Genomic environments scale the activities of diverse core promoters

Clarice K.Y. Hong and Barak A. Cohen

Genome Res. 2022 32: 85-96 originally published online December 27, 2021
Access the most recent version at doi:10.1101/gr.276025.121

Genomic environments scale the activities of diverse core promoters

Clarice K.Y. Hong[1,2] and Barak A. Cohen[1,2]

[1]*The Edison Family Center for Genome Sciences and Systems Biology, School of Medicine, Washington University in St. Louis, Saint Louis, Missouri 63110, USA;* [2]*Department of Genetics, School of Medicine, Washington University in St. Louis, Saint Louis, Missouri 63110, USA*

A classical model of gene regulation is that enhancers provide specificity whereas core promoters provide a modular site for the assembly of the basal transcriptional machinery. However, examples of core promoter specificity have led to an alternate hypothesis in which specificity is achieved by core promoters with different sequence motifs that respond differently to genomic environments containing different enhancers and chromatin landscapes. To distinguish between these models, we measured the activities of hundreds of diverse core promoters in four different genomic locations and, in a complementary experiment, six different core promoters at thousands of locations across the genome. Although genomic locations had large effects on expression, the intrinsic activities of different classes of promoters were preserved across genomic locations, suggesting that core promoters are modular regulatory elements whose activities are independently scaled up or down by different genomic locations. This scaling of promoter activities is nonlinear and depends on the genomic location and the strength of the core promoter. Our results support the classical model of regulation in which diverse core promoter motifs set the intrinsic strengths of core promoters, which are then amplified or dampened by the activities of their genomic environments.

[Supplemental material is available for this article.]

In the classical model of gene regulation, the core promoter serves as a universal platform for the assembly of the basal transcriptional machinery, whereas the specificity of expression is provided by distal enhancers and the chromatin landscape. However, some examples of core promoter specificity seem to challenge this model. Several studies suggest that different core promoters are specific for distinct sets of enhancers (Li and Noll 1994; Merli et al. 1996; Sharpe et al. 1998; Gehrig et al. 2009) and can even trap different enhancers at the same genomic location (Butler and Kadonaga 2001). Some transcription factors also preferentially activate core promoters containing specific motifs (Emami et al. 1995; Juven-Gershon et al. 2006; Parry et al. 2010; Haberle et al. 2019). More recently, a genome-wide massively parallel reporter assay (MPRA) showed that housekeeping and developmental core promoters respond to distinct classes of enhancers (Zabidi et al. 2015; Arnold et al. 2017), arguing that enhancer-promoter compatibility contributes to specificity in the genome. These data have led to an alternate model in which core promoters with different sequence elements respond specifically to the enhancers and chromatin features in distinct genomic environments, which we refer to as the "promoter compatibility" hypothesis. Determining whether the specificity of gene expression is governed by enhancers and chromatin features or by enhancer-promoter compatibility is crucial to understanding a variety of biological processes including cell type–specific regulatory programs and models of gene evolution.

The core promoter is the ~100-bp region around the transcription start site and is responsible for accurately positioning RNA polymerase II and binding general transcription factors (Roy and Singer 2015; Haberle and Stark 2018). It is now known that core promoters are a diverse set of sequences containing specific DNA sequence motifs, also termed core promoter elements or motifs. The most well-known core promoter motif is the TATA box, yet the TATA box is only present in 10%–20% of metazoan core promoters (Yang et al. 2007), suggesting that other motifs might have evolved for different functions. The different motifs have been associated with different functions; for example, the TATA box is often enriched in developmental promoters and show a "sharp" pattern of transcription initiation, whereas promoters with high CpG content tend to contain other less well-characterized motifs and are thought to be associated with a broader pattern of transcription initiation (Lenhard et al. 2012).

A strong prediction of the promoter compatibility hypothesis is that the relative strengths of different core promoters will change at different genomic locations because the distal enhancers and chromatin environments at different locations will be compatible with different types of core promoters (Fig. 1A). Here, we tested the promoter compatibility hypothesis by assaying hundreds of diverse core promoters at four different genomic locations and further extend our results genome-wide by assaying six core promoters across thousands of genomic locations.

Results

Measurement of diverse core promoter activities at different genomic locations

We first created a library in which diverse core promoters drive the expression of an mScarlet reporter gene. The library contains 676 133-bp core promoters spanning a variety of promoter features

Corresponding author: cohen@wustl.edu
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.276025.121.

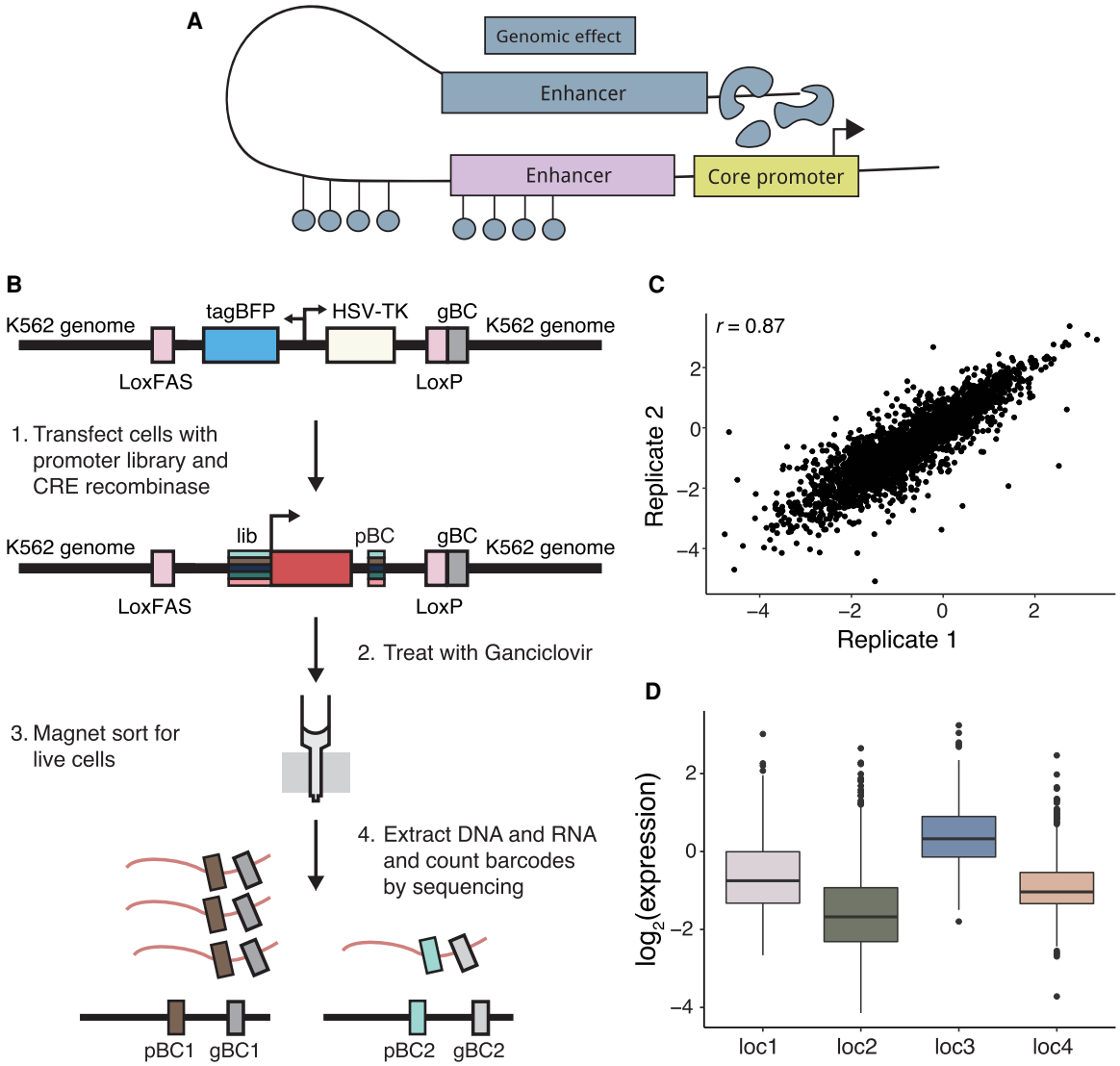

Figure 1. Measurements of a core promoter library at four genomic locations by patchMPRA. (*A*) Schematic of gene regulation by the core promoter, adjacent *cis*-regulatory sequences, and the genomic environment. (*B*) Schematic of patchMPRA method (see Methods for details). tagBFP: blue fluorescent protein; HSV-TK: herpes simplex virus thymidine kinase; gBC: genomic barcode; pBC: promoter barcode. (*C*) Reproducibility of core promoter measurements from independent patchMPRA transfections. (*D*) The expression of all core promoters in the library at each genomic location.

from Haberle et al. (2019). The core promoters are derived from endogenous promoter sequences and include the most common mammalian core promoter motifs (TATA, DPE, and TCT), CpG islands, and housekeeping (hk) and developmental (dev) promoters that do not contain any known core promoter motifs (Supplemental Table S1; Supplemental Data S1). To provide redundancy in the measurements, we included 10 copies of each individual core promoter in the library, each with a unique barcode (promoter BC; pBC) in the 3′ UTR. Because basal expression of the core promoters was expected to be weak, we included a common proximal enhancer directly upstream of the core promoters to boost expression (Methods).

