Genomic environments scale the activities of diverse core promoters

Authors

Clarice KY Hong[1,2], Barak A Cohen[1,2*]

Affiliations:

[1] The Edison Family Center for Genome Sciences and Systems Biology, School of Medicine, Washington University in St. Louis, Saint Louis, MO, USA.

[2] Department of Genetics, School of Medicine, Washington University in St. Louis, Saint Louis, MO, USA.

*Correspondence to: cohen@wustl.edu

Abstract:

One model for how cells integrate cis-regulatory information is that different classes of core promoters respond specifically to certain genomic environments. We tested this model using a genome-integrated massively parallel reporter assay (MPRA) to measure the activity of hundreds of diverse core promoters at four genomic locations and, in a complementary experiment, six core promoters at thousands of genomic locations. While genomic locations had large effects on expression, the relative strengths of core promoters were preserved across locations regardless of promoter class, suggesting that their intrinsic activities are scaled by diverse genomic environments. The extent of scaling depends on the genomic location and the strength of the core promoter, but not on its class. Our results support a modular genome in which genomic environments scale the activities of core promoters.

One Sentence Summary:

Genomic environments have consistent effects on gene expression that depend on the strength, but not the class of core promoter.

Main Text:

Cells carry out complex programs of gene expression by integrating information stored in locally acting cis-regulatory sequences (CRSs) and the genomic environment. We define the genomic environment as the distal enhancers in 3D space and the chromatin landscape at a genomic location. These properties influence a gene's local CRSs, which can be separated into two categories, the core promoter and its proximal regulatory sequences (Fig. 1A). The core promoter is the ~100bp region around the transcription start site which is responsible for accurately positioning RNA polymerase II and binding general transcription factors (TFs) (*1*, *2*). The adjacent regulatory elements, sometimes called proximal promoters or proximal enhancers (*2*, *3*), bind to TFs and modulate core promoter activity. How the cell integrates information from a gene's core promoter and its larger genomic environment is crucial to understanding how cell-type specific regulatory programs are achieved.

One model for how the cell integrates core promoter and genomic information is through the 'promoter compatibility' hypothesis. In this model, core promoters with different sequence elements respond specifically to distinct genomic environments containing different enhancers and chromatin features (*4*–*23*). Alternatively, core promoters and genomic environments could contribute independently to gene expression, and specificity is achieved via other mechanisms (*24*–*28*). A strong prediction of the promoter compatibility hypothesis is that the relative strengths of 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. Here, we tested the promoter compatibility hypothesis by assaying hundreds of core promoters at four different genomic locations and by assaying six core promoters across ~1000 genomic locations.

Results

Measurement of diverse core promoter activities at different genomic locations

We created a library of reporter genes driven by diverse core promoters. The library contains 676 core promoters spanning a variety of promoter features from Haberle *et al.* (*20*), including TATA, DPE and TCT motifs, CpG islands and housekeeping (hk) and developmental (dev) promoters (Table S1, Data S1). To provide redundancy in the measurements, we included ten 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 patchMPRA (parallel targeting of chromosome positions by MPRA), 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 (Fig. S1 and Table S2) (*29*). 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 (*29*). 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 the 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 (Figs. S2A-B). Biological replicates showed high reproducibility (average Pearson's r = 0.87) (Figs. 1C and S2C) and the locations of the landing pad had large effects on library expression that were consistent with previous studies (compare Fig.1D to Fig. S3; (*29*) indicating that the genomic environment is not drastically altered by a diverse core promoter library. 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 different classes of promoters will respond to the same genomic environment 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 and Fig. S4A). However, the magnitude of the genomic effect is not the same for all promoter classes. We focused on the hk/dev classification of promoters because there is the most evidence for separation of function of these two classes (*6*, *12*) and because there are sufficient numbers of promoters in each class for further analysis. We performed ANOVA to quantify the contribution of the genomic location and core promoters to gene expression and found that the genomic location has a larger effect on dev promoters than hk promoters (Fig. 2B). For hk promoters, genomic location and core promoters contribute ~25% and ~65% respectively. In contrast, for dev promoters, genomic location contributes ~55% while core promoters contribute only ~36%. This result suggests that genomic location has a larger effect on dev promoters than hk promoters.

We further 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. While promoter activities were highly correlated between genomic locations regardless of the class of promoter (Pearson's r = 0.74 - 0.9, Spearman's ρ = 0.72 - 0.88) (Fig. 2C), dev promoters were consistently less correlated than hk promoters (Fig. 2D). We further divided the promoters into smaller subclasses containing different motifs and/or CpG island classifications and showed that even within the hk or dev classes, each subclass had substantial differences in correlations between genomic locations (Fig. S4B). Taken together these results suggest that genomic environments scale the activities of all core promoters, but that the quantitative extent of scaling can differ between promoter classes.

Intrinsic promoter strength explains differences between promoter classes

A striking 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 and S5A). Thus, we tested if the different effects of genomic environments on hk and dev promoters was due to their differences in strength. 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 (~42% and ~46% respectively), but for weak promoters, genomic environments contribute ~73%, while core promoters contribute only ~15%. Weak promoters are also consistently less correlated than strong promoters (Fig. 3C). This analysis suggests that the difference between hk and dev in Fig. 2B is due to their differences in strength. Finally, we sampled a set of hk and dev promoters with similar average strengths (Fig. S5B) and compared their correlations across genomic locations. Using only this subset of promoters, correlations across genomic locations are comparable between hk and dev promoters (Fig. S5C). The differences in how genomic locations scale the activities of each core promoter subclass is also largely explained by the average strength of each promoter class (Fig. S5D). 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.

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 reflect 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 (r^2 = 0.59-0.76; Figs. 3D and S6A), indicating that the relative intrinsic activities of core promoters are preserved when integrated into the genome. Promoter activity on plasmids is also a good predictor of activity in the genome (adjusted r^2 = 0.78; Fig. S6B). 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

We next asked if the scaling of core promoters we observed at the four genomic locations holds at 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 (Table S3). We then measured their activities across the genome using the TRIP (Thousands of Reporters Integrated in Parallel) assay (Fig. S7A, *30*). Each core promoter was fused to a unique barcode (pBC) in its 3'UTR and cloned upstream of a reporter gene into a PiggyBac transposon vector for random delivery into the genome. TRIP libraries were generated by incorporating >10^5 random barcodes (tBCs) onto each core promoter reporter plasmid. After transposition, every genomic integration generates an mRNA with a pBC and tBC 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 in K562 cells. The replicates were highly correlated (Pearson's r^2 = 0.96, Fig. S7B). In total, we mapped 41,083 unique integrations in the genome, ranging between 6078-7418 integrations per promoter (Table S3, 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. 4A). 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 (Figs. 4B and S7C), which suggests that the effect of different genomic locations is to scale intrinsic promoter activities. To compare different promoters in the same genomic environment, we identified 1278 genomic regions in which at least 4 of the 6 promoters had integrated <5kb from each other (in separate cells) (Data S3). These genomic regions are located across the entire genome and span diverse chromHMM annotations (*31*, *32*) (Figs. S8A-B). Across these locations, expression consistently increases from the weakest (dev2) to strongest (hk1) promoter (Figs. 4C-D), 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 $r > 0.7$ (Fig. S8C), and a linear model assuming independent effects of genomic region and promoter explains ~54% of the variance in the data (Fig. S8D). Thus, measurements of integrated promoters across diverse genomic positions demonstrates that core promoter scaling is a genome-wide phenomenon.

Non-linear 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 (Fig. S8E). Thus, promoters with different strengths do not respond to differences in genomic environments in the same way.

Interestingly, the curves in Fig. 5A separate by the intrinsic strength of the core promoters and not by their hk or dev identity. To emphasize this point we calculated the correlations between the 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 (Fig. S9A). Integrations within 5kb of endogenous hk or dev promoters in K562 also showed no preference for hk or dev promoters respectively (Fig. S9B). 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 Fig. 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 (Fig. 5A). 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 while 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 gene expression, resulting in non-linear scaling of promoter activities across the genome.

Genomic environments with different strengths 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 datasets 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 (*33*, *34*). In general, cluster activity is correlated with chromatin marks associated with active transcription (H3K27ac, H3K4me3) and transcriptional activity (PolII binding, CAGE-seq) (Figs. 6A-C, S10A), while accessible chromatin (ATAC) and CpG methylation do not separate the clusters (Figs. S10B-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 (*35*, *36*). The SVMs performed well, with five-fold cross-validated AUCs ranging from 0.8 to 0.9 (Figs. 6D-F and S11A-C). Scrambling the cluster annotations led to essentially random predictions by the SVM (Figs. S11D-E). To further validate the model, we used the trained SVM to predict the cluster type of other TRIP integrations that were not in the 5kb region analysis. As expected, clusters that were predicted to be more active also showed higher expression (Fig. S11F). To identify the motifs that separate the clusters, we performed *de novo* motif enrichment and identified CG-rich sequences in the more active clusters (Fig. S11G). 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 find 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 is partially explained by the number of TFs binding in each cluster.

Discussion

We present a framework for dissecting the contributions of core promoters and genomic locations 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 their genomic environments. While promoter compatibility has been observed for specific promoter-genomic environment pairs (*4*–*8*), our results suggest that such interactions are relatively rare or have smaller effects than the effects of genomic scaling. 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 (*37*, *38*). 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.

Although core promoters are scaled across the genome, scaling is not a simple linear combination of genomic position effects and promoter effects (*29*). Instead, the scaling factors of strong and weak promoters change in different genomic environments (Fig. 5E). These data suggest that different core promoter sequence features set the strength of the promoter, which in turn determines how it interacts with the genomic environment. Our data is also consistent with recent simulations showing how promoters starting from different states can have different responses to increasing enhancer contact frequency (*39*). In the future, this relationship will allow us to predict gene expression by measuring core promoter strength and genomic environment activity independently.

References and Notes

1. A. L. Roy, D. S. Singer, Core promoters in transcription: old problem, new insights. *Trends Biochem. Sci.* **40**, 165–171 (2015).

2. V. Haberle, A. Stark, Eukaryotic core promoters and the functional basis of transcription initiation. *Nat. Rev. Mol. Cell Biol.* **19**, 621–637 (2018).

3. J. E. F. Butler, J. T. Kadonaga, The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev.* **16**, 2583–2592 (2002).

4. S. Ohtsuki, M. Levine, H. N. Cai, Different core promoters possess distinct regulatory activities in the Drosophila embryo. *Genes Dev.* **12**, 547–556 (1998).

5. J. E. F. Butler, J. T. Kadonaga, Enhancer–promoter specificity mediated by DPE or TATA core promoter motifs. *Genes Dev.* **15**, 2515–2519 (2001).

6. M. A. Zabidi, C. D. Arnold, K. Schernhuber, M. Pagani, M. Rath, O. Frank, A. Stark, Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*. **518**, 556–559 (2015).

7. X. Li, M. Noll, Compatibility between enhancers and promoters determines the transcriptional specificity of gooseberry and gooseberry neuro in the Drosophila embryo. *EMBO J.* **13**, 400–406 (1994).

8. C. Merli, D. E. Bergstrom, J. A. Cygan, R. K. Blackman, Promoter specificity mediates the independent regulation of neighboring genes. *Genes Dev.* **10**, 1260–1270 (1996).

9. F. C. Wefald, B. H. Devlin, R. S. Williams, Functional heterogeneity of mammalian TATA-box sequences revealed by interaction with a cell-specific enhancer. *Nature*. **344**, 260–262 (1990).

10. J. Sharpe, S. Nonchev, A. Gould, J. Whiting, R. Krumlauf, Selectivity, sharing and competitive interactions in the regulation of Hoxb genes. *EMBO J*. **17**, 1788–1798 (1998).

11. J. Gehrig, M. Reischl, É. Kalmár, M. Ferg, Y. Hadzhiev, A. Zaucker, C. Song, S. Schindler, U. Liebel, F. Müller, Automated high-throughput mapping of promoter-enhancer interactions in zebrafish embryos. *Nat. Methods*. **6**, 911–916 (2009).

12. C. D. Arnold, M. A. Zabidi, M. Pagani, M. Rath, K. Schernhuber, T. Kazmar, A. Stark, Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution. *Nat. Biotechnol.* **35**, 136–144 (2017).

13. K. H. Emami, W. W. Navarre, S. T. Smale, Core promoter specificities of the Sp1 and VP16 transcriptional activation domains. *Mol. Cell. Biol.* **15**, 5906–5916 (1995).

