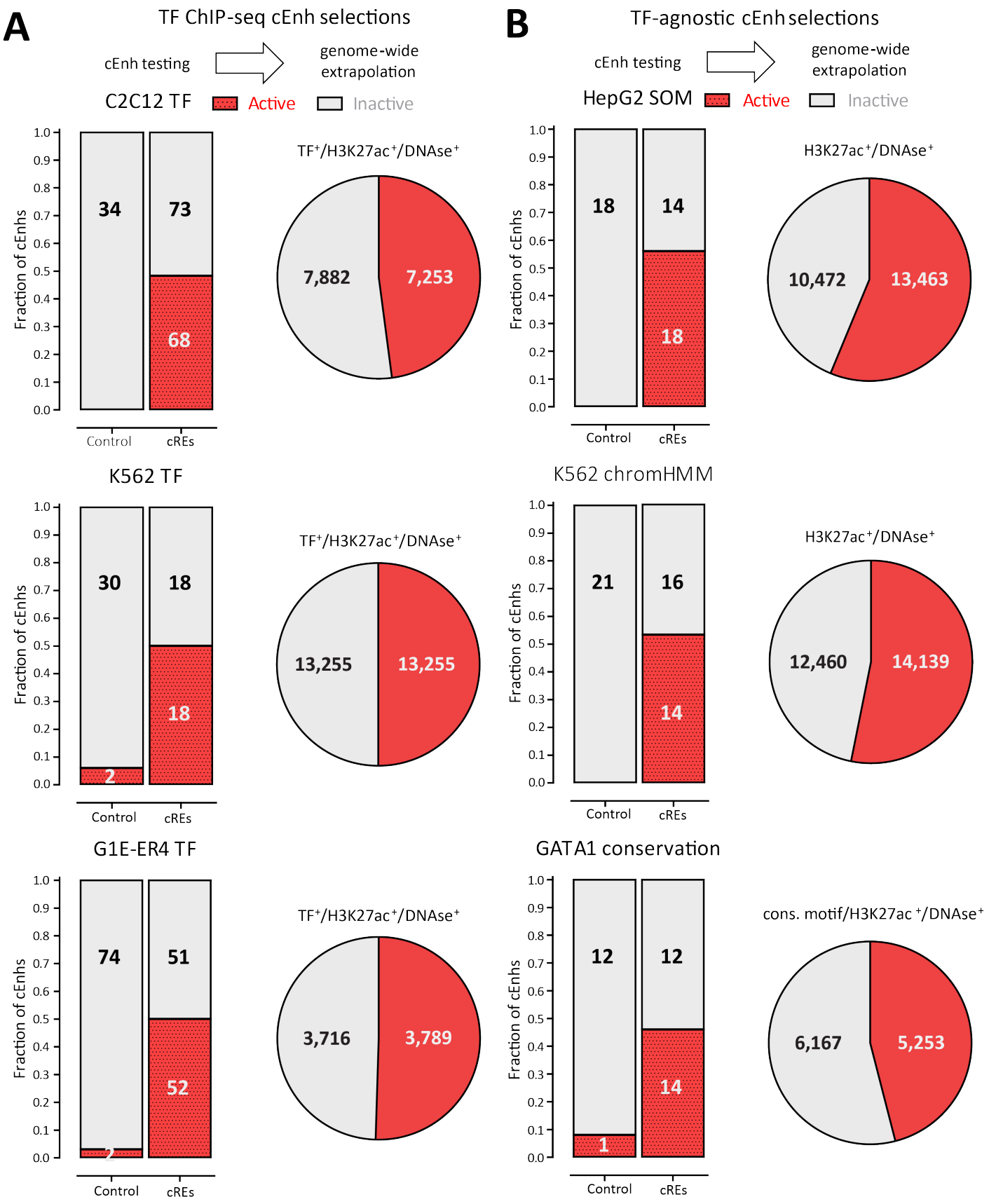

A
TF ChIP-seq cEnh selections
cEnh testing
genome-wide extrapolation
C2C12 TF
Active
Inactive
Fraction of cEnhs
34
73
68
Control
cREs
TF+/H3K27ac+/DNAse+
7,882
7,253
K562 TF
30
18
2
18
13,255
13,255
G1E-ER4 TF
74
51
2
52
3,716
3,789
B
TF-agnostic cEnh selections
HepG2 SOM
18
14
18
H3K27ac+/DNAse+
10,472
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

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 fraction of active constructs in negative controls and cEnhs are shown on the left. The expected number of active cEnhs genome-wide is extrapolated on the left based on the number of $TF^+/DNAse^+/H3K27ac^+$ regions in the genome; (B) TF-occupancy agnostic selections. Tested eEnhs selected using Self-Organizing Maps in HepG2 cells, chromHMM in K562 cells, and evolutionary conservation of GATA1 motifs in G1E cells were further subselected with the additional requirement of exhibiting DNAse hypersensitivity and the H3K27ac histone mark. The fraction of active constructs in negative controls and cEnhs are shown on the left. The expected number of active cEnhs genome-wide is extrapolated on the left based on the number of $DNAse^+/H3K27ac^+$ (for HepG2 SOM and K562 chromHMM selections) $DNAse^+/H3K27ac^+$ regions with a conserved GATA1 motif (for GATA1 conservation selections) in the genome.

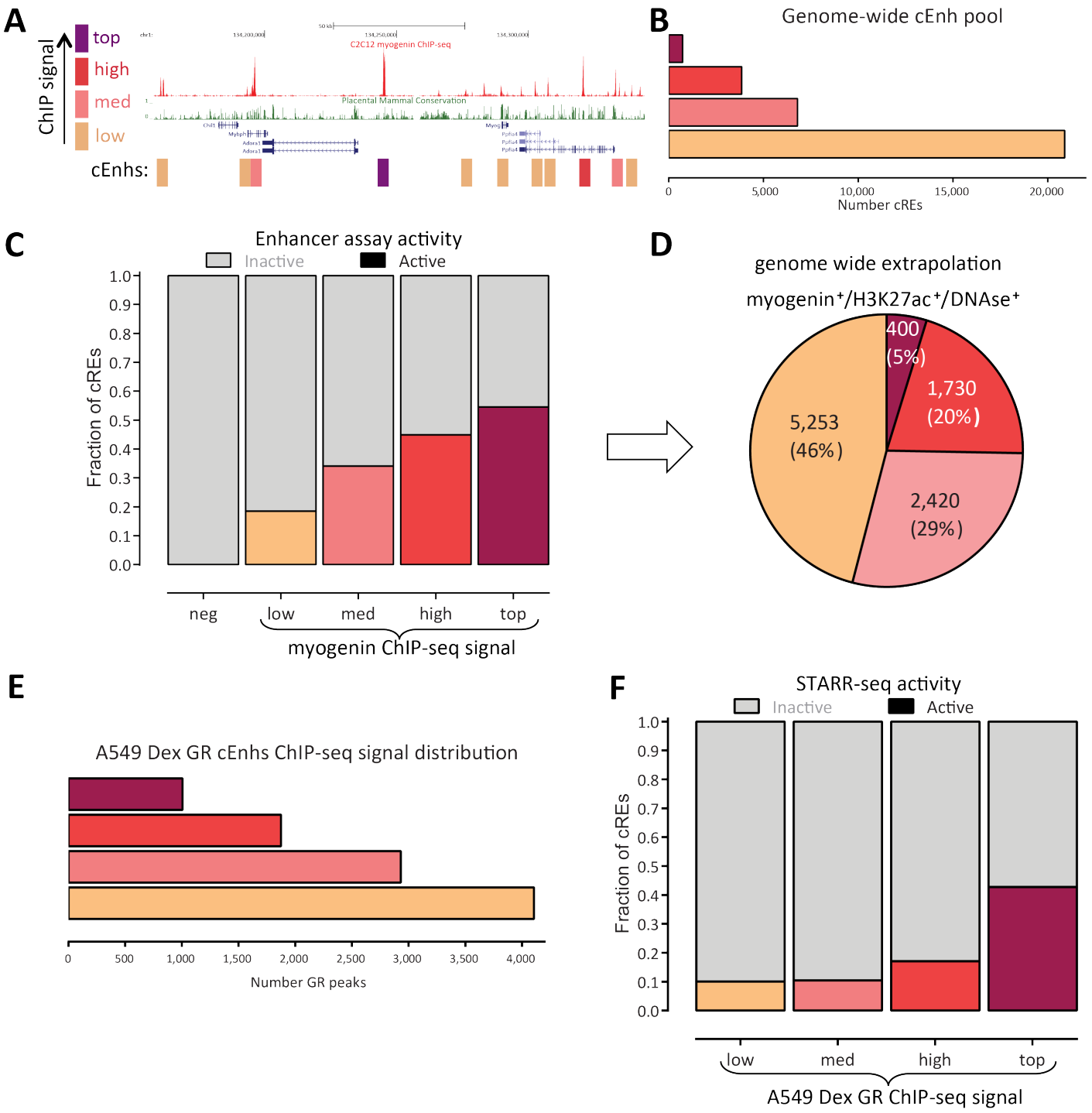


Figure 4: Enrichment of active cEnhs in different classes of cEnhs defined by the strength of their biochemical signatures. (A) cEnhs (rectangle boxes) belonging to different signal classes (based on ChIP-seq data for myogenin in C2C12 myocytes; "top": RPM ≥ 10; "high": RPM$\in [5, 10]$; "medium": RPM$\in [2.5, 5]$; "low" RPM≤ 2.5) in the neighborhood of the mouse *Myog* gene; (B) Genome-wide distribution of cEnhs in different signal classes based on ChIP-seq data for myogenin in C2C12 myocytes; (C) Fraction of active enhancers in different cEnh signal classes (based on ChIP-seq data for myogenin in C2C12 myocytes; "top": $n = 66$; "high": $n = 49$; "medium": $n = 45$; "low" $n = 27$) as well as in negative controls (with not myogenin occupancy; $n = 34$). Only cEnhs positive for myogenin, DNAse and H3K27ac were included; (D) Extrapolated number of active enhancers in C2C12 belonging to each signal strength class

(legend continued on next page)

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[?] . 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.

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.

Finally, we observe that biochemical marks can be decoupled from each other temporally, which can impact cEnh predictions based on their co-occurrence. A significant lag for the histone mark remodeling after sites are no longer accessed by TFs was previously noted in T-cell development **??**. For example, during muscle differentiation thousands of TF occupancy sites also exhibit DNAse hypersensitivity and p300 localization but are not yet robustly acetylated at H3K27, and conversely, in differentiated cells H3K27ac can remain for some time associated with sites previously TF-occupied DNAse hypersensitive sites even though they are no longer open.

discuss supplementary figure 19 here. note that GR likes to bind regions that are H3k27Ac orphaned as a possible rescue for these sites called "orphaned" by Rothenberg Cell 2012

Promoter-enhancer specificity[?]

integration in genome[? ? ?]

Methods

Except where otherwise stated, all analyses were performed using custom-written python scripts. The GENCODE