Using parallel targeting of chromosome positions by MPRA (patchMPRA), we measured the expression of the core promoter library in parallel at four genomic locations previously shown to have diverse expression levels and chromatin marks in K562 cell lines (Supplemental Table S2; Supplemental Fig. S1; Maricque et al. 2019). Each cell line contains a single "landing pad" at a different genomic location. Each landing pad has a unique genomic barcode (gBC) indicating its location in the genome and a pair of asymmetric Lox sites to facilitate site-specific recombination of the library. We pooled the four landing pad lines and integrated the library into the cells by cotransfection with CRE recombinase (Maricque et al. 2019). When a library member recombines into a landing pad, it produces a transcript with two unique barcodes in its 3′ UTR; a pBC specifying the core promoter and a gBC indicating its genomic location. By tabulating the pBC-gBC pairs in mRNAs from the pool, we obtained expression measurements for every core promoter at each genomic location in parallel (Fig. 1B).

We obtained reliable measurements of every core promoter at all four genomic locations. We recovered 70%–80% of all promoter barcodes and 99% of all promoters at all landing pads

(Supplemental Fig. S2A,B). The three biological replicates showed high reproducibility (average Pearson's $r = 0.87$) (Fig. 1C; Supplemental Fig. S2C,D), and the environments of the landing pads had large effects on library expression that were consistent with previous studies (cf. Fig. 1D and Supplemental Fig. S3A; Maricque et al. 2019), indicating that the genomic environment is not drastically altered by a diverse core promoter library. To ensure that the genomic environment effect is not driven by the expression of nearby promoters, we examined the expression of the endogenous genes that are closest to the landing pads and found that the expression of these genes does not correlate with the average expression in landing pads (Supplemental Fig. S2E). The data allowed us to compare the effects of the four genomic environments on the different classes of core promoters.

The effects of genomic locations on core promoters

The promoter compatibility hypothesis predicts that the same genomic environment will impact different classes of promoters differently. In contrast to this prediction, the genomic effect was similar on all promoter classes: more permissive genomic locations boosted the expression of all promoter classes regardless of their motif composition or their hk or dev designation (Fig. 2A). However, the magnitude of the genomic effect is not the same for all promoter classes. To quantify the contribution of the genomic location and core promoters to gene expression, we performed ANOVA on each class of promoters. In general, genomic locations have a larger effect on dev promoters than hk promoters regardless of their motif composition (Fig. 2B). Thus, we did not distinguish between the motif classes and focused on the hk and dev groupings for downstream analysis.

We next examined whether hk and dev core promoter activities are scaled by different genomic environments. We define scaling as the degree to which core promoter activities correlate between genomic locations. High correlations between genomic locations indicate that the rank order of core promoter activities is preserved across genomic locations. Whereas promoter activities were highly correlated between genomic locations regardless of the class of promoter (Pearson's $r = 0.74$–0.9, Spearman's $\rho = 0.72$–0.88) (Fig. 2C), dev promoters were consistently less correlated than hk promoters (Fig. 2D). Dividing the promoters into classes containing different motifs showed that each class also had substantial differences in correlations between genomic locations (Supplemental Fig. S3B). These results do not depend on the proximal enhancer immediately adjacent to the core promoter used to boost expression because replicate experiments at locations 1–3 without the enhancer yielded similar results (Supplemental Fig. S4A–C). The expression of libraries with and without the proximal enhancer is also largely correlated at locations 1–3 (Supplemental Fig. S4D), which suggests that scaling by different genomic locations does not depend on the proximal enhancer. Taken together, these results suggest that genomic environments scale the activities of all core promoters, but the quantitative extent of scaling can differ between promoter classes.

Intrinsic promoter strength explains differences between promoter classes

One difference between hk and dev promoters in our library is that they have different mean levels of expression—hk promoters are consistently stronger than dev promoters at all genomic locations (Fig. 2A; Supplemental Fig. S5A). Thus, any differences between hk

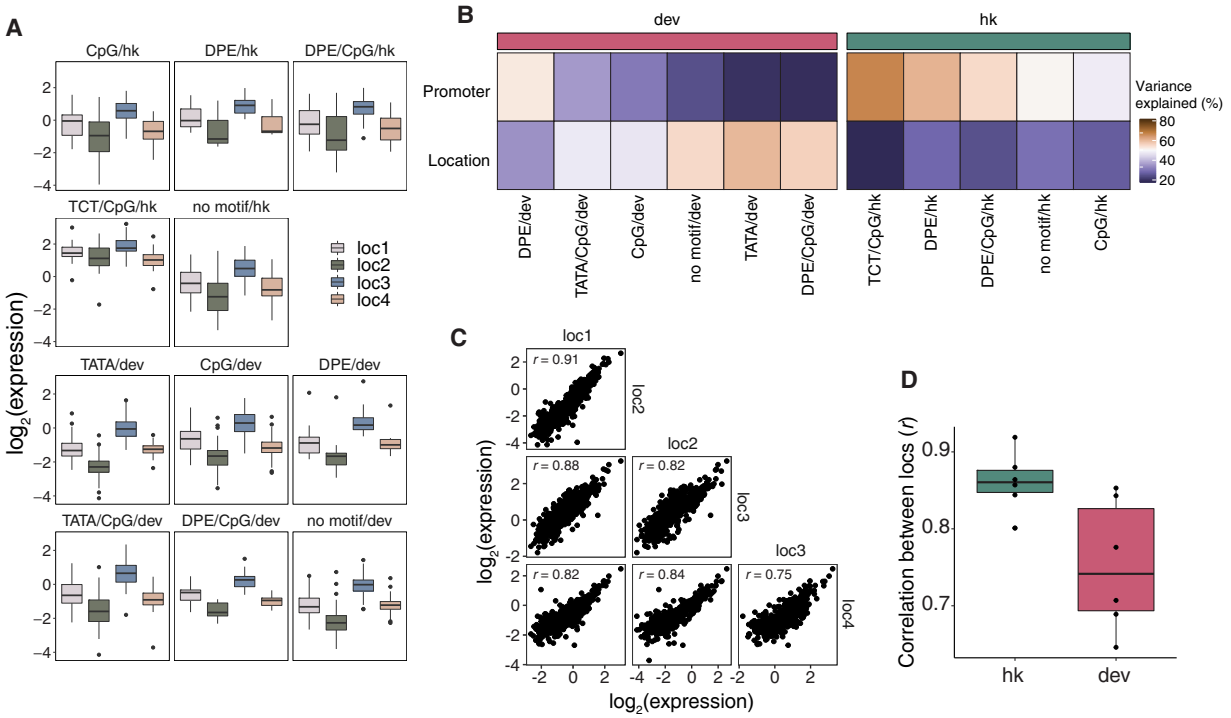

Figure 2. Effects of genomic locations on core promoter activity. (*A*) Expression of each class of core promoter motifs at each genomic location. (*B*) Amount of variance explained by core promoter and genomic location, respectively, using linear models fit on each class of core promoters separately. (*C*) Pairwise correlations (Pearson's *r*) of core promoter activity between the different genomic locations. (*D*) All pairwise correlations (Pearson's *r*) between genomic locations for hk and dev core promoters.

and dev promoters might be confounded by their difference in strength. To test if strength explains the differences between hk and dev promoters, we divided all core promoters into strong or weak bins based on their strengths and sampled equal numbers of hk and dev promoters within each bin to avoid confounding the results by hk/dev class. Plotting the effect of genomic position on strong and weak promoters showed that the direction of the effect was the same but that there were larger differences between genomic locations for weak promoters (Fig. 3A). We quantified the contributions of genomic locations and promoters within strong and weak bins, respectively, and found that the genomic environment has a larger impact on weak promoters compared to strong promoters (Fig. 3B). For strong promoters, genomic environments and core promoters contribute almost equally to gene expression (~33% and ~47%, respectively), but for weak promoters, genomic environments contribute ~61%, whereas core promoters contribute only ~12%. Weak promoters are also consistently less correlated than strong promoters (Fig. 3C). Again, assaying the library without an upstream proximal enhancer at locations 1–3 showed similar results (Supplemental Fig. S5B,C). Finally, we sampled sets of hk and dev promoters with similar average strengths (Supplemental Fig. S5D–F) and compared their correlations across genomic locations. In the strong and intermediate strength subsets, correlations across genomic locations are comparable between hk and dev promoters (Supplemental Fig. S5G,H). In the weaker subset of promoters, the correlations of hk promoters are weaker than the dev promoters, a result driven by low correlations in loc4, which might suggest that there are specific interactions between this subset of hk promoters and loc4 (Supplemental Fig. S5I). Thus, there may be additional interactions between weak promoters and other genomic locations that fall below our threshold of detection. Our data cannot rule out the possibility of extensive interactions between weak promoters and specific genomic locations. However, our modeling also shows that these specific interactions are small relative to the independent effects of genomic locations and core promoters (see Supplemental Fig. S6). The differences in how genomic locations scale the activities of each core promoter subclass are also largely explained by the average strength of each promoter class (Supplemental Fig. S5J). These data show that the observed differences between different promoter classes is a consequence of promoter strength rather than a feature of the hk/dev distinction, indicating that the strength of a promoter is a key determinant of its interactions with the genomic environment.

