Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press



Characterization of human transcription factor function and patterns of gene regulation in HepG2 cells

Belle A Moyers, E. Christopher Partridge, Mark Mackiewicz, et al.

Genome Res. published online October 18, 2023
Access the most recent version at doi:10.1101/gr.278205.123

P<P Published online October 18, 2023 in advance of the print journal.

Accepted Manuscript Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

Open Access Freely available online through the *Genome Research* Open Access option.

Creative Commons License This manuscript is Open Access.This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International license), as described at http://creativecommons.org/licenses/by/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**



To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

Published by Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

Characterization of human transcription factor function and patterns of gene regulation in HepG2 cells

Belle A. Moyers[1], E. Christopher Partridge[1], Mark Mackiewicz[1], Michael J. Betti[2], Roshan Darji[1], Sarah K. Meadows[1], Kimberly M. Newberry[1], Laurel A. Brandsmeier[1], Barbara J. Wold[3], Eric M. Mendenhall[1*], and Richard M. Myers[1*]

[1]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA

[2]Vanderbilt University Medical Center, Nashville, TN 37232, USA

[3]Merkin Institute for Translational Research, California Institute of Technology, Pasadena, California 91125, USA

*Corresponding Authors:

rmyers@hudsonalpha.org

emendenhall@hudsonalpha.org

Abstract

Transcription factors (TFs) are trans-acting proteins that bind *cis*-regulatory elements (CREs) in DNA to control gene expression. Here, we analyzed the genomic localization

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

profiles of 529 sequence-specific TFs and 151 cofactors and chromatin regulators in the human cancer cell line HepG2, for a total of 680 broadly-termed DNA-Associated Proteins (DAPs). We used this deep collection to model each TF's impact on gene expression, and identified a cohort of 26 candidate transcriptional repressors. We examine High Occupancy Target (HOT) sites in the context of three-dimensional genome organization and show biased motif placement in distal-promoter connections involving HOT sites. We also found a substantial number of closed chromatin regions with multiple DAPs bound and explored their properties, finding that a MAFF/MAFK TF pair correlates with transcriptional repression. Altogether, these analyses provide novel insights into the regulatory logic of the human cell line HepG2 genome and demonstrate the usefulness of large genomic analyses for elucidation of individual TF functions.

Introduction

Gene expression is regulated and modulated by the association, either direct or indirect, of various classes of proteins to DNA, including RNA polymerase and transcription-associated proteins, histone modifiers, and a broad suite of transcription factors (TFs) and associated cofactors. Together these DNA-associated proteins (DAPs) are encoded by approximately 10% of all protein-coding genes in the human genome (Vaquerizas et al. 2009; Lambert et al. 2018). DAPs are known to associate with DNA either through recognition of discrete small sequence motifs, by interactions with degenerate sequences having little complexity, or by cofactor recruitment. The most common assay for genome-wide identification of genomic binding or association sites for DAPs is chromatin

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

immunoprecipitation followed by high-throughput sequencing (ChIP-seq), which provides a statistically identified snapshot of regions, referred to as peaks (Johnson et al. 2007; Barski et al. 2007; Robertson et al. 2007; Kharchenko et al. 2008; Zhang et al. 2008; Savic et al. 2015; Meadows et al. 2020). For those TFs with DNA sequence specificity, associations occur with enough frequency to be detectable as a consistent DNA sequence motif through use of genome-wide binding data (Bailey et al. 2015) or *in vitro* molecular binding assays (Chai et al. 2011).

The Encyclopedia of DNA Elements (ENCODE) Consortium has completed and released 3,194 ChIP-seq datasets for 1,139 DAPs using both traditional antibody ChIP-seq and epitope-tagged ChIP-seq methods (The ENCODE Project Consortium 2012; The ENCODE Project Consortium et al. 2020; Partridge et al. 2020). The human liver cancer-derived cell line HepG2 currently has the largest number (n=814) of ENCODE released ChIP-seq datasets, some of which are repetitions of different ChIP-seq experiments with the same target for a total of 680 unique DAP targets. With this wealth of occupancy profiles for a single cell-type, the HepG2 ChIP-seq data allow for the assessment of biological roles of DAPs in a broad genomic context, including analyses of similarity and co-association frequency, association with regulatory region types, and impact on gene expression. These datasets provide the opportunity to explore the functional impact of individual TFs and associated proteins on gene expression and genome organization.

Here, we present an analysis of ChIP-seq data in the HepG2 that greatly expands on our previous work with this cell type (Partridge et al. 2020) including 492 ChIP-seq datasets not analyzed in that prior work, as well as a lentiviral Massively Parallel Reporter Assay (lentiMPRA, or MPRA) to functionally test elements. We provide an overview of this

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

resource and highlight novel findings with transcription factors and trans-regulatory proteins on *cis*-regulatory sequences, including patterns of TF genomic localization in the context of the three-dimensional organization of high-occupancy target (HOT) sites, and the association of TFs with closed chromatin regions that influence gene repression.

Results

It is estimated that there are 1,639 sequence-specific TFs encoded in the human genome (Lambert et al. 2018), only a subset of which is expressed in any given cell type. To gain a deeper understanding of gene regulatory mechanisms, we analyzed TF binding data, much of which we generated, in HepG2 cells, and leveraged the large number of TFs assayed in that cell line as the most comprehensive resource available. The expression level of any individual TF is not necessarily correlated with its biological significance; proteins can be expressed at a very low level and still perform important biological functions in a given context. Pragmatically, however, we have observed diminishing rates of success for ChIP-seq and epitope-tagged ChIP-seq datasets as the expression level of those TFs decrease (Meadows et al. 2020). Therefore, we identified all TFs in HepG2 cells that are expressed at levels of at least two transcripts per million (TPM), as measured by RNA-seq (ENCSR181ZGR). There are 895 TFs expressed at this level in HepG2 cells. We compiled the existing datasets produced from our lab and others from the ENCODE portal for 479 (53.5%) of these 895 TFs (**See Methods, Supplemental Table 1**). In addition to these 479 TFs, we also analyzed data for 50 TFs expressed at <2 TPM for which we were able to generate high quality ChIP-seq data despite their low

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

expression, and for 151 non-TF DAPs and nine histone marks, for a total of 680 unique ChIP-seq DAP targets in HepG2 and nine histone modifications. This expanded catalog of DAPs and associated gene regulatory datasets provides a rich resource to characterize and understand the functional impact of DAP binding on gene regulation.

DAP associations at candidate *Cis*-Regulatory Elements (cCREs) reveals interaction of DAP function and regulatory context

Transcription factors impact expression by associating with or binding to DNA, specifically at *cis*-regulatory elements (CREs). We therefore sought to determine which *cis*-regulatory elements are bound by TFs and the patterns of activity that those bound regions display. We examined *cis*-regulatory elements for the presence of at least one DAP peak. To do this, we used the Registry of Candidate *cis*-Regulatory Elements (V4 cCREs) derived from ENCODE data (The ENCODE Project Consortium et al. 2020, Moore JE, Pratt HE, Fan K, et al. in prep). These cCREs represent genomic regulatory elements across multiple human cell types and are derived from chromatin accessibility assays (DNase-seq and ATAC-seq), histone modifications and DAP-binding data. To filter for cCREs that are relevant in HepG2, we overlapped with HepG2 ATAC-seq data, generating a set of 318,567 HepG2 cCREs. Of these, 84.2% have at least one of the assayed DAPs associated, and those cCREs with no DAPs associated in HepG2 are largely distal enhancer-like sequences (**Figure 1A, Supplemental Figure 1, Supplemental Table 2**). We compared this pattern of binding to dinucleotide-matched control sequences and found that these regions are significantly more bound than controls (**Supplemental**

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure 2, Supplemental Table 2). As the number of associated DAPs increases at cCREs, the proportion of cCREs defined as “promoter-like” increases (**Supplemental Figure 3, Supplemental Figure 4, Supplemental Table 3**). We therefore conclude that coverage of cCREs with at least some subset of their associated DAPs is approaching completeness.

We also found 50,446 (15.8%) annotated cCREs overlapping with an ATAC-seq peak in HepG2 cells but with no DAP peaks in our dataset. Given their predicted regulatory activity and their open chromatin state in this cell type, we would expect that they should be bound by some DAP. At least three explanations are possible: 1) these cCREs are unbound by any DAP; 2) they are bound by DAPs that have not yet been assayed in HepG2 cells; and/or 3) that DAP binding was potentially missed as false negatives in the ChIP-seq assays. To measure functional activity of these elements (as well as for other analyses below), we performed a lentiviral Massively Parallel Reporter Assay (lentiMPRA, or MPRA) following established methods (Gordon et al. 2020). MPRAs functionally validate the regulatory activity of thousands of DNA elements simultaneously by insertion of DNA up- or downstream of a transcribed element (Klein et al. 2020). Our MPRA experiment contained 69,210 elements of 170 bp each, selected from various promoter and distal cCRE elements as well as from non-cCRE elements, as well as a set of synthetic, non-genomic elements with various numbers of TF motifs. We supplemented this dataset by also analyzing a publicly available HepG2 lentiMPRA dataset containing 139,877 elements of 200 bp each (Agarwal et al. 2023) representing all cCRE elements published under the previous ENCODE cCRE version 3 release. We found that only 32% of the elements in our lentiMPRA had any overlap with elements in the Agarwal MPRA,

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

indicating that these assays provided substantially different information. In both of these lentiMPRA datasets, the cCREs without DAP peaks had, on average, lower MPRA signal than did the elements with DAP peaks. However, these elements had significantly higher activity than negative control elements in our MPRA (**Figure 1B, Supplemental Table 4**), suggesting that some of these regions have regulatory activity and presumably have DAP associations that are not present in our dataset. We confirmed that this finding remained true using the Agarwal lentiMPRA (**Supplemental Figure 5, Supplemental Table 5**). To estimate how much additional cCRE coverage we are missing from the more than 400 DAPs expressed at a TPM of 2 or more in HepG2 cells but that have not yet been assayed, we performed a subsampling analysis to determine how the fraction of the cCREs covered changes with varying numbers of analyzed DAPs. We note that the first 200 DAPs appear to add a large amount of information, but for new DAPs beyond 200, there is a sharp decrease in new coverage. We extrapolate that assaying an additional 400 DAPs would result in total coverage of only 88% to 90% of HepG2 cCREs (**Supplemental Figure 6**). This suggests that adding more DAP datasets generated with ChIP-seq would not result in complete coverage of all HepG2 cCREs and may potentially be insufficient to detect trans-regulatory factors at some fraction of these sites. To further explore whether peak-calling in ChIP-seq misses some meaningful DAP associations, we built a gkm-svm model (Ghandi et al. 2016) for each DAP based on its bound sequences and applied these models to unbound regions. We found that 67% of cCRE regions without DAP peak calls had a strong score for at least one DAP's gkm-svm model (**Supplemental Figure 7, Supplemental Table 6**), suggesting that there are missed associations at these genomic regions. Finally, to explore this further, we used a sample

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

of 45 bigwig signal files from the ENCODE Project to determine if signal over these regions is meaningfully different. We found that in 51.1% of these bigwig files, unbound cCREs with gkm-svm scores at or above the 90th percentile had higher signal than unbound regions with gkm-svm scores at or below the 10th percentile (**Supplemental Figure 8**). This suggests that many of these regions may represent false negative peak calls for at least one TF. As an alternative to ChIP-seq, predictive computational methods for DAP binding may offer further insights (Ghandi et al. 2016; Schreiber et al. 2020).

We asked how DAPs work together and impact gene expression at different cCREs, specifically promoters and distal elements. As a measure of the functional impact of DAP associations, we assessed how varying numbers of DAP associations at cCREs impact potential gene expression by examining the correlation between numbers of DAPs and transcriptional activity in our lentiviral MPRA experiment (**Figure 1B, Supplemental Table 4**), and the MPRA from Agarwal et al. (2023) (**Supplemental Figure 5, Supplemental Table 5**). We observed a trend in both promoters and distal elements where expression increased as the number of bound DAPs increased. This supports earlier findings that MPRA activity correlates with number of DAPs bound at the endogenous element (Ramaker et al. 2020). Elements from distal regions have a much wider distribution of expression, but, on average, lower levels of reporter expression compared to promoters for each bin of DAP numbers. This is expected in the lentiMPRA assay system, in which test elements are directly upstream of the reporter open reading frame and thus ideal for promoter activity tests. However, distal elements with very large numbers of associated DAPs (>400) showed a sharp decrease in activity, while promoter elements with these numbers did not. As this is a novel finding, we confirmed the results

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

by using a second MPRA dataset (**Supplemental Figure 5, Supplemental Table 5**). Thus, distal elements with high number of TFs and associated cofactors show significantly lower activity in MPRA assays versus elements with fewer DAPs bound.

To further explore the distal elements with high numbers of DAPs yet low expression in the MPRA, and to uncover potential explanations for this finding, we examined these regions more broadly. We speculated that placing these regions in an artificial context close to the promoter of a reporter gene may not accurately recapitulate their endogenous looping to distal promoters to enhance gene expression. Therefore, we examined Activity-by-Contact (ABC) loop models (Fulco et al. 2019) in HepG2, an analysis that assigns a confidence score to an active loop, or connection between a promoter and a distal element using three-dimensional interaction metrics in the form of Hi-C data, as well as RNA-seq, ATAC-seq, and histone marks. Crucially, the ABC method does not consider DAP binding as a contribution to the score. We found that distal elements are increasingly likely to have an ABC connection to at least one promoter as the number of DAPs increases until reaching ~200 DAPs, with a substantial drop in connections for elements with more than 400 associated DAPs (**Figure 1C**). We compared this to dinucleotide-matched control sequences, and found that compared to controls these regions are enriched for ABC associations (**Supplemental Figure 9, Supplemental Table 7**). Nonetheless, we found that very highly-bound distal elements that do contain an ABC loop tended to be connected to highly expressed genes (**Figure 1D, Supplemental Table 8**). This suggests that these elements functionally enhance gene expression but not in a proximal MPRA context, and highlights the importance of regulatory context on DAP function. To explore these regions further, we identified all bound regions in the genome

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

with 201-400 DAPs bound and compared them to regions with 401+ DAPs bound. We found that regions with 401+ DAPs bound were on average larger (mean 1703 bp versus 1237 for 201-400), have a higher GC content (59.1% vs. 57.9%, $p=8.65 \times 10^{-4}$, Mann-Whitney *U*-test), and higher ATAC-seq signal (mean bigWig signal 538.4 vs. 443.75, $p<=2.2 \times 10^{-16}$, Mann-Whitney *U*-test) (**Supplemental Figure 10**).