14. S. K. Hansen, R. Tjian, TAFs and TFIIA mediate differential utilization of the tandem Adh promoters. *Cell*. **82**, 565–575 (1995).

15. B. Ren, T. Maniatis, Regulation of Drosophila Adh promoter switching by an initiator-targeted repression mechanism. *EMBO J*. **17**, 1076–1086 (1998).

16. T. Juven-Gershon, J.-Y. Hsu, J. T. Kadonaga, Caudal, a key developmental regulator, is a DPE-specific transcriptional factor. *Genes Dev*. **22**, 2823–2830 (2008).

17. F. J. van Werven, H. van Bakel, H. A. A. M. van Teeffelen, A. F. M. Altelaar, M. G. Koerkamp, A. J. R. Heck, F. C. P. Holstege, H. Th. M. Timmers, Cooperative action of NC2 and Mot1p to regulate TATA-binding protein function across the genome. *Genes Dev*. **22**, 2359–2369 (2008).

18. T. J. Parry, J. W. M. Theisen, J.-Y. Hsu, Y.-L. Wang, D. L. Corcoran, M. Eustice, U. Ohler, J. T. Kadonaga, The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. *Genes Dev*. **24**, 2013–2018 (2010).

19. M. Xu, E. Gonzalez-Hurtado, E. Martinez, Core promoter-specific gene regulation: TATA box selectivity and Initiator-dependent bi-directionality of serum response factor-activated transcription. *Biochim. Biophys. Acta*. **1859**, 553–563 (2016).

20. V. Haberle, C. D. Arnold, M. Pagani, M. Rath, K. Schernhuber, A. Stark, Transcriptional cofactors display specificity for distinct types of core promoters. *Nature*. **570**, 122–126 (2019).

21. M. J. Buck, J. D. Lieb, A chromatin-mediated mechanism for specification of conditional transcription factor targets. *Nat. Genet.* **38**, 1446–1451 (2006).

22. M. Chen, K. Licon, R. Otsuka, L. Pillus, T. Ideker, Decoupling Epigenetic and Genetic Effects through Systematic Analysis of Gene Position. *Cell Rep*. **3**, 128–137 (2013).

23. C. Leemans, M. C. H. van der Zwalm, L. Brueckner, F. Comoglio, T. van Schaik, L. Pagie, J. van Arensbergen, B. van Steensel, Promoter-Intrinsic and Local Chromatin Features Determine Gene Repression in LADs. *Cell*. **177**, 852-864.e14 (2019).

24. V. Narendra, P. P. Rocha, D. An, R. Raviram, J. A. Skok, E. O. Mazzoni, D. Reinberg, CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science*. **347**, 1017–1021 (2015).

25. J. E. Phillips-Cremins, M. E. G. Sauria, A. Sanyal, T. I. Gerasimova, B. R. Lajoie, J. S. K. Bell, C.-T. Ong, T. A. Hookway, C. Guo, Y. Sun, M. J. Bland, W. Wagstaff, S. Dalton, T. C. McDevitt, R. Sen, J. Dekker, J. Taylor, V. G. Corces, Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*. **153**, 1281–1295 (2013).

26. D. G. Lupiáñez, K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J. M. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S. A. Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, M. Spielmann, A. Visel, S. Mundlos, Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*. **161**, 1012–1025 (2015).

27. J. E. Phillips-Cremins, V. G. Corces, Chromatin insulators: linking genome organization to cellular function. *Mol. Cell*. **50**, 461–474 (2013).

28. D. Hnisz, D. S. Day, R. A. Young, Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. *Cell*. **167**, 1188–1200 (2016).

29. B. B. Maricque, H. G. Chaudhari, B. A. Cohen, A massively parallel reporter assay dissects the influence of chromatin structure on cis-regulatory activity. *Nat. Biotechnol.* **37**, 90–95 (2019).

30. W. Akhtar, J. de Jong, A. V. Pindyurin, L. Pagie, W. Meuleman, J. de Ridder, A. Berns, L. F. A. Wessels, M. van Lohuizen, B. van Steensel, Chromatin Position Effects Assayed by Thousands of Reporters Integrated in Parallel. *Cell*. **154**, 914–927 (2013).

31. J. Ernst, M. Kellis, Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* **28**, 817–825 (2010).

32. J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, B. E. Bernstein, Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*. **473**, 43–49 (2011).

33. M. Chen, K. Licon, R. Otsuka, L. Pillus, T. Ideker, Decoupling Epigenetic and Genetic Effects through Systematic Analysis of Gene Position. *Cell Rep.* **3**, 128–137 (2013).

34. M. Corrales, A. Rosado, R. Cortini, J. van Arensbergen, B. van Steensel, G. J. Filion, Clustering of Drosophila housekeeping promoters facilitates their expression. *Genome Res.* **27**, 1153–1161 (2017).

35. M. Ghandi, D. Lee, M. Mohammad-Noori, M. A. Beer, Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput. Biol.* **10**, e1003711 (2014).

36. M. Ghandi, M. Mohammad-Noori, N. Ghareghani, D. Lee, L. Garraway, M. A. Beer, gkmSVM: an R package for gapped-kmer SVM. *Bioinforma. Oxf. Engl.* **32**, 2205–2207 (2016).

37. S. B. Carroll, Evolution at Two Levels: On Genes and Form. *PLOS Biol.* **3**, e245 (2005).

38. B. Prud'homme, N. Gompel, S. B. Carroll, Emerging principles of regulatory evolution. *Proc. Natl. Acad. Sci.* **104**, 8605–8612 (2007).

39. J. Xiao, A. Hafner, A. N. Boettiger, *https://www.biorxiv.org/content/10.1101/2020.10.22.351395v1* (2020)

40. T. Juven-Gershon, S. Cheng, J. T. Kadonaga, Rational design of a super core promoter that enhances gene expression. *Nat. Methods*. **3**, 917–922 (2006).

41. Z. Qi, M. N. Wilkinson, X. Chen, S. Sankararaman, D. Mayhew, R. D. Mitra, An optimized, broadly applicable piggyBac transposon induction system. *Nucleic Acids Res.* **45**, e55 (2017).

42. S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Röder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, N. S. Bar, P. Batut, K. Bell, I. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, E. Falconnet, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, O. J. Luo, E. Park, K. Persaud, J. B. Preall, P. Ribeca, B. Risk, D. Robyr, M. Sammeth, L. Schaffer, L.-H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, X. Ruan, Y. Hayashizaki, J. Harrow, M. Gerstein, T. Hubbard, A. Reymond, S. E. Antonarakis, G. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigó, T. R. Gingeras, Landscape of transcription in human cells. *Nature*. **489**, 101–108 (2012).

43. E. Eisenberg, E. Y. Levanon, Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574 (2013).

44. Z. Gu, R. Eils, M. Schlesner, Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinforma. Oxf. Engl.* **32**, 2847–2849 (2016).

45. A. S. Hinrichs, D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey, R. A. Harte, F. Hsu, J. Hillman-Jackson, R. M. Kuhn, J. S. Pedersen, A. Pohl, B. J. Raney, K. R. Rosenbloom, A. Siepel, K. E. Smith, C. W. Sugnet, A. Sultan-Qurraie, D. J. Thomas, H. Trumbower, R. J. Weber, M. Weirauch, A. S. Zweig, D. Haussler, W. J. Kent, The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590-598 (2006).

46. M. Lawrence, W. Huber, H. Pagès, P. Aboyoun, M. Carlson, R. Gentleman, M. T. Morgan, V. J. Carey, Software for Computing and Annotating Genomic Ranges. *PLOS Comput. Biol.* **9**, e1003118 (2013).

47. M. D. Wilkerson, D. N. Hayes, ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*. **26**, 1572–1573 (2010).

48. Z. Gu, R. Eils, M. Schlesner, N. Ishaque, EnrichedHeatmap: an R/Bioconductor package for comprehensive visualization of genomic signal associations. *BMC Genomics*. **19**, 234 (2018).

49. M. Lizio, J. Harshbarger, H. Shimoji, J. Severin, T. Kasukawa, S. Sahin, I. Abugessaisa, S. Fukuda, F. Hori, S. Ishikawa-Kato, C. J. Mungall, E. Arner, J. K. Baillie, N. Bertin, H. Bono, M. de Hoon, A. D. Diehl, E. Dimont, T. C. Freeman, K. Fujieda, W. Hide, R. Kaliyaperumal, T. Katayama, T. Lassmann, T. F. Meehan, K. Nishikata, H. Ono, M. Rehli, A. Sandelin, E. A. Schultes, P. A. ‘t Hoen, Z. Tatum, M. Thompson, T. Toyoda, D. W. Wright, C. O. Daub, M. Itoh, P. Carninci, Y. Hayashizaki, A. R. Forrest, H. Kawaji, the FANTOM consortium, Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.* **16**, 22 (2015).

50. M. Lizio, I. Abugessaisa, S. Noguchi, A. Kondo, A. Hasegawa, C. C. Hon, M. de Hoon, J. Severin, S. Oki, Y. Hayashizaki, P. Carninci, T. Kasukawa, H. Kawaji, Update of the FANTOM web resource: expansion to provide additional transcriptome atlases. *Nucleic Acids Res.* **47**, D752–D758 (2019).

51. H. Pagès, BS genome: Software infrastructure for efficient representation of full genomes and their SNPs. *R package version 1.58.0* (2020).

52. R. C. McLeay, T. L. Bailey, Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics*. **11**, 165 (2010).

53. T. L. Bailey, DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*. **27**, 1653–1659 (2011).

54. C. E. Grant, T. L. Bailey, W. S. Noble, FIMO: scanning for occurrences of a given motif. *Bioinformatics*. **27**, 1017–1018 (2011).

Acknowledgments:

We are grateful to 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 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.

Funding:

National Institutes of Health grant R01GM092910 (BAC)

Author contributions:

Conceptualization/Methodology/Investigation/Visualization: CKYH, BAC

Funding acquisition/Supervision: BAC

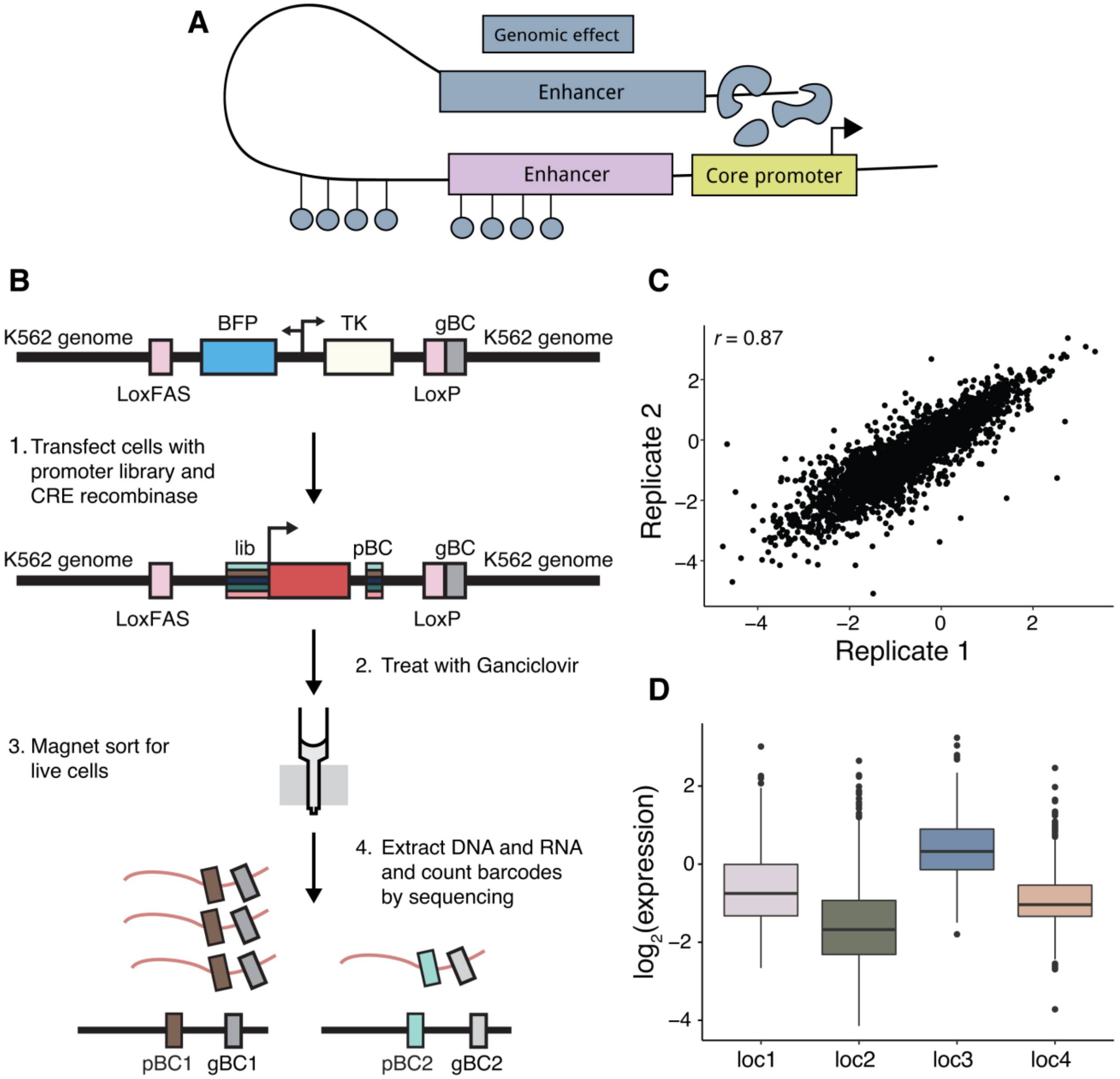
Writing: CKYH, BAC

Competing interests:

Authors declare that they have no competing interests.

