nature genetics

ARTICLES

https://doi.org/10.1038/s41588-020-0686-2

Transcription imparts architecture, function and logic to enhancer units

Nathaniel D. Tippens[1,2,3,4,6], Jin Liang[1,6], Alden King-Yung Leung[1,2], Shayne D. Wierbowski[1,2], Abdullah Ozer[3], James G. Booth[5], John T. Lis[3,4] ✉ and Haiyuan Yu[1,2,4] ✉

Distal enhancers play pivotal roles in development and disease yet remain one of the least understood regulatory elements. We used massively parallel reporter assays to perform functional comparisons of two leading enhancer models and find that gene-distal transcription start sites are robust predictors of active enhancers with higher resolution than histone modifications. We show that active enhancer units are precisely delineated by active transcription start sites, validate that these boundaries are sufficient for capturing enhancer function, and confirm that core promoter sequences are necessary for this activity. We assay adjacent enhancers and find that their joint activity is often driven by the stronger unit within the cluster. Finally, we validate these results through functional dissection of a distal enhancer cluster using CRISPR–Cas9 deletions. In summary, definition of high-resolution enhancer boundaries enables deconvolution of complex regulatory loci into modular units.

Since their identification in viral and mammalian genomes, enhancers have been defined primarily by their function: the ability to activate promoters independently of their distance and orientation[1–3]. More basic questions about the nature of enhancer elements are difficult to answer. What are the genomic features of active enhancers? How large are they? Classical examples such as the α- and β-globin locus control regions offer some clues: these locus control regions are predominantly driven by 400–900 base-pair (bp) DNase I hypersensitive sites (DHSs) harboring transcription factor binding and extensive noncoding transcription[4,5]. Similar properties were also observed from all enhancers identified from a recent CRISPR–Cas9 screen of the *MYC* locus[6]. Histone modifications such as H3K27ac[7] and H3K4me1 (ref. [8]) have been proposed to mark enhancers, although such predictors lack systematic comparison[9–11]. Similarly, genome annotation tools such as ChromHMM[12] have been developed using histone modifications to generate enhancer predictions averaging 600 bp in size.

The finding that transcription from distal enhancers is widespread and corresponds with activation[13,14] led to numerous hypotheses about the roles and functions of noncoding 'enhancer' RNAs (eRNAs). Many long noncoding RNAs (lncRNAs) were thought to facilitate gene regulatory functions, but systematic introduction of premature polyadenylation signals into lncRNAs demonstrated that most of their RNA sequences are dispensable; instead, recruitment of a transcription machinery drives their gene regulatory activity[15,16]. Recently, a 'molecular stirring' model was proposed wherein transcription increases molecular motion to facilitate enhancer–promoter interactions[17]. Similarly, we have proposed that the affinity of RNA Polymerase II (RNAPII) for common cofactors or subunits might facilitate enhancer–promoter interactions[18,19]. This model is supported by reports that the C-terminal domain (CTD) of RNAPII specifies active promoter localization through its affinity for other CTDs[20], as well as the low-complexity domain of Cyclin T1 (ref. [21]). If correct, these models suggest that transcription is required for distal enhancer function, challenging the commonplace methodology of using DHSs and histone marks to identify enhancers. Indeed, a large-scale study using capped analysis of gene expression data indicated that eRNAs are more specific predictors of enhancer function than histone modifications[22]. However, capped analysis of gene expression fails to detect most eRNAs[13] and therefore cannot be used to assess the important question of whether all active enhancers are transcribed[23]. If enhancer transcription could be shown to be a ubiquitous feature of functional enhancers, then this would imply a structural architecture within enhancer sequences that requires not only binding sites for sequence-specific transcription factors, but also well-positioned core promoter sequences for assembly of the pre-initiation complex[24].

Numerous high-throughput sequencing methods identify enhancers using either plasmid or integrated reporter constructs and are collectively known as massively parallel reporter assays (MPRAs). While these assays offer unprecedented throughput for surveying genome function, their technical biases and limitations are a focus of ongoing research and optimization[25–27]. For example, most published MPRAs have been limited to short synthetic sequences (50–150 bp), despite the precise size of genomic enhancers being unknown[11]. The development of self-transcribing active regulatory region sequencing (STARR-seq) circumvented this limitation with a simple cloning strategy to quantify genomic fragments as large as 1,500 bp by placing them into the 3′ untranslated region (3′ UTR) of a reporter gene[2]. After transfecting cells with the reporter library, enhancers drive their own RNA expression. Each candidate's enhancer activity is then defined as the ratio of messenger RNA to plasmid DNA, as quantified by Illumina sequencing.

In this study, we performed systematic functional comparisons of histone marks to transcription initiation patterns that are frequently observed at enhancers. We discovered that transcription initiation is found at essentially all active distal enhancers and validated a basic unit model for enhancer sequences delineated by their transcription start sites (TSS). Finally, we surveyed dozens of genomic TSS

[1]Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, NY, USA. [2]Department of Computational Biology, Cornell University, Ithaca, NY, USA. [3]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA. [4]Tri-Institutional Training Program in Computational Biology and Medicine, Cornell University, Ithaca, NY, USA. [5]Department of Statistics and Data Science, Cornell University, Ithaca, NY, USA. [6]These authors contributed equally: Nathaniel D. Tippens, Jin Liang. ✉e-mail: johnlis@cornell.edu; haiyuan.yu@cornell.edu