Modeling of TF effects on gene expression identifies putative trans-activators and repressors

We next asked if these numerous ChIP-seq datasets would allow us to predict and quantify each TF's activating or repressing behavior. TFs are a specific subset of DAPs that bind to DNA in a sequence-specific manner to regulate transcription (Lambert et al. 2018). Parsing out specific contributions of TFs to expression patterns is an ongoing effort in the genomics community. To explore functional effects of TF localization on gene expression, we created models based on the association of 529 TFs to the promoter region of genes. Linear modeling of gene expression based on promoter TF association offers a clear and interpretable effect of each TF as activating or repressing. Our reasoning for limiting this analysis to the promoter regions of annotated genes, as opposed to including all distal candidate enhancers, was to cleanly assign the TF binding event to a gene expression outcome. We built linear models based on 70% of promoters and used these to predict gene expression on the remaining 30% of promoters. We achieved accurate correlation between predictions and observation (**Figure 2A**), comparable to neural network models based on large amounts of sequence data

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

(Agarwal and Shendure 2020). Prediction accuracy was robust to promoter subtypes (CpG Island [CGI] promoters and non-CGI promoters) and was applicable across cell types (**Supplemental Figures 11-13**). Prediction accuracy was also moderately better than estimating expression levels on the number of associated TFs alone (**Supplemental Figure 14**). While we note that there are some cases where observed expression is high while predicted expression is low, as well as the converse, these are a minority of cases (~10%), and likely represent either the effects of factors which were not included in our ChIP-seq datasets, the role of distal enhancers, or the context-dependent effects of transcription factor function (see Discussion).

These findings show that modeling gene expression as a function of TF association at gene promoters can elucidate the functions of specific TFs. To determine a level of confidence in each TF's activating and repressing activity, we performed repeated subsampling of TFs and promoters for training and testing of models. For each TF, we

built 500 models based on a random 70%/30% split of promoters and repeated subsampling of 79 other TFs to analyze with the factor of interest. This ensured that a range of estimates for gene expression impact were gathered for each TF. We then identified those TFs with the largest positive and negative impact on gene expression, and assigned a significance based on the fraction of submodels in which the TF showed significant impact on gene expression (**Figure 2B**). The majority (n=359) of TFs we examined showed a positive impact on gene expression, reflecting the more common role of transcription factors as positive regulators of expression. These also included factors associated with active chromatin such as ASH2L, KMT2A and KMT2B **(Supplemental Table 9**). We also identified 26 TFs with a negative impact on gene

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

expression in the models that were significant in at least 50% of all submodels (**Supplemental Table 9**), indicating their consistent association with a lower level of expression when localized to promoters. These included confident prediction of a known repressor, REST (also known as NRSF), and several members of the repressive KRAB-ZNF family, including ZNF334, ZFP14 and ZNF140. This is consistent with a model in which these factors localize to regulatory elements to directly or indirectly decrease transcription of nearby genes. Many of the other factors have not been described as transcriptional repressors, including RREB1, ZNF660, and AHR, indicating we have identified novel putative transcriptional repressors. We explored the distribution of peaks for these 26 candidate repressors, and compared them to the top 26 candidate activators. We found that, in general, candidate repressors had fewer peaks than candidate activators (**Supplemental Figure 15**), consistent with a "Hit and Run" model of transcriptional repression (Shah et al. 2019). Comparing the distribution of peaks across cCRE types, we found that the majority of peaks for each type were found in Promoter-Like Signature, proximal Enhancer-Like Signature, and distal Enhancer-Like Signature regions, and that distribution of the fraction of peaks in these three region types did not differ significantly between top candidate activators and candidate repressors (p>0.05, Mann-Whitney U-test).

We next set out to confirm these predictions with functional data using our lentiMPRA dataset. We identified promoter elements in our lentiMPRA dataset whose genomic coordinates were bound by at least one of our 26 candidate repressors, the top 26 candidate activators, or a control set of 26 random TFs not found in either set. We compared the promoter regions bound with at least one of these TFs to promoter regions

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

bound by a similar number of TFs, but lacking the TF of interest. As the model predicted, regions with one of the 26 top predicted activators had higher activity in the MPRA compared to regions lacking one of these activators (**Figure 2C, Supplemental Table 10**). Correspondingly, activity of regions with a candidate repressive DAP was lower than the activity of regions bound by similar numbers of TFs **(Figure 2C, Supplemental Table 10)**, except in the case of 1-5 DAPs bound, which had only 10 observations in this dataset, so is likely a result of noise**.** These candidate repressor patterns held for most categories of the number of DAPs bound when looking at non-promoter elements (**Supplemental Figure 16, Supplemental Table 11**). We noted that REST was consistently the strongest candidate repressor, with other candidate repressors having a lower fraction of significance across submodels, a smaller repressive effect size, or both (**Figure 2B**), consistent with its well-characterized role in repression (Ballas and Mandel 2005; Ooi and Wood 2007; Huang et al. 1999). In order to confirm that other candidate repressors still showed a repressive effect, we performed additional analysis and found that the general repressive trend of these factors held when removing the effect of REST binding (**Supplemental Figure 17, Supplemental Table 12**). Lastly, the random set showed a pattern inconsistent with either case, with lower expression levels for low numbers of TFs bound and higher expression levels for high numbers of TFs bound (**Figure 2C, Supplemental Table 10**). Similar patterns were also found in the Agarwal lentiMPRA elements (**Supplemental Figures 18-19, Supplemental Tables 13-14**). We also assessed each of our top 26 candidate repressors and activators individually by performing a paired comparison of elements bound or not bound by the factor of interest and matched for the number of factors bound (**Supplemental Figure 20, Supplemental**

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

294 **Table 15**). We note that findings are generally concordant, with activators having a higher
295 distribution of differences than repressors. We explored the possibility that GC content
296 may confound these findings by comparing GC content differences between matched
297 sequences (**Supplemental Figure 21**), and found that the distributions were centered at
298 0, and balanced in either direction, suggesting that GC content could not explain the
299 differences observed. In some cases, predicted activators did have a lower activity level
300 than matched controls, but we note that these cases were almost universally nominally
301 significant (p<=0.05, * in **Supplemental Figure 20, Supplemental Table 15**), with
302 ZNF501, ZFX, ZFY, and ASH2L showing a greater degree of discordance with
303 predictions. This may be a reflection of the fact that the vast majority of factors are
304 predicted to have an activating effect in our models, indicating that replacement with
305 another factor is likely to result in a replacement with a stronger activator. These
306 observations generally confirm our predictions of TF activating and repressing activity.

307 To further demonstrate the utility of our model, we sought to direct binding of these
308 activating and repressing factors to elements using known sequence preferences to
309 activate or repress MPRA activity. We included in our MPRA a set of test elements with
310 known motifs from the JASPAR database for eight putative activating and five putative
311 repressing TFs. These motifs were inserted into two different promoter elements with
312 randomized insertion location, orientation, and spacing (**See Methods**). We normalized
313 reporter gene expression to control promoters to explore the impact of adding one, two,
314 or five motifs to the promoters, as it has previously been noted that the addition of multiple
315 motifs increases signal (Smith et al. 2013). We found that, of eight tested candidate
316 activators, five (ATF1, CEBPG, FOSL2, NFYC, NRF1) show the expected behavior when

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

a single motif is added, while three (MYC, ZNF317, ZNF331) show the opposite effect (**Figure 2D**, p-values in **Supplemental Table 16**). While this was unexpected, we note that for many TFs there are known context- or cobinding-dependencies of function, such as the finding of MYC mediating repression when binding with ZBTB17 (Walz et al. 2014). Additionally, we note that for ATF1 a single motif results in activation, while additional motifs result in lower signal than with only a single motif added. These results may be due to the artificial nature of the assay; in this case, insertion of multiple motifs did not lead to a simple interpretation. Of five tested candidate repressors, all showed the expected negative impact on expression. These general trends remained true when restricting to only one or the other of the two promoter elements (**Supplemental Figures 22-23**, p-values in **Supplemental Table 16**).

Analysis of HOT sites shows that ABC score increases with DAP occupancy

High-Occupancy Target (HOT) sites are regions of the genome with a high number of associated DAPs (Yip et al. 2012), and their biological meaning is an area of ongoing interest (Wreczycka et al. 2019; Hudaiberdiev and Ovcharenko 2023). In particular, several lines of evidence suggest that factor association can occur indirectly (Worsley Hunt and Wasserman 2014; Gordân et al. 2009; Nie et al. 2020), and this may contribute to HOT site formation. Given the large number of ChIP-seq datasets available in HepG2, we sought to provide new insight into the biology and genome organization that produce HOT sites. Though 38.5% of cCREs were bound by few DAPs, a substantial fraction was bound by large numbers of DAPs (**Supplemental Figure 3, Supplemental Table 3**). We

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

defined HOT sites as discrete regions with peaks called in 25% or more of the DAPs assayed, as suggested by our previous study (Ramaker et al. 2020). We identified 13,001 HOT sites in HepG2, representing only 1.7% of all discrete DAP-associated regions, but 52% of all DAP peaks across experiments in HepG2 cells. We note that HOT sites occur primarily in promoters or in proximal or distal enhancer-like signature elements (**Supplemental Figure 3, Supplemental Table 3**); 72.0% are in proximal promoter regions (i.e. overlapping PLS [Promoter-Like Signature] and pELS [proximal Enhancer-Like Signature] cCREs) and 27.5% are distal regions (i.e. overlapping dELS [distal Enhancer-Like Signature]) We found that as the number of factors increases, HOT sites become more enriched for promoters (**Supplemental Figure 4, Supplemental Table 3**).

Because of this association between HOT sites and promoters, we further explored the relationship between promoter activity and promoter type in these regions. In our MPRA, we observed a general positive correlation between the number of DAPs associated with a promoter and the activity of that promoter element in the MPRA (**Figure 1B Supplemental Table 4**), as has been previously reported (Partridge et al. 2020; Ramaker et al. 2020; Agarwal et al. 2023). Similar to other reports (Chen et al. 2014), we found that HOT promoter regions are primarily in CpG islands (6,431 of 7,411 promoter HOT sites) and frequently at annotated housekeeping genes (1,581 of 7,411 promoter HOT sites) (Hounkpe et al. 2021), and that the likelihood any given housekeeping gene promoter is a HOT site is very high (94.5%).

To help understand the function of HOT loci and to compare them to other DAP-associated regions in the genome, we performed a Principal Component Analysis (PCA) on a binary matrix of all regions bound by at least three DAPs, while noting which DAPs

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

are associated at each region. We confirmed that previously-noted associations between PCs 1 and 2 (Partridge et al. 2020) are maintained with the number of factors bound and proximal-distal distinctions, respectively (**Supplemental Figures 24-25).** We also noted that Activity-by-Contact (ABC) scores (Fulco et al. 2019) derived from high resolution intact Hi-C data for HepG2 showed relationships with both PC1 and PC2 (**Supplemental Figure 26**). The strength of connectivity as measured by ABC score increases along principal component 1 (number of factors bound in a region, Spearman's Rho = 0.2592, $p <= 2.2 \times 10^{-16}$), and decreases along principal component 2 (proximal versus distal elements, Spearman's Rho = -0.0724, $p <= 2.2 \times 10^{-16}$), suggesting that there are stronger loop interactions when a large number of DAPs are bound.

Motif placement in distal-promoter pairs suggests that 3D interactions contribute to HOT site formation

We next asked if the concordance in the number of factors bound with ABC score (**Supplemental Figure 27**, Spearman's Rho = 0.2591, $p <= 2.2 \times 10^{-16}$) could highlight distinct DAP binding patterns in promoter-distal interactions at HOT and non-HOT loci. We compared loci in the genome as either HOT or non-HOT, the latter having one or more associated DAP peaks but below the threshold of 25% of the DAPs in our dataset. We then observed the number of significant ABC connections that putative enhancers and promoters of each HOT classification (HOT promoter, non-HOT promoter, HOT putative enhancer, non-HOT putative enhancer) had with any other region (**Figure 3A**). We observed that HOT putative enhancers have a significantly larger number of

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

connections than do non-HOT putative enhancer, and that, conversely, HOT promoters tend to have fewer interactions than do non-HOT promoters. Because, as noted above, 87% of HOT promoters are CpG island promoters, which are a mix of ubiquitously expressed and tissue-specific genes, this potentially highlights a mechanism where DAP regulation is largely proximal for these promoters.

To determine whether connections in three-dimensional space led to peak calls for the same DAPs at both connected regions, we explored the fraction of DAP peak calls shared between a connected putative enhancer and promoter in cases where one or the other, both, or neither of the regions is a HOT site (**Supplemental Figure 28, Supplemental Table 17**). We note that the rate of shared DAP peaks is higher when both sites are HOT, and significantly lower when neither site is HOT **(Supplemental Figure 28**, green boxes), consistent with a model in which ChIP-seq peaks might be detected as a result of indirect association.