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

XXXX DETAILS XXXX

K562 cells

XXXX DETAILS XXXX

HepG2 cells

XXXX DETAILS XXXX

A549 cells

XXXX DETAILS XXXX

Functional assays

Cloning and DNA purification

XXXX DETAILS XXXX

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

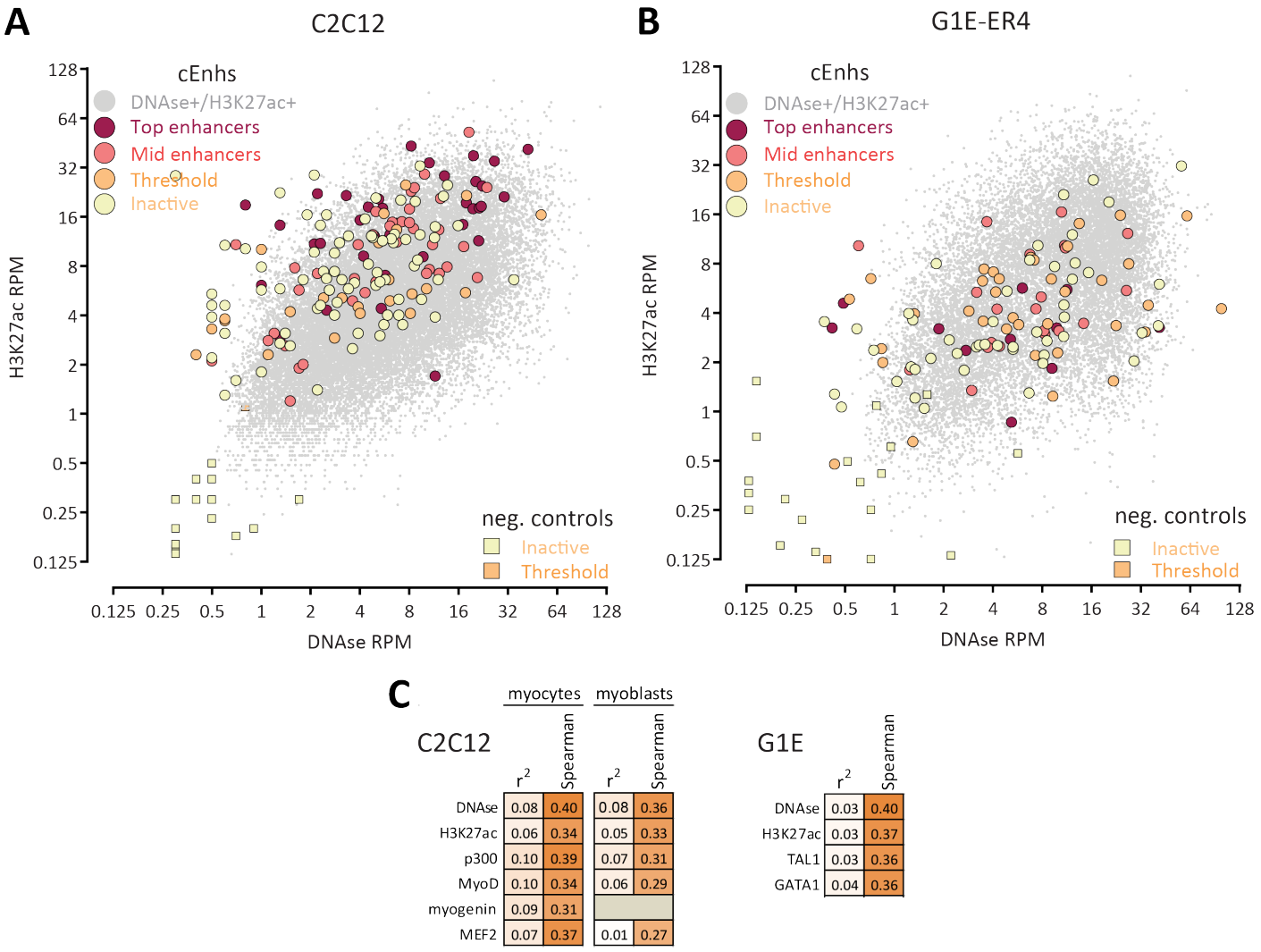


Figure 5: Absence of general strong correlation between biochemical signal strength and enhancer activity of cEnhs. (A) Distribution of tested cEnhs relative to the genome-wide DNAse and H3K27Ac signal distribution in C2C12 myocytes. Shown are DNAse and H3K27ac RPM values for all DNAse$^+$/H3K27ac$^+$ regions as well as for cEns tested for activity in C2C12 myocytes (outlined circles) and for occupancy negative control (outlined squares), with tested cEnhs separated into four classes based on their measured enhancer activity, from dark red (most active) to yellow (inactive). (B) Distribution of tested cEnhs relative to the genome-wide DNAse and H3K27Ac signal distribution in G1E-ER4 cells. Shown are DNAse and H3K27ac RPM values for all DNAse$^+$/H3K27ac$^+$ regions as well as for cEns tested for activity (outlined circles) and for occupancy negative control (outlined squares), with tested cEnhs separated into four classes based on their measured enhancer activity, from dark red (most active) to yellow (inactive). (C) Correlation between biochemical signals and measured enhancer activity in C2C12 and G1E cells. See also Supplementary Figures 8, 12, and 15 for more details.

C2C12 cells

Candidate REs as well as a set of 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 (Quiagen) and standardized to 30 ng/μL using fluorometry concentration measurements (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) 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 50ng 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 hour 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 trasnfection process was automated and carried out on a Tecan Freedom EVO 200 robot.

G1E/K562 cells

XXX DETAILS XXX

K562, HepG2 and A549 cells

XXX DETAILS XXX

Functional Assay Data processing

Luminometer data was ratioed relative to the basal promoter vector (relative assay activity). Active cREs were discriminated from inactive using a Z-score analysis that compared the population of test element technical replicate values to the set of negative control vectors. 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 $\times$ 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[?] . The following antibodies were used: α-myogenin (xxx), α-MyoD (xxx), α-MEF2 (xxx), α-p300 (xxx), α-E2A (xxx), α-H2B (xxx), and α-H3K27ac (xxx).

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

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

RNA-seq experiments

XXX C2C12 RiboMinus XXX

STARR-seq experiments

The STARR-seq experiments previously published by Vockley et al.[?] 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[?] .

Conservation analysis

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

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[?] (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[?] (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`). For H3K27ac control datasets were subjected to the same treatment (no background/control is available for DNAse data) The top 100,000 peaks from each replicate or pseudoreplicate (ranked by q-value) were then used as input into IDR[?] . 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[?] datasets (transcription factors), peak calling was carried out following the same procedure but using SPP[?] (version 1.10.1), using the top 300K 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

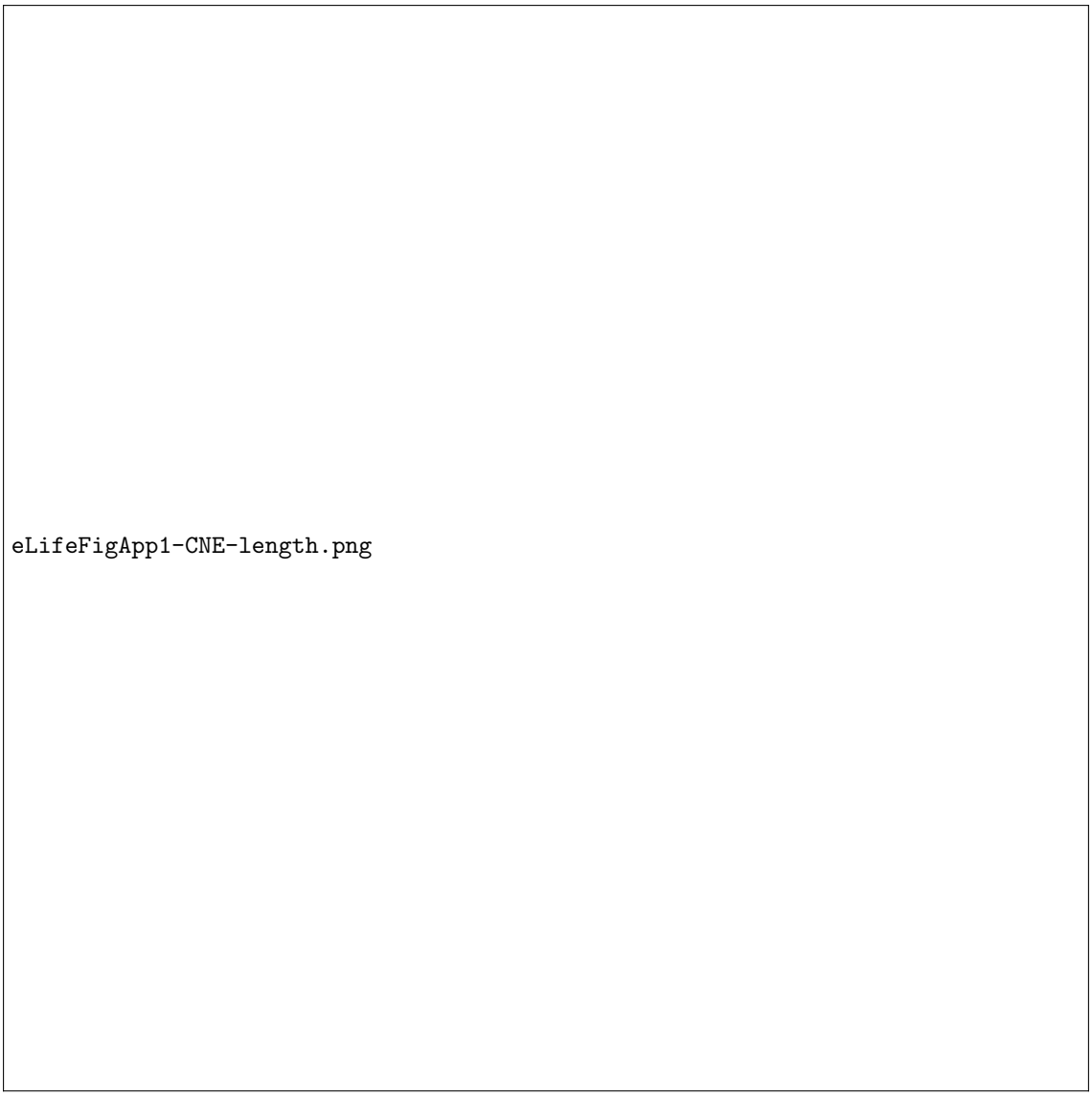
Acknowledgments

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.

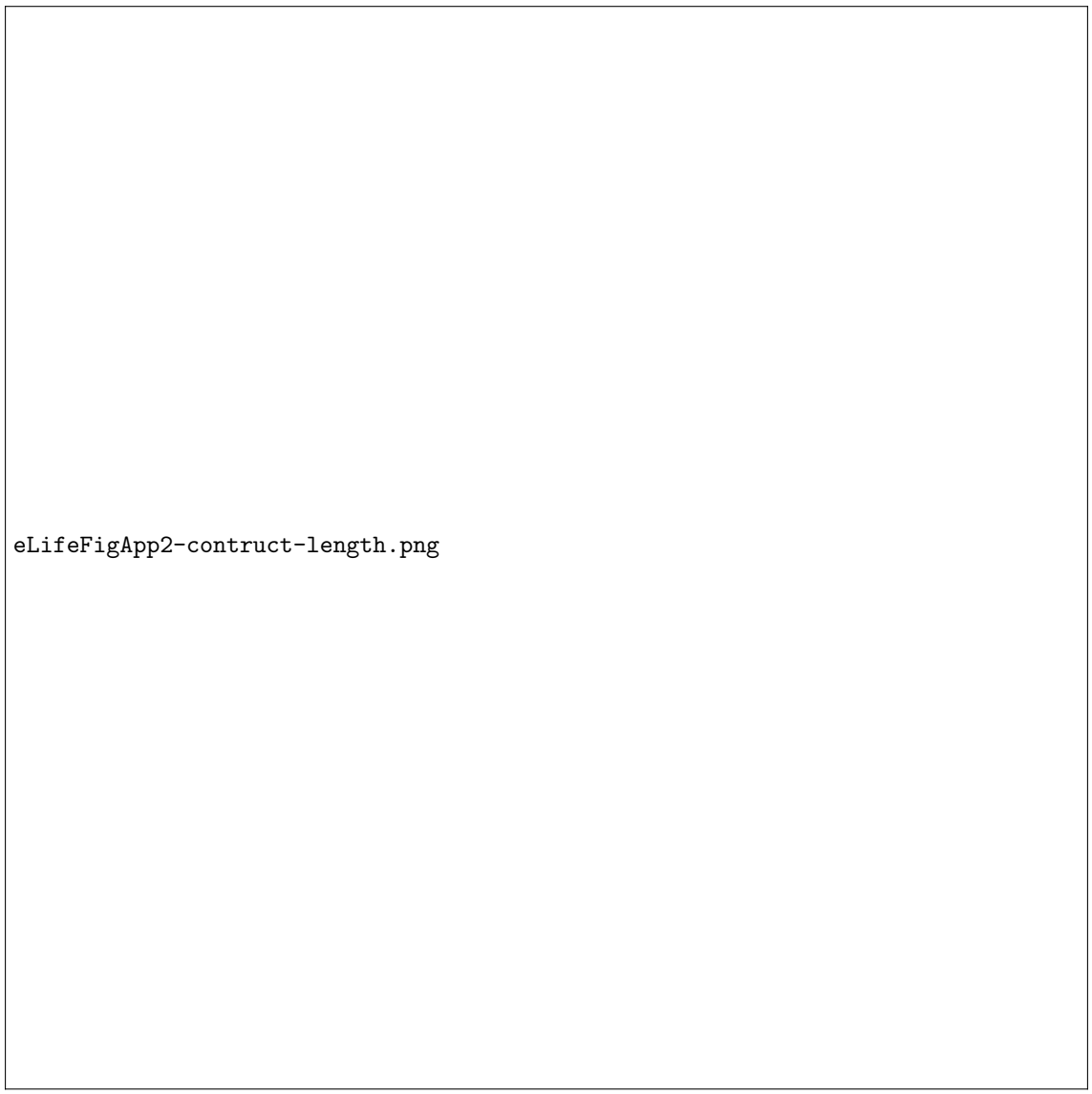
This material is based upon work supported by the National Science Foundation under Grant No. CNS-0521433.