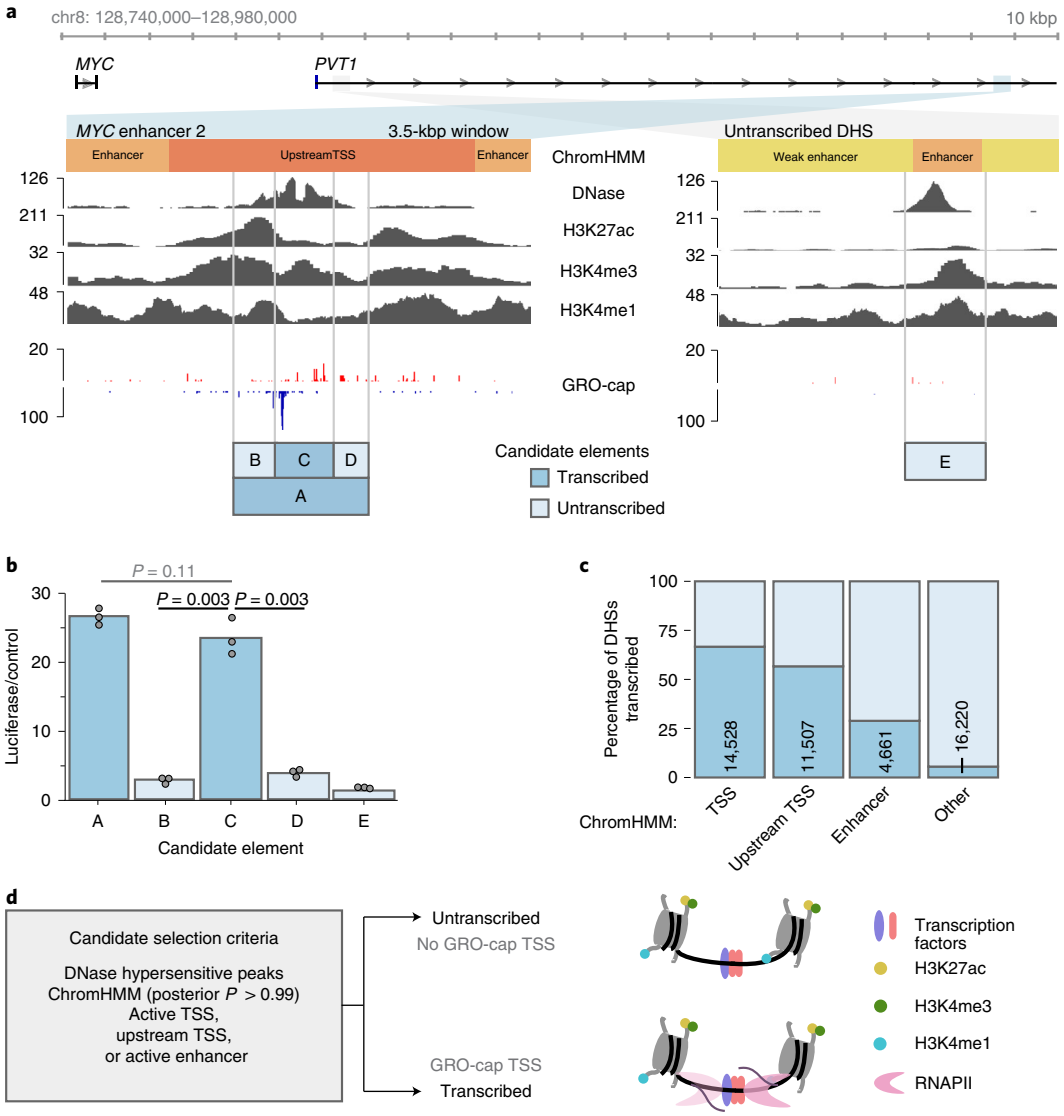


Fig. 1 | Divergent transcription identifies enhancer boundaries in high resolution. a, Features of two candidate regulatory elements in the *MYC* locus. Raw read counts are shown for each track and the 'candidate elements' track indicates the cloning boundaries used for the luciferase assays of tested sequences. **b**, Luciferase reporter activity for the regions indicated in **a** ($n = 3$ luciferase reactions). *P* values are from a one-sided *t*-test. **c**, The percentage of DHSs within each indicated ChromHMM class that are untranscribed (no GRO-cap TSS) versus transcribed (containing GRO-cap TSS). The number of transcribed DHSs are indicated. **d**, A schematic of candidate element selection using DNase hypersensitivity, ChromHMM and GRO-cap data. The molecular model illustrates DHSs sharing many features, with or without RNAPII transcription.

clusters with distal enhancer activity and revealed that their activity is primarily driven by a single dominant unit.

Results

Seven *MYC* enhancers that were recently identified by CRISPR–Cas9 interference exhibit many conventional features of active enhancer architecture[6]. For example, *MYC* enhancer 2 (element A) is a DHS and contains elevated levels of H3K27ac and H3K4me3 (Fig. 1a). It also contains a single divergent TSS pair. To test features critical for enhancer function, we subcloned element C from the larger element A previously verified by luciferase assays, as well as flanking sequences (elements B and D) for comparison. Notably, element C harbored virtually all observed distal enhancer activity in luciferase assays (Fig. 1b). A nearby site with similar DNase hypersensitivity and histone modifications that does not exhibit divergent transcription (element E) did not show enhancer activity. This example illustrates how divergent transcription may help localize active enhancer boundaries with high resolution, and avoid ambiguities derived from lower-resolution DNase hypersensitivity and chromatin immunoprecipitation (ChIP) profiles.

To generalize these results, we systematically sampled a larger set of candidate enhancers in K562 cells. This set was composed of DHSs from combinations of active ChromHMM classes[12] and transcription initiation classes defined by global run-on cap data[13] (GRO-cap; see Methods). Notably, most DHSs did not contain a GRO-cap TSS (86%). However, DHSs from the active enhancer, active TSS and upstream TSS ChromHMM classes were enriched for one or more GRO-cap TSS (Fig. 1c). We compared enhancer activity of transcribed and untranscribed DHSs from only high-confidence examples of these ChromHMM classes (Fig. 1d). Selected candidates ranged from 180 to 300 bp in size (Extended Data Fig. 1a).