Given this observation, we wanted to distinguish between two possibilities: the common DAP association is due to 1) independent direct binding of a DAP at multiple sites (either one molecule directly binding the two connected DNA regions simultaneously or two separate molecules binding, one at each region), or 2) direct DAP binding at one locus with an indirect pulldown of nearby connected chromatin in the ChIP-seq dataset. To answer this, we identified those DAPs for which a motif is known and quantified percentages of their peaks in promoters or putative enhancers that had a motif present. We found that if both the promoter and putative enhancer are HOT, the motif for the shared DAP is equally likely to be found in either locus of the interacting pair (**Figure 3B, Supplemental Table 18**). In contrast, if only one locus is a HOT site, there is a preference

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

for the motif of interest to be found within peaks in the non-HOT site rather than the HOT site (p=3.37 x 10^{-7} Kolmogorov-Smirnov test, **Figure 3B**, **Supplemental Table 19**). This trend remains true, although reduced in size, when restricting to cases of non-CGI promoters or CGI promoters (**Supplemental Figures 29-30, Supplemental Table 19**). We summarize this model in **Figure 3C**. For cases where a peak was observed in both the putative enhancer and the promoter for a given transcription factor and a motif occurs in one of those locations, we hypothesize that a given TF is more likely to be directly interacting with the DNA where its motif occurs, and a peak is found in regions without its motif due to indirect interactions. These observations are consistent with a model in which the large number of DAPs found bound at HOT sites can often result from indirect association of DAPs at a promoter-enhancer interaction, resulting in an apparent ChIP-seq peak when direct DAP-locus association is unlikely. We note that, among the factors analyzed, MYC and TP53 both contain strong evidence of intrinsically disordered regions based on DisProt annotations (Quaglia et al. 2022). Such disordered regions are thought to be involved in protein-protein interactions (Chen et al. 2022; Morris et al. 2021), and several studies have noted that the disordered regions of these proteins are involved in protein-protein interactions (Wells et al. 2008; Di Lello et al. 2006; Fladvad et al. 2005; McEwan et al. 1996; von der Lehr et al. 2003).

DAP localization outside of open chromatin regions suggests expansion of candidate CREs

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

We explored cases of multiple DAPs binding outside of annotated cCREs to assess whether these cases of binding had biological meaning. Early ChIP-seq efforts by ENCODE showed that more than 94% of DAP peaks occurred in open chromatin regions (Thurman et al. 2012). This is expected because many DAPs prefer binding to nucleosome-free DNA and cooperatively bind with complexes that displace nucleosomes or modify chromatin to make it more open, such as acetylation. The exceptions to this are DAPs associated with heterochromatin or lamin-associated regions, or pioneer factor TFs, which have been found to bind to closed chromatin and lead to nucleosome remodeling and histone modifications that make the region more accessible to additional DAPs (Becker et al. 2017; McCarthy et al. 2021). We expected to observe both of these groups of DAPs in our dataset, but also could find novel DAP associations with closed chromatin. We found that only 6.3% of DAP peaks occurred outside of any type of cCRE defined as having open chromatin, and 13.4% of peaks occurred outside of an ATAC-seq peak in HepG2. These observations are in agreement with prior analyses (Thurman et al. 2012).

For a conservative approach to DAP binding in closed chromatin regions, we limited our analysis to those sites bound by two or more DAPs and regions that were 700 bp or more from a cCRE associated with open chromatin (**Supplemental Table 20, Supplemental Figure 31**). We found 16,412 such closed chromatin regions bound by at least two DAPs. These DAP-bound closed chromatin regions had, on average, 2.8 associated DAPs. A majority (95.3%) were outside of an ATAC-seq peak in HepG2, suggesting that most were found in closed chromatin states (**Figure 4A**). Analysis of histone modifications at these regions revealed that 70.3% lack any histone signal, while another 15.5% of regions had

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

H3K9me3 signal, a mark of heterochromatin and association with the nuclear lamin. This suggests binding by factors outside of open chromatin cCRE regions is more common than expected, accounting for >5% of bound elements in HepG2 cells but occurring in several thousand independent locations far from cCREs. In contrast, a substantial set (12.6%) were associated with H3K4me1, a sign of poised or primed distal elements (Creyghton et al. 2010; Rada-Iglesias et al. 2011). As expected, we find peaks for one or more pioneer TFs at these poised sites, such as FOXA2, ATF7, JUND, CREB1, or CEBP in 41.8% of the H3K4me1 regions. There are several possible explanations for finding non-cCRE regions bound by several transcription factors. It is possible these genomic regions do not yet have sufficient evidence to be designated as cCREs but may be designated as such in subsequent versions of the Registry. Some of these regions might represent false positives from DAP ChIP-seq data, although we minimized this possibility by limiting to regions with two or more DAPs bound. We note that, while approximately 80% of these regions fall in repetitive sequence (**Supplemental Figure 32**), there is substantial overlap of other cCRE regions with repetitive sequence content (e.g. 38% of PLS and 57% of dELS overlap with repetitive sequence). We further note that the majority of these regions only show the presence of a few factors, with sMAF factors highly represented, suggesting that false positives due to repetitive sequence are not a consistent problem across ChIP-seq experiments. We speculate that these regions represent significant DAP genomic localization in closed chromatin. Together, these findings suggest that closed chromatin cCRE regions harbor important regulatory activity, and that TF binding should be considered in future efforts to identify novel candidate regulatory elements.

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

Binding of small MAF proteins at closed chromatin loci with active or repressive activity

In our analysis of the multiple DAPs bound at closed chromatin regions (**Figure 4A**), we found that 44.8% of bound non-cCRE regions showed association with MAFF, MAFK, or to a lesser extent MAFG. These three TFs make up a family of small MAF (sMAF) proteins. Unlike the related large MAF family of TFs, the sMAF proteins lack a transcriptional activation domain, and are known to act as transcriptional repressors when bound as sMAF/sMAF dimers (Motohashi et al. 2006; Katsuoka and Yamamoto 2016). The vast majority (98.6%) of sMAF co-associations in peaks, putatively indicating sMAF/sMAF dimer binding in these closed chromatin regions, consisted of only MAFF/MAFK, indicating these regions are strong candidates for sMAF repressed cis-elements. These regions all fall at least 700 bp away from annotated cCREs, and had a mean distance of 2164 bp from the nearest cCRE.

Based on these observations, we explored binding patterns and associated signals for sMAF proteins in our dataset. It is known that sMAFs can also heterodimerize with non-sMAF proteins such as NFE2 and act as transcriptional activators (Friling et al. 1990; Rushmore et al. 1991; Katsuoka and Yamamoto 2016). We examined global binding of all sMAF TFs in HepG2 cells and categorized them into two types of loci: 1) bound only by putative sMAF/sMAF dimers, indicated in our dataset by peak overlaps for two sMAF TFs; or 2) bound by a putative heterodimer composed of one sMAF with an activating TF, again assumed from peak overlaps. In this analysis, the former group consisted of 30,463

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

loci and the latter of 1,281 loci. The two groups have very distinct characteristics. The sMAF/sMAF dimers are found in regions with little to no ATAC-seq signal, while sMAF heterodimers show strong open chromatin (**Figure 4B**). The motif found at sMAF/sMAF dimers was the full T-MARE element, with a TGC on each end, while the sMAF heterodimers produced a motif of only the core TRE sequence (**Figure 4C**). As expected, the expression of the nearest genes to sMAF/sMAF dimers are significantly lower than those nearest to sMAF heterodimers or to a control set of DAP-bound regions (**Figure 4D**). We then asked if the sMAF bound regions show evidence of looping, possibly indicating a mechanism for direct repression of a connected gene. We used ENCODE Hi-C data and found almost no evidence for a loop originating at sMAF/sMAF dimers, while sMAF/co-factor bound loci loop as frequently as regions bound by any two random DAPs (**Figure 4E**). Thus, sMAF/sMAF dimer binding at closed chromatin appears to have a widespread repressive impact on gene expression through mechanisms other than looping, while the heterodimer behaves as an activating TF.

Discussion

In this study, we generated ChIP-seq data for nearly half of all DNA associated proteins expressed in HepG2 cells, including representatives of all major families of sequence specific transcription factors, and spanning the full range from highly expressed to nonabundant and from previously well-studied to largely unknown. This substantially larger and more biologically diverse catalog allowed us to seek specific insights into individual TF functions as well as evaluating more global patterns of genomic organization

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

518 and gene regulation. The vast majority of known ENCODE cCREs that are openly
519 accessible in HepG2 cells showed evidence of significant ChIP signal by at least one of
520 our 680 assayed DAPs, and most remaining open cCREs that scored as unbound were
521 computationally predicted to be highly preferred by at least one assayed DAP. We
522 demonstrate the power of this dataset by 1) identifying candidate trans-repressors; 2)
523 exploring binding patterns in distal-promoter interactions; 3) untangling the contributions
524 of such binding to HOT sites; 4) uncovering candidate cCREs that lack open chromatin
525 but are bound by DAPs; and 5) exploring vignettes of a specific class of TFs and their
526 activities.

527 We found that the majority of cCRE sequences are bound by a DAP which currently has
528 ChIP-seq data available (**Figure 1A, Supplemental Table 2**), and explored the likelihood
529 of covering apparently unbound cCREs would find a peak given ChIPing of additional
530 factors (**Supplemental Figure 6**). We find that many of these unbound regions are
531 predicted to have a high affinity for a DAP with existing ChIP-seq data (**Supplemental**
532 **Figure 7, Supplemental Table 6**), and that some regions may be bound, but have failed
533 to call a peak due to false negative error, based on bigWig signal (**Supplemental Figure**
534 **8**). Together, these findings suggest that better methods for determining DAP-DNA
535 associations may further elucidate regulatory information in existing data.

536 We tested functional activities of elements across a wide range of genomic elements with
537 a large lentiMPRA dataset and supplemented the resulting data with recently-published
538 independent functional data in the HepG2 cell line (Agarwal et al. 2023). Evidence for
539 candidate 26 DAP repressor elements emerged, and this finding was robust to removal
540 of a major canonical repressor, REST. While ranges in **Figure 2B** were large, this is

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

541 expected as many transcription factors display context-dependent regulatory behavior.
542 This is highlighted by our discussion of sMAF genes and their differing behavior based
543 on binding partners in **Figure 4**. Another example of this kind of behavior is in the RFX
544 family, which have a dimerization domain used for homo- and heterodimerization both
545 within the RFX family (Sugiaman-Trapman et al. 2018; Reith et al. 1990; Aftab et al. 2008;
546 Morotomi-Yano et al. 2002) and with other TFs (Caretti et al. 2000). Their interactions
547 have been hypothesized to be important for transcriptional activity (Reith et al. 1994),
548 including activation and possible repression (Iwama et al. 1999; Zhu et al. 2000). Our
549 study highlights the possibility that RFX1 and RFX3 appear to have a repressive effect in
550 HepG2 cells (**Figure 2A**). Further study of this family and a deeper examination of TF
551 cooperation and competition will reveal further nuance of individual TF activity. This study,
552 however, has provided broad hypotheses for the function of hundreds of TFs. We also
553 note that the ability to detect repressors, in particular, is complicated by the possibility of
554 hit-and-run repressors (Shah et al. 2019), and thus a factor displaying repressive activity
555 may not have been captured by ChIP-seq. The complicated nature of TF behavior
556 outlined above may explain in part the unexpected behavior of some predicted activator
557 motifs in our MPRA seen in **Figure 2D**, such as MYC, ZNF317, and ZNF331.

558 The lentiMPRA functional data was also highly useful for supporting analyses derived
559 from the correlative data represented by the ChIP-/CETCh-seq experiments. This data
560 supported the general notion that genomic elements with DAP binding are more likely to
561 be functional, and indeed that there is a correlation between the number of bound DAPs
562 and the activity of these elements.

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

We also found evidence that HOT sites in promoters represent functional binding rather than mere hyper-ChIP-able regions (Wreczycka et al. 2019), as the average element activity continued to increase with genomic binding signal as factors are added, even for highly-bound regions. By contrast, the increase in element activity seen with increasingly HOT promoters was not true for highly-bound distal elements (**Figure 1B Supplemental Table 4**). A possible explanation for this result is found in the analysis of ABC connections, ChIP-seq peaks, and motif locations explored in **Figure 3**. We found that, for a given HOT region, when making an active connection with a non-HOT region, specific explanatory motifs are more likely to occur in the non-HOT region. We also find that HOT putative enhancers are more likely to have a higher number of connections than non-HOT putative enhancers (**Figure 3A**). It is known that a single enhancer can regulate multiple genes (Fulco et al. 2016; Han et al. 2018), and this finding may help distinguish interesting properties about multi-target versus single-target enhancers. Finally, as the number of factors bound at a putative enhancer increases, the expression of genes with ABC support continues to increase (**Figure 1D, Supplemental Table 8**). These observations are consistent with a model in which highly active enhancers regulating multiple promoters have apparent ChIP-seq peaks due to close 3-dimensional association with promoters at which factors are directly bound. In this model, when removed from their larger genomic context, these distal regions would have far fewer factors directly bound than ChIP-seq data would suggest, leading to a lower overall activity. A key caveat in making element activity correlations with TF occupancy, open chromatin elements and 3D interactions is that the element sizes and context for these is very different from the MPRA assays. MPRA elements are about 170 bp while typical

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

586 conserved distal elements are 450 bp, and active promoter regions are often larger than
587 that. In the native chromosome, the entire element is operating, and is doing so with the
588 option to interact with other nearby elements. Thus, MPRA assays, with their impressively
589 large data outputs, are a valuable foundation to support future multi-element and larger-
590 sized element efforts.

591 While **Figure 3C** highlights one possible model for understanding HOT sites as partially
592 a function of indirect association due to 3D interactions, we note that there are many other
593 mechanisms that may explain large numbers of factors binding at the same location.
594 These may include the known protein-protein interactions through dimerization domains,
595 such as those known for the sMAF and RFX families, or the noted condensation of factors
596 through the phase-transition of their activation domains, as recently noted (Boija et al.
597 2018).