Data and materials availability:

All data are available in the main text or the supplementary materials.



455 **Fig. 1.**

Measurements of 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). BFP: blue fluorescent protein; TK: thymidine kinase; gBC: genomic barcode; pBC: promoter barcode. 460 **(C)** Reproducibility of core promoter measurements from independent patchMPRA transfections. **(D)** The expression of all core promoters in the library at each genomic location.

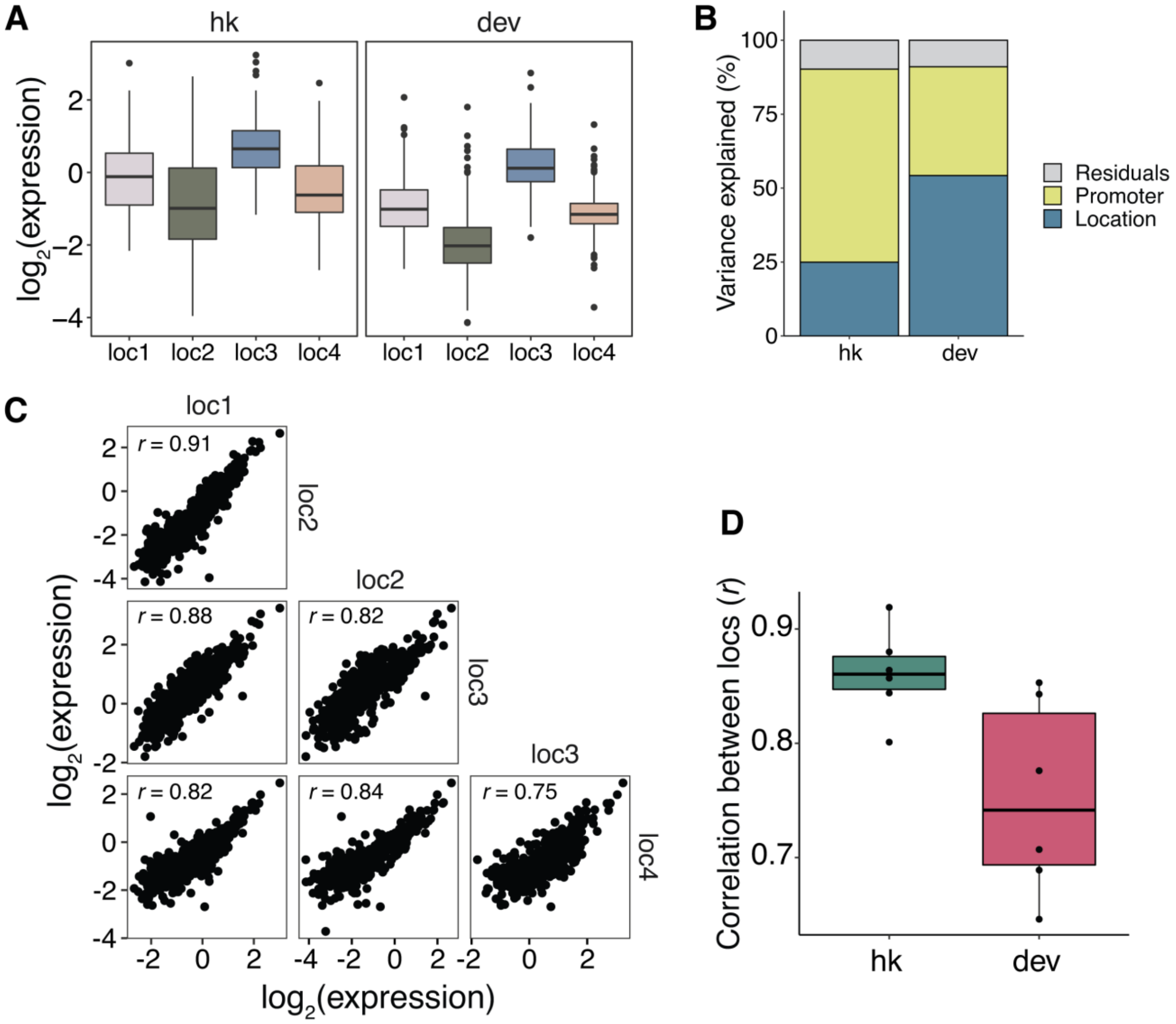


Fig. 2.

465 **Effects of genomic locations on core promoter activity. (A)** Expression of hk and dev core promoters at each genomic location. **(B)** Amount of variance explained by core promoter and genomic location respectively using linear models fit on hk and dev 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

470 core promoters.

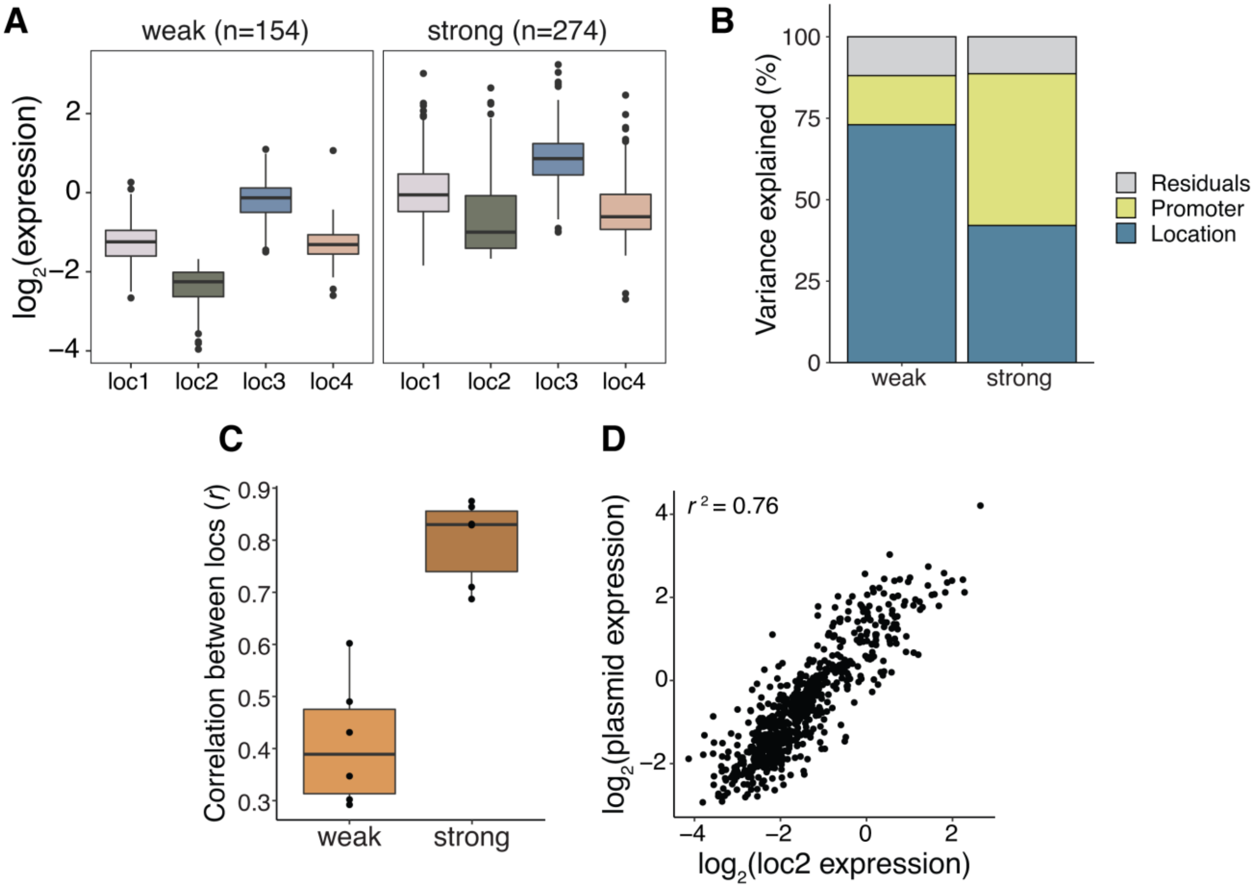


Fig. 3.

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.

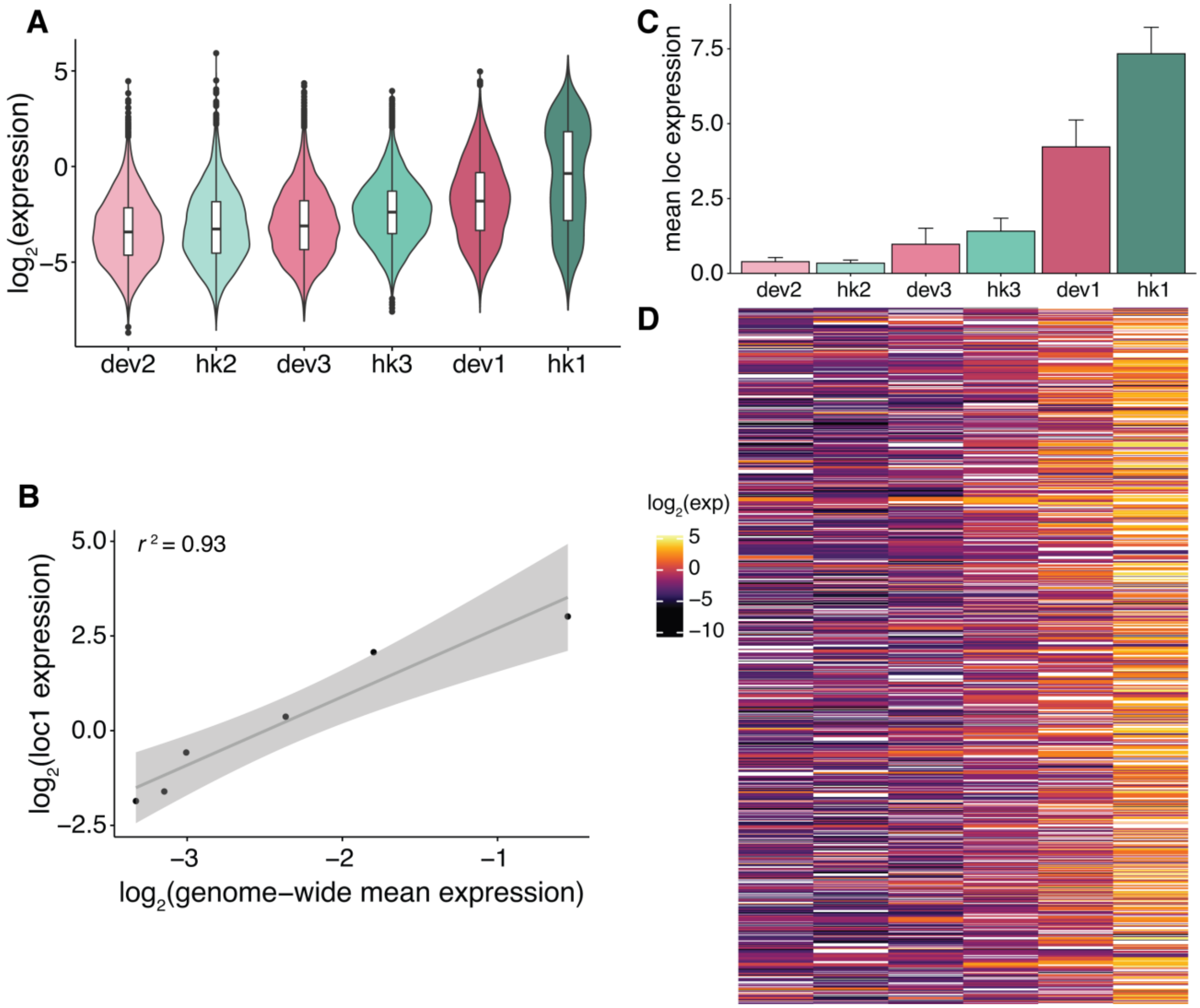


Fig. 4.