To further probe how much promoter strength contributes to interactions with the genomic environment, we divided all the promoters into the four bins of expression levels based on loc2. We then fit models with and without interactions between genomic environment and expression bin. As expected, a simple linear model without accounting for interactions explains ~80% of the variance, and the addition of the interaction explains an additional 5% (Supplemental Fig. S6A). To test whether there are any specific interactions between landing pads and individual promoters, we added an additional interaction term between the landing pads and promoters. However, this interaction only explains 3% of the residual variance (Supplemental Fig. S6A), suggesting there are no specific interactions between landing pads and promoters with large effects.

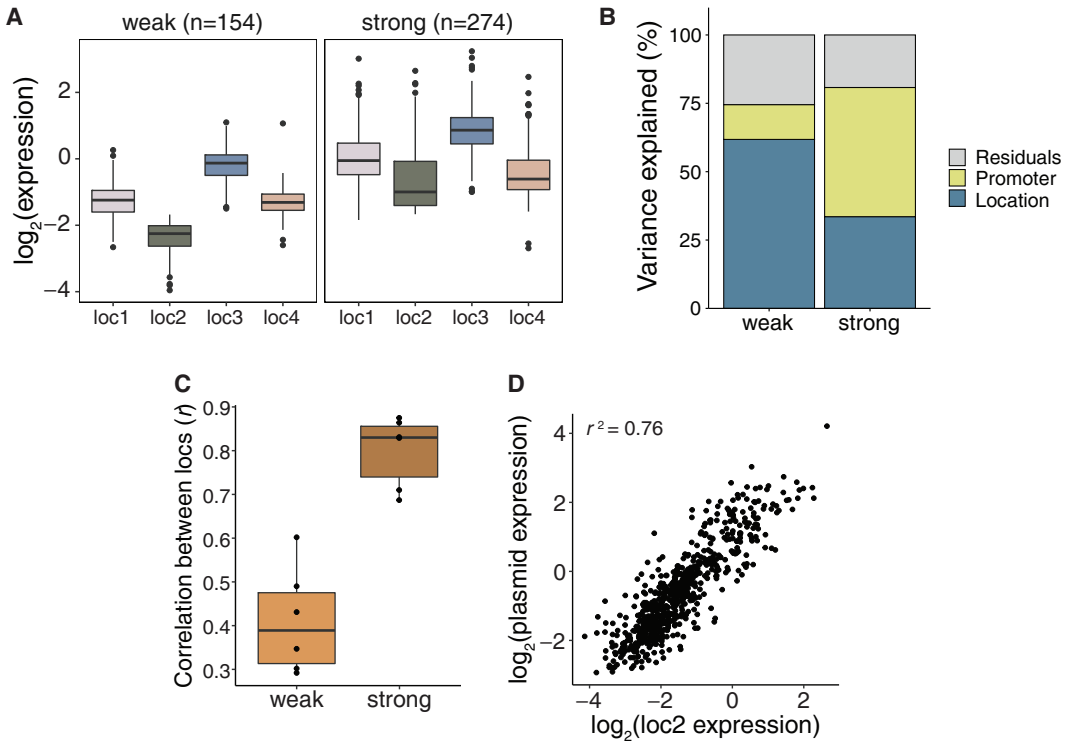

Figure 3. Intrinsic promoter strength explains differences between promoter classes. (*A*) The effect of genomic location on the expression of weak and strong core promoters. (*B*) Amount of variance explained by core promoters and genomic locations, respectively, using linear models fit on weak and strong promoters separately. (*C*) All pairwise correlations (Pearson's *r*) between genomic locations for weak and strong core promoters. (*D*) Correlation (Pearson's *r*) between promoter activity measured on plasmids and promoter activity at loc2.

We also tested whether specific genomic environments restricted the expression of specific promoters, causing them to "drop out" of our analyses at certain landing pad locations. Such interactions would be rare because we recovered measurements for 99% of our promoters at each landing pad (Supplemental Fig. S2B). For promoters that did drop out of our analyses, we found that most were only lost from a single landing pad, with only 15 promoters being lost from two or more landing pads (Supplemental Fig. S6B). The loss of these promoters can be explained by their low abundance in the cloned plasmid library, where promoters lost from more landing pads are present at lower levels in the library (Supplemental Fig. S6C). Thus, there does not appear to be any systematic or specific restriction of expression by different genomic environments.

Given the importance of the interaction between promoter strength and genomic location, we next asked if core promoter strengths, as measured in the genome, reflect the promoters' intrinsic activities. If this is true, then the measurements in the genome should correlate with measurements on plasmids, assuming that plasmids represent a neutral environment that reflects the intrinsic activities of core promoters. Thus, we performed an episomal MPRA on the core promoter library in K562 cells. The plasmid measurements are well-correlated with expression at each genomic location (Pearson's $r^2 = 0.59$–0.76) (Fig. 3D; Supplemental Fig. S7A), indicating that the relative intrinsic activities of core promoters are preserved when integrated into the genome. We were also able to predict activity in the genome using activity on plasmids (adjusted $r^2 = 0.72$) (Supplemental Fig. S7B). These results demonstrate that genomic locations scale the intrinsic activities of strong and weak promoters to different extents, suggesting that the main role of diverse core promoter motifs is to set the intrinsic strength of the promoter rather than direct specific interactions with the genomic environment.

Core promoter scaling is a genome-wide phenomenon

To extend the results we observed at four genomic locations to diverse locations across the genome, we selected six core promoters (three hk and three dev) spanning a range of expression levels and motifs within each class (Fig. 4A; Supplemental Table S3). We note that, while hk1 drives constitutive expression of the ribosomal gene *RPS27* across multiple cell types, this promoter contains a TATA box, a motif that is generally enriched in developmental promoters (Duan et al. 2002; Zabidi et al. 2015). We then measured their activities genome-wide using the Thousands of Reporters Integrated in Parallel (TRIP) assay (Akhtar et al. 2013) in K562 cells (Supplemental Fig. S8A). Each core promoter was cloned upstream of a reporter gene with a unique promoter barcode (pBC) in its 3′ UTR into a PiggyBac transposon vector for random delivery into the genome. No upstream proximal enhancer was included in these constructs. TRIP libraries were generated by incorporating $>10^5$ random barcodes (tBCs) onto each core promoter reporter plasmid. After transposition, every genomic integration contains a unique pBC and tBC pair specifying the identity of the core promoter and its location in the genome, respectively. This double barcoding strategy allowed us to pool promoter libraries into a single TRIP experiment. The replicates were highly correlated (Pearson's $r^2 = 0.96$) (Supplemental Fig. S8B). In total, we mapped 41,083 unique integrations in the genome, ranging between 6078 and 7418 integrations per promoter (Supplemental Table S3; Supplemental Data S2).

Genomic positions have large effects on core promoter activities, with expression ranging more than 1000-fold for the same promoter across genomic locations (Fig. 4B). However, even with these large effects of genomic location, the rank order of promoter strengths is preserved across locations and correlates with mean expression in the landing pads (Fig. 4C; Supplemental Fig. S8C), which suggests that the effect of different genomic locations is to scale intrinsic promoter activities. Because PiggyBac is known to have a preference for H3K27-acetylated regions (Yoshida et al. 2017; Moudgil et al. 2020), we grouped the integrations by their locations into three groups: H3K27ac regions ($n = 14{,}275$), within 50 kb of a H3K27ac region ($n = 21{,}623$), or far away from H3K27ac regions ($n = 5185$) (Supplemental Fig. S8D). Integrations that are far from H3K27ac regions are generally weak, consistent with the idea that these locations are less permissive for expression. However, the rank order of promoters in these regions is the same as integrations in the other locations. Furthermore, integrations in or near H3K27ac regions span the entire >1000-fold dynamic range of our library, suggesting that there is still substantial diversity within H3K27ac regions. Taken together, these data indicate that core promoters are scaled by diverse genomic environments.

To compare different promoters in the same genomic environment, we identified 1278 genomic regions in which at least four of the six promoters had integrated <5 kb from each other (in separate cells) (Supplemental Data S3). These genomic regions are located across the entire genome and span diverse ChromHMM annotations (Supplemental Fig. S9A,B; Ernst and Kellis 2010; Ernst et al. 2011). Across these locations, expression consistently increases from the weakest (dev2) to strongest (hk1) promoter (Fig. 4D,E), showing that the relative strengths of core promoters are preserved across >1000 genomic locations with 1000-fold differences in expression. The expression of the promoters in each region also correlates well with expression in the landing pads, with >60% of locations having $r > 0.7$ (Supplemental Fig. S9C), and a linear model assuming independent effects of genomic region and promoter explains ~54% of the variance in the data (Supplemental Fig. S9D). Thus, measurements of integrated promoters across diverse genomic positions demonstrate that core promoter scaling is a genome-wide phenomenon.