References

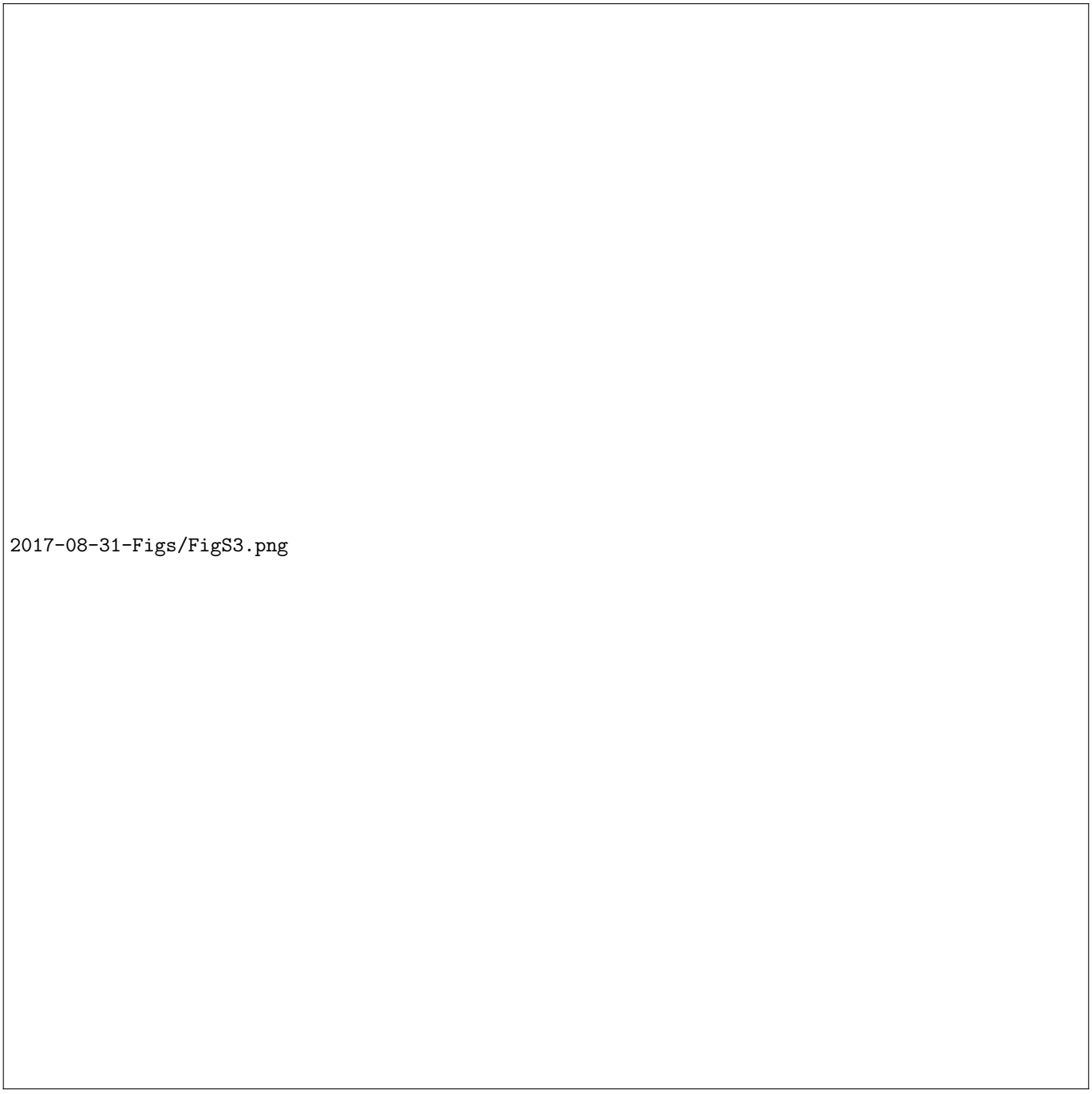
Supplementary Materials



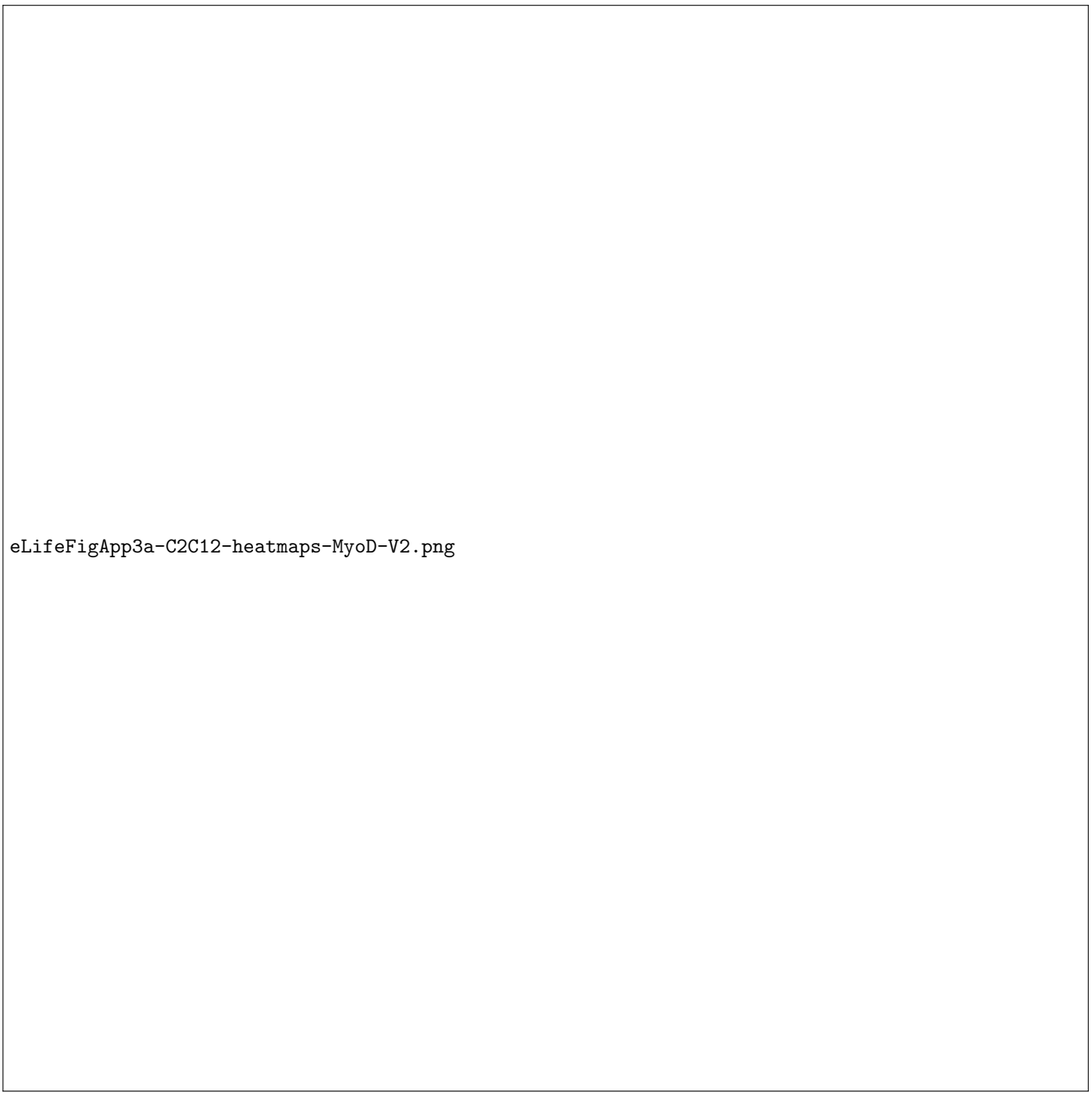
Supplementary Figure 1: The length of thousands of conserved noncoding elements in mammalian genomes greatly exceeds the size range of MPRA constructs. (A) The length distribution of conserved noncoding regions in the human genome. The `phastCons100way` conservation track for the `hg20` version of the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



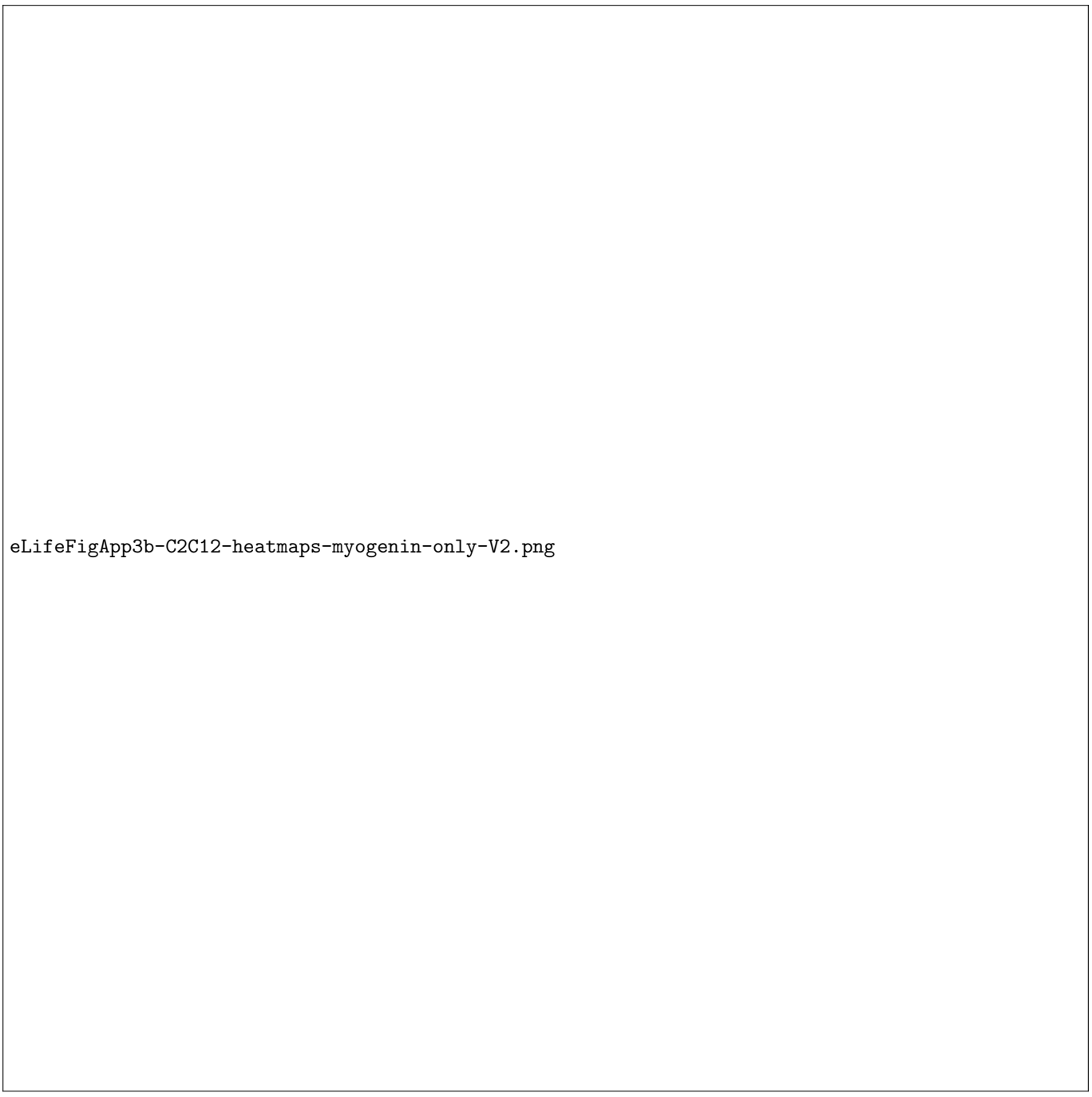
Supplementary Figure 2: Length distribution of functional assays constructs used to test cREs in this study. (A) Distribution of functional assay construct lengths tested in this study in C2C12 cells. (B) Distribution of functional assay construct lengths tested in this study in G1E cells. (C) Distribution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

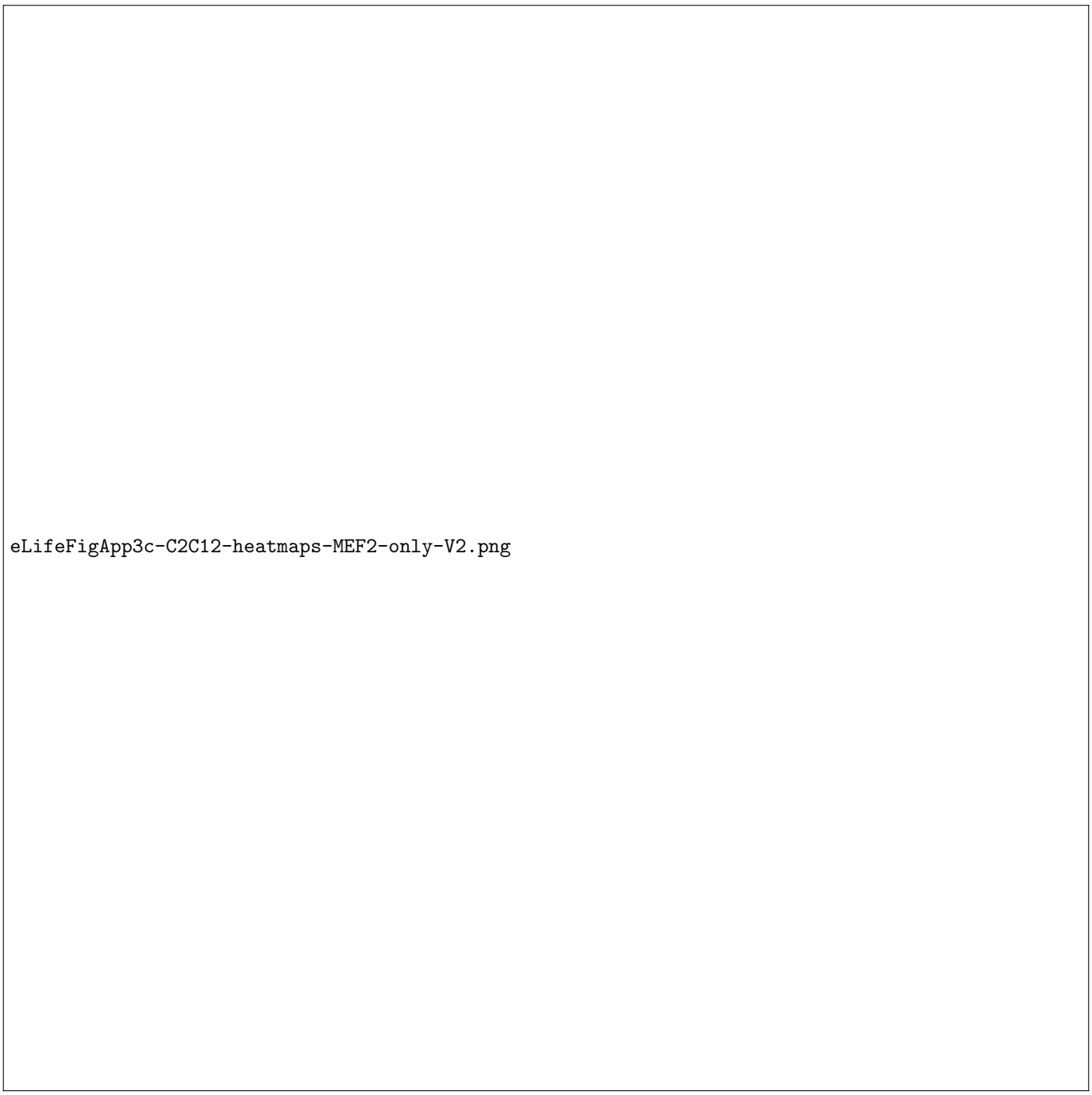


Supplementary Figure 3: Distribution of biochemical signal in tested cEnhs and genome-wide. Shown is the distribution of ChIP-seq or DNAse-seq RPM values for the set of cEnhs tested and for the genome-wide set of cEnh with similar biochemical signatures shown in Figure 3.