Core promoter scaling is a genome-wide phenomenon. (A) 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. **(B)** 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. **(C)** Mean expression of each core promoter from four genomic locations as measured by patchMPRA. Error bars represent the SEM. **(D)** Heatmap of expression of each core promoter (column) at each genomic region (row) that has ≥4 different integrated promoters. White boxes represent NA values.

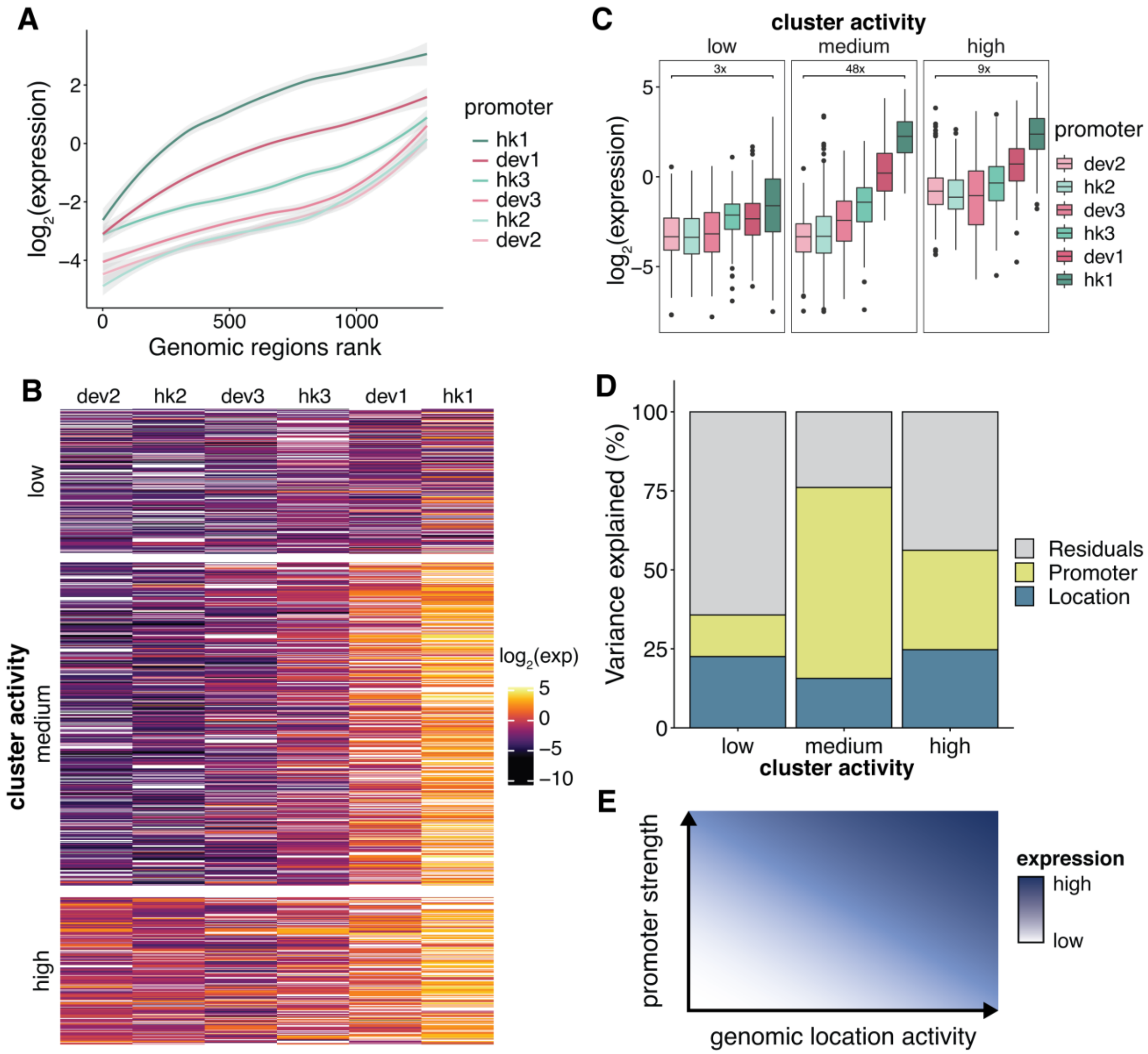


Fig. 5.

Non-linear 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)** Heatmap in Fig. 4D split into 3 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

495

using linear models fit on each genomic cluster respectively. **(E)** Summary model of the relationship between core promoter strength and genomic environment activity.

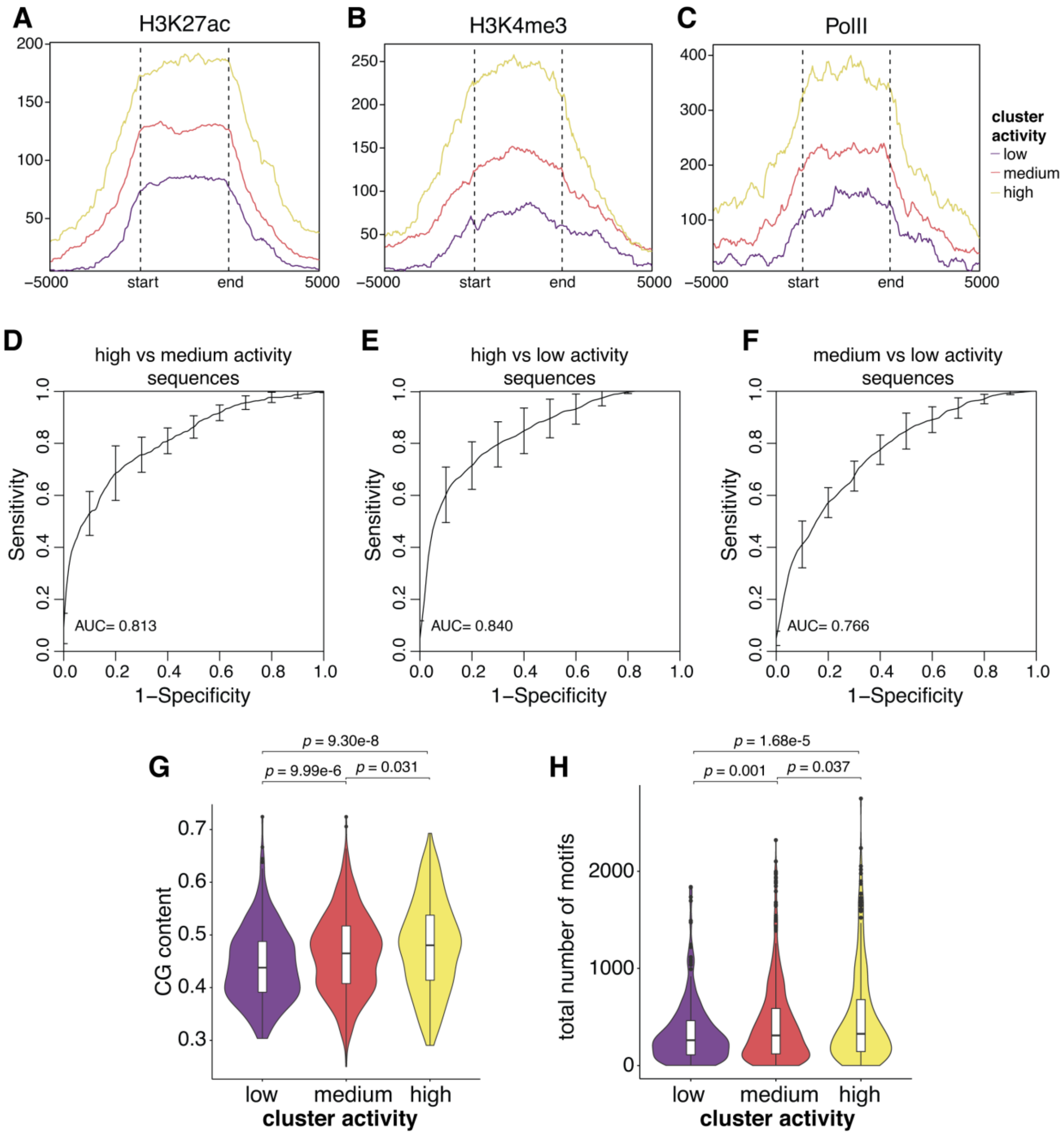


Fig. 6: Genomic clusters have different chromatin states and sequence features. (A, B, C) Metaplots of H3K27ac, H3K4me3 and PolII levels respectively in each genomic cluster. The start and end marks the boundaries of each genomic region, which are determined by the first 505 and last integration in the region. The x-axis extends +/- 5kb around each genomic region. **(D, E, F)** Performance of gkmSVM used to classify sequences from different genomic clusters.

Receiver-operating characteristics (ROC) curves were generated using five-fold 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.

Materials and Methods

Library design

We obtained a set of 6916 core promoter sequences from Haberle et al. (*20*) and selected 672 sequences for our library. Each promoter is 133bp long and centered on the major transcription start site (TSS). We selected the sequences to contain diverse core promoter types and expression patterns (description in Table S1, sequences in Data S1) using the designations obtained from Haberle *et al.*. We also included the super core promoter (SCP1), as well as versions of SCP1 with TATA and DPE single and double mutants (*40*).

patchMPRA library cloning

The core promoter library was synthesized by Agilent technologies through a limited licensing agreement as 200bp oligonucleotides including flanking sequences for cloning. Each element in the library contained 10 barcodes for redundancy, leading to a total of 6760 oligonucleotides. We selected a plasmid with a single strong CRS from the pGL transfer library of our previous patchMPRA experiment (*29*) and replaced the hsp68-dsRed construct with the synthesised promoter library including its corresponding BCs. We then inserted an mScarlet reporter gene between the promoter and barcodes.

patchMPRA

We replaced the HygTK-GFP cassette in the original landing pad cell lines from Maricque *et al.* (*29*) with a reporter expressing both TK (thymidine kinase) and BFP. The new cassette contains a functional TK gene, allowing for negative selection of cells that do not have a library member integrated.

K562 cells were maintained in Iscove's Modified Dulbecco′s Medium (IMDM) + 10% FBS + 1% non-essential amino acids + 1% penicillin/streptomycin. To integrate the library into the genome, we co-transfected the library and CRE recombinase (pBS185 CMV-Cre, Addgene 11916) into 4 K562 ‘landing pad’ cell lines expressing the thymidine kinase (TK) gene (landing pad details in Table S2). For each replicate, we transfected 32μg library with 32μg CRE recombinase into 9.6 million total cells using the Neon Transfection System (Life Technologies). We performed 3 separate transfections representing 3 biological replicates. After 3 days, we treated the cells with 2mM ganciclovir to kill the cells that did not successfully integrate a library element. Cells were treated every day for 4 days. 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 was harvested from the cells using the TRIzol reagent (Life Technologies). The RNA was treated with two rounds of DNase using the Rigorous DNase treatment procedure in the Turbo DNase protocol (Ambion), and cDNA was synthesised with oligo-dT primers using the SuperScript IV First Strand Synthesis System (Invitrogen). The barcodes were then amplified from the cDNA and genomic DNA (gDNA) using the Q5 High Fidelity 2X Master Mix (New England Biolabs) with primers specific to our reporter gene (CPL1-2; Table S4). We performed 32 PCRs per cDNA biological replicate and 48 PCRs per gDNA biological replicate, then pooled the PCRs of each replicate for PCR purification. 4ng from each replicate was then further amplified with 2 rounds of PCR to add Illumina sequencing adapters (CPL3-6; Table S4). Barcodes were sequenced on the Illumina NextSeq platform.

Episomal MPRA

We replaced the mScarlet reporter in the patchMPRA library with a tdTomato reporter gene between the promoter and pBC. To ensure that the 3'UTR from the episomal library matches that of the patchMPRA, we further subcloned the library into the landing pad lentiviral vector.