598 We found that, for closed chromatin regions bound by any two (or more) factors,
599 sMAF/sMAF dimers are by far the most common binding pair. It has been reported that
600 sMAF factors can bind as repressors, or co-bind with other DAPs to activate transcription
601 (Katsuoka and Yamamoto 2016). We here find evidence here that sMAF/sMAF dimers
602 play a role in repression through mechanisms that do not appear to involve looping.
603 Whether or not they are bound in known annotated cCRE regions, these dimers occur in
604 closed chromatin, and nearby genes show a much reduced level of expression when
605 compared to genes near sMAF heterodimers with an activator, or other co-bound pairs of
606 TFs. These observations may also be related to the previously-mentioned hit-and-run
607 mechanism of repression (Shah et al. 2019). If this is the case, then sMAF proteins may

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

be only responsible for maintenance of repressed state, while another factor may have induced the repressed state.

These analyses demonstrate the power of large ChIP-seq datasets, coupled with functional assays, in a specific cellular context. This includes the ability to mine specific factor behaviors and corresponding cis-element classes that are obscured by aggregating binding data across cell and tissue types. This recognizes that many TFs have target sites and activities (enhancing versus repressing, for example) that depend on cell type because of isoform coding differences and post-translational activity differences. Furthermore, bound DNA elements are typically compound sequence structures whose integrated functional output is a nonlinear combination of the action of any given bound factor. This means that analyses at the level of element subtypes within the environment of a single cell type gives important leverage.

There remain hundreds of sequence-specific TFs (Lambert et al. 2018) for which high-quality ChIP-seq datasets are not yet available. Having such datasets in a context with a large number of other factors for which there is ChIP-seq data will allow for a more refined and concrete understanding of these factors' functions. Promising tools are being developed to predict the activity of factors (Schreiber et al. 2020; Avsec et al. 2021), yet these tools require high-quality datasets in defined contexts for accurate training. This underlies the need for continued work in generating and analyzing ChIP-seq datasets. The current effort is limited by the fact that the vast majority of ChIP-seq data is in cancer cell line contexts, which affects the identity of factors to be assayed as well as the relevance of these findings to non-cancer biology. As methods for ChIP-seq continue to be improved, and as complementary improvements are made in high-throughput cis-

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

element assays, we expect a richer and more predictive understanding of the cis-trans code that controls genome activity.

Methods

Data Analysis

Data analysis was performed using R versions 3.6.1 and 4.1.0 (2010), as noted in appropriate scripts.

ChIP-seq datasets

We downloaded processed Optimal Peak Calls of all HepG2 ChIP-seq datasets available on the ENCODE portal on July 6, 2021, and supplemented the datasets with those performed locally at the HudsonAlpha Institute for Biotechnology which were not uploaded to the ENCODE portal. In-house data was processed using the same ENCODE ChIP-seq Processing Pipeline to be consistent with data from the ENCODE portal.

Selection of preferred ChIP-seq datasets

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

For DAPs which were represented by multiple ChIP-seq datasets, we identified a preferred dataset based on the number and severity of Audits present on the ENCODE portal. We identified a list of unacceptable audits as follows:

- Failed IDR
- Low Read Depth
- Control Low Read Depth
- Extremely Low Read Depth
- Poor Library Analysis
- Severe Bottlenecking

For this set of audits, in cases of multiple datasets for the same DAP, we identified the number of audits contained in each dataset. If one dataset contained fewer of these audits than the others, we chose the dataset with the smallest number of audits, including zero. If more than one dataset had this smallest number of audits, compared the audits within each dataset to determine if a given dataset had a less severe set of audits (e.g. Low Read Depth versus Extremely Low Read Depth), or a smaller number of audits overall. Ranking of the severity of audits was in the order of the list above, from most severe to least severe. Finally, if datasets were not distinguishable, we chose the dataset with the largest number of peaks. In cases where a dataset was available on the ENCODE portal, such dataset was always preferred to those not from the portal.

The same process was used to identify K562 ChIP-seq datasets for comparison in binding-expression models.

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

Gene expression data

We used RNA-seq datasets from the ENCODE portal under accession ENCSR181ZGR for HepG2 and ENCSR885DVH for K562. We took the counts table of the two replicates and took the average TPM as the TPM for a given transcript. For genes with multiple transcription start sites, we identified the transcript with the highest TPM for the purposes of selecting an isoform.

cCRE catalog

The V4 cCRE human dataset was downloaded from ENCODE portal under accession ENCSR800VNX.

ATAC-seq Datasets

ATAC seq peaks were acquired from the ENCODE portal under accession ENCSR291GJU on February 5th 2021.

HOT sites

Using only those DAP datasets (excluding histone marks) which were Preferred (see above), we restricted peaks to the central 100 bp. We merged all peaks and determined the number of DAPs contributing to each merged region. We then identified HOT sites

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

as those which had at least 25% of all factors bound. As we had 680 unique DAPs in our dataset, HOT sites are defined as locations with >=170 unique DAPs bound.

Gkm-svm model construction and use

HOT sites and the ENCODE exclusion list were removed from peaks, and peaks were then sorted by $-\log_{10}$ pvalue. Up to the top 10,000 peaks were then extracted. Null sequences were generated using the nullseq_generate script from gkm-SVM with the parameters -G -x 2 -m 1000 -r 1 (Ghandi et al. 2016). We ran gkmtrain to produce a model of sequence preference with -l 8 -k 5 -d 2. We ran this for all datasets. Once made, we applied these models to the regions of the cCRE catalog to which no DAP ChIP-seq dataset had evidence of association using gkmpredict.

Activity by Contact predictions

HepG2 intact Hi-C dataset ENCSR888DEJ was used to produce Activity-By-Contact (ABC) predictions and was provided to us by Jesse Engreitz.

Hi-C Data Loop Calls

Hi-C loops for the non-cCRE sMAF analyses in **Figure 4** were called with the Juicer pipeline (Durand et al. 2016). 500 bp resolution calls were used.

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

Motif Derivation

For each dataset, we removed all peaks found in HOT sites or the ENCODE exclusion list, sorted the remaining peaks by -$\log_{10}$ qvalue, and determined the number of peaks remaining. If there were at least 1000 peaks remaining, we then restricted peak width to the +/- 50 bp surrounding the peak center. Using the top 500 peaks of the dataset, we used MEME version 5.1.0's meme-chip method to derive motifs using the following arguments: -dna -meme-mod zoops -meme-nmotifs 5 -meme-minw 6 -meme-maxw 50 -spomao-skip -fimo-skip. We then performed several tests to identify those motifs which were sufficiently enriched in our data. Namely:

1. The motif's e-value as determined by MEME must be 0.05 or less.
2. We identify the top 501-1,000 peaks in the dataset, and extract the sequence +/- 150 bp from the center of each peak, and determine the number of times the motif is observed based on FIMO (Grant et al. 2011). We then perform the same method on null-matched sequences, sampling many times to determine a mean and standard deviation of the number of times a motif is observed. We take the *z*-score of the observed number of motifs in the test set, and calculate a p-value based on the standard normal distribution. If the pvalue is <= 0.00001, the motif passes this test.
3. For all peaks past the first 500, using the sequence +/- 150 bp and determine the number of times the motif is observed based on FIMO. For those same peaks, we also extract an additional set of control sequences 150 bp upstream and 150 bp

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

730 downstream of the test regions, and determine the number of times the motif is
731 observed based on FIMO. If the number of test observations is >= 1.25 times the
732 number of control observation and the motif is found in at least 10% of test
733 sequences, the motif passes this test.
734 4. Finally, we determine if the e-value of the motif found in meme-chip's centrimo
735 tests is <=0.05.

736 If the motif passes tests 1, 2, and either 3 or 4, then the motif was considered real for the
737 datasets in question.

738

739 *ABC-HOT relationships with regard to motif placement.*

740 Using ABC predictions, we determined, for each putative enhancer and promoter region,
741 whether the putative enhancer, promoter, both, or neither overlapped with a HOT site.
742 For each category (Both HOT, Promoter HOT, Enhancer HOT, Neither HOT), we then
743 determined, for each factor with a derived motif (see above), which distal-promoter
744 interactions had a peak for that factor in both the putative enhancer and the promoter.
745 For each of those cases, we then determined whether that factor's motif was found within
746 the TF's peak in the putative enhancer or promoter. We identified the fraction of cases in
747 which the TF was in the putative enhancer and the fraction in the promoter, and then
748 determined the ratio by taking the log(fraction_promoter/fraction_enhancer). We then
749 created a distribution plot of these ratios for each of the 4 categories of HOT interactions.
750 Kolmogrov-Smirnov tests were used to identify significant differences among the
751 distributions. When considering all promoters, we restricted to cases of factors that had

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

752 at least 100 instances of shared peaks (i.e. the peak for the TF was found in both the
753 putative enhancer and the promoter of a pair) to avoid concerns of noise overly influencing
754 log-ratio distributions. For restrictions to only CGI and non-CGI promoters, there were too
755 few TFs with 100 shared peaks observations to perform statistical tests, so we used all
756 available TF data.

757

758 *CGI and non-CGI promoters*

759 CGI annotations in promoters were taken from (Illingworth et al. 2010) and coordinates
760 were converted from hg19 to hg38 using the command-line version of the liftOver tool of
761 the UCSC Genome Browser (Hinrichs et al. 2006).

762

763 *Binding-expression models*

764 We identified transcription start sites for genes based on RefSeq transcription start sites
765 acquired April 12, 2019. For genes with multiple transcription start sites, we identified the
766 transcript with the highest expression level (see above) and used that TSS for the gene.
767 We took the +/- 500 bp upstream of the gene's TSS, and determined which TFs showed
768 association with the promoter based on overlap with peaks of the gene. We then
769 constructed a linear model by training on 70% of genes and testing on 30% of genes, with
770 10x cross-validation to determine the ability of the linear model to predict gene
771 expression.

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

Once this was determined, we then wanted to investigate the estimated impact of each TF on gene expression. We first subsampled different numbers of transcription factors to predict expression levels, taking different sample sizes of TFs 500 times, building a model based on 70% of TSS, and testing on the remaining 30%. We found that little improvement was to be found for sample sizes of more than 80TFs (**Supplemental Figure 33**). Therefore, for every transcription factor, we performed 500 samples, subsampling the factor and an additional 79 factors, and built a model of gene expression based upon factor binding. For each subsampled model, we determined the estimate of the factor of interest and whether or not it was found to be significant in that model. This demonstrated the average ability of a factor to determine expression, independent of the other factors included in the model.

MPRA data and processing

Processed MPRA data was acquired from (Agarwal et al. 2023) and are available under ENCODE accessions ENCSR632EPR and ENCSR463IRX at https://www.encodeproject.org/ . To determine which factors were considered bound to the corresponding genomic context, we identified the summit of each peak and expanded to +/-25 bp in each direction. Those peaks which were fully contained by the hg38 genomic coordinates of these elements were considered bound to the genomic region of the element.

Lentiviral MPRA data was produced by following the protocol in Gordon et al 2020, with minor modifications. Oligos (n=138,420, **Supplemental Table 21**) were synthesized by

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

Twist Bioscience, and were composed of 170 nucleotides test element with 15 nucleotides 5' flanking sequence and 15 nucleotides 3' flanking sequence (**Supplemental Table 21**). First-round amplification of the oligo library was performed as described with 5BC-AG-f01 and 5BC-AG-r01, but a separate first-round amplification was performed in parallel with custom primers 5BC-AG-f01B (5'-CTCACTCAGCCTGCATTTCTGCCAGGGCCCGCTCTAGACCTGCAGGTCGGTTCAC GCAATG-3') and 5BC-AG-r01B (5'-GCTTTCGCTTAGCGATGTGTTCACTTTGCACAGTACCGGATTGCCAAGCTGGAAGT CGAGCTTCCATTATATACCCTCTAGTGAGGACCGGATCAACT-3'). These two reactions gave one library with test elements in forward orientation and one library with test elements in reverse orientation. These libraries were combined in equal quantities before second-round PCR amplification. For barcode association sequencing, we modified the read 1 and read 2 sequencing primers for our dual orientation library to (read 1) 5'- CTGCATTTCTGCCAGGGCCCGCTCTAGACCTGCAGG-3' and (read 2) 5'-TGGAAGTCGAGCTTCCATTATATACCCTCTAGTG-3'. All other primers and protocol steps followed the published method. We produced lentivirus in 293FT cells using established protocols and filtered media (0.45 uM) to obtain virus. We seeded HepG2 cells in 6-well plates (1 million cells per well), and 24 hours post-seeding spinfected virus into HepG2 cells using established protocols (2000 RPM, 1.5 hours). After 48 hours incubation, we extracted RNA and DNA from cells using Norgen Total RNA purification kit and Qiagen DNeasy Blood & Tissue kit, respectively; we purified three replicates, using 8 wells per replicate (16 RNA columns per replicate). All sequencing runs were performed on NextSeq high output flowcells. We obtained 151,069,132 reads for barcode

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

association, 454,643,305 total reads for the three RNA libraries, and 167,815,184 total reads for the three DNA libraries (**Supplemental Table 22)**. Signal was calculated by determining the normalized number of reads corresponding to each element in the RNA and the DNA libraries. We observed a mean of 91.74 barcodes per test element. We confirmed a strong concordance among replicates (**Supplemental Figure 34**) and that there was sufficient barcode representation for elements **(Supplemental Figure 35**). Promoter elements were identified by overlapping with GENCODE human promoter annotations (Frankish et al. 2021). For Binding-expression analyses, we used two sequences, H_046 and ENH_HMM_B_1, and inserted either 1, 2, or 5 instances of established motifs for a given factor taken from the JASPAR database (Castro-Mondragon et al. 2022). Location and orientation of inserted motifs were randomized. In the case of multiple motif insertions, minimum spacing was also randomized, but always greater than 5 bp.