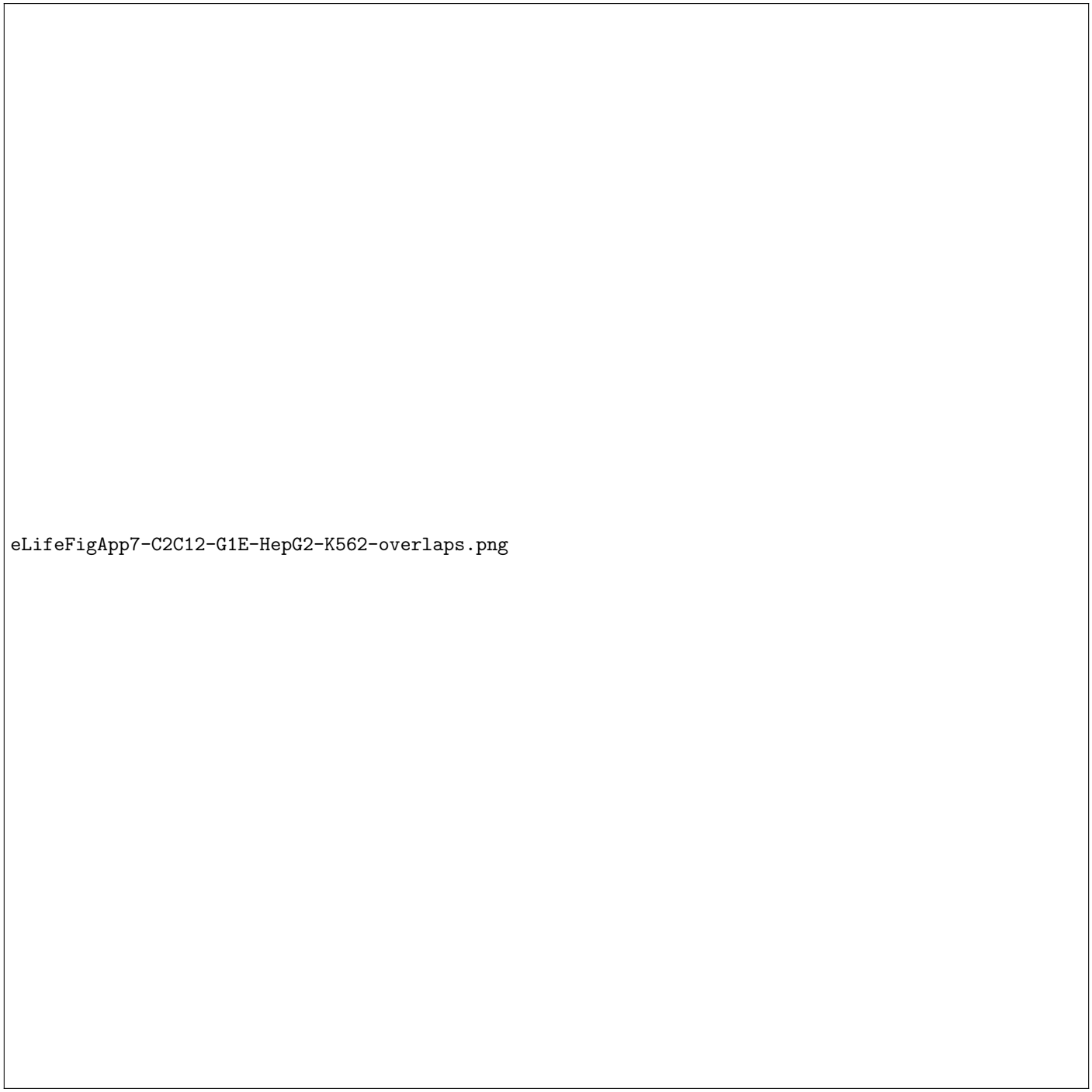


eLifeFigApp3a-C2C12-heatmaps-MyoD-V2.png

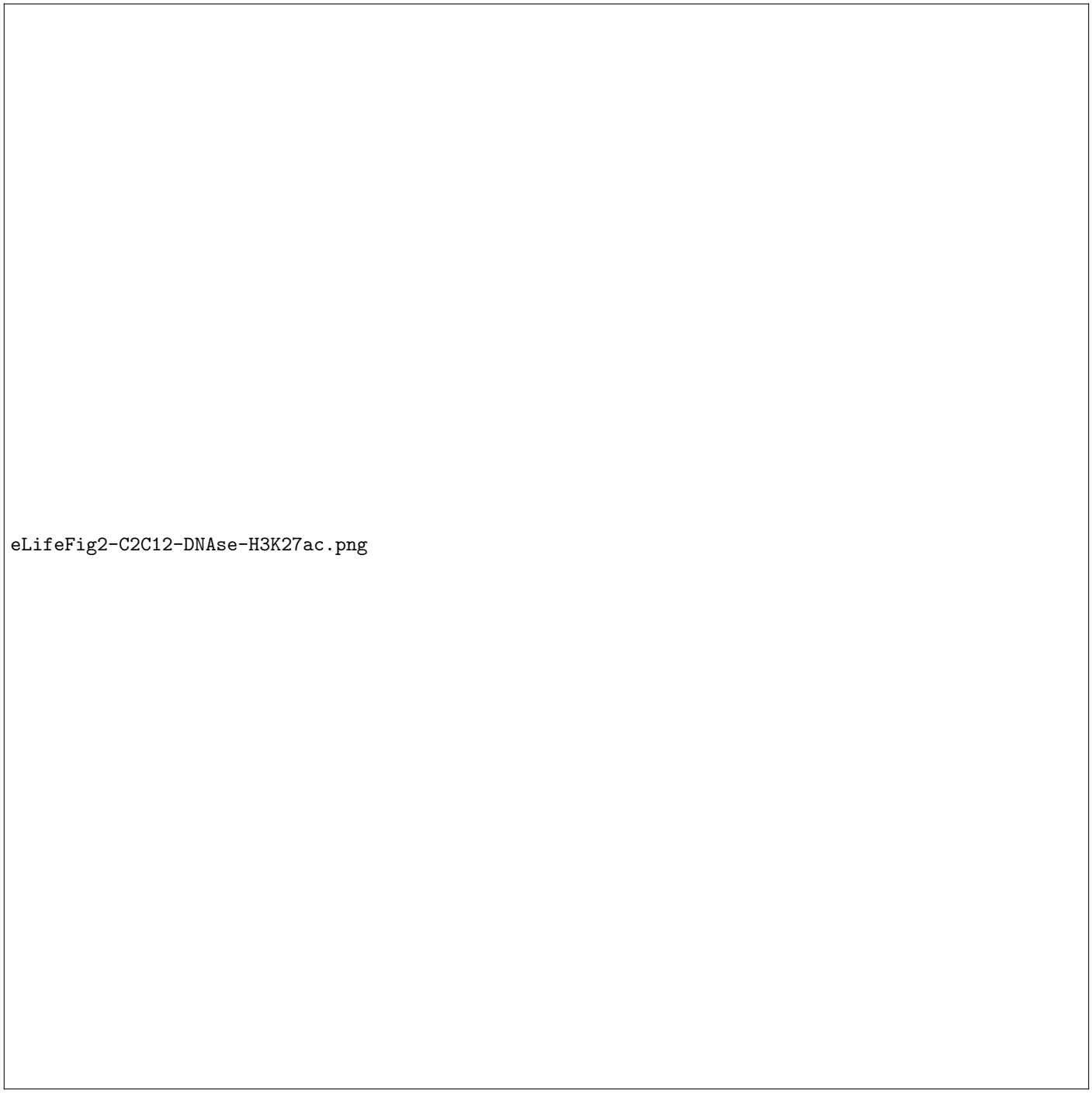

eLifeFigApp3b-C2C12-heatmaps-myogenin-only-V2.png



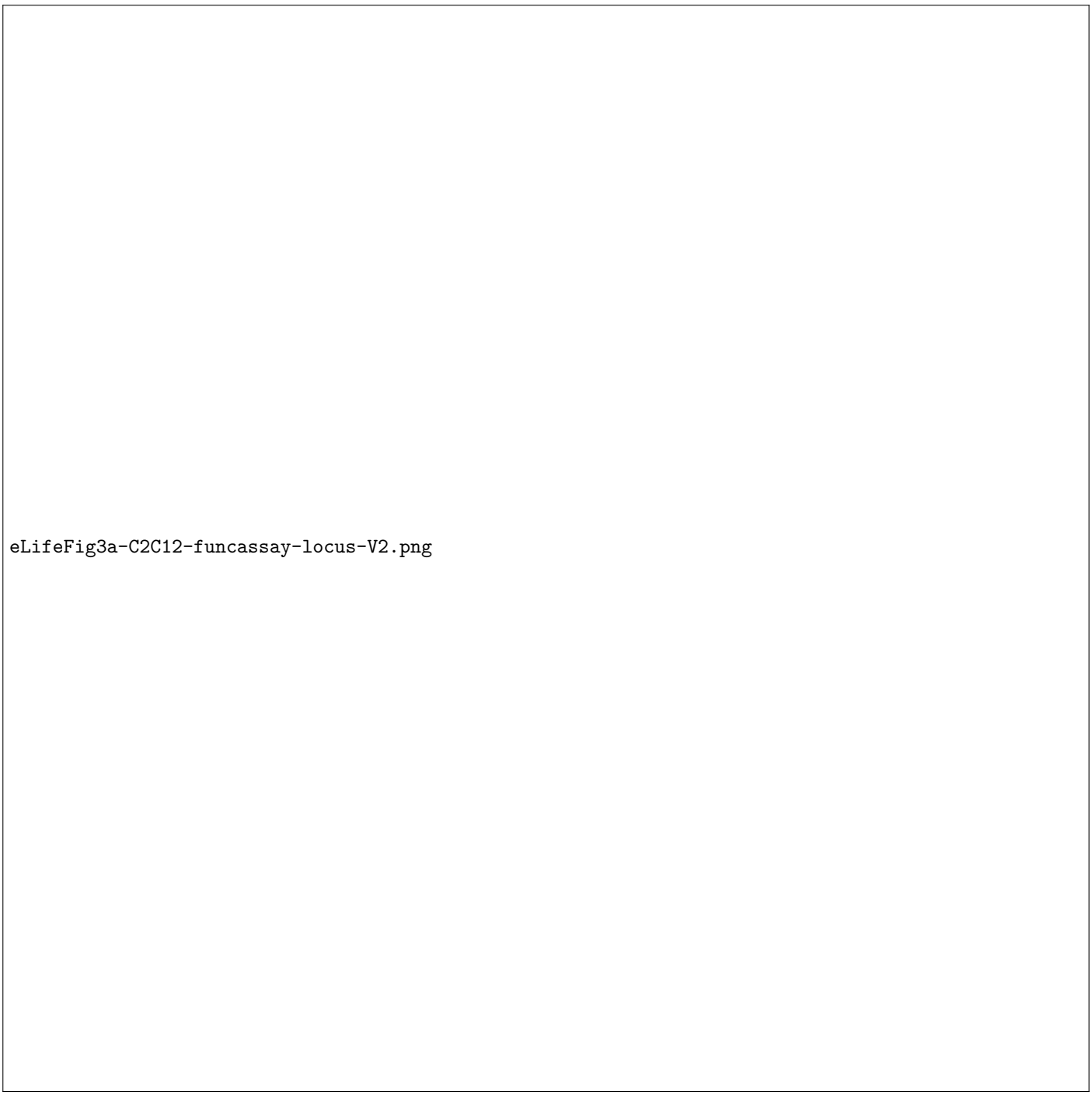
Supplementary Figure 4: Regulatory landscape of muscle differentiation. DNAse-seq and ChIP-seq experiments against H3K27ac, p300, the MRFs MyoD and myogenin, and cofactors (MEF2, E2A/TCF3, HEB/TCF12, and Pbx1) in undifferentiated (myoblast, or "MB") and differentiated (myocyte, or "MC") C2C12 cells were 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.