For the MPRA, we transfected the library into K562 cells using the Neon Transfection System (Life Technologies). We performed 2 biological replicates, transfecting 2.4 million cells with 10µg of library per replicate. After 24h, 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 CPL2 and CPL7 (Table S4) with the Q5 High Fidelity 2X Master Mix (New England Biolabs). We performed 4 PCRs per replicate from cDNA. For DNA normalisation, we performed the same PCR (2 PCRs per replicate; 2 replicates) on the plasmid library. The PCRs from the same replicates were then pooled and purified. 4ng from each replicate was then further amplified with 2 rounds of PCRs to add Illumina sequencing adapters (CPL3-6; Table S4). Barcodes were sequenced on the Illumina NextSeq platform.

TRIP library cloning

We performed TRIP according to the published protocol with some modifications (*30*). Each selected promoter was amplified from the promoter library (CPL8-19; Table S4) and cloned into a PiggyBac vector with a unique barcode that identifies the promoter (pBC). Importantly, the promoter and reporter to be integrated is located between two parts of a split-GFP reporter gene (gift from Robi Mitra lab) (*41*). When the promoter-reporter-barcode construct is integrated into the genome, the split-GFP combines to produce functional GFP, allowing us to sort for cells that have successfully integrated the promoters. Each construct was then randomly barcoded by digesting the plasmid with XbaI followed by HiFi assembly (New England Biolabs) with a single-stranded oligo containing 16 random N's (TRIP barcodes; tBC) and homology arms to the plasmid (CPL20; Table S4). Since 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 Mitra lab) was co-transfected into wild-type K562 cells at a 1:1 ratio using the Neon Transfection System (Life Technologies). In total, we transfected 4.8 million cells with 16µg each of library and transposase. The cells were sorted after 24 hours for GFP-positive cells to enrich for cells that have integrated the reporters. After a week, the cells were sorted into 4 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 (Life Technologies). 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 CPL7 and CPL21 (Table S4). We performed 4 PCRs per pool from cDNA and gDNA respectively using the Q5 High Fidelity 2X Master Mix (New England Biolabs), then pooled the PCRs and purified them. 4ng from each replicate was then further amplified with 2 rounds of PCRs to add Illumina sequencing adapters (CPL22-23, CPL5-6; 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 hours. The digestions were purified and self-ligated at 4°C for another 16 hours. After purifying the ligations, we performed inverse PCR to amplify the barcodes with the associated genomic DNA region (primers CPL24-25; Table S4). We did 8 PCRs per pool, purified them and used 4ng of each pool for a further 2 rounds of PCRs to add Illumina sequencing adapters (CPL26-28, CPL6; 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 both the pBC and gBC in the proper sequence context were included in subsequent analysis. 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 Data S4 (patchMPRA) and Data S5 (episomal MPRA).

TRIP data processing

We obtained approximately 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 3 hamming distance 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 4 independent pools were combined in all analyses. Expression values can be found in 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 Data S2.

TRIP data analysis

We downloaded a list of expressed genes in K562 cells using whole cell long polyA RNA-seq data generated by ENCODE (*42*) from the EMBL Expression Atlas. We then designated the genes as hk or dev based on the list of hk genes obtained from Eisenberg and Levanon (*43*). Using the locations of these promoters (GENCODE Release 36, GRCh38.p13) we identified TRIP integrations located within 5kb 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 4 different promoters integrated within 5kb of each other (Full list of regions in 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 heatmaps were generated using the ComplexHeatmap package in R (*44*). To determine the diversity of the identified 5kb regions, we downloaded the 15-state segmentation for K562 (hg19) from the ENCODE portal and performed a liftover to hg38 using the UCSC liftover tool (*45*). We then overlapped the 5kb regions with chromHMM regions using a minimum overlap of 200bp using the Genomic Ranges R packages (*46*).

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 5kb genomic regions, we ran k-means clustering on the imputed data using the ConsensusClusterPlus package in R (*47*). 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 datasets (full list of sources in Table S5). For CpG methylation, we downloaded both replicates and used the averaged signal from both replicates. For H3K27ac, H3K4me3, PolII, CpG methylation and ATAC-seq, we used the EnrichedHeatmap package in R (*48*) to draw the metaplots for each cluster extending 5kb upstream and downstream of each genomic region. For CAGE-seq, we downloaded the hg19 dataset from the FANTOM5 consortium (*49*, *50*) and converted it to hg38 using the UCSC liftover tool (*45*). Because the signal was relatively sparse across genomic locations, we plotted the total CAGE signal across each genomic region.

Sequence features analysis

We obtained the sequences of each region using the BSgenome package in R (*51*). For the gapped k-mer predictions, we used the gkmSVM R package (*36*) with word length = 10 and number of informative columns = 6. We used AME for motif enrichment analysis (*52*), DREME for *de novo* motif discovery (*53*) and FIMO to determine the number of motifs per sequence (*54*), 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 polyA RNA-seq data generated by ENCODE (*42*) downloaded from the EMBL Expression Atlas.

To predict the type of genomic region of other integrations not in the defined 5kb regions, we obtained genomic sequences of the 1kb flanking region around the integration (500bp upstream and 500bp 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. Variance explained by each term was calculated with one-way ANOVAs of the respective models.

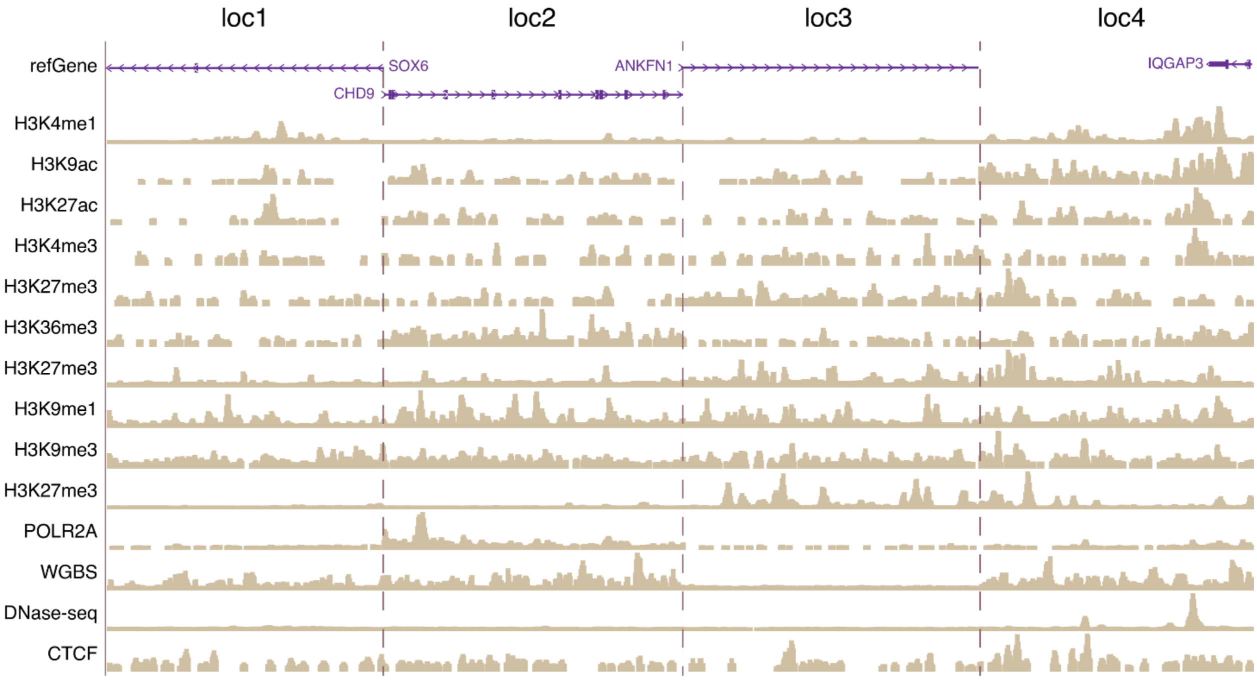


Fig. S1.

700 Landing pad locations have diverse chromatin marks and transcriptional activity.

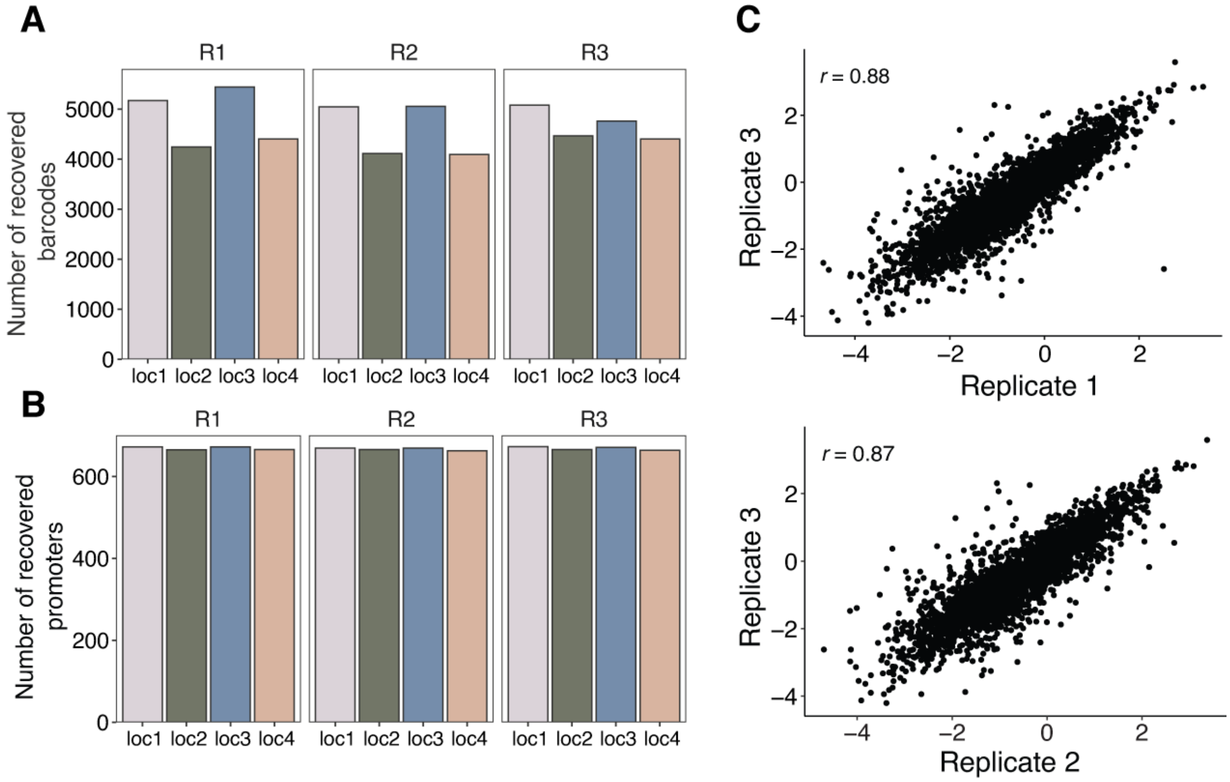
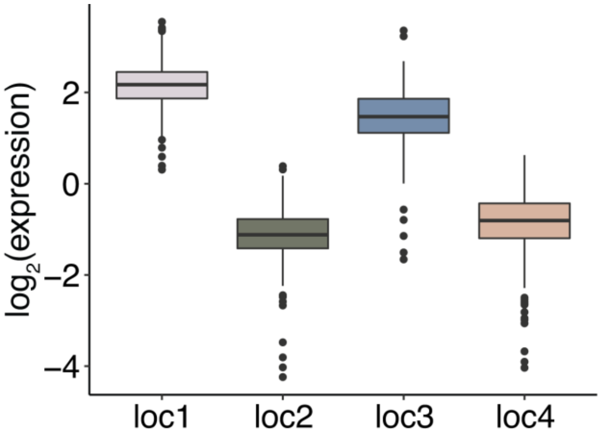