Nonlinear scaling of core promoters by genomic environments

We next explored the relationship between core promoter strength and genomic environments in the TRIP data. We ranked the TRIP genomic regions based on mean promoter expression and plotted the expression of the promoters (Fig. 5A). As expected, all six core promoters increase expression as genomic environments become more permissive. However, the rates at which their expression changes are different for strong and weak promoters. In less permissive regions, strong promoters increase rapidly but then level off in more permissive regions. In contrast, weak promoters increase slowly in less permissive regions and then sharply in more permissive regions. To ensure that hk1 expression in activating regions is not saturated due to the dynamic range limits of TRIP, we tested hk1 with an upstream enhancer, and it was expressed at still higher levels (Supplemental Fig. S9E). Thus, promoters with different strengths do not respond to differences in genomic environments in the same way.

In agreement with our results from the patchMPRA experiment above, the curves in Figure 5A separate by the intrinsic strength of the core promoters and not by their hk or dev identity. To illustrate this point, we calculated the correlations between the

A

Name	Gene	Motifs/Features
hk1	*RPS27*	TATA, CpG, TCT
hk2	*SGCE*	CpG
hk3	*SLC66A1*	DPE, Motif5
dev1	*NLRP1*	DPE
dev2	*CFAP298*	No known motif
dev3	*RNF223*	TATA

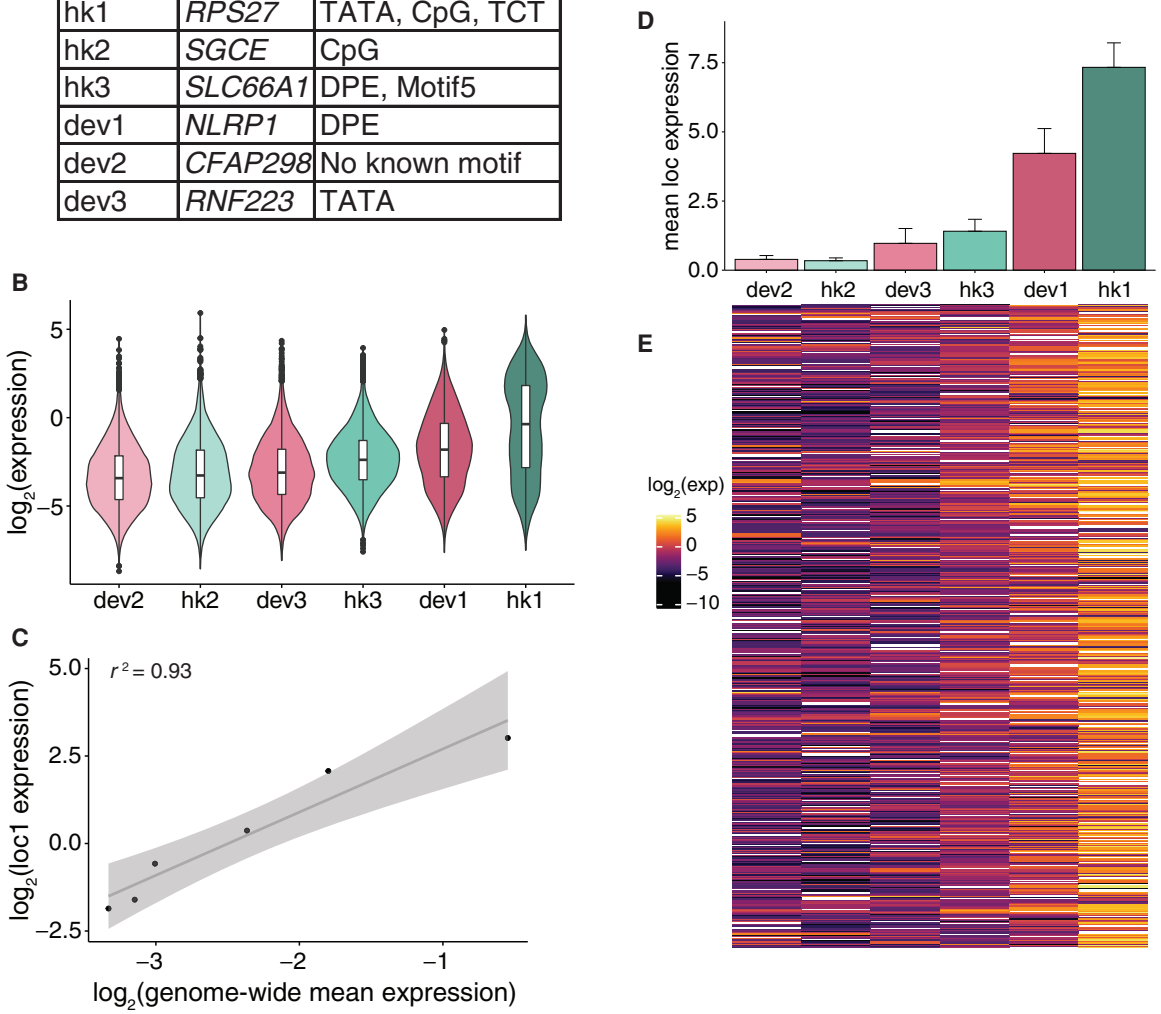

Figure 4. Core promoter scaling is a genome-wide phenomenon. (*A*) Features of core promoters selected for TRIP experiments. (*B*) Expression of each core promoter across all mapped genomic locations sorted by increasing means measured by TRIP. Blue-green denotes hk promoters and pink denotes dev promoters. (*C*) Correlation (Pearson's *r*) between mean expression of each core promoter genome-wide (measured by TRIP) and loc1. The shaded region around the fitted line represents the 95% confidence interval. (*D*) Mean expression of each core promoter from four genomic locations as measured by patchMPRA. Error bars represent the SEM. (*E*) Heat map of expression of each core promoter (column) at each genomic region (row) that has ≥4 different integrated promoters. White boxes represent NA values.

curves of each promoter and show that the promoters cluster based on their intrinsic strengths, with the stronger promoters (dev1 and hk1) in one cluster and the others in another (Supplemental Fig. S10A). Integrations within 5 kb of endogenous hk or dev promoters in K562 also showed no preference for hk or dev promoters, respectively (Supplemental Fig. S10B). This result again highlights that a promoter's strength, not class, determines its interaction with genomic environments.

The differences in the way core promoters respond to genomic environments in Figure 5A also demonstrate that genomic environments do not scale promoter activities linearly. Although the rank order of core promoters is preserved across the genome, the fold change between strong and weak core promoters is different in different parts of the genome. To quantify the effects of different genomic environments, we identified three clusters of TRIP genomic regions that appear to have different levels of activity (Fig. 5B). While the clusters are defined by their average differences in core promoter expression, the extent of scaling is also different in each cluster (Fig. 5C). This difference in scaling is due to differences in the contributions of genomic location and promoter effects in the three clusters. In regions of the genome with low activity, genomic location contributes ~23% to gene expression whereas core promoters contribute only ~12%. In the cluster with high activity, genomic location also contributes about ~24%, but core promoters contribute ~31%, suggesting that differences in expression at these locations depend more on core promoter strength. In the cluster with medium activity, the core promoter contribution is much larger, explaining ~64% of the variance compared to ~16% by genomic location (Fig. 5D). Thus, the strength of the genomic environment determines how much it will contribute to