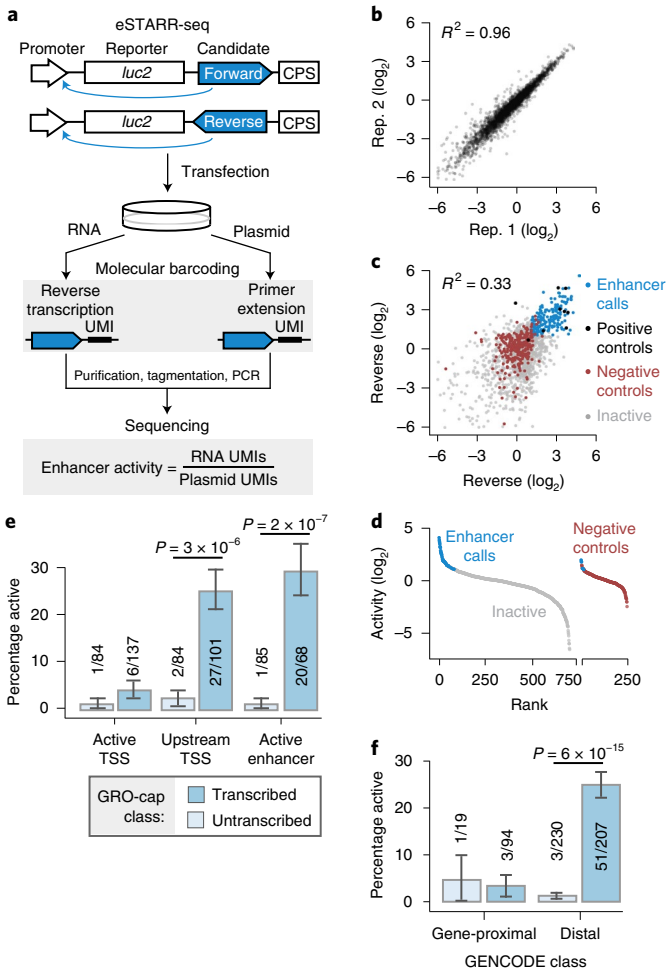


Fig. 2 | Transcription marks active eSTARR-seq enhancers. a, Outline of eSTARR-seq. Each candidate is inserted in the forward or reverse orientation before the cleavage/polyadenylation signal (CPS) of a reporter gene. After transfection, RNA and plasmids are purified separately. The addition of UMIs occurs during reverse transcription for RNA or primer extension for plasmids. After sequencing, enhancer activity is estimated by the ratio of RNA to plasmid UMIs. **b**, eSTARR-seq is highly reproducible between biological replicates. **c**, Comparison of activity from forward versus reverse cloning orientations. Data points are shown as $\log_2$ fold change versus negative controls. Positive controls are known *MYC* or viral enhancers (black). Negative controls are human ORFs (red). Elements with significantly elevated activity in both orientations are called enhancers (blue). Remaining candidates are called inactive (gray). **d**, Summary of enhancer calls from **c** after averaging forward and reverse activities. The empirical FDR is 2.4% (6 out of 243 negative controls misidentified as enhancers). **e**,**f**, Within each ChromHMM (**e**) or distance (**f**) class, the percentage of active enhancers identified by eSTARR-seq is indicated. Protein-coding gene annotations are from GENCODE. The error bars indicate the s.e.m. calculated for a sample of binary trials, centered on the observed success rate. *P* values are from a two-sided Fisher's exact test.

Divergent transcription marks active enhancer elements. To test hundreds of candidate enhancer sequences across broad length scales, we adapted STARR-seq for use with sequence-verified candidate elements, which we call element-STARR-seq (eSTARR-seq; Fig. 2a and Extended Data Fig. 1b,c). We cloned every candidate sequence in both forward and reverse orientations within the 3′ UTR of the reporter gene to distinguish sequences that may regulate mRNA stability. We added unique molecular identifiers (UMIs, 12 nucleotides (nt)) to the reverse transcription primer for removal of PCR duplicates and tagmentation before Illumina sequencing to circumvent length limitations and minimize bias (Fig. 2a; Methods). As in other MPRAs, enhancer activity is quantified as the ratio of mRNA to transfected DNA (after de-duplication with UMIs). eSTARR-seq improves agreement with luciferase data compared with conventional STARR-seq (Extended Data Fig. 1b), probably because UMIs increase the dynamic range, and is highly reproducible from true biological replicates (Fig. 2b). We note that more recent human STARR-seq protocols may track luciferase more robustly[26]. Finally, we measured the relationship between fragment size and reporter activity (Extended Data Fig. 1c) using negative control sequences. We selected human ORFs unlikely to destabilize mRNA or harbor distal enhancer activity as negative controls (Methods). In conclusion, eSTARR-seq enables robust quantification of enhancer activity while minimizing PCR, size and orientation biases.

Enhancer activity is known to be orientation-independent[1,3], whereas mRNA stability is affected by strand-specific RNA sequences. Thus, we required candidates to exhibit significantly higher reporter activity than controls in both forward and reverse cloning orientations to be classified as an enhancer (Fig. 2c; Methods). Only 2.5% (6 out of 243) of negative controls met these criteria, confirming very few false-positive enhancer calls (Fig. 2d).

Comparing transcribed and untranscribed DHSs revealed that essentially all eSTARR-seq enhancers were found in transcribed DHSs, although rarely within the active TSS class (Fig. 2e). Within the upstream TSS and active enhancer ChromHMM classes, 25–30% of transcribed candidates exhibited significant enhancer activity. By contrast, only approximately 2% of untranscribed candidates exhibited significant enhancer activity, in line with the false-positive rate estimated from the negative controls (2.5%; see Fig. 2d). GRO-cap provides similar predictive performance without ChromHMM after using a 500-bp distance cutoff from GENCODE annotations to distinguish gene promoters from distal enhancers (Fig. 2f). We further confirmed these results with the standard STARR-seq promoter, the mammalian synthetic core promoter (SCP1; Extended Data Fig. 2). Our results strengthen previous associations between transcription and enhancer activity[10,22,28,29], provide compelling evidence that essentially all active enhancers are transcribed and suggest a functional role for transcription from active enhancers.

Transcription delineates regulatory sequence architecture. Given the striking co-occurrence of transcription initiation and active enhancer elements, we revisited the model that promoters and enhancers share a universal architecture[13,30] (Fig. 3a). Classic studies defined minimal 'core promoter' sequences that coordinate assembly of the pre-initiation complex; in this study, we defined core promoters as beginning 32 bp upstream of the TSS (the location of transcription factor II D ($TF_{II}D$) binding to the TATA box motif when present) and ending at the RNAPII pause site (≤60 bp beyond the TSS[19]). Two distinct core promoters were found up to 240 bp apart (that is, 300 bp between TSS) and may help position the −1 and +1 nucleosomes[31]. By contrast, the 'upstream region' contains regulatory transcription factor motifs that may activate one or both core promoters.

To illustrate similarities in architecture at both promoters and enhancers genome-wide, we plotted motif densities relative to the stronger TSS (or 'maximum TSS' from the pair) at both gene-proximal and gene-distal TSS pairs (Fig. 3b). Briefly, we sorted divergent TSS pairs by width and computed motif densities around all pairs containing a motif from −400 to +100 bp from the maximum TSS (see Methods). Interestingly, some motifs were well aligned to TSS, especially those known to recruit and position TFIID. Similar to the well-known TATA box bound by TBP (maximum motif density at −32 bp), SP1 (ref. [24]) (at −53 bp) and STAT2

(ref. [32]) (−5 bp) show striking TSS alignment and are known to recruit TFIID. Systematic classification of core promoter sequences is particularly important since <10% of human TSS contain a TATA box and recent reports demonstrated how core promoters respond differently to coactivators and distal enhancers[24,33,34]. However, most motifs appeared dispersed throughout the 'upstream region' between divergent TSS, as illustrated by PU.1, JUND and GATA1 (Fig. 3b). By contrast, the CTCF and ZNF143 motifs are found near the weaker TSS. Notably, CTCF and ZNF143 have been implicated in facilitating distal loop interactions, reinforcing the idea that similar motif alignments identify similar regulatory roles. Whereas ChIP sequencing (ChIP–seq) analyses can only reveal central and core promoter binding transcription factors[13], sequence motif analyses reveal more nuanced spatial preferences within these elements[35].