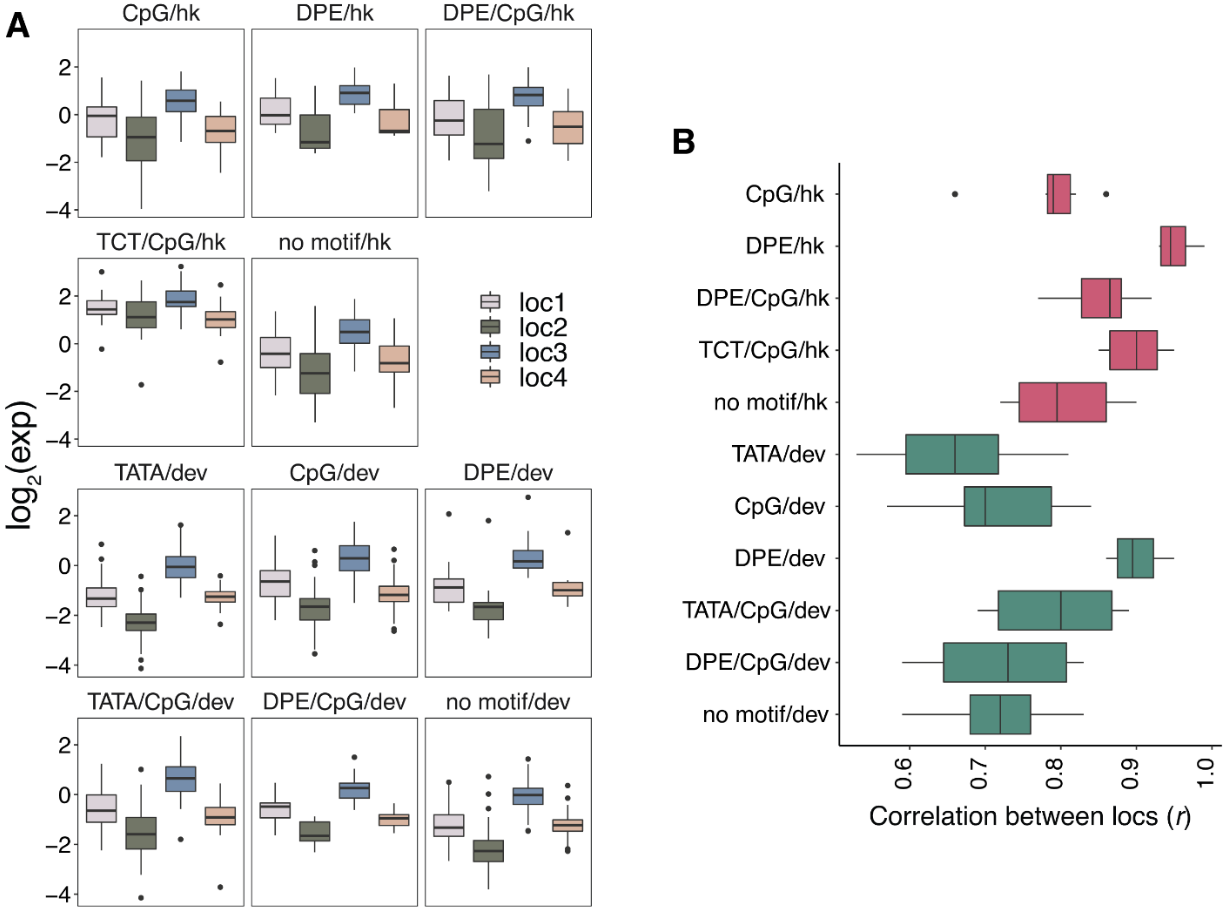
Fig. S2.

patchMPRA measurements are reproducible. (A) Number of promoter barcodes recovered from each location per biological replicate. **(B)** Number of promoters recovered from each location per biological replicate. **(C)** Reproducibility of core promoter measurements from independent patchMPRA transfections. 705



710 **Fig. S3.**

Expression of a library of proximal enhancers at each genomic location *(29)*.



715 **Fig. S4.**

Effect of genomic locations on core promoters. (A) Effect of genomic position and **(B)** all pairwise correlations (Pearson's *r*) between genomic locations for core promoters with different motifs within each class.

720

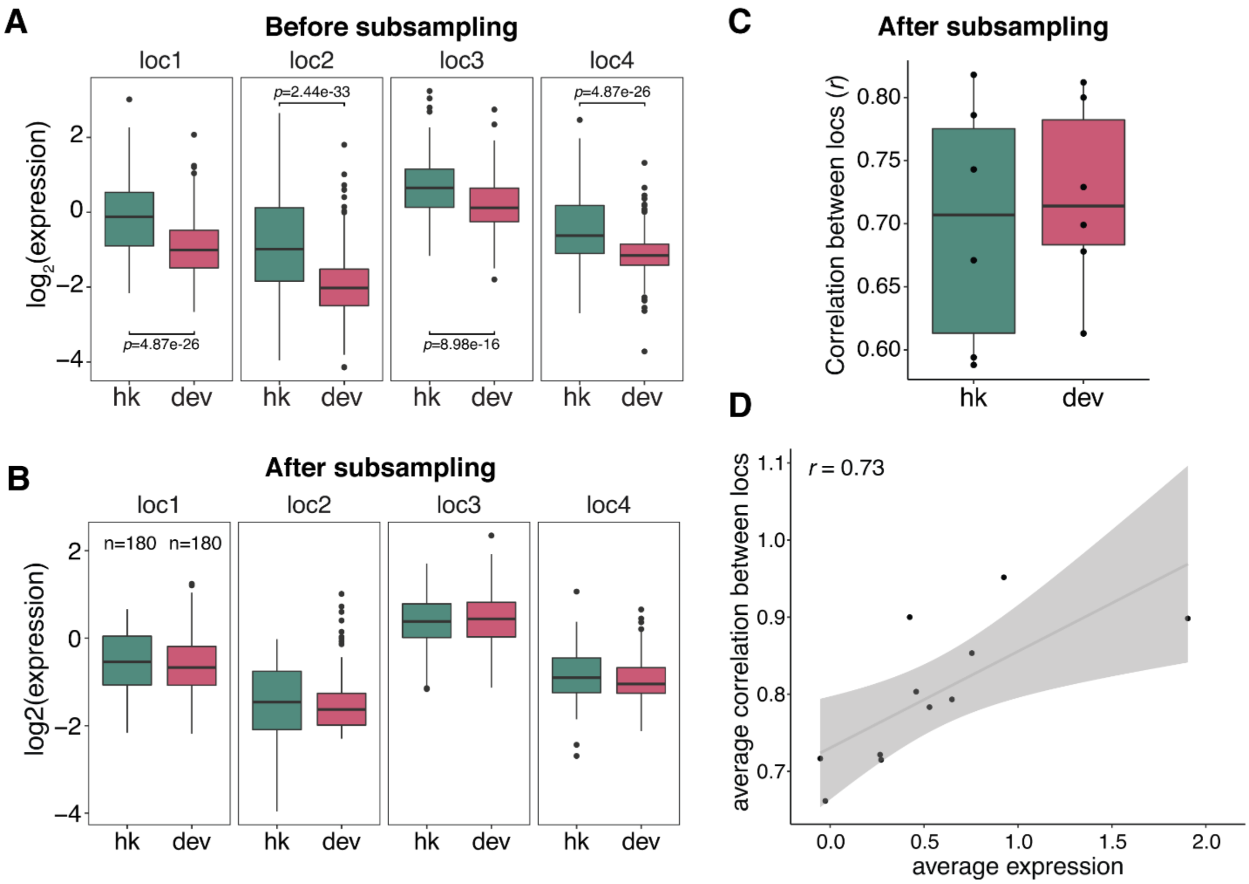


Fig. S5.

Intrinsic promoter strength explains differences between classes of core promoters. (A) Expression of all hk and dev promoters at each genomic location. *p*-values were calculated by Student’s *t*-tests. **(B)** Expression of hk and dev promoters at each genomic location after sampling promoters such that the two classes have equivalent average strengths. n indicates number of promoters sampled from each class. **(C)** All pairwise correlations (Pearson's *r*) between genomic locations for subsampled hk and dev core promoters. **(D)** The pairwise correlations of core promoters with different motifs (from Fig. S4B) are explained by the average expression of each group.

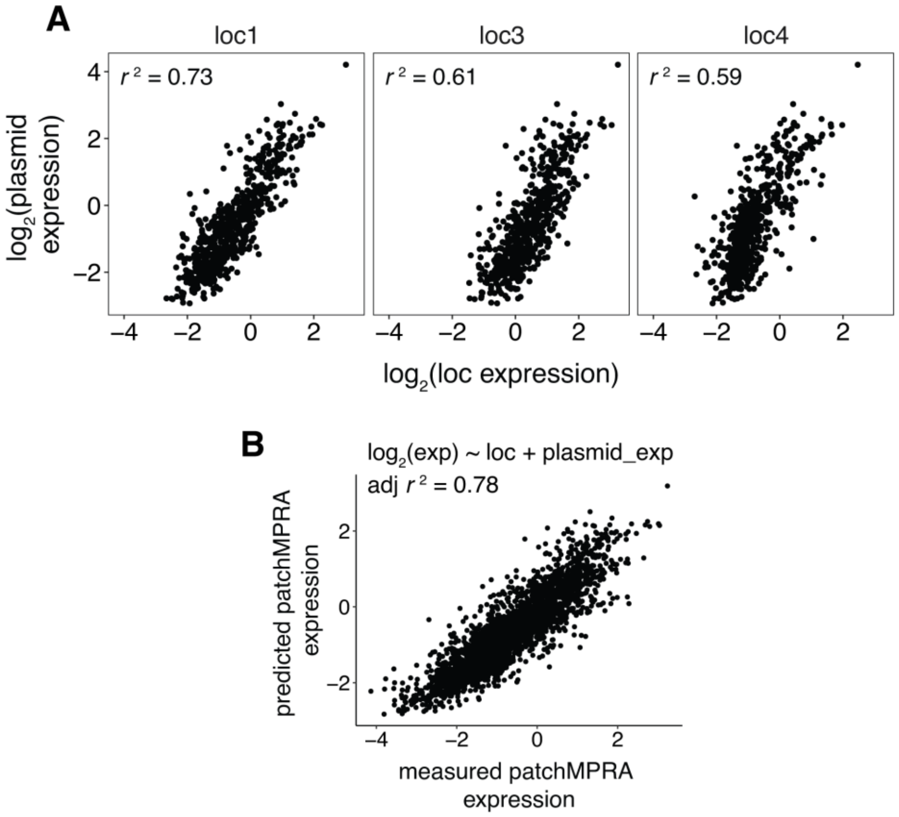


Fig. S6.

Core promoter activities in the genome reflect the promoters' intrinsic activity. (A) Correlations between expression of core promoter library measured on plasmids and at the indicated genomic location by patchMPRA. **(B)** Correlation between measured expression by patchMPRA and predicted expression by a linear model using core promoter intrinsic activity measured on plasmids.

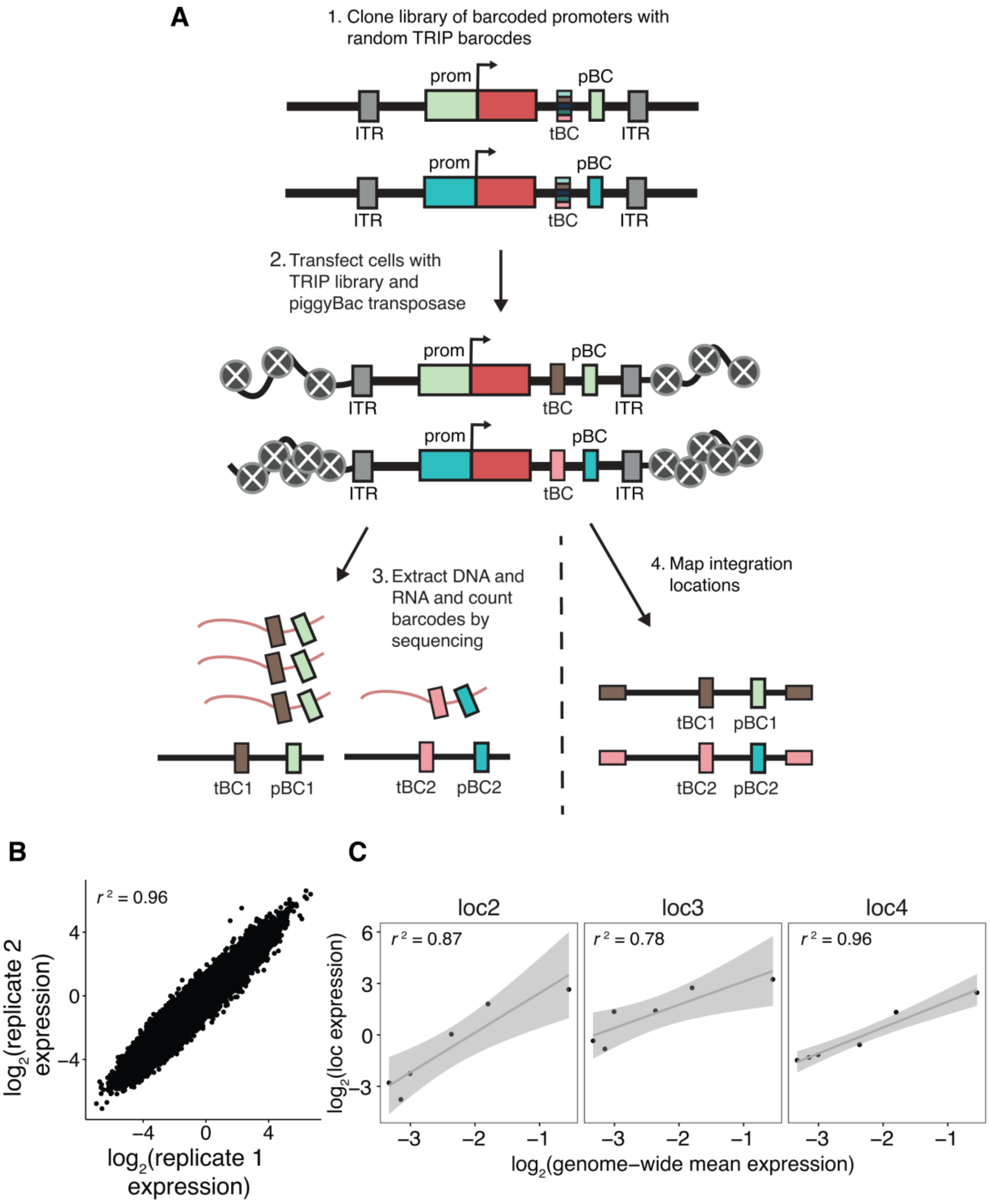


Fig. S7.