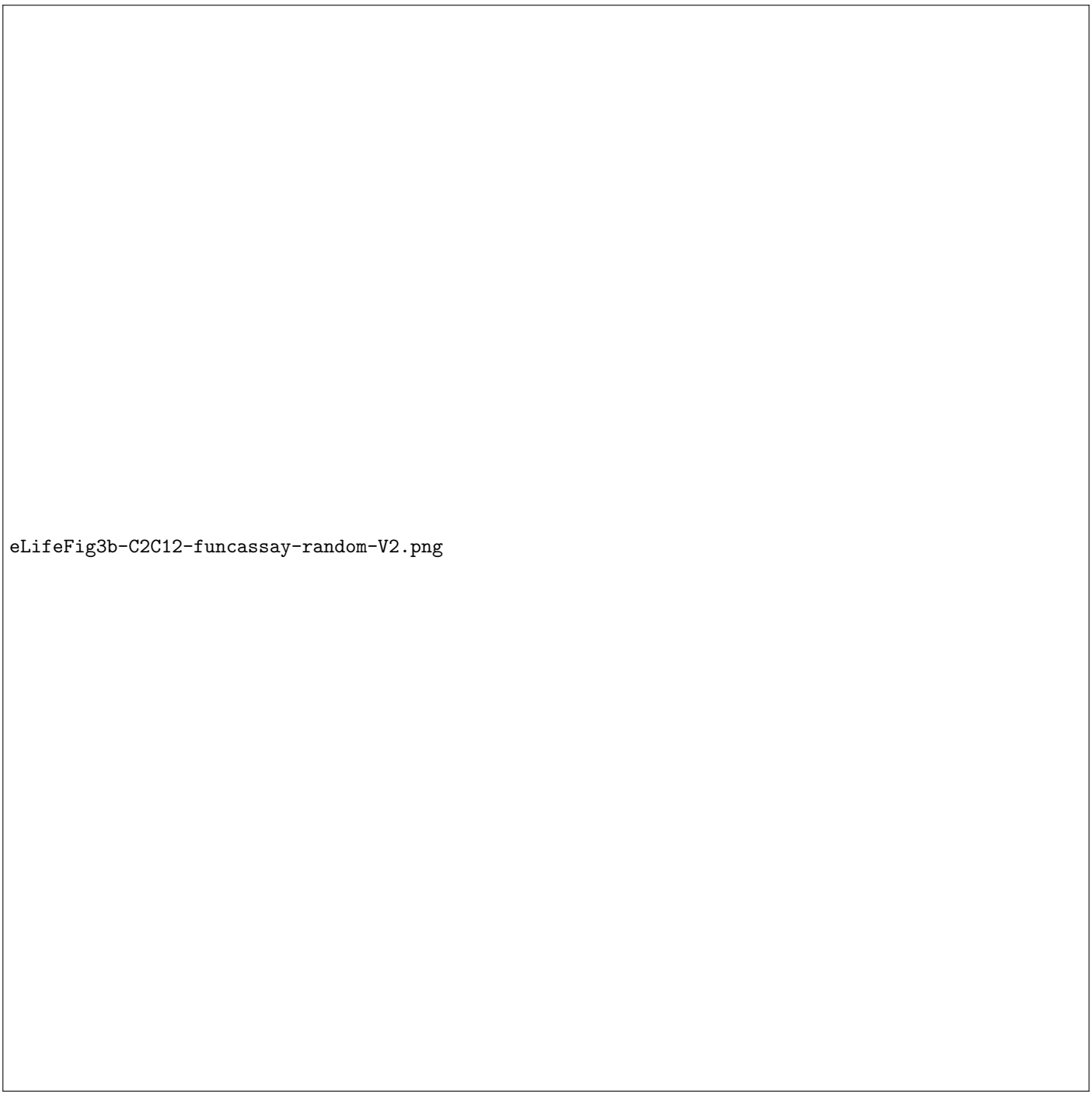


Supplementary Figure 5: . finish caption

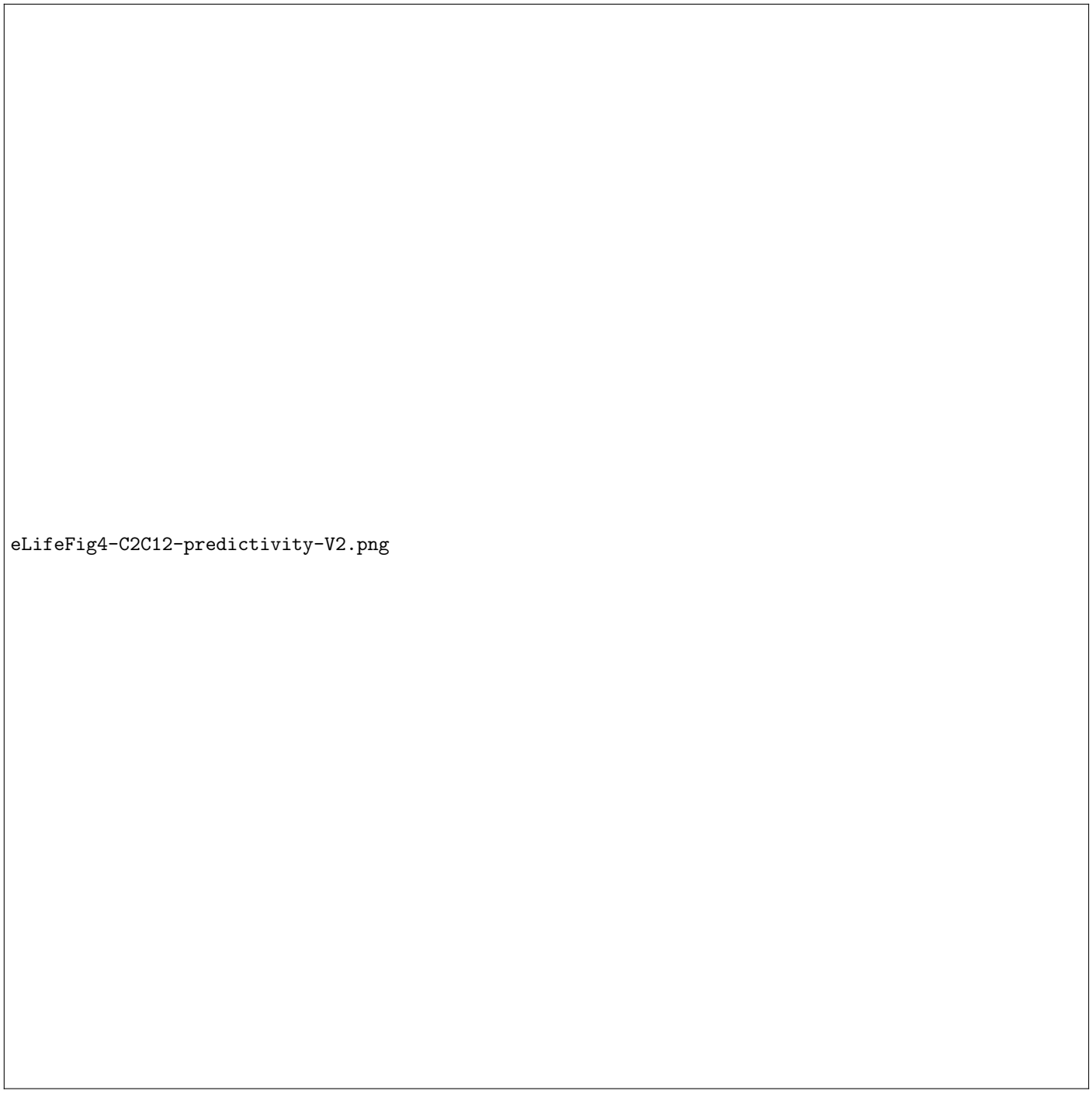


Supplementary Figure 6: Relationship between DNAse hypersensitivity and H3K27 acetylation during muscle differentiation. (A) Overlap between DNAse hypersensitive and H3K27ac-positive promoter-proximal regions in C2C12 myoblasts; (B) Overlap between DNAse hypersensitive and H3K27ac-positive promoter-proximal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