PCA Analyses

For exploration the relationship between PCAs of a binding matrix and various genomic features, protocol was followed as in (Partridge et al. 2020). Briefly, we constructed a matrix of DAPs bound across 2 kb genomic bins, restricted to bins with at least 3 DAPs bound, and performed principal components analysis on the resulting matrix using the princomp() function in R.

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

838 *Non-cCRE DAP dimer identification*

839 To identify putative dimers in non-cCRE regions, we restricted peaks of all factors to the
840 summit of the peak +/-50 bp. We then identified all cases of two factors' peaks
841 overlapping by at least 50 bp. Putative sMAF/sMAF dimers were identified by first finding
842 all such overlaps, and then removing any instances of similar overlap with any other
843 factor. sMAF cofactors were identified as listed in (Katsuoka and Yamamoto 2016), and
844 putative sMAF/cofactor dimers were identified by first removing all cases of putative
845 sMAF/sMAF dimers and then identifying all cases where a sMAF peak had at least 50 bp
846 overlap with a cofactor peak. Finally, putative sMAF/other and other/other dimers were
847 identified by first removing all cases of putative sMAF/sMAF and sMAF/cofactor dimers,
848 and then identifying all cases of 50 bp overlap between a sMAF peak and other factor or
849 other/other peaks, respectively.

850 ATAC-seq profiles over putative dimer regions were generated using deepTools (Ramírez
851 et al. 2016). Heatmap in **Figure 4A** was generated using the ComplexHeatmap package
852 (Gu et al. 2016).

853

854 *Repetitive Sequence Identification*

855 Repetitive sequences were acquired from http://genome.ucsc.edu/cgi-bin/hgTables by
856 selecting the "Repeats" group, designating GRCh38/hg38, and selecting the output
857 format as "BED", and saving the resulting bed file.

858

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

Data Access

All raw and processed MPRA and ChIP-seq sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/). under accession numbers GSE235360 and GSE235477, respectively. Code is provided as Supplemental Code and both code and relevant data for the creation of plots are availabe at GitHub (https://github.com/bmoyers/Moyers_et_al_2023_HepG2_TF/).

Competing Interest Statement

We have no competing interests to disclose.

Acknowledgements

We thank Jessie Engreitz for providing ABC model predictions (Fulco et al. 2019) specific to HepG2 data for use in this manuscript, and Jill Moore and Zhiping Weng for providing the V4 cCRE annotations. We additionally thank Sara Cooper, Greg Cooper, and Nick Cochran for helpful conversations and valuable feedback. We also thank the ENCODE Consortium for providing v4 cCRE calls to us for analyses as well as access to ChIP-seq and ATAC-seq data. Special thanks to Cathleen Shaw for assistance in producing figures (**Figure 3C** was made with BioRender). This work was funded by NIH Grant UM1HG009411 to R.M.M. and E.M.M. and by funds from The HudsonAlpha Institute for

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

Biotechnology and by NIH UM1HG009443; the Bren Chair and Caltech Merkin Institute to BJW.

References

Aftab S, Semenec L, Chu JS-C, Chen N. 2008. Identification and characterization of novel human tissue-specific RFX transcription factors. *BMC Evol Biol* **8**: 226.

Agarwal V, Inoue F, Schubach M, Martin BK, Dash PM, Zhang Z, Sohota A, Noble WS, Yardimci GG, Kircher M, et al. 2023. *Massively parallel characterization of transcriptional regulatory elements in three diverse human cell types*. Genomics http://biorxiv.org/lookup/doi/10.1101/2023.03.05.531189 (Accessed March 30, 2023).

Agarwal V, Shendure J. 2020. Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. *Cell Rep* **31**: 107663.

Bailey TL, Johnson J, Grant CE, Noble WS. 2015. The MEME Suite. *Nucleic Acids Res* **43**: W39-49.

Ballas N, Mandel G. 2005. The many faces of REST oversee epigenetic programming of neuronal genes. *Curr Opin Neurobiol* **15**: 500–506.

Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129**: 823–837.

Becker JS, McCarthy RL, Sidoli S, Donahue G, Kaeding KE, He Z, Lin S, Garcia BA, Zaret KS. 2017. Genomic and Proteomic Resolution of Heterochromatin and Its Restriction of Alternate Fate Genes. *Mol Cell* **68**: 1023-1037.e15.

Boija A, Klein IA, Sabari BR, Dall'Agnese A, Coffey EL, Zamudio AV, Li CH, Shrinivas K, Manteiga JC, Hannett NM, et al. 2018. Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell* **175**: 1842-1855.e16.

Caretti G, Cocchiarella F, Sidoli C, Villard J, Peretti M, Reith W, Mantovani R. 2000. Dissection of functional NF-Y-RFX cooperative interactions on the MHC class II Ea promoter. *J Mol Biol* **302**: 539–552.

Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Lemma RB, Turchi L, Blanc-Mathieu R, Lucas J, Boddie P, Khan A, Manosalva Pérez N, et al. 2022. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **50**: D165–D173.

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

911 Chai C, Xie Z, Grotewold E. 2011. SELEX (Systematic Evolution of Ligands by EXponential
912 Enrichment), as a powerful tool for deciphering the protein-DNA interaction space.
913 *Methods Mol Biol* **754**: 249–258.

914 Chen R, Li X, Yang Y, Song X, Wang C, Qiao D. 2022. Prediction of protein-protein interaction
915 sites in intrinsically disordered proteins. *Front Mol Biosci* **9**: 985022.

916 Chen RA-J, Stempor P, Down TA, Zeiser E, Feuer SK, Ahringer J. 2014. Extreme HOT regions
917 are CpG-dense promoters in C. elegans and humans. *Genome Res* **24**: 1138–1146.

918 Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato
919 MA, Frampton GM, Sharp PA, et al. 2010. Histone H3K27ac separates active from
920 poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**:
921 21931–21936.

922 Di Lello P, Jenkins LMM, Jones TN, Nguyen BD, Hara T, Yamaguchi H, Dikeakos JD, Appella
923 E, Legault P, Omichinski JG. 2006. Structure of the Tfb1/p53 complex: Insights into the
924 interaction between the p62/Tfb1 subunit of TFIIH and the activation domain of p53. *Mol*
925 *Cell* **22**: 731–740.

926 Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL. 2016. Juicer
927 Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell*
928 *Syst* **3**: 95–98.

929 Fladvad M, Zhou K, Moshref A, Pursglove S, Säfsten P, Sunnerhagen M. 2005. N and C-
930 terminal sub-regions in the c-Myc transactivation region and their joint role in creating
931 versatility in folding and binding. *J Mol Biol* **346**: 175–189.

932 Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, Sisu C, Wright JC,
933 Armstrong J, Barnes I, et al. 2021. GENCODE 2021. *Nucleic Acids Res* **49**: D916–D923.

934 Friling RS, Bensimon A, Tichauer Y, Daniel V. 1990. Xenobiotic-inducible expression of
935 murine glutathione S-transferase Ya subunit gene is controlled by an electrophile-
936 responsive element. *Proc Natl Acad Sci U S A* **87**: 6258–6262.

937 Fulco CP, Munschauer M, Anyoha R, Munson G, Grossman SR, Perez EM, Kane M, Cleary B,
938 Lander ES, Engreitz JM. 2016. Systematic mapping of functional enhancer-promoter
939 connections with CRISPR interference. *Science* **354**: 769–773.

940 Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, Grossman SR, Anyoha
941 R, Doughty BR, Patwardhan TA, et al. 2019. Activity-by-contact model of enhancer-
942 promoter regulation from thousands of CRISPR perturbations. *Nat Genet* **51**: 1664–1669.

943 Ghandi M, Mohammad-Noori M, Ghareghani N, Lee D, Garraway L, Beer MA. 2016.
944 gkmSVM: an R package for gapped-kmer SVM. *Bioinformatics* **32**: 2205–2207.

945 Gordân R, Hartemink AJ, Bulyk ML. 2009. Distinguishing direct versus indirect transcription
946 factor-DNA interactions. *Genome Res* **19**: 2090–2100.

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

947 Gordon MG, Inoue F, Martin B, Schubach M, Agarwal V, Whalen S, Feng S, Zhao J, Ashuach
948 T, Ziffra R, et al. 2020. lentiMPRA and MPRAflow for high-throughput functional
949 characterization of gene regulatory elements. *Nat Protoc* **15**: 2387–2412.

950 Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif.
951 *Bioinformatics* **27**: 1017–1018.

952 Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in
953 multidimensional genomic data. *Bioinformatics* **32**: 2847–2849.

954 Han R, Li L, Ugalde AP, Tal A, Manber Z, Barbera EP, Chiara VD, Elkon R, Agami R. 2018.
955 Functional CRISPR screen identifies AP1-associated enhancer regulating FOXF1 to
956 modulate oncogene-induced senescence. *Genome Biol* **19**: 118.

957 Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey
958 TS, Harte RA, Hsu F, et al. 2006. The UCSC Genome Browser Database: update 2006.
959 *Nucleic Acids Res* **34**: D590-598.

960 Hounkpe BW, Chenou F, de Lima F, De Paula EV. 2021. HRT Atlas v1.0 database: redefining
961 human and mouse housekeeping genes and candidate reference transcripts by mining
962 massive RNA-seq datasets. *Nucleic Acids Res* **49**: D947–D955.

963 Huang Y, Myers SJ, Dingledine R. 1999. Transcriptional repression by REST: recruitment of
964 Sin3A and histone deacetylase to neuronal genes. *Nat Neurosci* **2**: 867–872.

965 Hudaiberdiev S, Ovcharenko I. 2023. *Sequence characteristics and an accurate model of high-*
966 *occupancy target loci in the human genome*. Genomics
967 http://biorxiv.org/lookup/doi/10.1101/2023.02.05.527203 (Accessed April 20, 2023).

968 Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr ARW, James KD, Turner DJ, Smith C,
969 Harrison DJ, Andrews R, Bird AP. 2010. Orphan CpG islands identify numerous
970 conserved promoters in the mammalian genome. *PLoS Genet* **6**: e1001134.

971 Iwama A, Pan J, Zhang P, Reith W, Mach B, Tenen DG, Sun Z. 1999. Dimeric RFX proteins
972 contribute to the activity and lineage specificity of the interleukin-5 receptor alpha
973 promoter through activation and repression domains. *Mol Cell Biol* **19**: 3940–3950.

974 Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-
975 DNA interactions. *Science* **316**: 1497–1502.

976 Katsuoka F, Yamamoto M. 2016. Small Maf proteins (MafF, MafG, MafK): History, structure
977 and function. *Gene* **586**: 197–205.

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

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

980 Klein JC, Agarwal V, Inoue F, Keith A, Martin B, Kircher M, Ahituv N, Shendure J. 2020. A
981 systematic evaluation of the design and context dependencies of massively parallel
982 reporter assays. *Nat Methods* **17**: 1083–1091.

983 Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR,
984 Weirauch MT. 2018. The Human Transcription Factors. *Cell* **172**: 650–665.

985 McCarthy RL, Kaeding KE, Keller SH, Zhong Y, Xu L, Hsieh A, Hou Y, Donahue G, Becker
986 JS, Alberto O, et al. 2021. Diverse heterochromatin-associated proteins repress distinct
987 classes of genes and repetitive elements. *Nat Cell Biol* **23**: 905–914.

988 McEwan IJ, Dahlman-Wright K, Ford J, Wright AP. 1996. Functional interaction of the c-Myc
989 transactivation domain with the TATA binding protein: evidence for an induced fit model
990 of transactivation domain folding. *Biochemistry* **35**: 9584–9593.

991 Meadows SK, Brandsmeier LA, Newberry KM, Betti MJ, Nesmith AS, Mackiewicz M,
992 Partridge EC, Mendenhall EM, Myers RM. 2020. Epitope Tagging ChIP-Seq of DNA
993 Binding Proteins Using CETCh-Seq. *Methods Mol Biol* **2117**: 3–34.

994 Morotomi-Yano K, Yano K, Saito H, Sun Z, Iwama A, Miki Y. 2002. Human regulatory factor
995 X 4 (RFX4) is a testis-specific dimeric DNA-binding protein that cooperates with other
996 human RFX members. *J Biol Chem* **277**: 836–842.

997 Morris OM, Torpey JH, Isaacson RL. 2021. Intrinsically disordered proteins: modes of binding
998 with emphasis on disordered domains. *Open Biol* **11**: 210222.

999 Motohashi H, Katsuoka F, Miyoshi C, Uchimura Y, Saitoh H, Francastel C, Engel JD,
1000 Yamamoto M. 2006. MafG sumoylation is required for active transcriptional repression.
1001 *Mol Cell Biol* **26**: 4652–4663.

1002 Nie Y, Shu C, Sun X. 2020. Cooperative binding of transcription factors in the human genome.
1003 *Genomics* **112**: 3427–3434.

1004 Ooi L, Wood IC. 2007. Chromatin crosstalk in development and disease: lessons from REST.
1005 *Nat Rev Genet* **8**: 544–554.

1006 Partridge EC, Chhetri SB, Prokop JW, Ramaker RC, Jansen CS, Goh S-T, Mackiewicz M,
1007 Newberry KM, Brandsmeier LA, Meadows SK, et al. 2020. Occupancy maps of 208
1008 chromatin-associated proteins in one human cell type. *Nature* **583**: 720–728.

1009 Quaglia F, Mészáros B, Salladini E, Hatos A, Pancsa R, Chemes LB, Pajkos M, Lazar T, Peña-
1010 Díaz S, Santos J, et al. 2022. DisProt in 2022: improved quality and accessibility of
1011 protein intrinsic disorder annotation. *Nucleic Acids Res* **50**: D480–D487.