Measurements of six core promoters at thousands of genomic locations by TRIP. (A)

745 Schematic of TRIP experiment. tBC: TRIP barcode; pBC: promoter barcode; ITR: inverted terminal repeats. **(B)** Reproducibility between measurements from independent DNA and RNA

extractions. **(C)** Correlations between mean expression of core promoters measured by TRIP and at the indicated genomic location by patchMPRA.

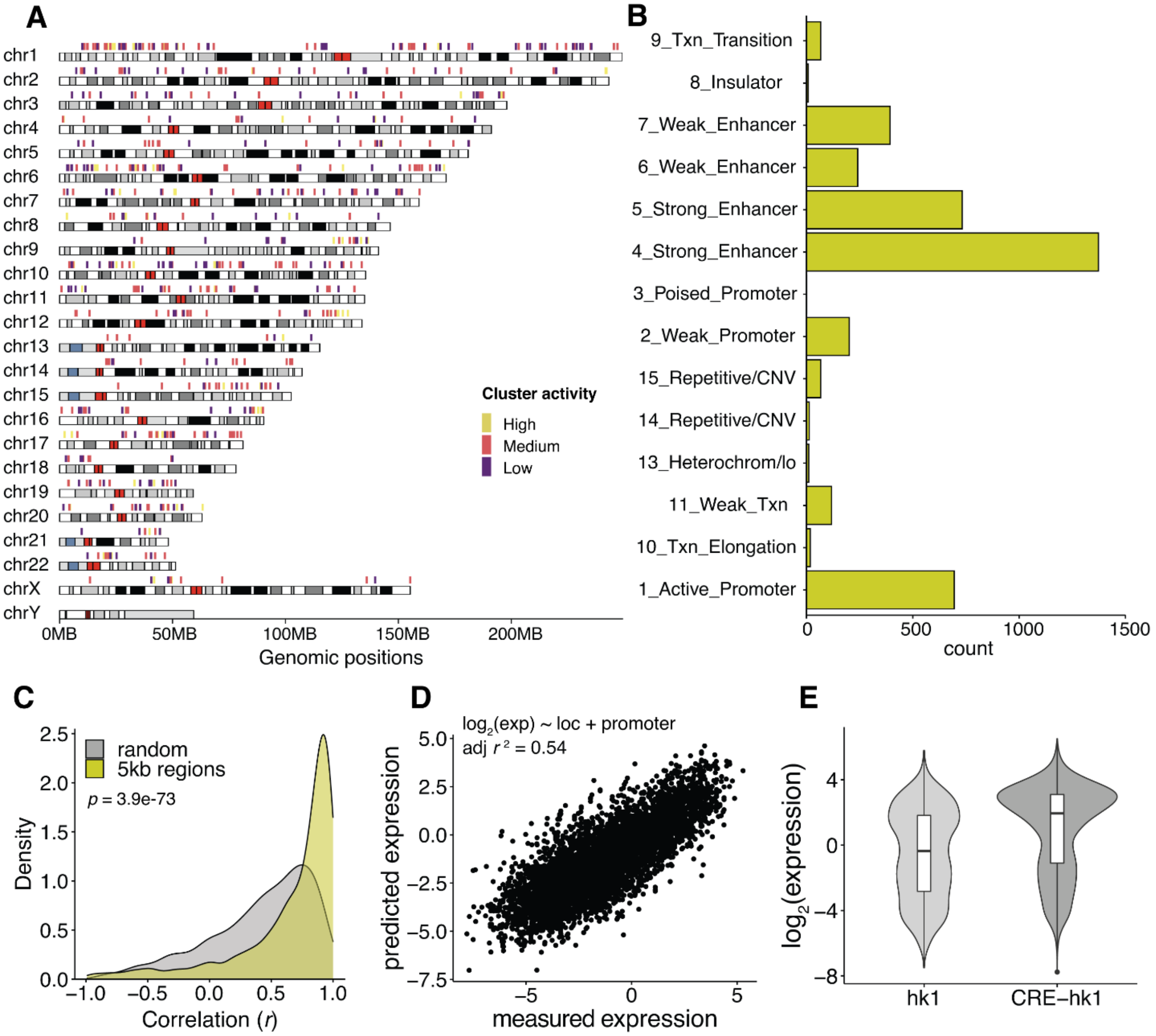
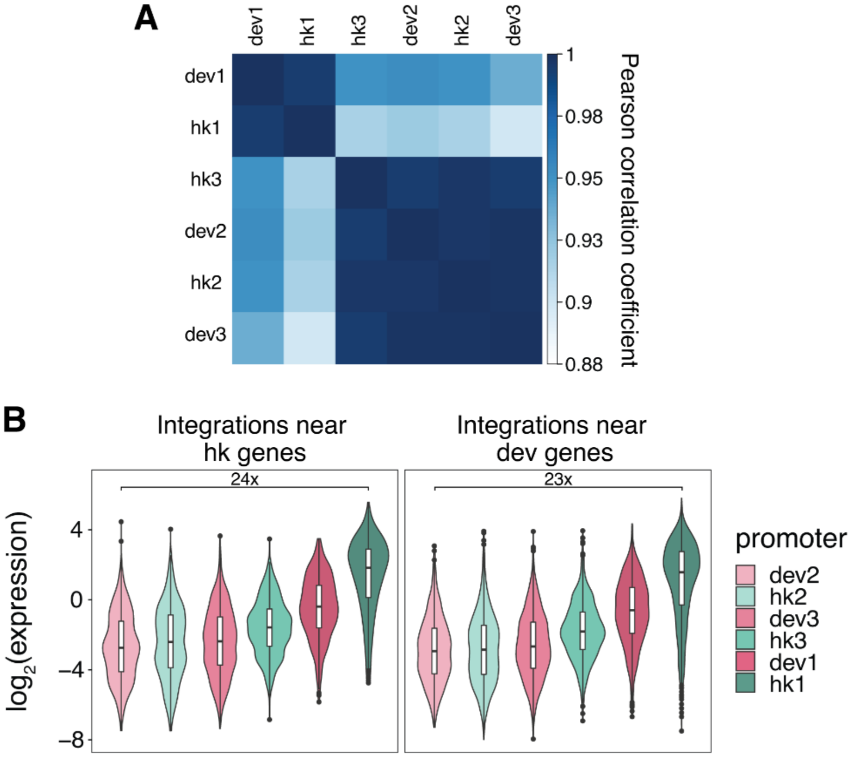


Fig. S8.

Core promoter scaling is a genome-wide phenomenon. (A) Regions with ≥4 different promoters integrated within 5kb of each other are located across the genome. Cluster activity was designated by the analysis in Fig. 5B. **(B)** Distribution of chromHMM annotations of defined 5kb regions. **(C)** For each defined 5kb region, correlations (Pearson's *r*) between core promoter activity measured by TRIP and by patchMPRA were calculated and all correlations were plotted as a density plot. As a comparison, we randomly grouped promoters without considering their integration locations and calculated the correlations for each group. The *p* value was calculated using the Mann–Whitney *U* test. **(D)** Correlation between measured expression by TRIP and predicted expression using a model assuming independence between genomic environments and core promoters. **(E)** Expression of all integrations of hk1 and hk1 with an upstream *cis*-regulatory enhancer (CRE-hk1).



765

Fig. S9.

Promoter strength, not class, determines its interaction with the genomic environment. (A) Correlation coefficients between curves fitted on each promoter in Fig. 5A. **(B)** Hk and dev integrated core promoters behave similarly near endogenous hk or dev promoters.

770

775

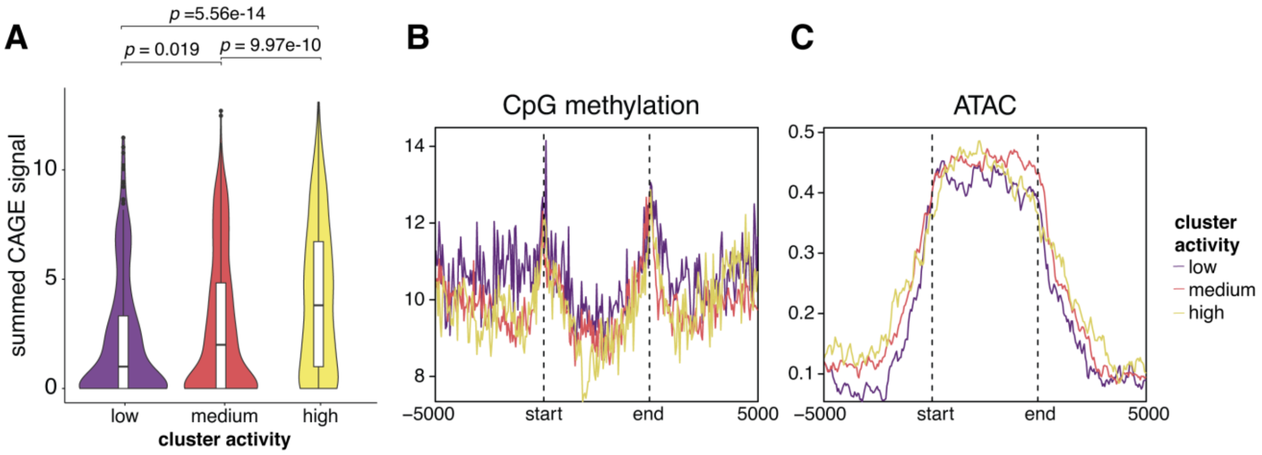


Fig. S10.

Epigenomic signatures of genomic clusters with different activities. (A) CAGE-seq signal

780 was calculated for each genomic region, and the summed signals were plotted for each cluster. *p* values were calculated by Student's *t*-tests. **(B, C)** Metaplots of CpG methylation and ATAC-seq signals respectively in each genomic cluster. The start and end mark the boundaries of each genomic region, which are determined by the first and last integration in the region. The x-axis extends +/- 5kb around each genomic region.