We retested a subset of elements after adding sequence context on each side to test whether core promoter boundaries are sufficient to capture enhancer activity (TSS + 60 bp versus TSS + 200 bp). Importantly, adding sequence context affected enhancer activity less than testing identical fragments in differing orientations (Fig. 3c $R^2 = 0.53$ compared with Fig. 2c $R^2 = 0.33$). This indicates that enhancer activity appears to be generally captured with sequences extending 60 bp beyond divergent TSS, thus providing a basic unit definition of enhancers. In summary, we validated a boundary definition of individual enhancer units and revealed motif alignments that might help decipher regulatory function[34–36].

Enhancers require core promoters for activity. Next, we sought to determine whether all components of the divergent TSS model (Fig. 3a) are necessary to drive distal enhancer activity. Previous studies found significant conservation of core promoter sequences at distal enhancers[22], but this conservation could be driven by selection for promoter function[15,23]. We reasoned that if transcription is spurious or unimportant to enhancer activity, core promoter sequences should be dispensable. To test this hypothesis, we recloned 13 eSTARR-seq enhancers to 'delete' (by omission) each of their core promoter regions, defined as −35 to +60 bp from the TSS (Fig. 4a). Since each enhancer contains a divergent pair of TSS, we compared the effect of deleting either the maximum TSS (defined from the GRO-cap signal) or the weaker 'minimum TSS'. Deletion of a TSS resulted in at least twofold reduced activity from 9 out of 13 enhancers (Fig. 4b,c). Interestingly, these enhancers could depend on the maximum or minimum TSS, or both. These results demonstrate that core promoter regions significantly contribute to enhancer activity.

Next, we compared enhancer TSS to the gene-proximal TSS included in our study. eSTARR-seq enhancer TSS produce significantly less GRO-cap signal than promoters, but there is not enough separation between the populations for this feature alone to distinguish them (Fig. 4d,e). Additionally, the divergent TSS within eSTARR-seq enhancers are not significantly less directional than gene promoters, as quantified by the ratio between maximum and minimum TSS signal (Fig. 4f). Together, these results demonstrate that enhancers' core promoter regions contribute to function but are not easily distinguishable from gene promoter TSS.

Comparison to a genome-scale STARR-seq dataset. To confirm our findings, we reanalyzed the 'High-resolution Dissection of Regulatory Activity' (HiDRA) dataset[37], which uses the STARR-seq assay on analysis of transposase-accessible chromatin (ATAC–seq) fragments. This impressively comprehensive dataset from GM12878 cells quantifies enhancer activity from 100–600-bp fragments enriched within DHSs, thus dissecting potential enhancer elements genome-wide. Given our observations of pronounced orientation effects in STARR-seq assays (Fig. 2c), we attempted to remove this bias wherever possible. Unfortunately, most HiDRA fragments (87%) did not share ≥90% overlap with a fragment tested in the opposite orientation (Extended Data Fig. 3a). We assessed orientation bias across all 763,373 fragment pairs tested in both orientations and found very little agreement across orientations (Extended Data Fig. 3b; HiDRA $R^2 = 0.07$). Interestingly, HiDRA fragments that contain a DHS exhibited less orientation bias (Extended Data Fig. 4a; $R^2 = 0.38$), closely matching our eSTARR-seq results ($R^2 = 0.33$; Fig. 2c).

Importantly, accounting for orientation bias in STARR-seq datasets has substantial impact on enhancer identification. While 93% of HiDRA fragment pairs appeared inactive (Extended Data Fig. 3b, quadrant I), the 7% of fragment pairs with elevated RNA/DNA signal (quadrants II–IV) are dominated by orientation bias (quadrants II and III): only 19% of these fragment pairs exhibited elevated activity in both cloning orientations (quadrant IV; Extended Data Fig. 3c). This is true even when only considering fragments that span a DHS, with 71.2% of enhancers exhibiting orientation dependence ($n = 580$ out of 827 enhancer fragment pairs; Extended Data Fig. 4a). Interestingly, most transcribed DHSs showed enrichment for orientation-dependent activity (Extended Data Fig. 4b). When using a stringent orientation-independent enhancer criterion, HiDRA identified only 0.22% of tested fragments as enhancers, although this should be greatly improved by selection of larger fragments to increase capture of whole elements.

GM12878 HiDRA fragments containing enhancer units defined by divergent TSS were most enriched in the active enhancer ChromHMM category (Extended Data Fig. 3d), confirming our observations in K562 cells (Fig. 2d). To determine if one or both core promoter sequences are necessary for enhancer activity, we evaluated the fraction of HiDRA enhancers around unpaired GRO-cap TSS. At unpaired TSS, the upstream and core promoter regions can be easily separated for functional analysis (Extended Data Fig. 3e). Strikingly, we observed little enrichment for orientation-independent enhancers from upstream or TSS regions alone, while activity was enriched within fragments containing both the TSS and upstream regions (Extended Data Fig. 3e). These results demonstrate that core promoter sequences within TSS regions are necessary for distal enhancer activity and strongly suggest a functional role for RNAPII recruitment to enhancers. Our findings are reminiscent of recent dissections of promoter activity and provide strong support for similar architectures at promoters and enhancers[13,30], although they each exhibit clearly distinct functions (Fig. 2e and Extended Data Fig. 3d,e).

Since TSS functionally contribute to enhancer activity, we directly compared enhancer activity to transcription levels. We found no correlation between GRO-cap signal and eSTARR-seq activity (Extended Data Fig. 5a), although we caution that this analysis compared different contexts (genomic and episomal). We also compared enhancer TSS histone modifications to those of gene promoters. As expected, enhancers identified from the eSTARR or HiDRA datasets exhibited elevated H3K4me1 and H3K27ac, but reduced H3K4me3 levels (Extended Data Fig. 5b,c, top). To estimate if these differences might be explained by transcriptional activity, we computed the ratio between each histone modification and transcription measured by GRO-cap. Interestingly, the H3K4me3 to transcription ratio did not differ between promoters and enhancers, whereas H3K27ac and H3K4me1 ratios were higher at enhancers than promoters (Extended Data Fig. 5b,c, bottom). Together, these results suggest a complex relationship between histone modifications, transcription and enhancer activity.