1012 Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. 2011. A unique
1013 chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**:
1014 279–283.

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

Ramaker RC, Hardigan AA, Goh S-T, Partridge EC, Wold B, Cooper SJ, Myers RM. 2020. Dissecting the regulatory activity and sequence content of loci with exceptional numbers of transcription factor associations. *Genome Res* **30**: 939–950.

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**: W160-165.

Reith W, Herrero-Sanchez C, Kobr M, Silacci P, Berte C, Barras E, Fey S, Mach B. 1990. MHC class II regulatory factor RFX has a novel DNA-binding domain and a functionally independent dimerization domain. *Genes Dev* **4**: 1528–1540.

Reith W, Kobr M, Emery P, Durand B, Siegrist CA, Mach B. 1994. Cooperative binding between factors RFX and X2bp to the X and X2 boxes of MHC class II promoters. *J Biol Chem* **269**: 20020–20025.

Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, et al. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* **4**: 651–657.

Rushmore TH, Morton MR, Pickett CB. 1991. The antioxidant responsive element. Activation by oxidative stress and identification of the DNA consensus sequence required for functional activity. *J Biol Chem* **266**: 11632–11639.

Savic D, Partridge EC, Newberry KM, Smith SB, Meadows SK, Roberts BS, Mackiewicz M, Mendenhall EM, Myers RM. 2015. CETCh-seq: CRISPR epitope tagging ChIP-seq of DNA-binding proteins. *Genome Res* **25**: 1581–1589.

Schreiber J, Bilmes J, Noble WS. 2020. Completing the ENCODE3 compendium yields accurate imputations across a variety of assays and human biosamples. *Genome Biol* **21**: 82.

Shah M, Funnell APW, Quinlan KGR, Crossley M. 2019. Hit and Run Transcriptional Repressors Are Difficult to Catch in the Act. *Bioessays* **41**: e1900041.

Smith RP, Taher L, Patwardhan RP, Kim MJ, Inoue F, Shendure J, Ovcharenko I, Ahituv N. 2013. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat Genet* **45**: 1021–1028.

Sugiaman-Trapman D, Vitezic M, Jouhilahti E-M, Mathelier A, Lauter G, Misra S, Daub CO, Kere J, Swoboda P. 2018. Characterization of the human RFX transcription factor family by regulatory and target gene analysis. *BMC Genomics* **19**: 181.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

1049 The ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N,
1050 Adrian J, Kawli T, Davis CA, Dobin A, et al. 2020. Expanded encyclopaedias of DNA
1051 elements in the human and mouse genomes. *Nature* **583**: 699–710.

1052 Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC,
1053 Stergachis AB, Wang H, Vernot B, et al. 2012. The accessible chromatin landscape of the
1054 human genome. *Nature* **489**: 75–82.

1055 Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. 2009. A census of human
1056 transcription factors: function, expression and evolution. *Nat Rev Genet* **10**: 252–263.

1057 von der Lehr N, Johansson S, Wu S, Bahram F, Castell A, Cetinkaya C, Hydbring P, Weidung I,
1058 Nakayama K, Nakayama KI, et al. 2003. The F-box protein Skp2 participates in c-Myc
1059 proteosomal degradation and acts as a cofactor for c-Myc-regulated transcription. *Mol*
1060 *Cell* **11**: 1189–1200.

1061 Walz S, Lorenzin F, Morton J, Wiese KE, von Eyss B, Herold S, Rycak L, Dumay-Odelot H,
1062 Karim S, Bartkuhn M, et al. 2014. Activation and repression by oncogenic MYC shape
1063 tumour-specific gene expression profiles. *Nature* **511**: 483–487.

1064 Wells M, Tidow H, Rutherford TJ, Markwick P, Jensen MR, Mylonas E, Svergun DI,
1065 Blackledge M, Fersht AR. 2008. Structure of tumor suppressor p53 and its intrinsically
1066 disordered N-terminal transactivation domain. *Proc Natl Acad Sci U S A* **105**: 5762–
1067 5767.

1068 Wickham H. 2009. *Ggplot2: elegant graphics for data analysis*. Springer, New York.

1069 Worsley Hunt R, Wasserman WW. 2014. Non-targeted transcription factors motifs are a
1070 systemic component of ChIP-seq datasets. *Genome Biol* **15**: 412.

1071 Wreczycka K, Franke V, Uyar B, Wurmus R, Bulut S, Tursun B, Akalin A. 2019. HOT or not:
1072 examining the basis of high-occupancy target regions. *Nucleic Acids Res* **47**: 5735–5745.

1073 Yip KY, Cheng C, Bhardwaj N, Brown JB, Leng J, Kundaje A, Rozowsky J, Birney E, Bickel P,
1074 Snyder M, et al. 2012. Classification of human genomic regions based on experimentally
1075 determined binding sites of more than 100 transcription-related factors. *Genome Biol* **13**:
1076 R48.

1077 Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM,
1078 Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*
1079 **9**: R137.

1080 Zhu XS, Linhoff MW, Li G, Chin KC, Maity SN, Ting JP. 2000. Transcriptional scaffold: CIITA
1081 interacts with NF-Y, RFX, and CREB to cause stereospecific regulation of the class II
1082 major histocompatibility complex promoter. *Mol Cell Biol* **20**: 6051–6061.

1083 2010. *R a language and environment for statistical computing: reference index*. R Foundation
1084 for Statistical Computing, Vienna.

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

1085

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure 1. Genomic properties and activity of DAP-bound regions in genomic and reporter contexts. **A.** The majority of cCREs of each type are bound by an assayed DAP. Bars show the number of sites of each cCRE class (x-axis) with at least one DAP association ("bound") and those with none in our dataset ("unbound") when restricted to those overlapping with an ATAC-seq peak in HepG2. PLS = promoter-like signature, pELS = proximal enhancer-like signature, dELS = distal enhancer-like signature, CA-H3K4me3 = chromatin accessible H3K4me3 region, CA-CTCF = chromatin accessible CTCF bound region, CA-TF = chromatin accessible TF bound region, TF = TF bound region lacking chromatin accessibility, CA = chromatin accessibility only. **B**. Promoter elements from locally-performed lentiMPRA experiment require fewer DAPs binding for high activity in lentiMPRA than distal elements. Boxes show MPRA signal (natural log of normalized RNA reads over normalized DNA reads) of promoter elements as a function of binned number of DAPs (x-axis) with a peak in the genomic region. Promoters are defined as elements whose bounds overlapped with a 200 bp region centered on GENCODE TSSs. Distal elements are defined as elements at least 5 kb from annotated TSSs. Positive and Negative control elements are plotted for comparison. In this figure and in panel D, boxes represent 25-75% quartiles with line indicating median, whiskers extend to +/-1.5*IQR (inter-quartile range) past the boxes, and when present points are observations falling outside of this range. Unpaired *t*-tests were used to identify significant differences in the means between distal and promoter element activity in each category. *=0.05, *=0.0001, ***<=2.2 x 10^{-16} **C**. The fraction of distal loci with an ABC connection as a function of binned number of DAPs at a distal element. **D.** Expression of genes genome-wide increases as the number of factors bound and

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

connected distal elements increases. Y-axis: natural log expression distribution of the ABC-supported gene as a function of binned number of DAPs at a distal element. Unpaired *t*-tests were used to identify significant differences in the means between the expression of a given category compared to expression in the 0 category. *=0.05.

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure 2. Effects of specific TFs on gene expression and in MPRA assays. **A**. Observed (x-axis) versus predicted (y-axis) natural log of gene expression as measured by transcripts per million. A linear model was constructed based on binding of TFs at a gene's TSS +/- 500 bp. Training and testing were performed on a 70/30 split of all genes. Pearson's correlation=0.77, $p<=2.2 \times 10^{-16}$. Blue line was generated from geom_smooth in the ggplot2 package (Wickham 2009). **B**. Boxplot shows distribution of linear model estimates (x-axis) for select transcription factors (y-axis) from submodels. 500 submodels with unique subsets of randomized TFs (n=79) were constructed for each TF, and estimates for the focal TF were recorded. Colors closer to blue indicate that the focal TF was significant in a higher proportion of submodels, and colors closer to pink indicate that the focal TF was significant in a lower proportion of submodels. For this panel and panels C and D, boxes represent 25-75% quartiles with line indicating median, whiskers extend to +/-1.5*IQR (inter-quartile range) past the boxes, and points are observations falling outside of this range. Unpaired *t*-tests were used to identify significant differences in the means between bound and unbound sequences in each group. *=0.05, **=0.0001, ***<=2. 2×10^{-16}. **C**. Boxes show MPRA signal (natural log of normalized RNA reads over normalized DNA reads) (y-axis) as a function of binned number of DAPs (x-axis) for promoter regions either bound by one of the top factors identified in the linear model as an Activator (blue), Repressor (red), or randomly-selected TF (purple), compared to regions which were not bound by one of those TFs for each group (grey), demonstrating activating, repressing, and uncertain activity for each respective group of TFs, respectively. **D**. Boxes show MPRA signal as in panel C (y-axis) for motifs inserted into enhancer sequences at various intervals (x-

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

axis). A group of candidate activators (x-axis, green line) and candidate repressors (x-axis, red line) were selected and one (green), two (red), or five (brown) motifs were inserted. Control ratio was based on the sequence without any motif insertions. P-values for this figure are available in **Supplemental Table 16**.

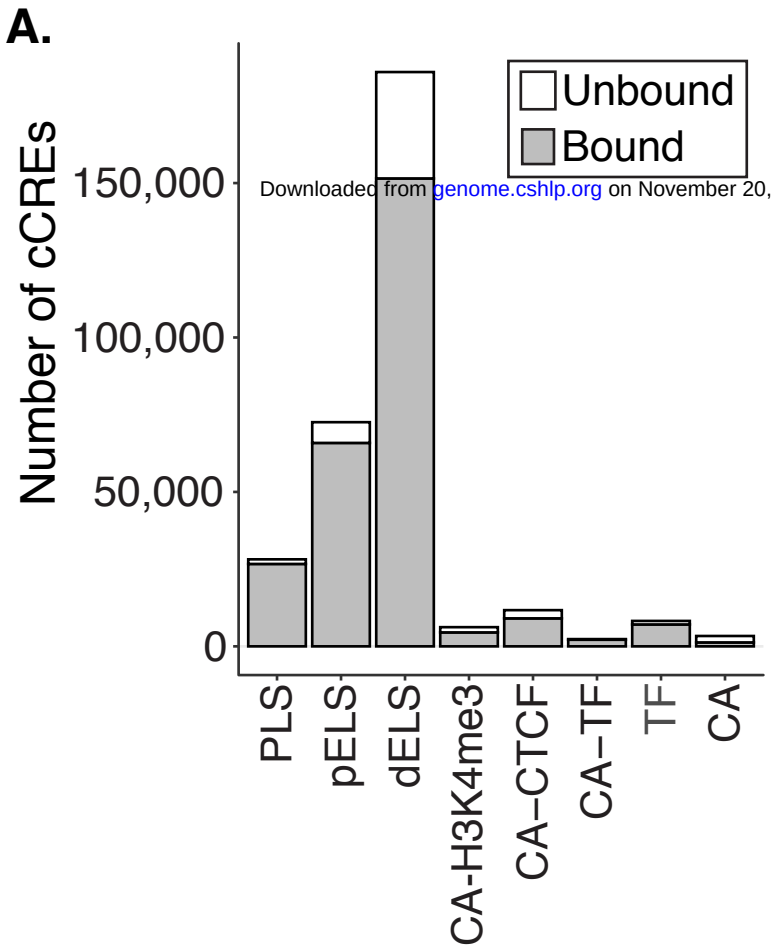
Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