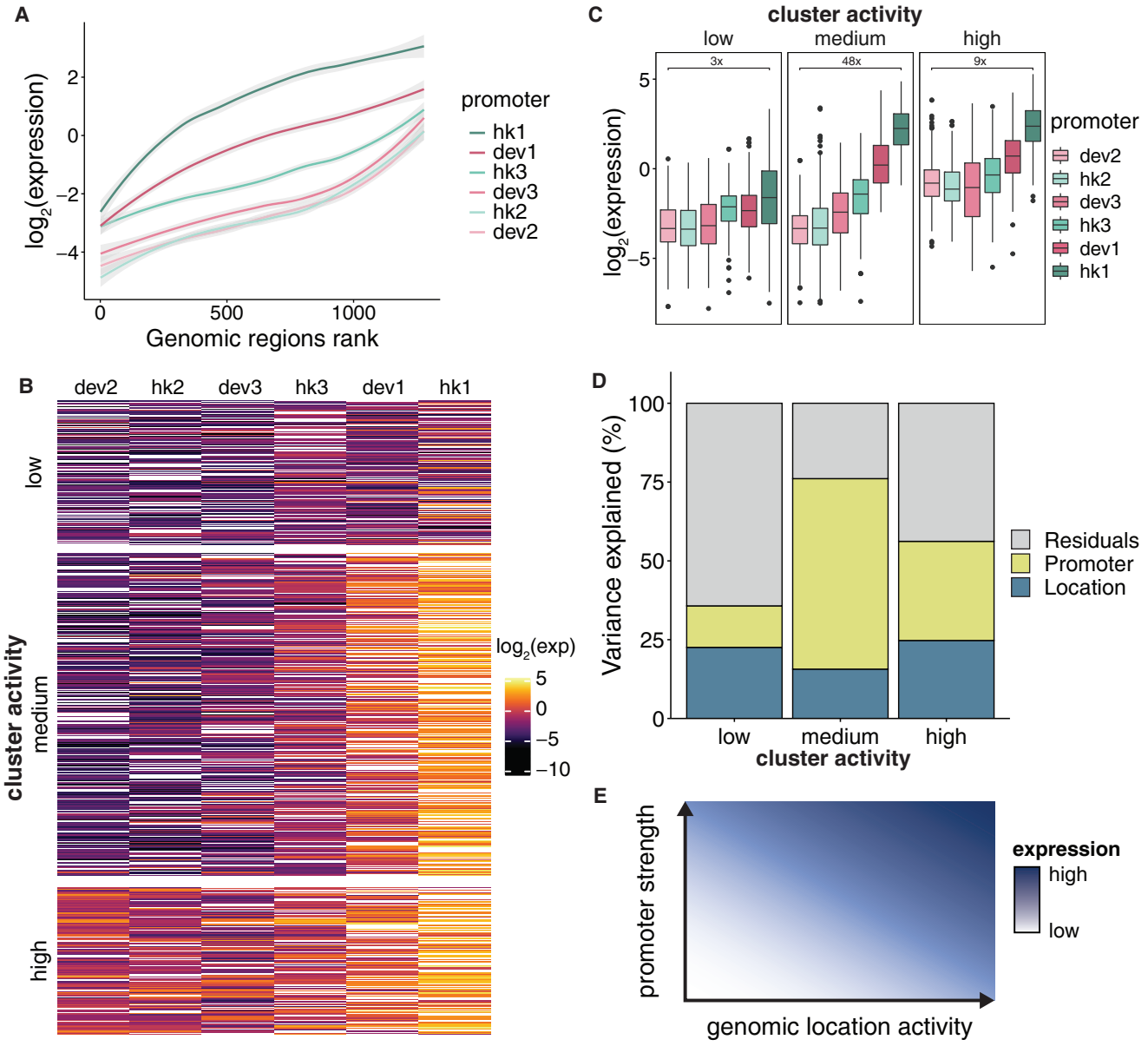

Figure 5. Nonlinear scaling of core promoters by genomic environments. (*A*) Genomic regions defined by TRIP were sorted by the mean expression of the promoters in each region. The shaded region around the fitted line represents the 95% confidence interval. (*B*) Heat map in Figure 4E split into three clusters by *k*-means clustering. Clusters were assigned different activity levels based on the overall expression in the cluster. (*C*) Expression of core promoters in each genomic cluster. (*D*) Amount of variance explained by core promoters and genomic locations, respectively, using linear models fit on each genomic cluster, respectively. (*E*) Summary model of the relationship between core promoter strength and genomic environment activity.

gene expression, resulting in nonlinear scaling of promoter activities across the genome. This is in contrast with the linear scaling we previously observed using a library of proximal enhancers (Maricque et al. 2019), suggesting that core promoters and proximal enhancers may interact with the genomic environment in different ways.

Genomic clusters have different chromatin states and sequence features

Finally, we asked what features of each cluster distinguish them from each other by overlapping our genomic regions with existing epigenomic data sets and sequence features. Previous studies have shown that reporter genes integrated into the genome tend to take on the chromatin state of the integration site (Chen et al. 2013; Corrales et al. 2017). In general, cluster activity is correlated with chromatin marks associated with active transcription (H3K27ac, H3K4me3) and transcriptional activity (PolII binding, CAGE-seq) (Fig. 6A–C; Supplemental Fig. S11A), while accessible chromatin (ATAC) and CpG methylation do not separate the clusters (Supplemental Fig. S11B,C). This suggests that the three clusters are mainly distinguished by their level of transcriptional activity. We also used sequence features to classify the clusters using gapped *k*-mer SVMs comparing two clusters at a time (Ghandi

et al. 2014, 2016). The SVMs performed well, with fivefold cross-validated AUCs ranging from 0.8 to 0.9 (Fig. 6D–F; Supplemental Fig. S11D–F). Scrambling the cluster annotations led to essentially random predictions by the SVM (Supplemental Fig. S11G,H). To further validate the model, we used the trained SVM to predict the cluster type of other TRIP integrations that were not in the 5-kb region analysis. As expected, clusters that were predicted to be more active also showed higher expression (Supplemental Fig. S11I). To identify the motifs that separate the clusters, we performed de novo motif enrichment and identified CG-rich sequences in the more active clusters (Supplemental Fig. S11J,K). Similarly, the CG content of each sequence increases from low to high activity clusters on average (Fig. 6G). Motif enrichment using known TF position weight matrices did not identify any obvious enriched TF motifs, suggesting that the clusters are not defined by any single TF. However, when we scanned each sequence for known TF motifs, we found that sequences in more active clusters have more TF motifs than less active clusters on average (Fig. 6H). This result suggests that the differences between clusters are partially explained by the number of TFs binding in each cluster.

Discussion

Gene expression results from the integration of multiple inputs including the core promoter, chromatin environment, distal enhancers, and the surrounding transcription factor concentrations. Here, we present a framework for dissecting the contributions of core promoters and the surrounding genomic environments to gene expression. Using this framework, we found that the intrinsic activities of core promoters are preserved across diverse genomic locations and are consistent with their activities on plasmids. Contrary to the promoter compatibility hypothesis, hk and dev promoters scale similarly across genomic locations when normalized for differences in strength. These results suggest a general lack of specificity between core promoters and the chromatin landscape/enhancers in their genomic environments, which is consistent with the classical idea of core promoters as passive sequences for the assembly of basal transcriptional machinery. Although promoter compatibility has been observed for specific promoter-genomic environment pairs (Li and Noll 1994; Merli et al. 1996; Ohtsuki et al. 1998; Butler and Kadonaga 2001; Zabidi et al. 2015), our results suggest that such interactions are relatively rare or have smaller effects than the effects of genomic scaling. Our results are also consistent with recent work showing that enhancers and promoters are broadly compatible and combine multiplicatively to control gene expression (Bergman et al. 2021). In this model, sequence-specific or protein-specific interactions between core promoters and genomic environments contribute less to gene expression than the independent effects of core promoters and genomic environments. This model suggests a modular genome compatible with the evolution of gene expression by genome rearrangements (Carroll 2005; Prud'homme et al. 2007). In a modular genome, core promoters will function in new genomic locations without having to evolve the machinery for a new set of specific interactions at each location.

Unlike our previous results with *cis*-regulatory sequences upstream of the core promoter, scaling is not a simple linear combination of genomic position effects and promoter effects (Maricque et al. 2019). In a linear relationship, the genomic environment scales the activity of local promoters such that the rank order and quantitative differences between promoters are always preserved. This occurs when the contribution of genomic and promoter effects remains constant across genomic locations and promoters. Instead, we find that the quantitative differences between strong and weak promoters change in different genomic environments (Fig. 5E), suggesting that genomic environments scale core promoter activities in a nonlinear manner. Such nonlinear scaling is characteristic of the thermodynamics of dose/response curves, which follow a sigmoidal relationship with three phases. At low and high input levels, increasing input has little to no effect on output because the output levels are below detection or saturated, respectively. However, in the linear range, increases in input have large effects on output. We speculate that different core promoter sequence features set the strength of the promoter, causing them to start at different points of the dose/response curve. This in turn determines how the promoter interacts with the genomic environment.

Our data are also consistent with recent simulations showing how promoters starting from different states representing different promoter strengths can have different responses to increasing enhancer contact frequency, giving the appearance of enhancer-promoter specificity (Xiao et al. 2021). This result may also explain the apparent differences between our data and previous results showing enhancer-promoter specificity (Butler and Kadonaga 2001; Zabidi et al. 2015; Arnold et al. 2017). We hypothesize that enhancer-promoter specificity is not governed by biochemical differences in transcription factor usage between hk and dev promoters; instead, the observed differences may be due to differences in intrinsic strengths in the hk and dev promoters used. Alternatively, the differences might be due to technical differences between the episomal MPRAs used previously and the genome-integrated assays used here, or to biological differences between *Drosophila* and human cells. Controlling the intrinsic strengths of different classes of promoters will be important for testing enhancer-promoter specificity. In the future, the nonlinear relationship between promoter strength and genomic position effects will help us to predict gene expression by measuring core promoter strength and genomic environment activity independently.

Methods

Library design

We obtained a set of 6916 core promoter sequences from Haberle et al. (2019) and selected 672 sequences for our library. Each promoter is 133 bp long and centered on the major transcription start site (TSS). We selected the sequences to contain diverse core promoter types and expression patterns (Supplemental Table S1; Supplemental Data S1) using the designations obtained from Haberle et al. (2019). We also included the super core promoter (SCP1), as well as versions of SCP1 with TATA and DPE single and double mutants (Juven-Gershon et al. 2006). A library of oligonucleotides (oligos) encoding the selected core promoters was synthesized by Agilent Technologies through a limited licensing agreement. Each oligo in the library is 200 bp and contains a core promoter, a unique barcode that specifies the identity of the promoter, and flanking sequences for subcloning. An example oligo is shown here:

CCTTACACGGAGTGGATA-SpeI-core promoter-HindIII-NheI-XbaI-12bp barcode-SalI-CATAACTTCGTATAATGT

Each promoter is present 10 times in the library, each time with a different unique barcode, to provide redundancy in the measurements. In total, the oligo pool contains 6760 unique sequences. The barcodes were randomly selected from barcode lists generated by the FREE barcodes software (Hawkins et al. 2018).