Dissection of compact enhancer clusters with eSTARR-seq. Many gene-distal TSS are found in dense regulatory clusters that have complex histone modification patterns[19], implying widespread clustering of basic enhancer units. To explore how individual enhancer units might cooperate within these clusters, we fit a model to predict the enhancer activity of a cluster from the activities of

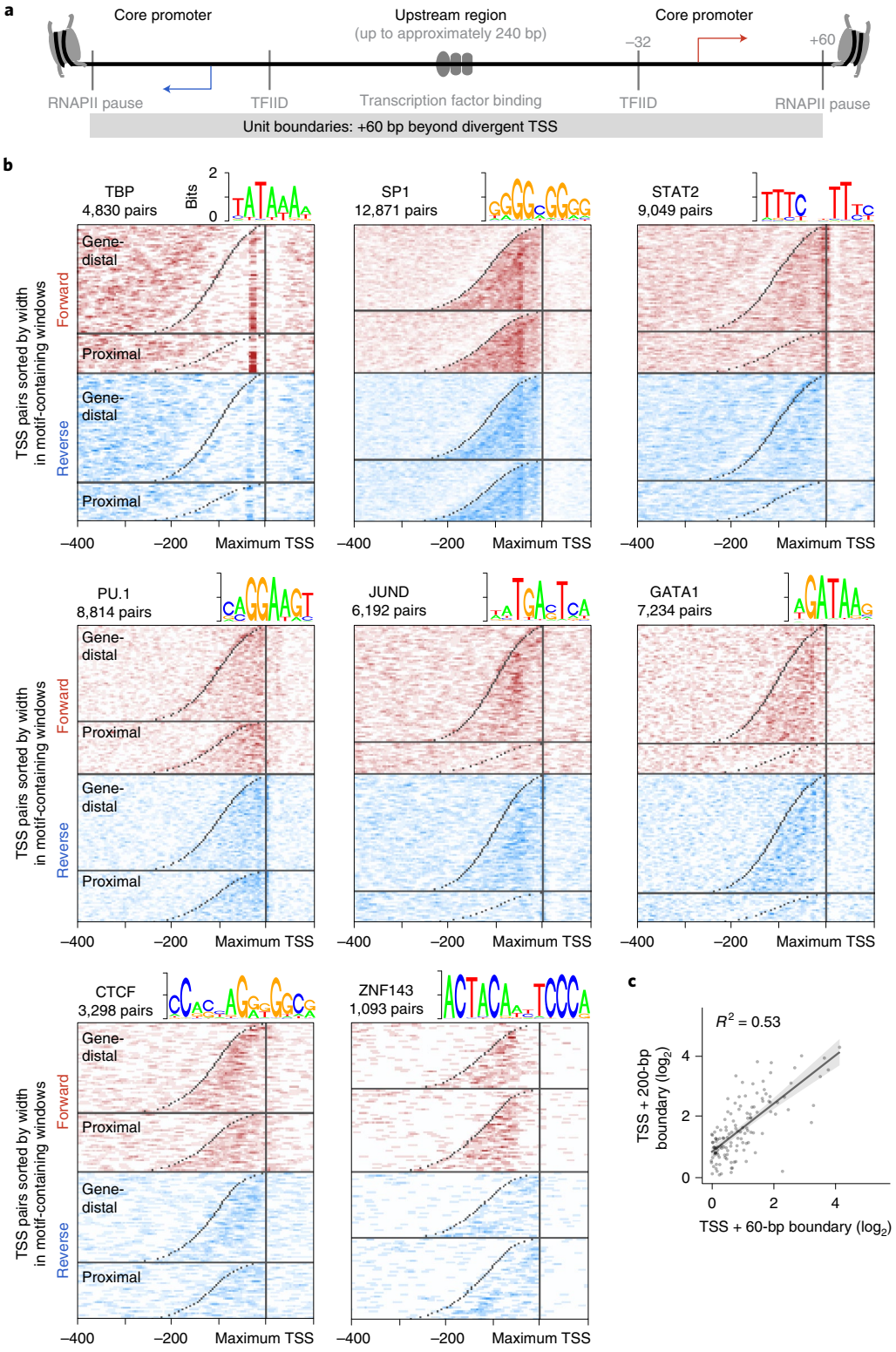


Fig. 3 | Enhancer unit boundaries reveal sequence architecture. a, Illustration of a unified model for regulatory sequence architecture of promoters and enhancers. Core promoter motifs (TBP, SP1, STAT2) surround an upstream region containing transcription factor motifs. We defined core promoters as the region from TFIID binding 32 bp upstream of each TSS, to the RNAPII pause sites at +60 bp from each TSS. **b**, Divergent TSS pairs were sorted by width and aligned to the maximum TSS. TSS pairs were also divided by GENCODE class (gene-distal versus proximal). The heatmaps indicate transcription factor motif densities from pairs containing at least 1 motif within −400 to +100 bp of the maximum TSS. Motifs are shown in both forward (red) and reverse (blue) orientations relative to the maximum TSS. TSS positions are marked in gray. **c**, Comparison of enhancer activities for the same set of elements using TSS + 60 bp and TSS + 200 bp cloning boundaries. The overlay shows linear regression with the 95% confidence interval shaded gray ($n = 93$ candidate element pairs).

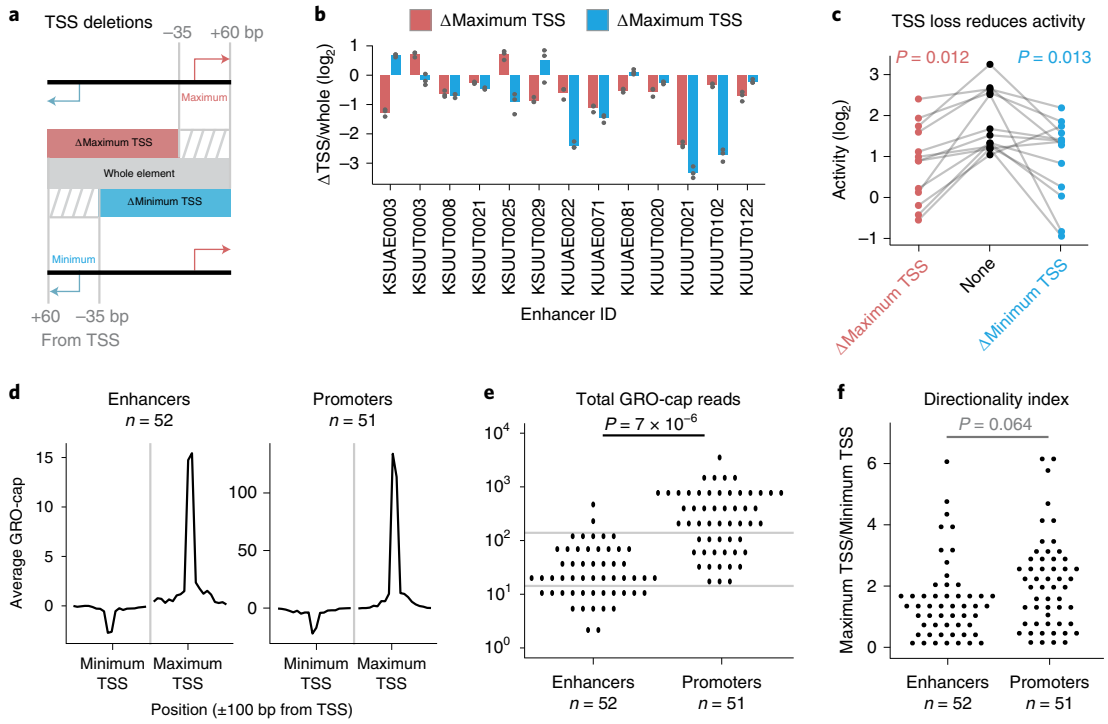


Fig. 4 | Function and features of enhancer TSS. a, Boundary definitions for whole elements (gray box) and TSS deletions (red and blue boxes). The stripes indicate 'deleted' regions. **b**, Change in eSTARR-seq activity after deleting either the maximum TSS (red) or minimum TSS (blue; $n=3$ transfections). **c**, Plot of element activities after TSS deletion ($n=13$ enhancers). *P* values are from a one-sided paired *t*-test. **d**, Average profiles of GRO-cap signal from eSTARR-called enhancers versus promoters. Note the tenfold difference in the *y* axis scales. **e**,**f**, Dot plot of TSS signal (**e**) and directionality index (**f**) at enhancers versus promoters. The gray lines emphasize substantial overlap between enhancer and promoter distributions. *P* values are from a one-sided *t*-test.