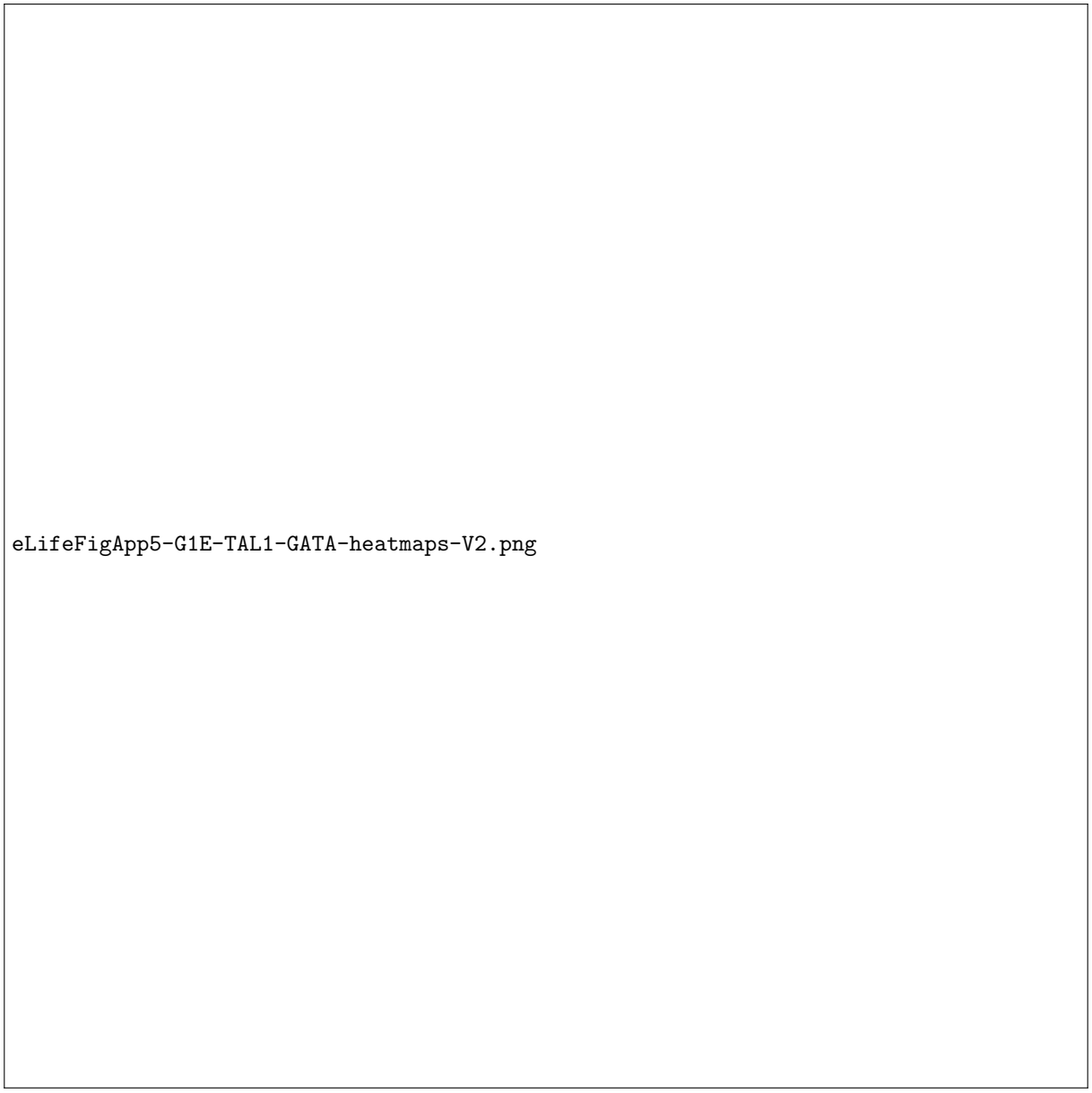

eLifeFig3a-C2C12-funcassay-locus-V2.png


eLifeFig3b-C2C12-funcassay-random-V2.png

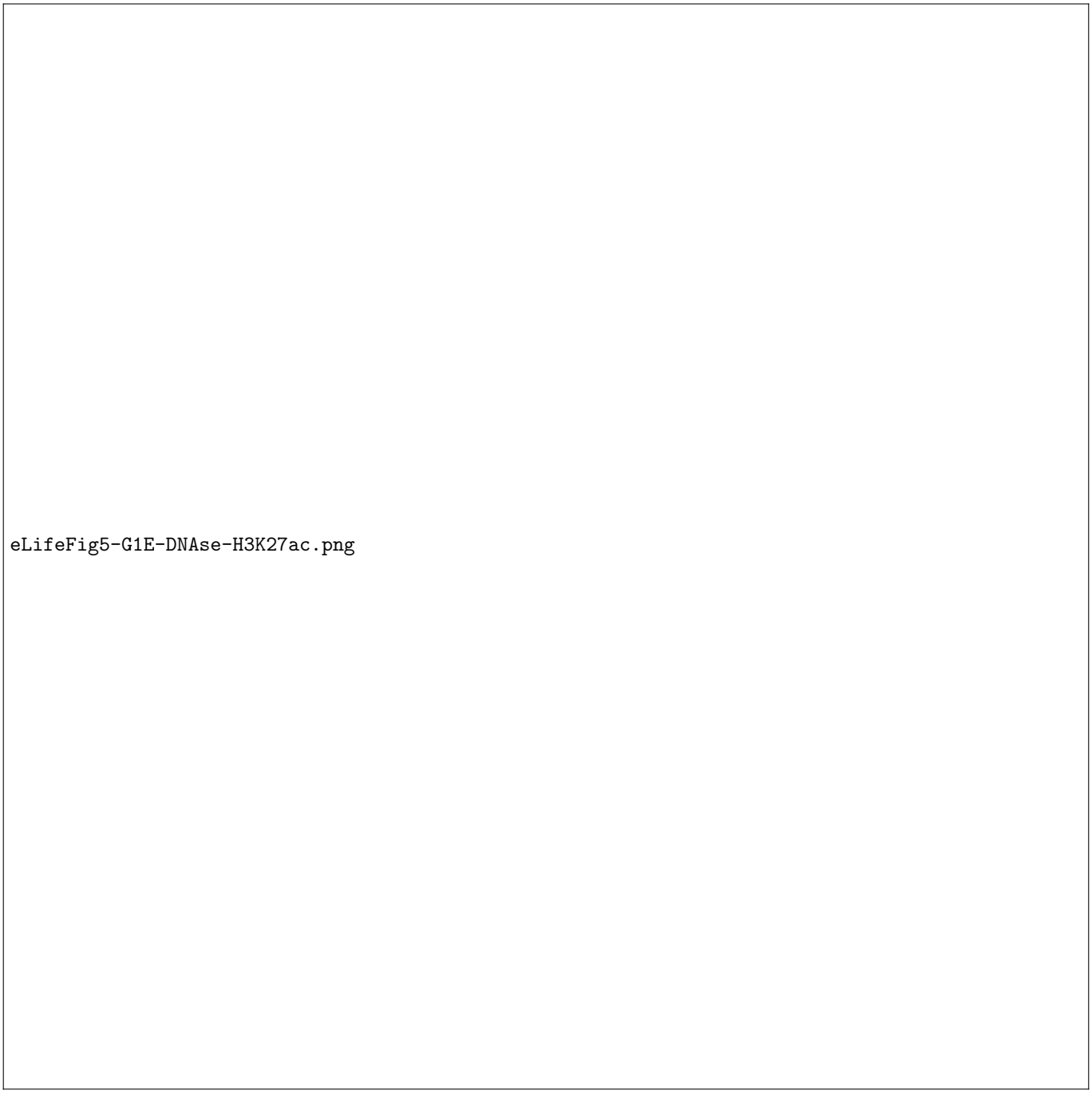
Supplementary Figure 7: Functional assay testing of cRE regulatory activity in C2C12 cells. Fold activity in myocytes (top) and myoblasts (bottom) across biological replicates ($n = 4$) and technical replicates ($n = 4$ for each biological replicate) is shown. Candidate REs were sorted first by their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



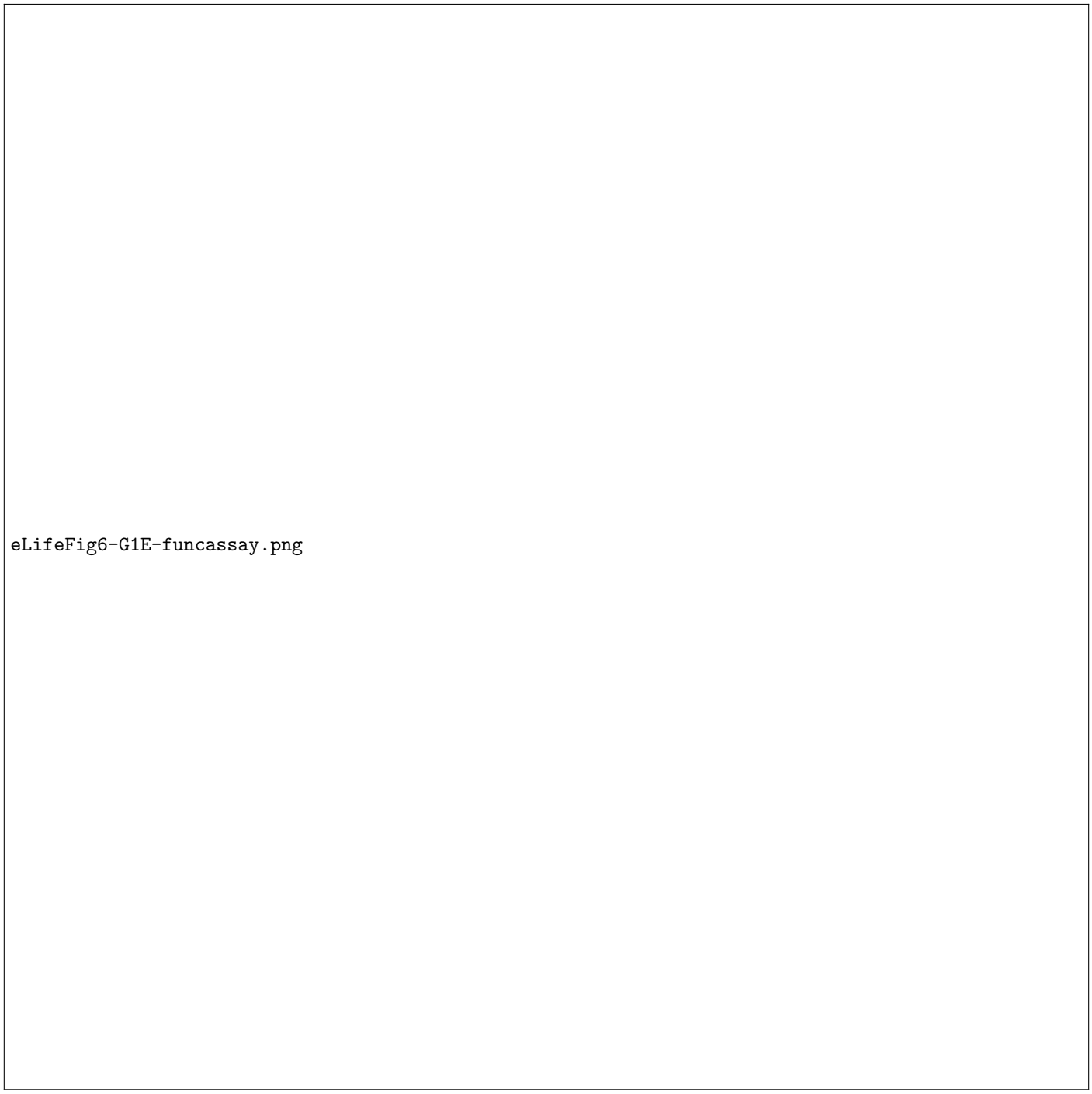
Supplementary Figure 8 *(preceding page)*: Correlation between regulatory activity and biochemical marks in C2C12 cells. (A and B) Correlation between fold activity and DNAse hypersensitivity, H3K27ac, p300, myogenin, MyoD and MEF2 occupancy in myoblasts and myocytes; (C) ROC curves showing biochemical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



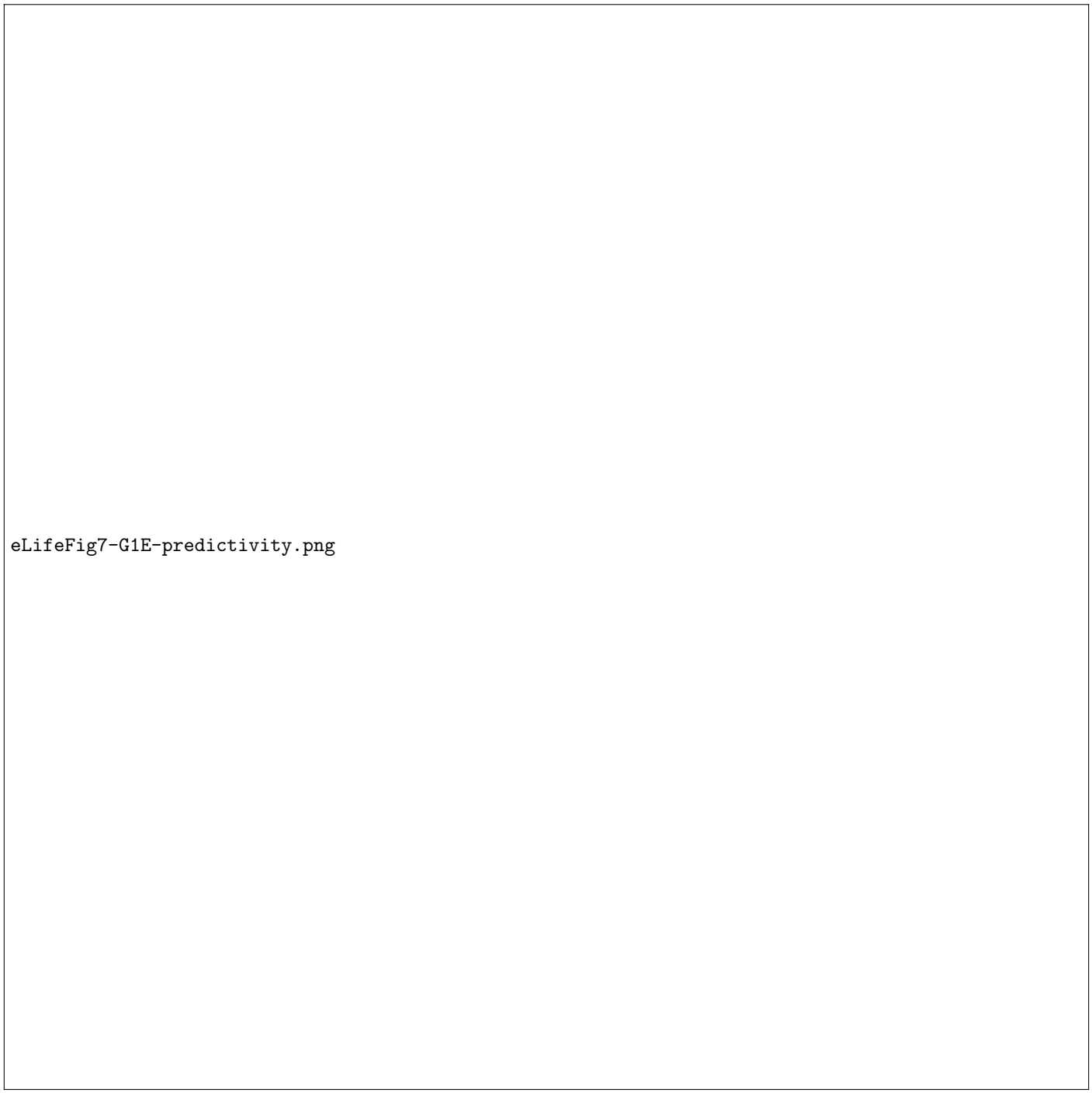
Supplementary Figure 9 *(preceding page)*: Regulatory landscape of eryhtroid differentiation. DNAse-seq and ChIP-seq experiments against H3K27ac, GATA1, TAL1 and GATA2 G1E and G1E-ER4 were analyzed. Sites were split into subgroups depending on GATA1 and TAL1 occupancy (IDR=0.05), then sorted by ChIP-seq 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.