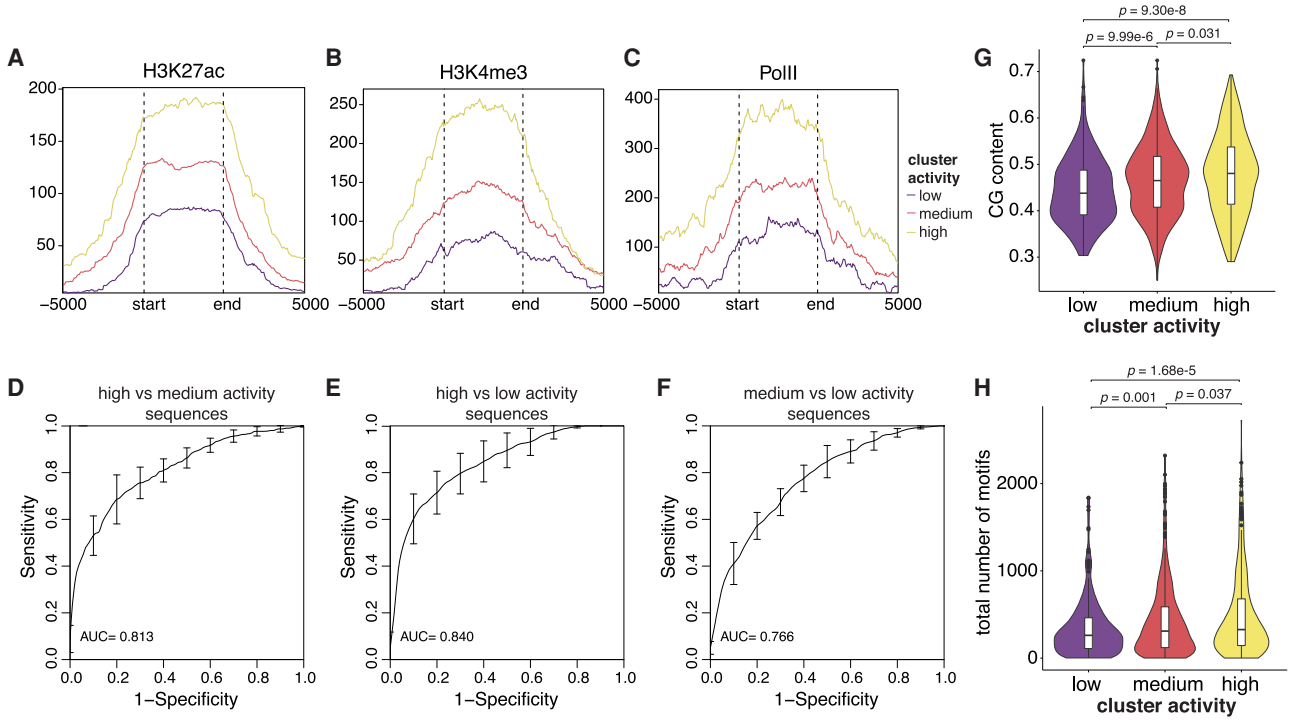

Figure 6. Genomic clusters have different chromatin states and sequence features. (*A*–*C*) Metaplots of H3K27ac, H3K4me3, and PolII levels, respectively, in each genomic cluster. The start and end marks the boundaries of each genomic region, which are determined by the first and last integration in the region. The *x*-axis extends ±5 kb around each genomic region. (*D*–*F*) Performance of gkmSVM used to classify sequences from different genomic clusters. Receiver-operating characteristics (ROCs) curves were generated using fivefold cross-validation. (*G*) The GC fraction of each genomic region was calculated and plotted for each cluster. (*H*) Number of TF binding sites in each genomic region was calculated and plotted for each cluster. *P* values were calculated by Student's *t*-tests.

patchMPRA library cloning

We selected a single plasmid from a previous patchMPRA library (Maricque et al. 2019) to serve as the backbone of our promoter library. This plasmid contains a single enhancer and drove robust expression in a previous experiment. The enhancer contains motifs for FOS/JUN and MAF transcription factors, and its full sequence is TGCCCCCCTTCTTCCTATGTCTGATGGAGTTTCCTCTCTAAG TAGCCATTTTATTCTGCTGACTCACCCTCTAACTCCCGGTCTTA TTCCATCCTGCCTCAGGGTCTGTGGTGTAGTCATAGCAC.

To create our library (representative vector in Supplemental Data S4; pCPL1), we first removed the hsp68-dsRed construct from the selected plasmid with HindIII and XhoI. We then amplified the oligo pool using primers CPL1 and CPL2 (Supplemental Table S4) and inserted it into the digested backbone using HiFi DNA Assembly (New England Biolabs). Next, we digested the library with HindIII and XbaI and inserted an mScarlet fluorophore between the promoter and barcodes. To test the library without an upstream enhancer, we also cloned the library into a vector backbone that does not contain an enhancer (representative vector in Supplemental Data S5; pCPL2). The backbone was digested with SpeI and XhoI, and the oligo pool was amplified with primers CPL2 and CPL3 (Supplemental Table S4). The fragments were assembled using HiFi DNA Assembly (New England Biolabs), and the mScarlet fluorophore was inserted in the same way as described above.

patchMPRA

We replaced the HygTK-GFP cassette in the original landing pad cell lines from Maricque et al. (2019) with a reporter expressing both HSV-TK (herpes simplex virus thymidine kinase) and the monomeric blue fluorescent protein tagBFP. The new cassette contains a functional HSV-TK gene, allowing for negative selection of cells that do not have a library member integrated (pCPL3) (Supplemental Data S6).

K562 cells were maintained in Iscove's Modified Dulbecco's Medium (IMDM) + 10% FBS + 1% nonessential amino acids + 1% penicillin/streptomycin. To integrate the library into the genome, we cotransfected the library and CRE recombinase (pBS185 CMV-Cre, Addgene 11916) into four K562 "landing pad" cell lines expressing the HSV-TK gene (landing pad details in Supplemental Table S2). For each replicate, we transfected 32 µg library with 32 µg CRE recombinase into 9.6 million total cells using the Neon Transfection System (Thermo Fisher Scientific). We performed three separate transfections representing three biological replicates. After three days, we treated the cells with 2 mM ganciclovir to kill the cells that did not successfully integrate a library element. Cells were treated every day for 4 d. We then selected for live cells using the MACS Dead Cell Removal kit (Miltenyi Biotech), and the cells were allowed to grow until there were sufficient cells for DNA/RNA extractions (about 10 million cells).

DNA and RNA were harvested from the cells using the TRIzol reagent (Invitrogen). The RNA was treated with two rounds of DNase using the Rigorous DNase treatment procedure in the Turbo DNase protocol (Ambion), and cDNA was synthesized with Oligo(dT) primers using the SuperScript IV First Strand Synthesis System (Invitrogen). The barcodes were then amplified from cDNA and genomic DNA (gDNA) using Q5 High Fidelity 2X Master Mix (New England Biolabs) with primers specific to

our reporter gene (CPL4-5) (Supplemental Table S4). We performed 32 PCR reactions per cDNA biological replicate and 48 PCR reactions per gDNA biological replicate, then pooled the PCRs of each replicate for PCR purification. Four nanograms from each replicate were then further amplified with two rounds of PCR to add Illumina sequencing adapters (CPL6-9) (Supplemental Table S4). Barcodes were sequenced on the Illumina NextSeq platform.

Episomal MPRA

We first digested the patchMPRA library with HindIII and XbaI to replace the mScarlet fluorophore with a tdTomato between the promoter and pBC. We then subcloned the promoter library with the tdTomato fluorophore into the landing pad lentiviral vector from Maricque et al. (2019) to ensure that the 3′ UTR from the episomal library matches that of the patchMPRA experiment. Briefly, the lentiviral vector and patchMPRA library were digested with XhoI/SpeI and NheI/SalI, respectively, and the library was ligated into the lentiviral backbone with T4 DNA ligase (New England Biolabs).

For the MPRA, we transfected the library into K562 cells using the Neon Transfection System (Thermo Fisher Scientific). We performed two biological replicates, transfecting 2.4 million cells with 10 μg of library per replicate. After 24 h, we harvested RNA from the cells using the PureLink RNA Mini kit (Invitrogen). The RNA was treated with DNase and converted to cDNA in the same way as the patchMPRA library above. We then amplified barcodes from cDNA using primers CPL5 and CPL10 (Supplemental Table S4) with Q5 High Fidelity 2X Master Mix (New England Biolabs). We performed four PCR reactions per replicate from cDNA. For DNA normalization, we performed the same PCR (two PCR reactions per replicate; two replicates) on the plasmid library. The PCRs from the same replicates were then pooled and purified. Four nanograms from each replicate were then further amplified with two rounds of PCR to add Illumina sequencing adapters (CPL6-9) (Supplemental Table S4). Barcodes were sequenced on the Illumina NextSeq platform.