its units (Fig. 5a). One hundred clusters and associated units were successfully cloned so that their enhancer activity could be quantified independently within the same experiment; 45% of clusters showed significant enhancer activity compared with negative controls (Extended Data Fig. 6a), and predominantly contained a single active sub-element (Extended Data Fig. 6b).

We fit a linear model to predict cluster activities (interaction model; Fig. 5b) from the activities of the observed units (e_1 and e_2, where $e_1 > e_2$) and an interaction term ($e_1 \times e_2$). Strikingly, this analysis revealed significant covariance between cluster activity and the unit with higher activity (e_1, $P=0.01$), but not the unit with lower activity (e_2). Indeed, including only the unit with higher activity (maximum model) explains 77.2% of the observed variance (Fig. 5b), which was not significantly less than the interaction model ($P=0.31$). This suggests that genomic enhancer clusters are predominantly driven by a single active unit but afforded little insights into cooperativity between multiple active units.

To directly assess cooperativity between active units, we generated synthetic pairs made by randomly fusing eSTARR-seq active enhancer units (Fig. 5c). We developed a pooled strand overlap extension PCR strategy to fuse units into random pairs linked with a constant 25-bp sequence. This method generated 188 fusions, 69 of which were pairs of active enhancer units (Extended Data Fig. 7a). Individual units were retested in the same pool as the fused sequences and their eSTARR-seq activities agreed well with previous measurements (Extended Data Fig. 7b). Surprisingly, the interaction model including both units still did not find statistically significant predictive power from the weaker unit and failed to outperform the maximum model (Fig. 5d; $P=0.28$), demonstrating that proximity to a stronger enhancer effectively abolishes the activity of weaker enhancers. The maximum model explains 49.2% of the variance among active enhancer pairs and 39.2% of the variance among all enhancer-containing pairs ($n=86$; Extended Data Fig. 7c). As expected, the maximum model does not perform well for pairs lacking any enhancer activity, explaining only 17.6% of the variance ($n=33$; Extended Data Fig. 7d). These results demonstrate that immediate proximity of enhancer units in DNA often allows only the strongest enhancer to function and may therefore be used to select for the maximum activity from neighboring enhancer units.

Dissection of the endogenous *NMU* enhancer cluster. We sought to test our TSS-based definition of enhancer boundaries in the genomic context by targeting the distal enhancer of *NMU* ('e*NMU*'), which was reported to exhibit a large effect after homozygous deletion without impeding cell growth[38]. Published datasets revealed elevated levels of DNase hypersensitivity, H3K27ac, H3K4me3 and H3K4me1 at this element, and we identified two candidate enhancer units based on the pattern of GRO-cap TSS (Fig. 6a). Episomal luciferase assays suggested similar behavior as other genomic clusters we previously dissected with eSTARR-seq (Fig. 5b): a single dominant unit (e_1) driving the activity of the cluster (Fig. 6b). To confirm this behavior in the genomic context, we transiently transfected K562 cells with plasmids expressing Cas9 and pairs of guide RNAs targeting the boundaries of each indicated candidate element. We obtained e*NMU* deletion lines as controls[38] and established new clonal lines for genotyping by genomic PCR to ensure successful homozygous deletions (Extended Data Fig. 8). To estimate effect size from each clone, we performed quantitative PCR (qPCR) with reverse transcription and computed *NMU* expression compared to wild-type (WT) cells (Fig. 6c). We also computed *NMU* expression relative to e*NMU* deletion (Δe*NMU*; Fig. 6c, right axis) to directly