785

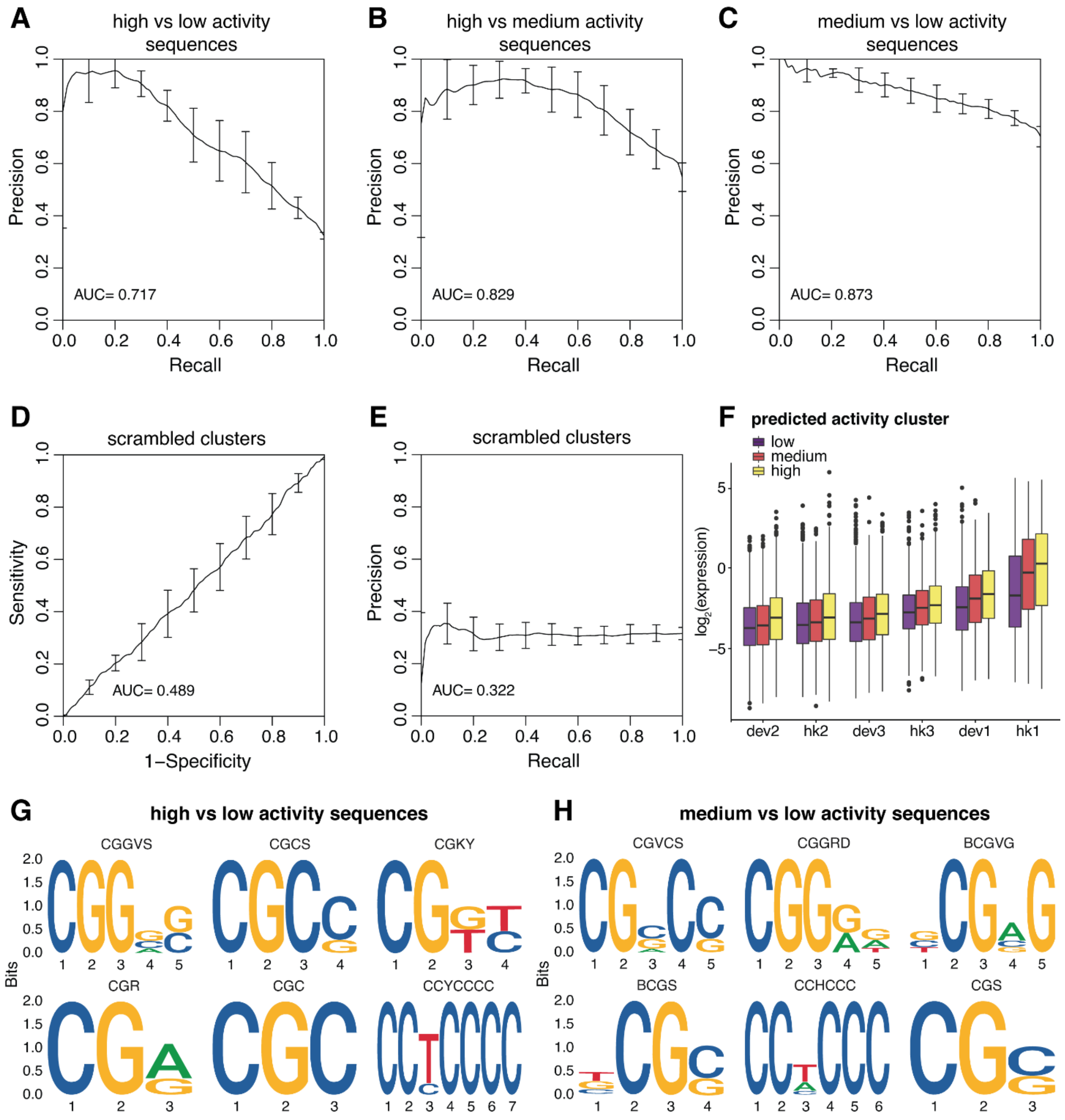


Fig. S11.

Sequence features of genomic clusters with different activities. (A, B, C) Performance of

790 gkmSVM used to classify sequences from different genomic clusters. Precision-recall curves (PRCs) were generated using five-fold cross-validation. **(D, E)** Performance of gkmSVM on sequences with scrambled cluster assignments. **(F)** TRIP integrations that were not included in the 5kb genomic region analysis were assigned to a cluster based on their sequence features from

the gkmSVM, and the expression of each promoter was plotted based on their predicted clusters.

(G, H) Top 6 motifs identified by *de novo* motif finding comparing high/low and medium/low activity sequences respectively.

Table S1.

Composition of promoter classes in the core promoter library.

Promoter class	**hk/dev**	**Number**
TATA-box	dev	100
CpG island	hk	100
CpG island	dev	100
TCT	hk	2
DPE	hk	8
DPE	dev	13
No known motif	hk	100
No known motif	dev	100
TATA-box & CpG island	dev	63
TATA-box & DPE	dev	2
DPE & CpG island	dev	14
DPE & CpG island	hk	43
TCT & CpG island	hk	22
SCP (and mutants)	-	4
Total		676

800

Table S2.

Locations of four landing pads in patchMPRA.

Name	**Chr**	**Location (hg19)**	**Annotation**	**Name in Maricque *et al.* (*29*)**
loc1	chr11	16,258,750	Sox6 Intron	LP3
loc2	chr16	53,275,015	CHD9 Intron	LP4
loc3	chr17	56,426,171	Intergenic	LP5
loc4	chr1	156,489,766	Intergenic	LP6

Table S3.

Promoters selected for TRIP.

Name	Oligo_id	Motifs/ Features	Number of TRIP integrations	Sequence
hk1	chr1_153963171_153963304_+	TATA, CpG, TCT	6032	GCACAAGATCCTTGCGTCATTTCCTGTAGTGTGCTCTATATAAGGGGCAGGATTTCCGCTTTCGCTCCTTTCCGGCGGTGACGACCTACGCACACGAGAACATGCCTGTGAGTGCTTTGGTCCAGGTTTCGGC
hk2	chr7_94285368_94285501_-	CpG	7157	GGGATGCTGATGCTGAACTGGCCAAGCTGGGAGGGAAGAAGAAAGGGAGGGGAGGGGAGAATCGAGGACGGACGGCCTAGCCAGGCCAAGAATGCAATTGCCCCGGTGGTGGGAGCTGGGAGACCCCTGTGCT
hk3	chr1_19638753_19638886_+	DPE	6851	GGGCGGGGCCTGCGGTTCCCGCGGGGGCGGTGGCGCGCGGTCAGCTGACCCGGCGGGCTTGACCCAGAAGCTGGGCCCTGGCGGCGGATCTGGACGTGGTGAGCCGGACCGGGGGCAGGTGGCAAACTTCAC
dev1	chr17_5522677_5522810_-	DPE	6328	CTCGCGATAGTGAGTGAGTTCCCACGAGATCTGATGGTTTTATAAGGGGCTTCCCCGTTACTCAGCACTTCTTCTCTCCTGCCGCCATTTGAAGGACGTGTCTGCTTCCACTCCTGCCGTGATTGTCAGCTTC
dev2	chr21_33976756_33976889_-	No known motif	7079	TATCTCCCGATCCTCACTGCCATCTGTGCTGCCAGCATTGGGCTCTTTCTCCTTTGAGAATTCTTTGCACTTCATTGTACTCCATGCTCAGTGCTGCTCACCGTCTGCTTTATAATACAGGCCACGGTGTGCT
dev3	chr1_1009619_1009752_-	TATA	7364	CTGAGGCTTGCGGCCACACCCTTGGCCCATAGGGTATAAATAGACCTGCTTGGGAGCCCACACCCAGCAACTCACACCTGCCTCAGACCAGAGCTCTGTGCGGGTGACGGCGCACGCATTCCTTGTGTCCCCG

Table S4.

Primers used in this study.

Name	Sequence	Description
CPL1	CCCCGTAATGCAGAAGAAGA	Amplify barcodes from integrated promoter library for Illumina sequencing.
CPL2	GCAGCGTATCCACATAGCGTAAAAG	Amplify barcodes from integrated promoter library for Illumina sequencing.
CPL3	CTTTCCCTACACGACGCTCTTCCGATCT(N$_{1-4}$)CATGGACGAGCTGTACAAGTAATCTAGA	Add first round of adapters to promoter library amplified barcodes. Variable numbers of Ns included to phase the library for Illumina sequencing.
CPL4	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT(N$_{0-3}$)GCGGCCGCTTTAGGATCC	Add first round of adapters to promoter library amplified barcodes. Variable numbers of Ns included to phase the library for Illumina sequencing.
CPL5	AATGATACGGCGACCACCGAGATCTACACNNNNNNACACTCTTTCCCTACACGACGCT	Add second round of adapters to promoter library amplified barcodes. N's indicated variable sequences for indexing.
CPL6	CAAGCAGAAGACGGCATACGAGATNNNNNNNNNGTGACTGGAGTTCAGACGTG	Add second round of adapters to amplified barcodes. N's indicated variable sequences for indexing.
CPL7	ACCATCTACATGGCCAAGAAGC	Amplify barcodes from episomal and TRIP library for Illumina sequencing.
CPL8	ATATCAGGCGCGCCAAGCTTGGATCCGCACAAGATCCTTGCGTC	Amplify hk1 from promoter library with homology to backbone for HiFi assembly.
CPL9	TCCTCGCCCTTGCTCACCATCCTAGGGCCGAAACCTGGACCAAA	Amplify hk1 from promoter library with homology to backbone for HiFi assembly.
CPL10	ATATCAGGCGCGCCAAGCTTGGATCCGGGATGCTGATGCTGAACTGG	Amplify hk2 from promoter library with homology to backbone for HiFi assembly.
CPL11	TCCTCGCCCTTGCTCACCATCCTAGGAGCACAGGGGTCTCCCAG	Amplify hk2 from promoter library with homology to backbone for HiFi assembly.
CPL12	ATATCAGGCGCGCCAAGCTTGGATCCGGGCGGGGCCTGCGGTTC	Amplify hk3 from promoter library with homology to backbone for HiFi assembly.
CPL13	TCCTCGCCCTTGCTCACCATCCTAGGGTGAAGTTTGCCACCTGCCCCCG	Amplify hk3 from promoter library with homology to backbone for HiFi assembly.
CPL14	ATATCAGGCGCGCCAAGCTTGGATCCCTCGCGATAGTGAGTGAG	Amplify dev1 from promoter library with homology to backbone for HiFi assembly.

CPL15	TCCTCGCCCTTGCTCACCATCCTAGGGAAGCTGACAATCACGGC	Amplify dev1 from promoter library with homology to backbone for HiFi assembly.
CPL16	ATATCAGGCGCGCCAAGCTTGGATCCTATCTCCCGATCCTCACTGCCA	Amplify dev2 from promoter library with homology to backbone for HiFi assembly.
CPL17	TCCTCGCCCTTGCTCACCATCCTAGGAGCACACCGTGGCCTGTA	Amplify dev2 from promoter library with homology to backbone for HiFi assembly.
CPL18	ATATCAGGCGCGCCAAGCTTGGATCCCTGAGGCTTGCGGCCACA	Amplify dev3 from promoter library with homology to backbone for HiFi assembly.
CPL19	TCCTCGCCCTTGCTCACCATCCTAGGCGGGGACACAAGGAATGCGTG	Amplify dev3 from promoter library with homology to backbone for HiFi assembly.
CPL20	GCTCTATAAGTAAGAGCTCTCGCTTCGAGTCTAGANNNNNNNNNNNNNNNNGATCACTCGAGTTGTGGCCGGCCCTT	Oligo for adding random barcodes to TRIP library by HiFi assembly,
CPL21	AACGCCAGGGTTTTCCCAA	Amplify barcodes from TRIP library for Illumina sequencing.
CPL22	CTTTCCCTACACGACGCTCTTCCGATCT(N_{1-4})CTCGCTTCGAGTCTAGA	Add first round of adapters to amplified TRIP barcodes. Variable numbers of Ns included to phase the library for Illumina sequencing.
CPL23	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCCAGGGTTTTCCCAAC	Add first round of adapters to amplified TRIP barcodes.
CPL24	CGCATGATTATCTTTAACGTACGTCAC	Amplify TRIP barcode and associated genomic region by inverse PCR.
CPL25	GCCAGGGTTTTCCCAAC	Amplify TRIP barcode and associated genomic region by inverse PCR.
CPL26	ACGACGCTCTTCCGATCTGCTCGAT(N_{0-3})GTACGTCACAATATGATTATCTTTCTAG	Add first round of adapters to TRIP amplified barcodes. Variable numbers of Ns included to phase the library for Illumina sequencing.
CPL27	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCCAGGGTTTTCCCAAC	Add first round of adapters to TRIP amplified barcodes. Variable numbers of Ns included to phase the library for Illumina sequencing.
CPL28	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT	Add second round of adapters to TRIP amplified barcodes. N's indicated variable sequences for indexing.

Table S5.

Sources of epigenome datasets used in this study.

Data	**Source Experiment**	**Source File**
H3K27ac ChIP-seq	ENCSR000AKP	ENCFF437DPT
H3K4me3 ChIP-seq	ENCSR000EWA	ENCFF916MPM
PolII ChIP-seq	ENCSR388QZF	ENCFF285MBX
CAGE	FANTOM5	chronic%20myelogenous%20leukemia%20cell%20line%3aK562.CNhs11250.10454106G4.hg19.ctss.bed.gz
CpG methylation	ENCSR765JPC	ENCFF867JRG; ENCFF721JMB
ATAC-seq	ENCSR868FGK	ENCFF698MIQ