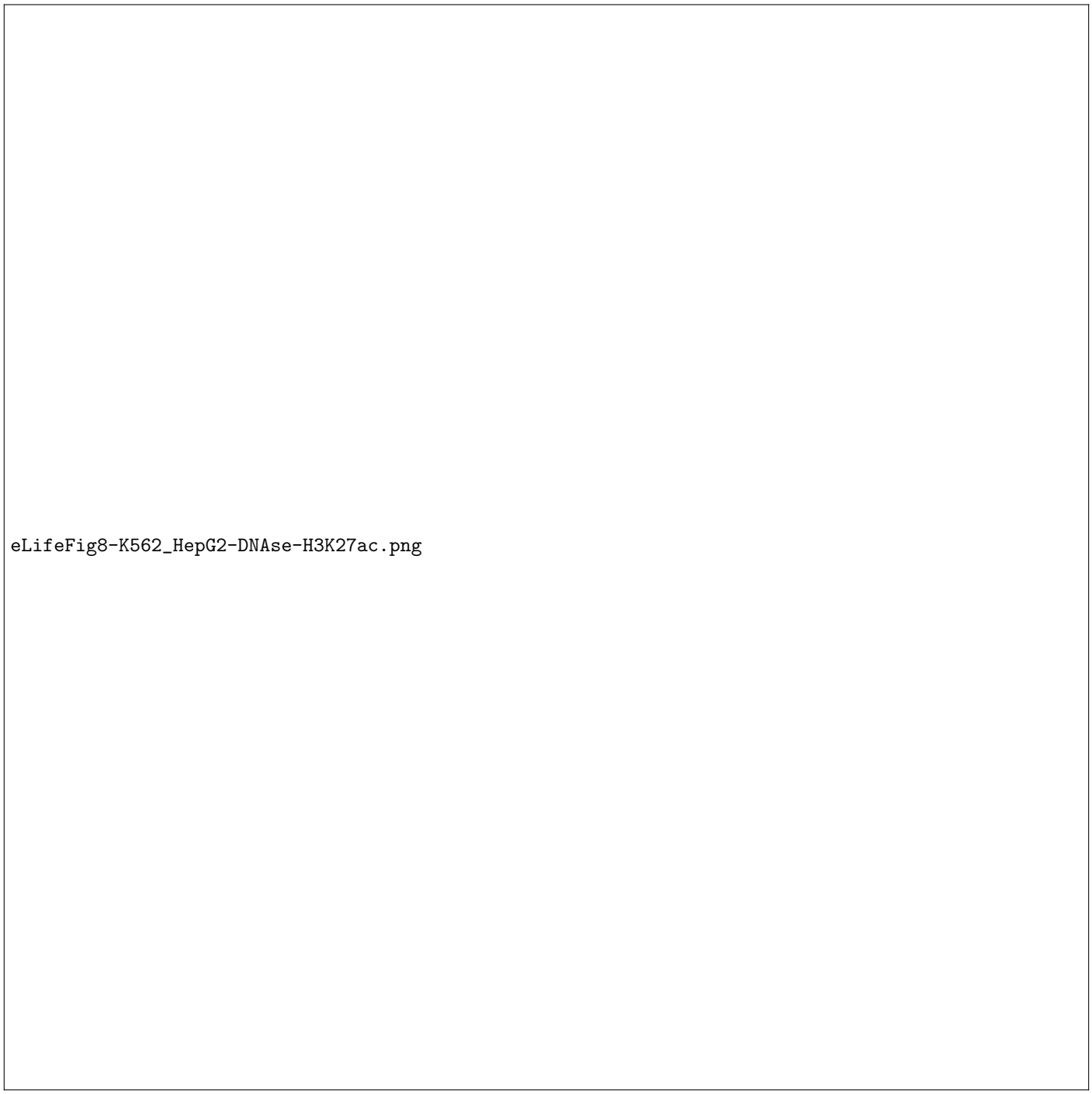
Supplementary Figure 10: Relationship between DNAse hypersensitivity and H3K27 acetylation during erythroid differentiation. (A) Overlap between DNAse hypersensitive and H3K27ac-positive promoter-proximal regions in G1E cells; (B) Overlap between DNAse hypersensitive and H3K27ac-positive promoter-proximal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



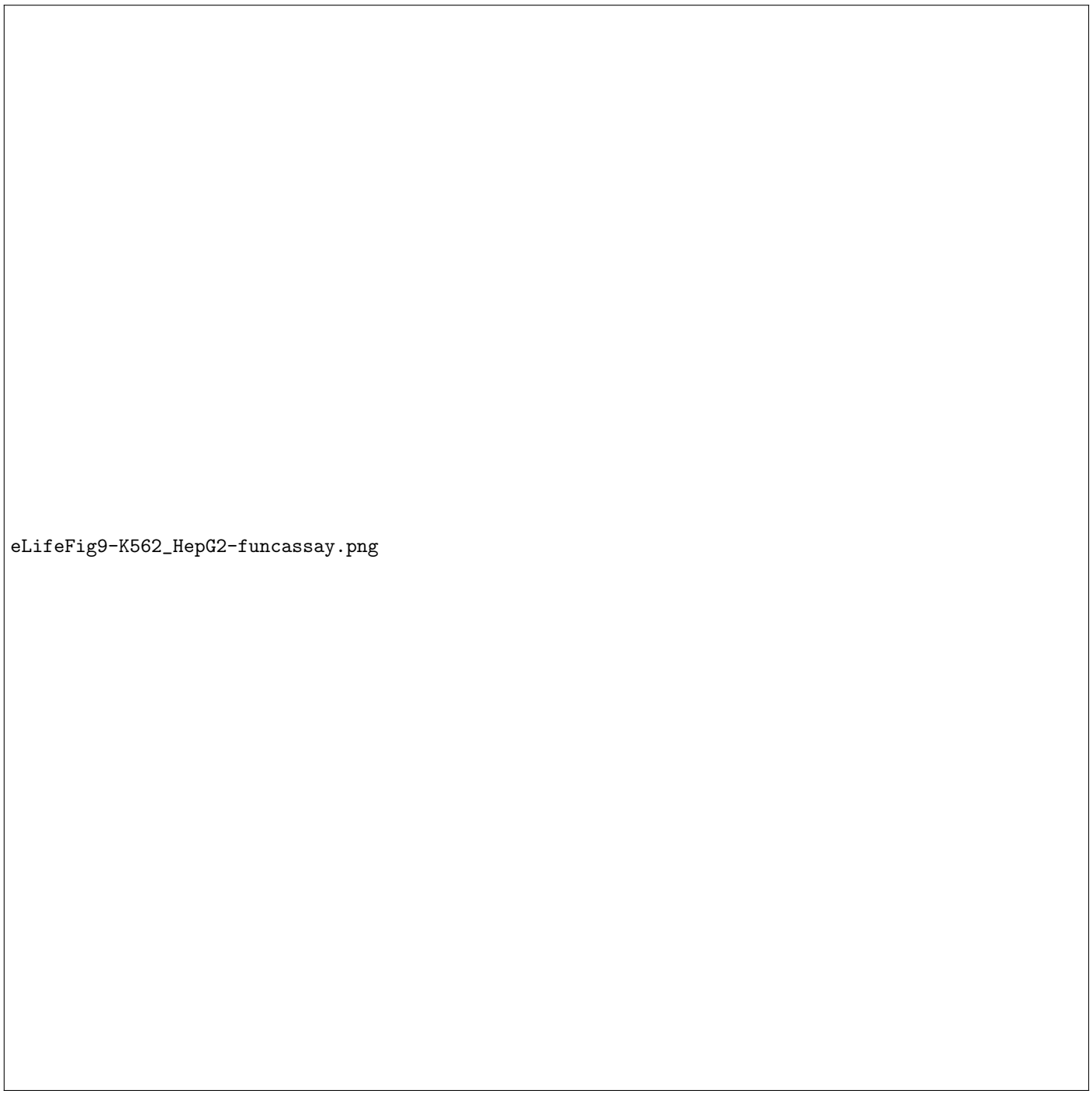
Supplementary Figure 11 *(preceding page)*: Functional assay testing of the regulatory activity of erythroid cREs. Fold activity in K562 cells across biological replicates ($n \in [1, 9]$) and technical replicates ($n = 4$ for each biological replicate) is shown. Candidate REs were sorted first by their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



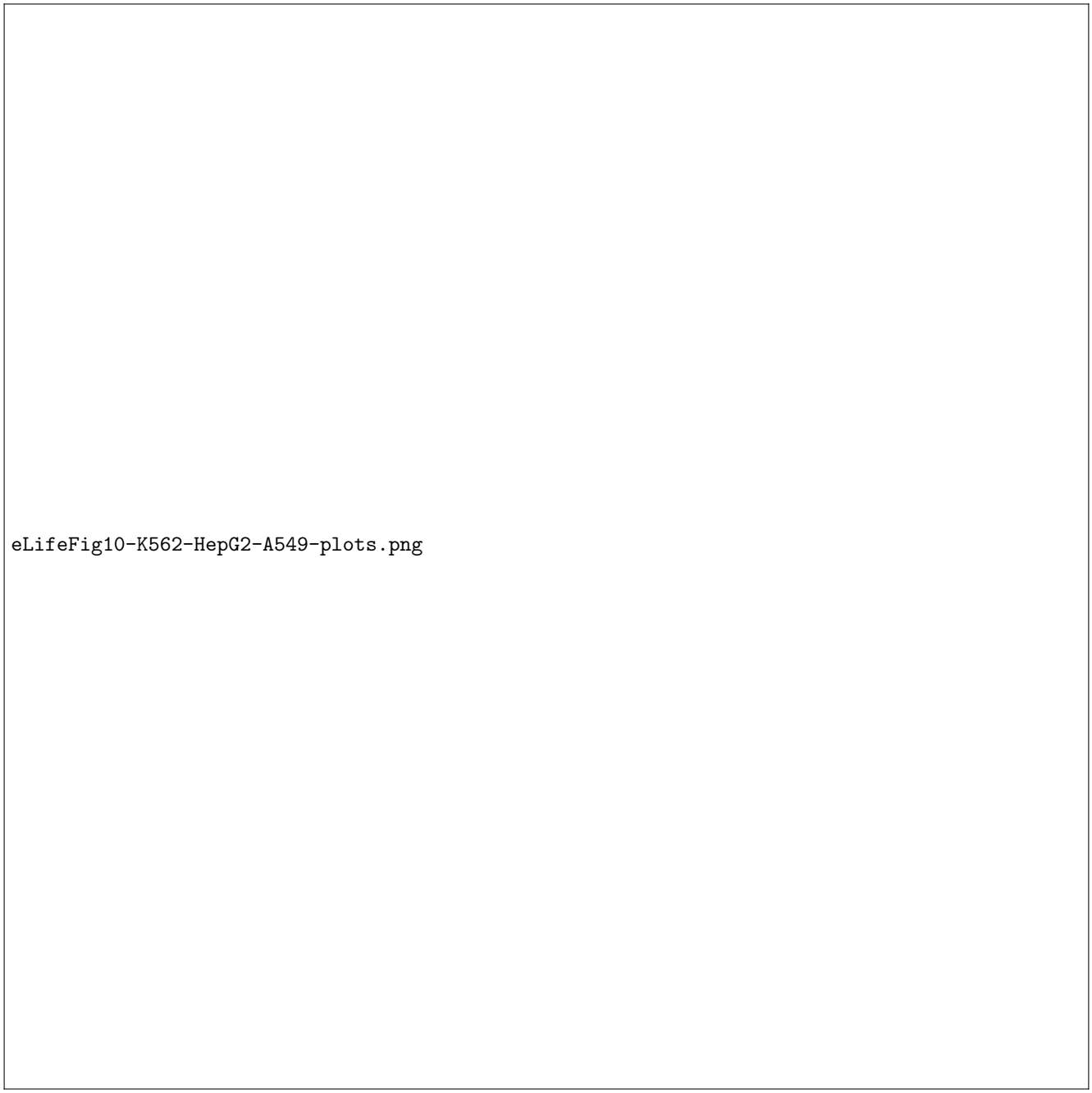
Supplementary Figure 12: Correlation between regulatory activity and biochemical marks in erythroid cells. (A and B) Correlation between fold activity in K562 cells and DNAse hypersensitivity, H3K27ac, TAL1, and GATA1 occupancy in G1E and G1E-ER4 cells; (C) ROC curves showing biochemical mark predictivity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



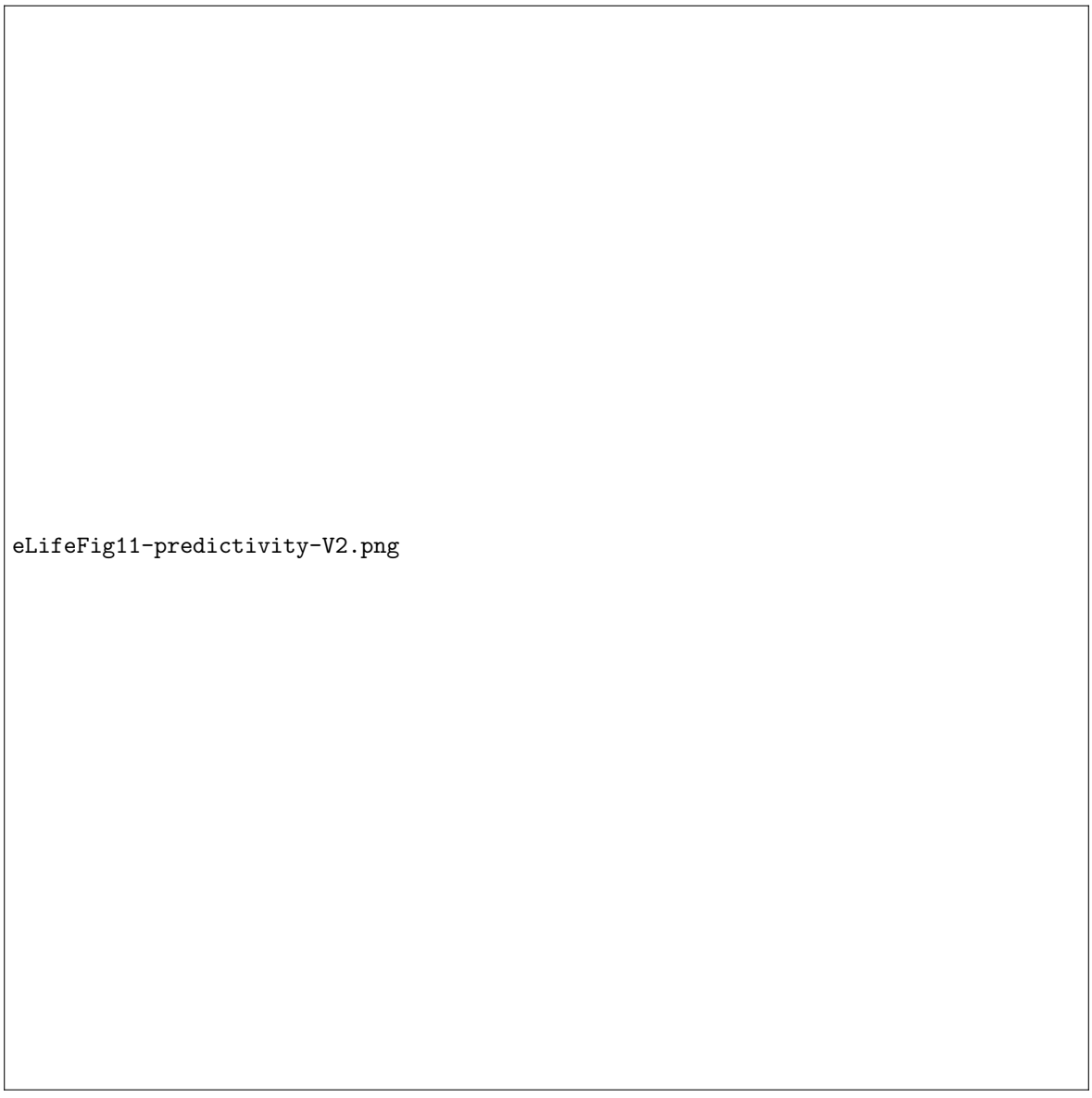
Supplementary Figure 13: Relationship between DNAse hypersensitivity and H3K27 acetylation in immortalized human cell lines. (A) Overlap between DNAse hypersensitive and H3K27ac-positive promoter-proximal regions in K562 cells; (B) Overlap between DNAse hypersensitive and H3K27ac-positive distal regions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