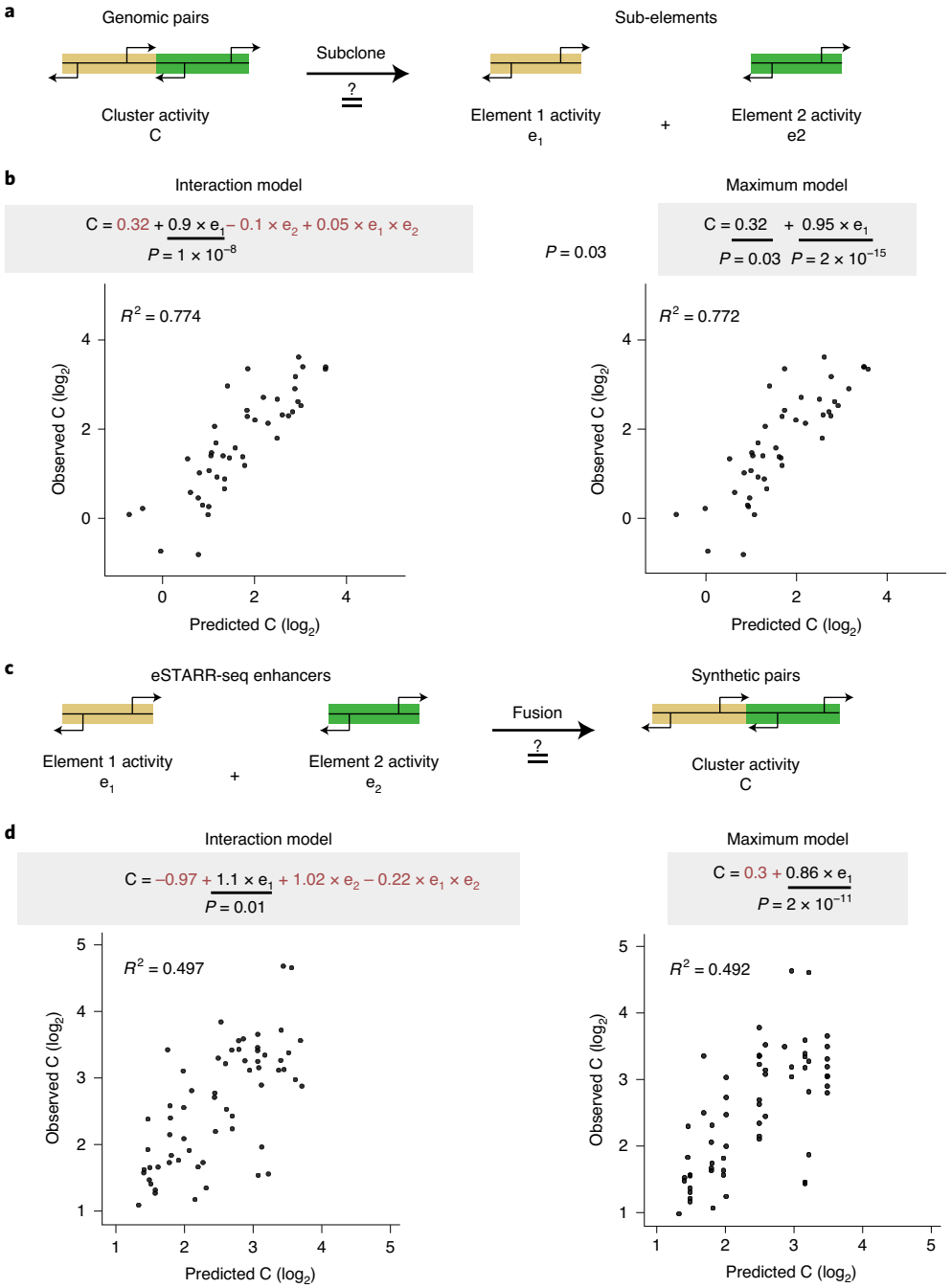


Fig. 5 | Functional dissection of adjacent enhancers. a, Dissection of genomic TSS clusters into individual sub-elements to quantify enhancer cooperativity. **b**, Two linear models were fitted to eSTARR-seq measurements of full clusters (C) and individual enhancers within the cluster (e_1 and e_2). The interaction model includes both individual enhancers and an interaction term, while the maximum model only considers the stronger sub-element (chosen to be e_1). Fitted equations are shown with significant covariates underlined and nonsignificant covariates colored red. The interaction model was a linear regression with 42 d.f. ($F = 40.1$). The maximum model was linear regression with 44 d.f. ($F = 144$). Comparing both models with a one-way analysis of variance (ANOVA), $F = 1.93$ and $P = 0.158$, thus indicating similar performance. **c**, Schematic illustrating fusion of active enhancer sequences into synthetic enhancer pairs. **d**, Fitting of same linear models as **b** to enhancer activities of individual elements and their synthetic fusion (as shown in **c**). The interaction model was linear regression with 62 d.f. ($F = 23$). The maximum model was linear regression with 64 d.f. ($F = 67$). Comparing both models with a one-way ANOVA, $F = 0.997$ and $P = 0.375$, thus indicating similar performance.

estimate endogenous enhancer activity. From this perspective, WT e*NMU* drives *NMU* expression almost 10,000×, as reported previously[38]. Deletion of the full cluster C (ΔC) or the stronger unit (Δe_1) revealed complete loss of enhancer activity, confirming that TSS boundaries define enhancer units within dense TSS clusters. Surprisingly, e_2 deletion (Δe_2) resulted in 3–5% of WT *NMU* expression, indicating that e_1 alone cannot fully recapitulate activity. e_1 maintained enhancer function in the absence of e_2 (100× over Δe*NMU*), confirming its role as the 'dominant' enhancer within this cluster, but nevertheless exhibited multiplicative cooperativity[39]

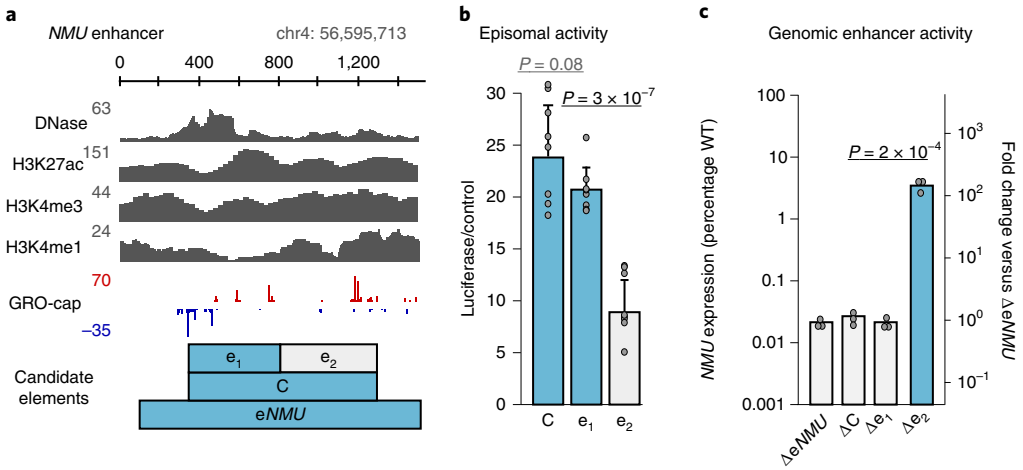


Fig. 6 | Dissection of the *NMU* enhancer. a, Dissection of the TSS cluster within the *NMU* enhancer (e*NMU*). Cluster 'C' contains two distinct candidate sub-elements, e_1 and e_2. The presence of e_1 is indicated with blue throughout the figure. **b**, Normalized luciferase activity of the candidate cluster and sub-elements using the *MYC* promoter ($n = 5$ luciferase reactions). **c**, Quantification of *NMU* expression from the indicated homozygous Cas9 deletion clones ($n = 3$ PCR replicates). Representative Δe*NMU* and Δe_2 expression clones are shown from $n = 5$ clonal lines; ΔC and Δe_1 are from $n = 1$ clonal line. All error bars indicate s.d. centered on the mean. All *P* values are from a two-sided *t*-test.

with e_2 not detected by episomal assays. These results validate enhancer unit boundaries defined by TSS, confirm that a dominant unit often drives activity within dense enhancer clusters[40] and identify important differences between episomal and genomic reporter assays.

Discussion

Although transcription and histone modifications are closely correlated[8,11,13], we find that histone marks offer lower resolution for defining active enhancers compared to transcription initiation patterns provided by GRO-cap[13,41]. We further demonstrate that TSS are useful anchors in revealing motif positioning within enhancers and enable dissection of regulatory clusters into individual units.