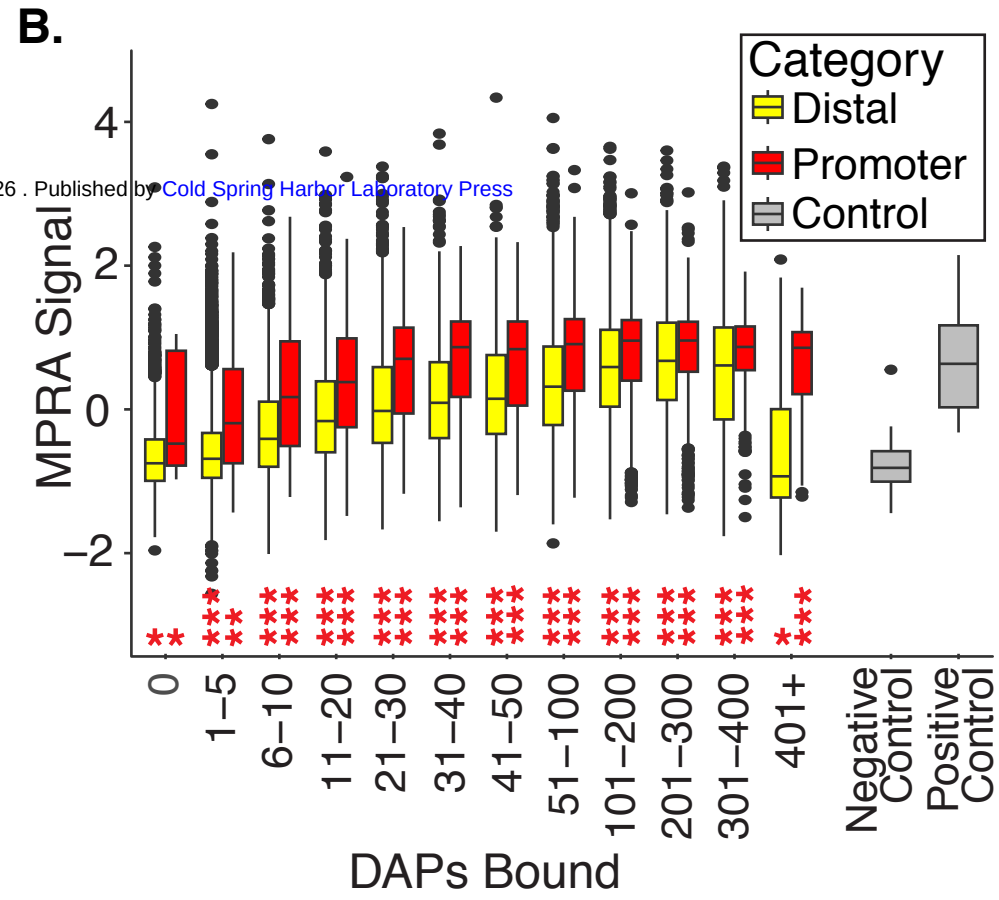
Figure 3. Looping contributes to shared TF binding between putative enhancers (labeled "Enhancer" for plot simplicity) and promoters at HOT sites. **A**. Boxplot shows the number of loops (y-axis) for putative enhancers and promoters when they are HOT or non-HOT (x-axis). Given that at least one loop is present, HOT putative enhancers have more loops than non-HOT putative enhancers, and non-HOT promoters have more loops than HOT promoters (*t*-test *p*-value <=2. 2 x 10^{-16}). Boxes represent 25-75% quartiles with line indicating median, whiskers extend to +/-1.5*IQR (inter-quartile range) past the boxes, and points are observations falling outside of this range. **B**. Density plot of the natural log of fraction promoters with the relevant motif over fraction of putative enhancers with the relevant motif for loops in which a DAP's peak is found in both the putative enhancer and the promoter. Blue indicates that both putative enhancer and promoter are HOT. Red indicates that only the promoter is HOT. Orange indicates that only the putative enhancer is HOT. Grey denotes that neither are HOT. Kolmogorov-Smirnov tests were performed to categorize differences between distributions, and are presented in **Supplemental Table 19**. **C**. Model of motif and TF placement in looping scenarios. Placement of transcription factors represent hypothetical placement in this model based on the occurrence of motifs, and not necessarily an actual example of specific transcription factor binding at specific promoters and putative enhancers based on peak locations observed in datasets.

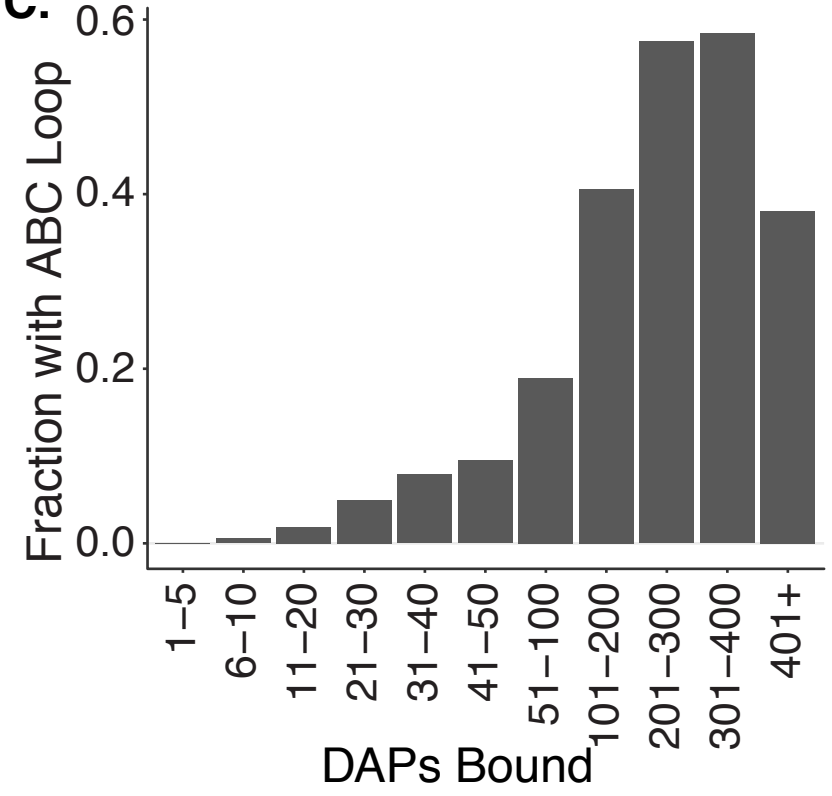
Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

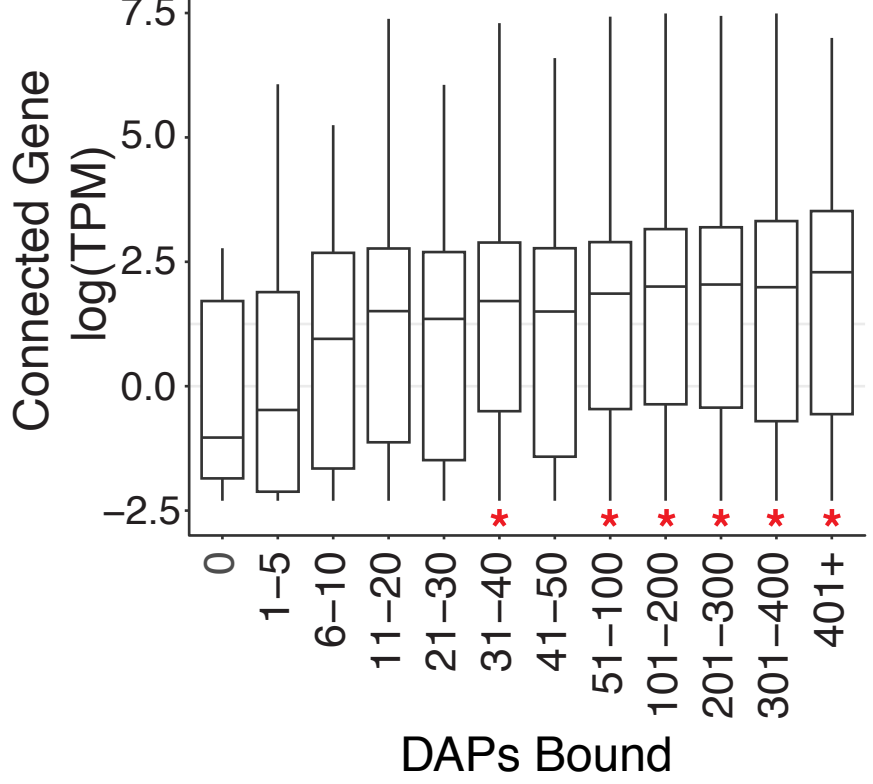
Figure 4. Small MAF proteins bind widely to non-cCRE regions in a sequence-directed manner and show wide-spread influence on nearby gene expression. **A.** Heatmap of non-cCRE regions with binding of various transcription factor and histone marks. MAFF and MAFK show the largest degree of binding of any such mark. **B.** ATAC-seq signal (y-axis) at regions bound by various TF dimer sets +/- 500 bp (x-axis). sMAF/sMAF (green) dimers show markedly lower ATAC-seq signal compared to other groups, and sMAF heterodimers with known activating cofactors (blue) show high openness. sMAF heterodimers with non-cofactors (red) show minimal openness, and heterodimers made of other TFs (orange) show high but broad openness. **C**. Motifs generated via MEME using regions bound by noted dimer sets. sMAF/sMAF dimers (top) show more flanking structure outside of the core MAF motif than do heterodimers of sMAF and non-sMAF proteins. **D.** Boxplot shows the distribution of natural log of expression levels as measured by TPM (y-axis) for genes closest to each putative dimer set (x-axis). Putative sMAF/sMAF dimers show much lower expression profiles than do other groups. Boxes represent 25-75% quartiles with line indicating median, whiskers extend to +/-1.5*IQR (inter-quartile range) past the boxes, and points are observations falling outside of this range. **E.** Bars show the fraction (y-axis) of putative dimer sets (x-axis) which have evidence for binding in a looped region, based on Hi-C data. sMAF/sMAF dimers show a significantly lower rate of binding in looped regions than do other dimer types.

Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

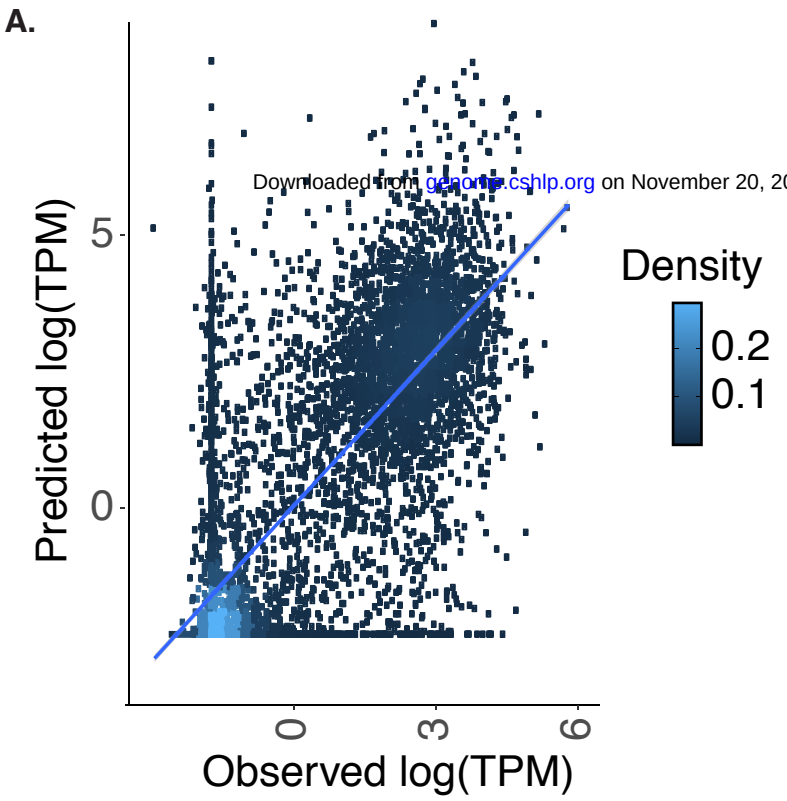


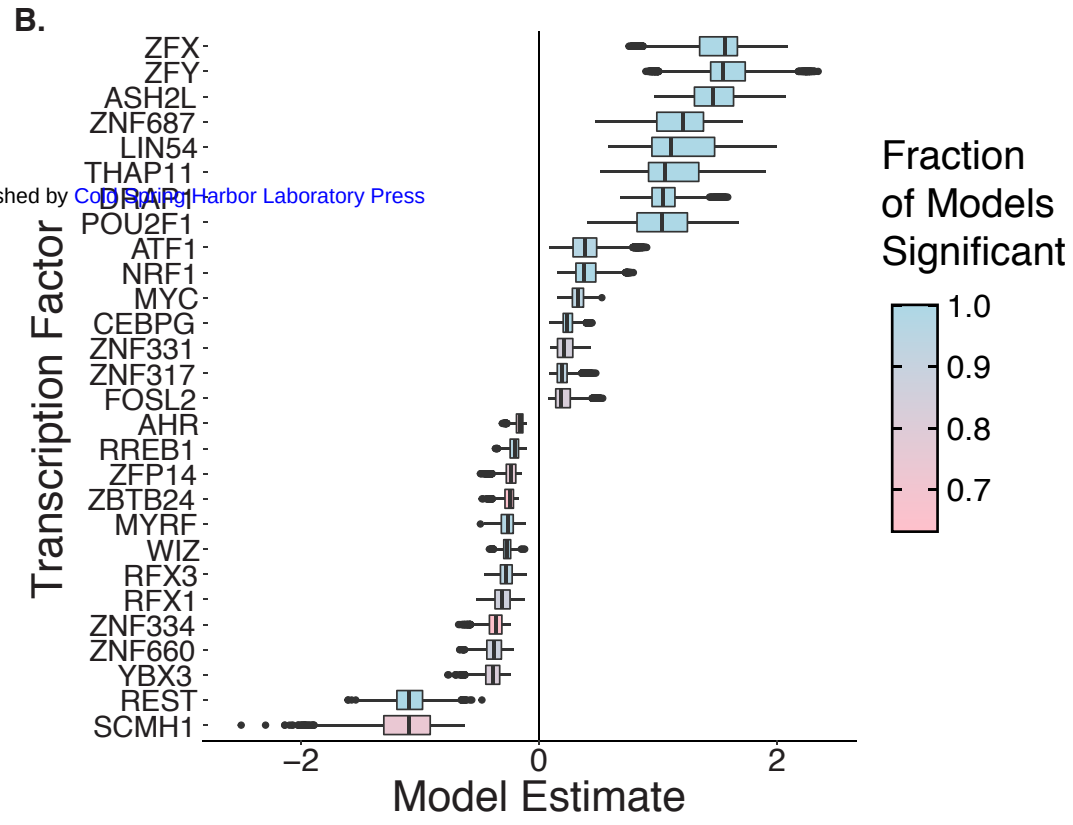


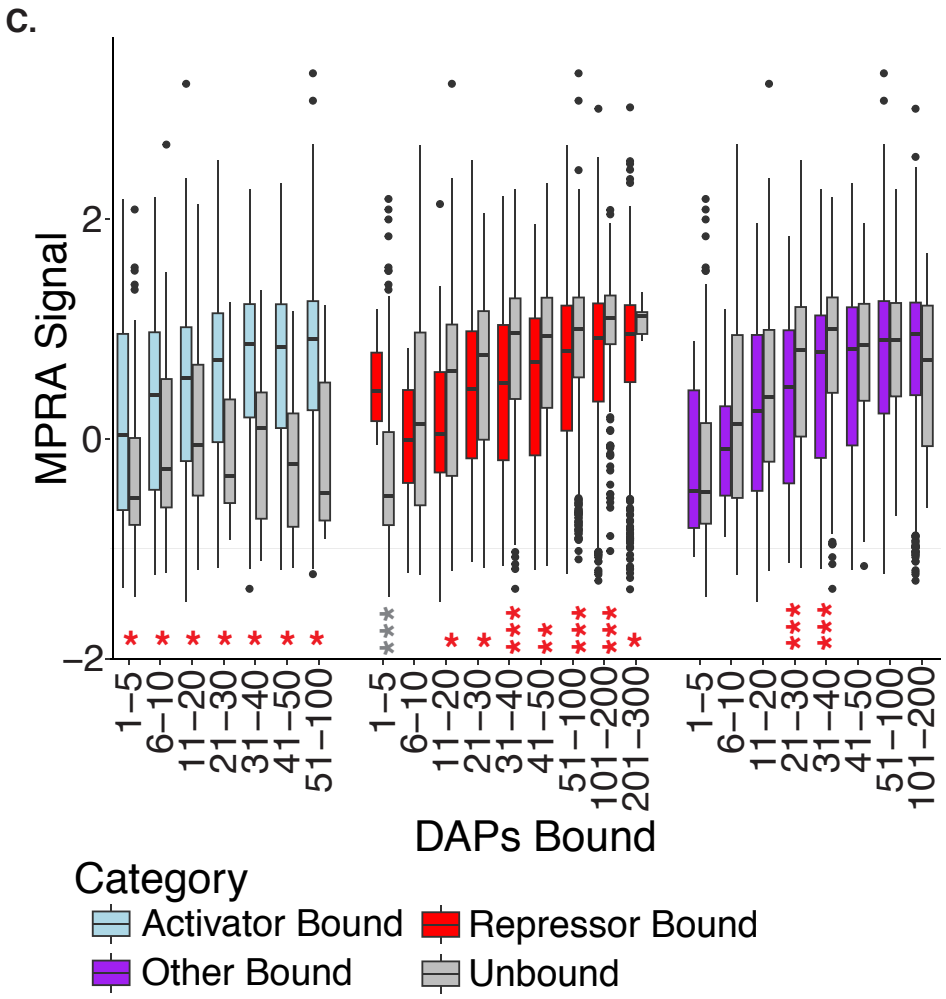


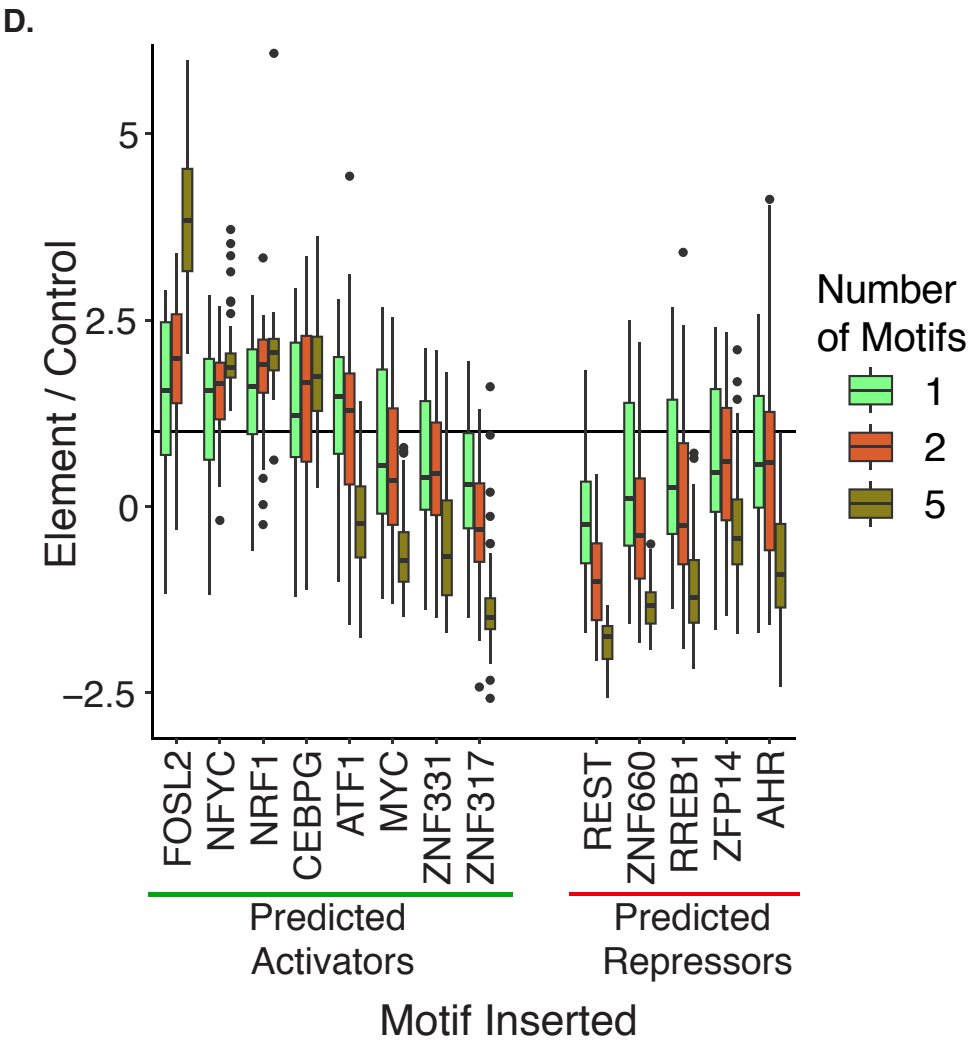


Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press



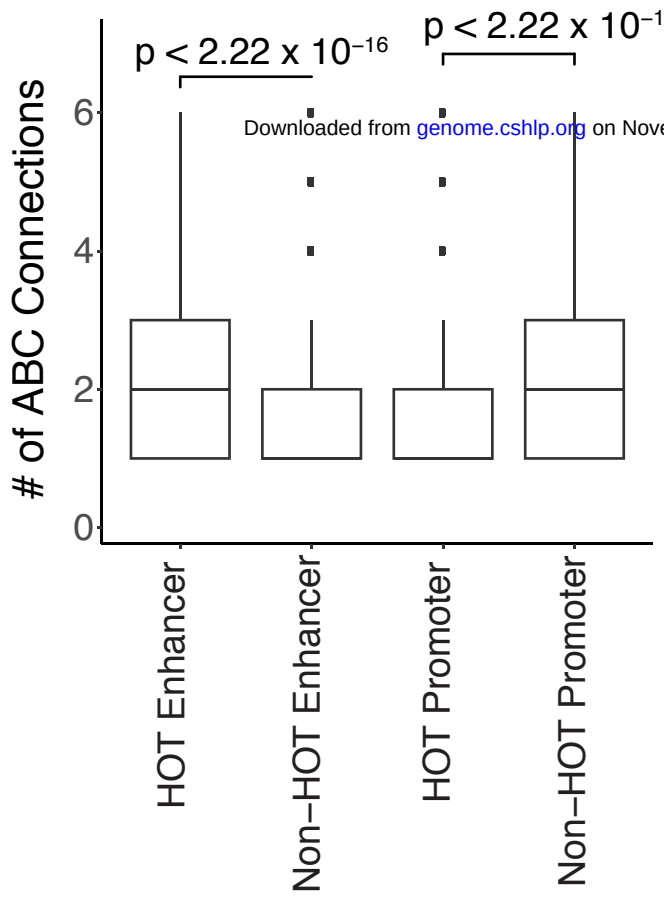




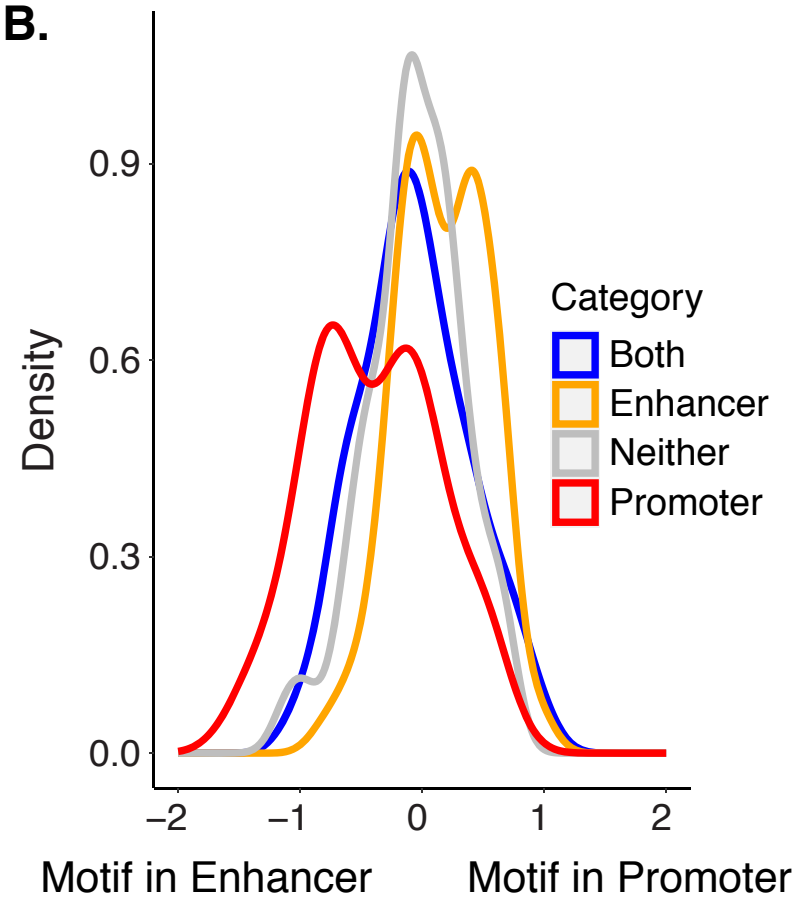


Downloaded from genome.cshlp.org on November 20, 2026 . Published by Cold Spring Harbor Laboratory Press

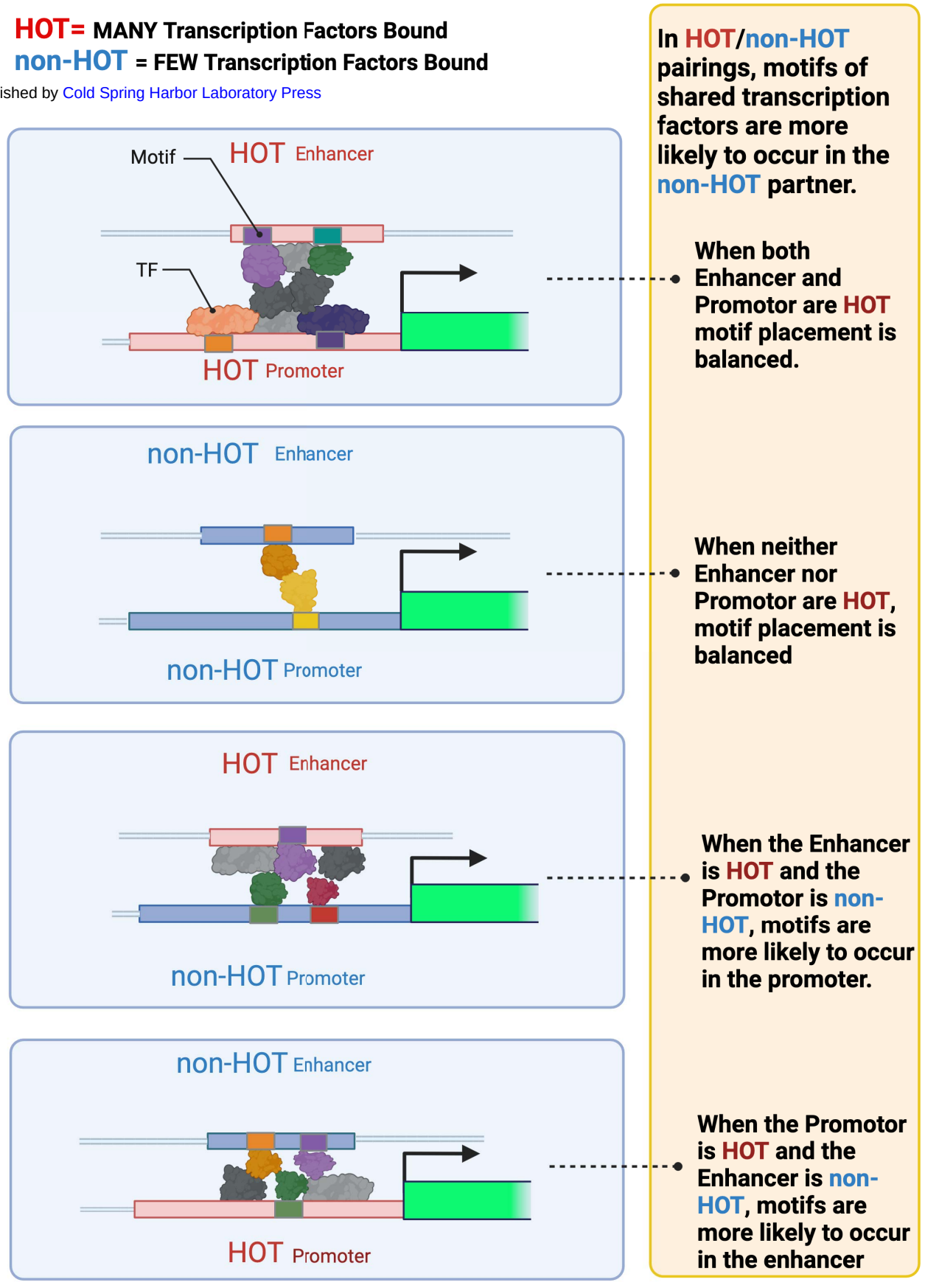
A.



B.



C.



Downloaded from genome.cshlp.org on November 20, 2026 - Published by Cold Spring Harbor Laboratory Press