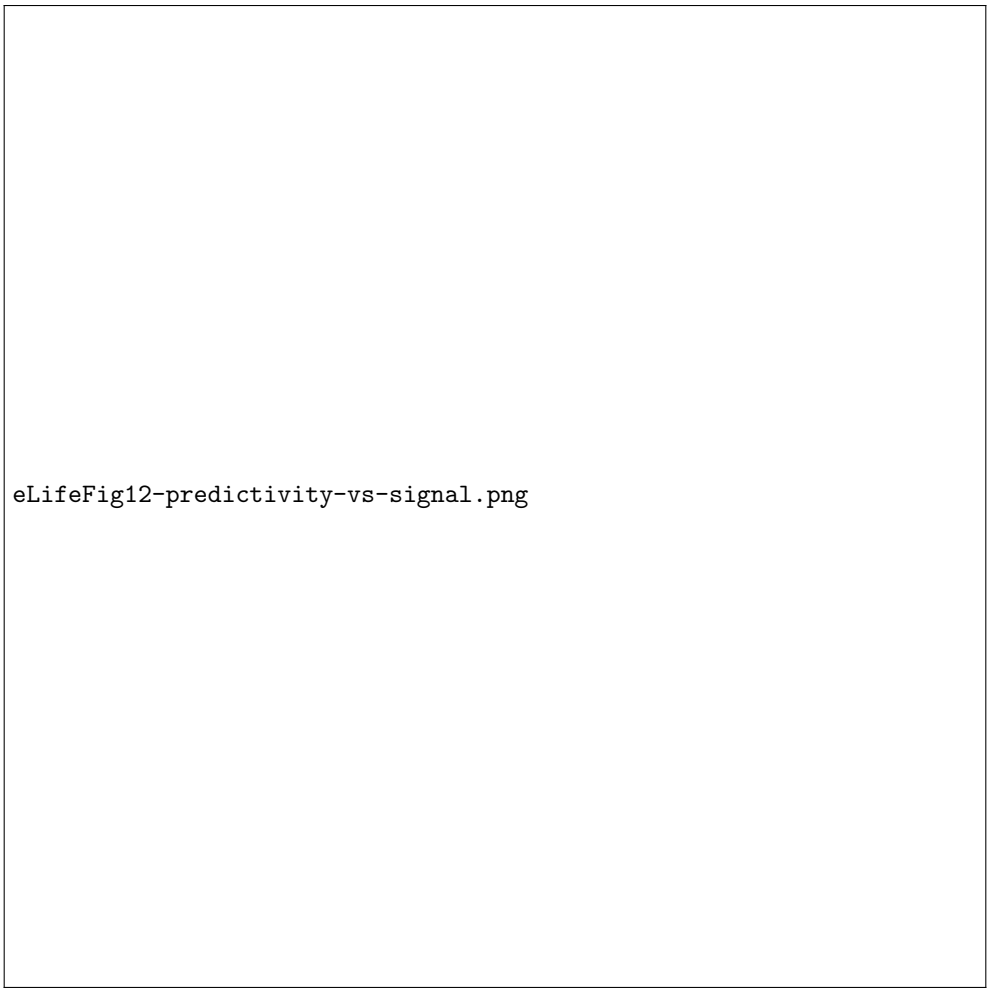
Supplementary Figure 14: Functional assay testing of cRE regulatory activity in human immortalized cell lines. Fold activity across biological replicates (n = ???) and technical replicates (n = ??? for each biological replicate) is shown. Candidate REs were sorted first by their DNAse status and then 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.


eLifeFig10-K562-HepG2-A549-plots.png

Supplementary Figure 15 *(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.



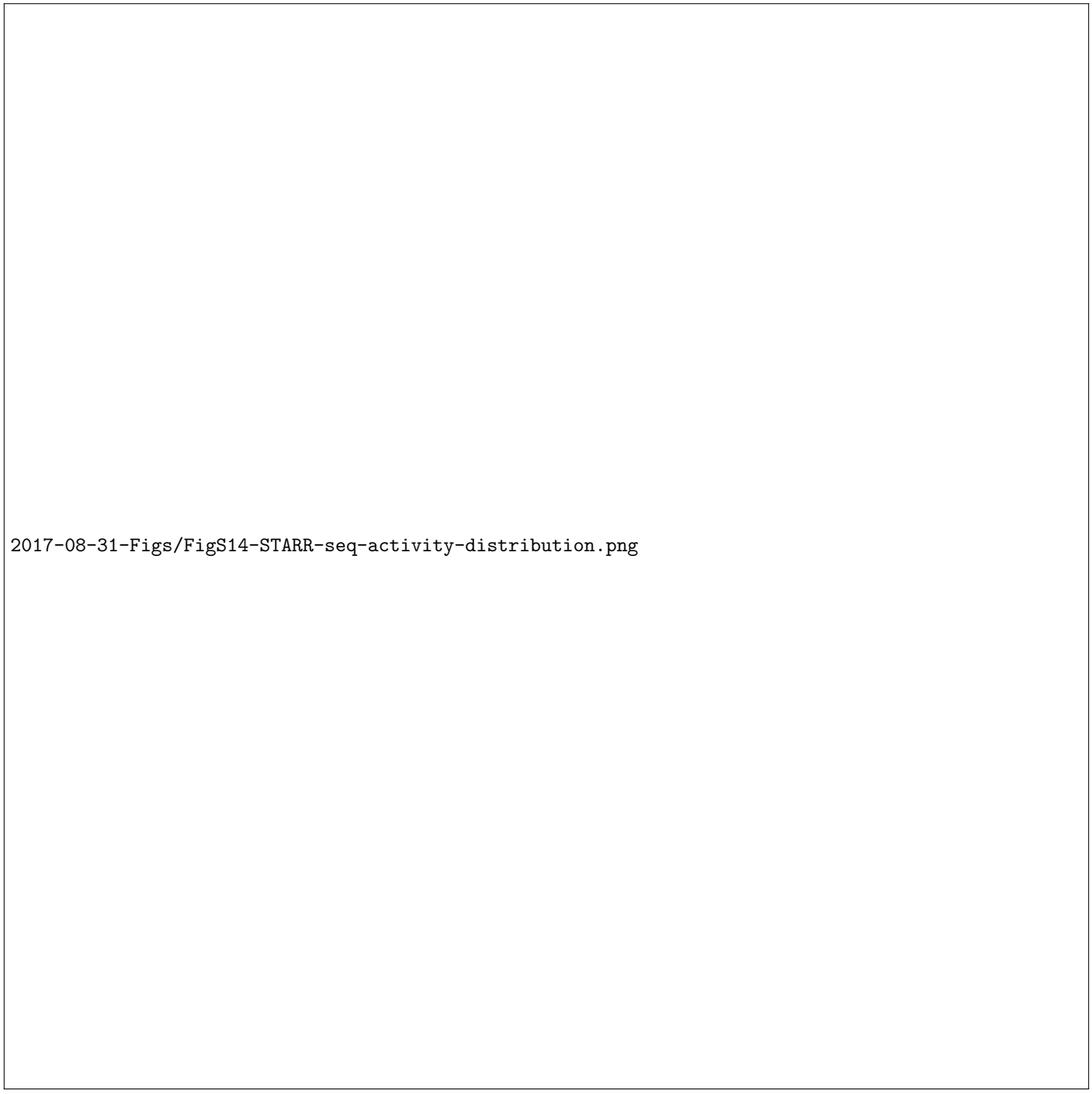
Supplementary Figure 16: CAPTION. (A) CAPTION GOES HERE



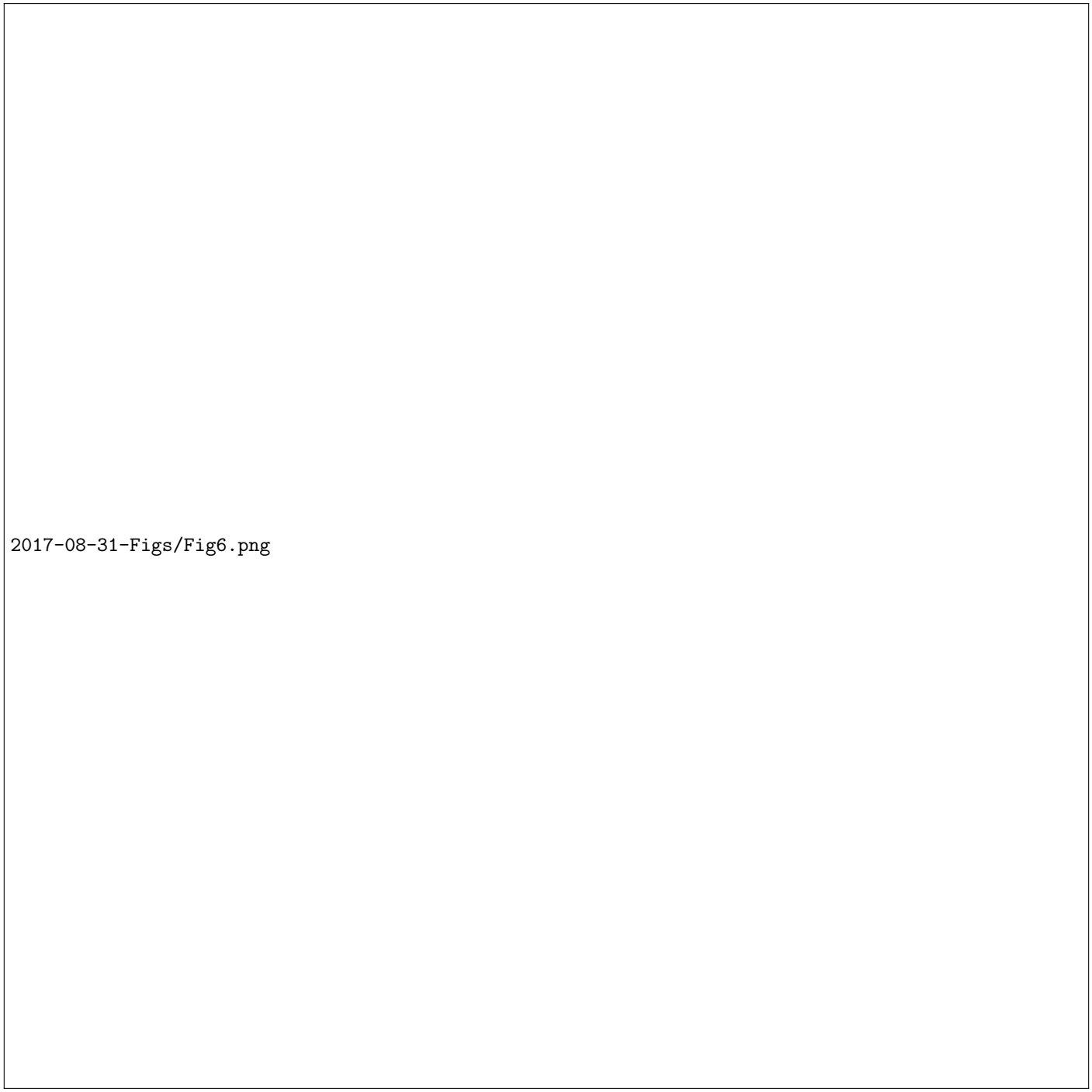
Supplementary Figure 17: **CAPTION** . (A) CAPTION GOES HERE

eLifeFigApp6-A549-heatmaps-V2.png

Supplementary Figure 19: . finish caption



Supplementary Figure 20: 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.



Supplementary Figure 21: . (A) (B) .