TRIP library cloning

We performed TRIP according to the published protocol (Akhtar et al. 2013) with some modifications. We first digested the PiggyBac vector (PBSsplitGFP, gift from Robi Mitra lab) (Qi et al. 2017) with BamHI and NotI. Each selected promoter was amplified from the promoter library (CPL11-22) (Supplemental Table S4) and assembled into the vector with a tdTomato fluorophore and the neuropilin 1 poly(A) sequence (Akhtar et al. 2013) using HiFi DNA Assembly (New England Biolabs). We then added a unique barcode that identifies the promoter (pBC) to each promoter construct using the Q5 Site-Directed Mutagenesis kit (New England Biolabs). A second random barcode was added to each promoter construct by digesting with XbaI followed by HiFi DNA Assembly (New England Biolabs) with a single-stranded oligo containing 16 random Ns (TRIP barcodes; tBC) and homology arms to the plasmid (CPL23) (Supplemental Table S4). The components of the final library are shown in Supplemental Figure S8A, and a representative vector is in Supplemental Data S9 (pCPL4). The PiggyBac ITRs, promoter, and tdTomato reporter cassette are located between two parts of a split-GFP reporter gene which is driven by a separate EF1a promoter. When the barcoded reporter cassette is integrated into the genome, the split-GFP remaining on the plasmid combines to produce functional GFP, allowing us to sort for cells that have successfully integrated the promoters (Supplemental Fig. S8A). Because each promoter is uniquely barcoded, we combined all the promoters into a single library for subsequent TRIP experiments.

TRIP

The TRIP library and piggyBac transposase (gift from Robi Mitra lab) were cotransfected into wild-type K562 cells at a 1:1 ratio using the Neon Transfection System (Thermo Fisher Scientific). In total, we transfected 4.8 million cells with 16 μg each of library and transposase. The cells were sorted after 24 h for GFP-positive cells to enrich for cells that have integrated the reporters. After a week, the cells were sorted into four pools of 7000 cells each to ensure that each pBC-tBC pair is only integrated once in each pool. The pools were then allowed to grow until there were sufficient cells for DNA/RNA extractions.

We harvested DNA and RNA from the cells using the TRIzol reagent (Invitrogen). The RNA was treated with DNase and converted to cDNA in the same way as the patchMPRA library above. We then amplified barcodes from cDNA and gDNA using primers CPL10 and CPL24 (Supplemental Table S4). We performed four PCRs per pool from cDNA and gDNA, respectively, using Q5 High Fidelity 2X Master Mix (New England Biolabs), then pooled the PCRs and purified them. Four nanograms from each replicate were then further amplified with two rounds of PCR to add Illumina sequencing adapters (CPL25-26, CPL8-9) (Supplemental Table S4). Barcodes were sequenced on the Illumina NextSeq platform.

To map the locations of TRIP integrations, we digested gDNA with a combination of AvrII, NheI, SpeI, and XbaI for 16 h. The digestions were purified and self-ligated at 4°C for another 16 h. After purifying the ligations, we performed inverse PCR to amplify the barcodes with the associated genomic DNA region (primers CPL24-25) (Supplemental Table S4). We did eight PCRs per pool, purified them, and used 4 ng of each pool for a further two rounds of PCR to add Illumina sequencing adapters (CPL29-31, CPL9) (Supplemental Table S4). The library was then sequenced on the Illumina NextSeq platform.

patchMPRA and episomal MPRA data processing

For patchMPRA, we obtained approximately 11–13 million reads per DNA or RNA replicate from sequencing. For episomal MPRA, we obtained approximately 500,000 reads per DNA or RNA replicate. Reads that contained the barcodes in the proper sequence context were included in subsequent analysis. The pBCs were then decoded using the FREE barcodes software (Hawkins et al. 2018), and the expression of each barcode pair was calculated as $\log_2$(RNA/DNA). We averaged the expression of barcodes corresponding to the same promoter within each replicate to get promoter expression per replicate, then averaged across replicates for subsequent downstream analysis. Expression values can be found in Supplemental Data S7 (patchMPRA) and Supplemental Data S8 (episomal MPRA).

TRIP data processing

We obtained ~14–25 million reads per DNA or RNA pool from sequencing. Reads that contained both the tBC and pBC in the proper sequence context were included in subsequent analysis. We further filtered tBCs such that they are at least three hamming distances apart from every other barcode to account for mutations that occurred during PCR and sequencing. The expression of each BC pair was calculated as $\log_2$(RNA/DNA). We added a pseudocount to the RNA counts to include barcode pairs that had DNA but no RNA reads. Data from the four independent pools were

combined in all analyses. Expression values can be found in Supplemental Data S2.

For the locations of TRIP integrations, reads containing each barcode pair were matched with the sequence of its integration site. The integration site sequences were then aligned to hg38 using BWA with default parameters. Only barcodes that mapped to a unique location were kept for downstream analyses. The mapped integration locations can be found in Supplemental Data S3.

TRIP data analysis

We downloaded a list of expressed genes in K562 cells using whole-cell long poly(A) RNA-seq data generated by ENCODE (Djebali et al. 2012) from the EMBL-EBI Expression Atlas (https://www.ebi.ac.uk/gxa/home). We then designated the genes as hk or dev based on the list of hk genes obtained from Eisenberg and Levanon (2013). Using the locations of these promoters (GENCODE Release 36, GRCh38.p13), we identified TRIP integrations located within 5 kb of either hk or dev promoters and plotted the expression of these integrations separately.

To increase the resolution of the analysis, we identified genomic regions where at least four different promoters integrated within 5 kb of each other (full list of regions in Supplemental Data S3). For regions in which the same promoter integrated more than once, we used the median expression of that promoter. This yielded 1268 genomic regions. All heat maps were generated using the ComplexHeatmap package in R (R Core Team 2010; Gu et al. 2016). To determine the diversity of the identified 5-kb regions, we downloaded the 15-state segmentation for K562 (hg19) from the ENCODE portal and converted the genomic coordinates to hg38 using the UCSC liftOver tool (Hinrichs et al. 2006). We then overlapped the 5-kb regions with ChromHMM regions using a minimum overlap of 200 bp using the Genomic Ranges R packages (Lawrence et al. 2013).

To rank and cluster the regions, we first imputed missing values using the mean of the promoter across all locations. We then used the means of each region to rank the clusters and plotted the smoothed expression of each promoter. To cluster the 5-kb genomic regions, we ran *k*-means clustering on the imputed data using the ConsensusClusterPlus package in R (R Core Team 2010; Wilkerson and Hayes 2010). The imputed data was only used for ranking and clustering and not downstream analysis.

Epigenome data analysis

For the cluster metaplots, we considered the boundaries of each genomic region as the locations of the first and last integrations in each region. We then downloaded various K562 epigenome data sets (full list of sources in Supplemental Table S5). For CpG methylation, we downloaded both replicates and used the averaged signal from both replicates. For H3K27ac, H3K4me3, PolII, and CpG methylation and ATAC-seq, we used the EnrichedHeatmap package in R (Gu et al. 2018) to draw the metaplots for each cluster extending 5 kb upstream of and downstream from each genomic region. For CAGE-seq, we downloaded the hg19 data set from the FANTOM5 consortium (Lizio et al. 2015, 2019) and converted it to hg38 using the UCSC liftOver tool (Hinrichs et al. 2006). Because the signal was relatively sparse across genomic locations, we plotted the total CAGE signal across each genomic region.

Sequence feature analysis

We obtained the sequences of each region using the BSgenome package in R (https://rdrr.io/bioc/BSgenome/). For the gapped *k*-mer predictions, we used the gkmSVM R package (Ghandi et al. 2016) with word length = 10 and number of informative columns = 6. We used AME for motif enrichment analysis (McLeay and Bailey 2010), DREME for de novo motif discovery (Bailey 2011), and FIMO to determine the number of motifs per sequence (Grant et al. 2011), from MEME suite 5.0.4. For all motif analyses, we limited analysis to expressed transcription factors (FPKM ≥1) in K562 from whole-cell long poly(A) RNA-seq data generated by ENCODE (Djebali et al. 2012) downloaded from the EMBL-EBI Expression Atlas.

To predict the type of genomic region of other integrations not in the defined 5-kb regions, we obtained genomic sequences of the 1-kb flanking region around the integration (500 bp upstream and 500 bp downstream). We then used the trained gkmSVM kernels to calculate the weights of each flanking region and assigned the integrations into low, medium, or high activity clusters based on their weights. Only integrations that could be confidently assigned were included.

Modeling

We fit $\log_2$ expression values with linear models of core promoter and genomic location activities using the lm function in R (https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/lm). Variance explained by each term was calculated with one-way ANOVAs of the respective models.

Data access

All raw and processed 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 number GSE173678. The code used to process the data can be found in the Supplemental Code and at GitHub (https://github.com/claricehong/core_promoters_2021).

Competing interest statement

B.A.C. is on the Scientific Advisory Board of Patch Biosciences.

Acknowledgments

We thank Vanja Haberle and Alexander Stark for providing us with sequences from their core promoter library and helpful suggestions for which promoters to select. We also thank Max Staller and Robi Mitra for providing us with plasmids for our experiments. We also thank Ting Wang, Brett Maricque, and members of the Cohen Lab for their helpful comments and critical feedback on the manuscript; Jessica Hoisington-Lopez and MariaLynn Crosby in the DNA Sequencing Innovation Lab for assistance with high-throughput sequencing, and the Genome Engineering and iPSC Center for kindly allowing us to use their flow cytometer for cell sorting. This work was supported by grants to B.A.C. from the National Institutes of Health, National Institute of General Medical Sciences (R01GM092910).

Author contributions: C.K.Y.H. and B.A.C. conceived and designed the project. C.K.Y.H. designed and conducted all experiments and analyses. C.K.Y.H. and B.A.C. wrote the manuscript.

References

Akhtar W, de Jong J, Pindyurin AV, Pagie L, Meuleman W, de Ridder J, Berns A, Wessels LFA, van Lohuizen M, van Steensel B. 2013. Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell* **154:** 914–927. doi:10.1016/j.cell.2013.07.018

Arnold CD, Zabidi MA, Pagani M, Rath M, Schernhuber K, Kazmar T, Stark A. 2017. Genome-wide assessment of sequence-intrinsic enhancer

responsiveness at single-base-pair resolution. *Nat Biotechnol* **35:** 136–144. doi:10.1038/nbt.3739
Bailey TL. 2011. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27:** 1653–1659. doi:10.1093/bioinformatics/btr261
Bergman DT, Jones TR, Liu V, Siraj L, Kang HY, Nasser J, Kane M, Nguyen TH, Grossman SR, Fulco CP, et al. 2021. Compatibility logic of human enhancer and promoter sequences. bioRxiv doi:10.1101/2021.10.23.462170v1
Butler JEF, Kadonaga JT. 2001. Enhancer–promoter specificity mediated by DPE or TATA core promoter motifs. *Genes Dev* **15:** 2515–2519. doi:10.1101/gad.924301
Carroll SB. 2005. Evolution at two levels: on genes and form. *PLoS Biol* **3:** e245. doi:10.1371/journal.pbio.0030245
Chen M, Licon K, Otsuka R, Pillus L, Ideker T. 2013. Decoupling epigenetic and genetic effects through systematic analysis of gene position. *Cell Rep* **3:** 128–137. doi:10.1016/j.celrep.2012.12.003
Corrales M, Rosado A, Cortini R, van Arensbergen J, van Steensel B, Filion GJ. 2017. Clustering of *Drosophila* housekeeping promoters facilitates their expression. *Genome Res* **27:** 1153–1161. doi:10.1101/gr.211433.116
Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. 2012. Landscape of transcription in human cells. *Nature* **489:** 101–108. doi:10.1038/nature11233
Duan Z-J, Fang X, Rohde A, Han H, Stamatoyannopoulos G, Li Q. 2002. Developmental specificity of recruitment of TBP to the TATA box of the human γ-globin gene. *Proc Natl Acad Sci* **99:** 5509–5514. doi:10.1073/pnas.072084499
Eisenberg E, Levanon EY. 2013. Human housekeeping genes, revisited. *Trends Genet* **29:** 569–574. doi:10.1016/j.tig.2013.05.010
Emami KH, Navarre WW, Smale ST. 1995. Core promoter specificities of the Sp1 and VP16 transcriptional activation domains. *Mol Cell Biol* **15:** 5906–5916. doi:10.1128/MCB.15.11.5906
Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28:** 817–825. doi:10.1038/nbt.1662
Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473:** 43–49. doi:10.1038/nature09906
Gehrig J, Reischl M, Kalmár É, Ferg M, Hadzhiev Y, Zaucker A, Song C, Schindler S, Liebel U, Müller F. 2009. Automated high-throughput mapping of promoter-enhancer interactions in zebrafish embryos. *Nat Methods* **6:** 911–916. doi:10.1038/nmeth.1396
Ghandi M, Lee D, Mohammad-Noori M, Beer MA. 2014. Enhanced regulatory sequence prediction using gapped *k*-mer features. *PLoS Comput Biol* **10:** e1003711. doi:10.1371/journal.pcbi.1003711
Ghandi M, Mohammad-Noori M, Ghareghani N, Lee D, Garraway L, Beer MA. 2016. gkmSVM: an R package for gapped-kmer SVM. *Bioinforma Oxf Engl* **32:** 2205–2207. doi:10.1093/bioinformatics/btw203
Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27:** 1017–1018. doi:10.1093/bioinformatics/btr064
Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinforma Oxf Engl* **32:** 2847–2849. doi:10.1093/bioinformatics/btw313
Gu Z, Eils R, Schlesner M, Ishaque N. 2018. EnrichedHeatmap: an R/Bioconductor package for comprehensive visualization of genomic signal associations. *BMC Genomics* **19:** 234. doi:10.1186/s12864-018-4625-x
Haberle V, Stark A. 2018. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol* **19:** 621–637. doi:10.1038/s41580-018-0028-8
Haberle V, Arnold CD, Pagani M, Rath M, Schernhuber K, Stark A. 2019. Transcriptional cofactors display specificity for distinct types of core promoters. *Nature* **570:** 122–126. doi:10.1038/s41586-019-1210-7
Hawkins JA, Jones SK, Finkelstein IJ, Press WH. 2018. Indel-correcting DNA barcodes for high-throughput sequencing. *Proc Natl Acad Sci* **115:** E6217–E6226. doi:10.1073/pnas.1802640115
Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, et al. 2006. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* **34:** D590–D598. doi:10.1093/nar/gkj144
Juven-Gershon T, Cheng S, Kadonaga JT. 2006. Rational design of a super core promoter that enhances gene expression. *Nat Methods* **3:** 917–922. doi:10.1038/nmeth937
Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9:** e1003118. doi:10.1371/journal.pcbi.1003118
Lenhard B, Sandelin A, Carninci P. 2012. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet* **13:** 233–245. doi:10.1038/nrg3163
Li X, Noll M. 1994. Compatibility between enhancers and promoters determines the transcriptional specificity of gooseberry and gooseberry neuro in the *Drosophila* embryo. *EMBO J* **13:** 400–406. doi:10.1002/j.1460-2075.1994.tb06274.x
Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, Abugessaisa I, Fukuda S, Hori F, Ishikawa-Kato S, et al. 2015. Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol* **16:** 22. doi:10.1186/s13059-014-0560-6
Lizio M, Abugessaisa I, Noguchi S, Kondo A, Hasegawa A, Hon CC, de Hoon M, Severin J, Oki S, Hayashizaki Y, et al. 2019. Update of the FANTOM web resource: expansion to provide additional transcriptome atlases. *Nucleic Acids Res* **47:** D752–D758. doi:10.1093/nar/gky1099
Maricque BB, Chaudhari HG, Cohen BA. 2019. A massively parallel reporter assay dissects the influence of chromatin structure on *cis*-regulatory activity. *Nat Biotechnol* **37:** 90–95. doi:10.1038/nbt.4285
McLeay RC, Bailey TL. 2010. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics* **11:** 165. doi:10.1186/1471-2105-11-165
Merli C, Bergstrom DE, Cygan JA, Blackman RK. 1996. Promoter specificity mediates the independent regulation of neighboring genes. *Genes Dev* **10:** 1260–1270. doi:10.1101/gad.10.10.1260
Moudgil A, Wilkinson MN, Chen X, He J, Cammack AJ, Vasek MJ, Lagunas T, Qi Z, Lalli MA, Guo C, et al. 2020. Self-reporting transposons enable simultaneous readout of gene expression and transcription factor binding in single cells. *Cell* **182:** 992–1008.e21. doi:10.1016/j.cell.2020.06.037
Ohtsuki S, Levine M, Cai HN. 1998. Different core promoters possess distinct regulatory activities in the *Drosophila* embryo. *Genes Dev* **12:** 547–556. doi:10.1101/gad.12.4.547
Parry TJ, Theisen JWM, Hsu J-Y, Wang Y-L, Corcoran DL, Eustice M, Ohler U, Kadonaga JT. 2010. The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. *Genes Dev* **24:** 2013–2018. doi:10.1101/gad.1951110
Prud'homme B, Gompel N, Carroll SB. 2007. Emerging principles of regulatory evolution. *Proc Natl Acad Sci* **104:** 8605–8612. doi:10.1073/pnas.0700488104
Qi Z, Wilkinson MN, Chen X, Sankararaman S, Mayhew D, Mitra RD. 2017. An optimized, broadly applicable piggyBac transposon induction system. *Nucleic Acids Res* **45:** e55. doi:10.1093/nar/gkw1290
R Core Team. 2010. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.
Roy AL, Singer DS. 2015. Core promoters in transcription: old problem, new insights. *Trends Biochem Sci* **40:** 165–171. doi:10.1016/j.tibs.2015.01.007
Sharpe J, Nonchev S, Gould A, Whiting J, Krumlauf R. 1998. Selectivity, sharing and competitive interactions in the regulation of *Hoxb* genes. *EMBO J* **17:** 1788–1798. doi:10.1093/emboj/17.6.1788
Wilkerson MD, Hayes DN. 2010. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26:** 1572–1573. doi:10.1093/bioinformatics/btq170
Xiao JY, Hafner A, Boettiger AN. 2021. How subtle changes in 3D structure can create large changes in transcription. *eLife* **10:** e64320. doi:10.7554/eLife.64320
Yang C, Bolotin E, Jiang T, Sladek FM, Martinez E. 2007. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene* **389:** 52–65. doi:10.1016/j.gene.2006.09.029
Yoshida J, Akagi K, Misawa R, Kokubu C, Takeda J, Horie K. 2017. Chromatin states shape insertion profiles of the piggyBac, Tol2 and Sleeping Beauty transposons and murine leukemia virus. *Sci Rep* **7:** 43613. doi:10.1038/srep43613
Zabidi MA, Arnold CD, Schernhuber K, Pagani M, Rath M, Frank O, Stark A. 2015. Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518:** 556–559. doi:10.1038/nature13994

Received July 21, 2021; accepted in revised form November 15, 2021.