Previous analyses of conserved enhancers across species found widespread transcription factor motif rearrangements that did not impact function, leading to a 'flexible' sequence model for enhancers that was only evaluated with promoter-proximal MPRAs[42,43]. Using the distal enhancer design of STARR-seq, we find that enhancer activity requires at least one core promoter in addition to transcription factor binding in the flexible upstream region, suggesting a functional role for RNAPII recruitment at enhancers. Likewise, recent analyses of population variants affecting gene-distal GRO-cap TSS suggest that core promoter mutations in distal enhancers can disrupt enhancer function[28]. The requirement for core promoters at enhancers is particularly intriguing given reports that core promoters confer specificity for enhancers and coactivators[24,33,34]; this suggests enhancers could target promoters by recruiting a similar core promoter machinery. Additionally, RNAPII pausing at enhancers[10] may facilitate distal interactions through the affinity of the CTD for other CTDs[20], resulting in coordinated pause release at promoters and associated enhancers by recruitment of P-TEFb kinase[44]. Further analysis of regulatory architectures at promoters and enhancers may expand the lexicon for noncoding elements beyond individual transcription factor motifs and clarify enhancer–promoter interaction specificities and mechanisms.

eSTARR-seq resulted in a relatively modest validation rate of approximately 25% for gene-distal GRO-cap candidate elements. We reason that this might be explained by low reporter sensitivity or the need to screen different promoter types[33]. Additionally, it is unlikely that all elements exhibiting bidirectional transcription carry enhancer activity: consistent with previous studies[2,26,29], we find few human promoters with distal enhancer activity, despite their bidirectional transcription. This observation highlights remaining questions about the distinguishing features of these two regulatory elements. In general, promoters and enhancers have been reported to differ in guanine-cytosine content and transcription factor recruitment preferences, but such rules lack specificity[30]. Core promoter sequence features might help distinguish enhancers from promoters, particularly if RNAPII itself reads a regulatory code during pausing or early elongation. For example, RNAPII pausing is sequence-dependent[19,45] and is substantially longer-lived at promoters than enhancers[10]. Stable RNAPII pausing at promoters may provide time to recruit distal regulatory complexes by colocalization with the unstable RNAPII pausing seen at enhancers. Finally, transcriptional burst size is thought to be encoded within core promoter sequences[46]. Promoters may undergo selection for larger burst sizes, whereas enhancers maximize burst frequency to drive distal gene activation[47].

Genomic enhancer clusters have recently been dissected resulting in different models of their cooperativity[40,48,49]. Analysis of these datasets demonstrated that both reports are consistent with multiplicative generalized linear models[39] although statistical power was greatly constrained by sample size. While these studies assessed cooperativity over significant distances (2–50 kb), we assayed dozens of adjacent enhancer pairs (≤600 bp apart) and fitted a single multiplicative (or log-additive) linear model to explain their cumulative activity. Our episomal dataset surveys a larger number of clusters and indicates that a single active unit often drives cluster activity. We validated this dominant unit model at the e*NMU* cluster, where deletion of the e_1 unit abolishes all enhancer activity. Although e_2 is unable to enhance *NMU* expression without e_1, it exhibits multiplicative amplification of e_1 (20× increase). We speculate that this may be mechanistically explained by a 5′ splice site that can dramatically boost enhancer activity[15], or hierarchical behavior[40] where the accessibility and/or transcription of e_2 depends on e_1. A recent report of TSS 'switching' within developmental enhancer clusters[50] underscores the need for further TSS-based interrogation of enhancer units. If confirmed on a larger scale, TSS-based enhancer definition can reduce complex regulatory programs into simple, modular units.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-020-0686-2.

Received: 12 August 2019; Accepted: 28 July 2020;
Published online: 21 September 2020

References

1. Serfling, E., Jasin, M. & Schaffner, W. Enhancers and eukaryotic gene transcription. *Trends Genet.* **1**, 224–230 (1985).
2. Arnold, C. D. et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
3. Canver, M. C. et al. *BCL11A* enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527**, 192–197 (2015).
4. Tuan, D., Solomon, W., Li, Q. & London, I. M. The 'beta-like-globin' gene domain in human erythroid cells. *Proc. Natl Acad. Sci. USA* **82**, 6384–6388 (1985).
5. Orkin, S. H. Regulation of globin gene expression in erythroid cells. *Eur. J. Biochem.* **231**, 271–281 (1995).
6. Fulco, C. P. et al. Systematic mapping of functional enhancer–promoter connections with CRISPR interference. *Science* **354**, 769–773 (2016).
7. Creyghton, M. P. et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl Acad. Sci. USA* **107**, 21931–21936 (2010).
8. Heintzman, N. D. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).
9. Dorighi, K. M. et al. Mll3 and Mll4 facilitate enhancer RNA synthesis and transcription from promoters independently of H3K4 monomethylation. *Mol. Cell* **66**, 568–576.e4 (2017).
10. Henriques, T. et al. Widespread transcriptional pausing and elongation control at enhancers. *Genes Dev.* **32**, 26–41 (2018).
11. Kellis, M. et al. Defining functional DNA elements in the human genome. *Proc. Natl Acad. Sci. USA* **111**, 6131–6138 (2014).
12. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
13. Core, L. J. et al. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46**, 1311–1320 (2014).
14. Kim, T.-K. et al. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).
15. Engreitz, J. M. et al. Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539**, 452–455 (2016).
16. Joung, J. et al. Genome-scale activation screen identifies a lncRNA locus regulating a gene neighbourhood. *Nature* **548**, 343–346 (2017).
17. Gu, B. et al. Transcription-coupled changes in nuclear mobility of mammalian *cis*-regulatory elements. *Science* **359**, 1050–1055 (2018).
18. Tippens, N. D., Vihervaara, A. & Lis, J. T. Enhancer transcription: what, where, when, and why? *Genes Dev.* **32**, 1–3 (2018).
19. Tome, J. M., Tippens, N. D. & Lis, J. T. Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers. *Nat. Genet.* **50**, 1533–1541 (2018).
20. Lu, F., Portz, B. & Gilmour, D. S. The C-terminal domain of RNA polymerase II is a multivalent targeting sequence that supports *Drosophila* development with only consensus heptads. *Mol. Cell* **73**, 1232–1242.e4 (2019).
21. Lu, H. et al. Phase-separation mechanism for C-terminal hyperphosphorylation of RNA polymerase II. *Nature* **558**, 318–323 (2018).
22. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
23. Andersson, R. Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. *Bioessays* **37**, 314–323 (2015).
24. Vo Ngoc, L., Wang, Y. L., Kassavetis, G. A. & Kadonaga, J. T. The punctilious RNA polymerase II core promoter. *Genes Dev.* **31**, 1289–1301 (2017).
25. Inoue, F. et al. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* **27**, 38–52 (2017).
26. Muerdter, F. et al. Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat. Methods* **15**, 141–149 (2017).
27. Klein, J. et al. A systematic evaluation of the design, orientation, and sequence context dependencies of massively parallel reporter assays. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/576405v1 (2019).
28. Kristjándsóttir, K. et al. Population-scale study of eRNA transcription reveals bipartite functional enhancer architecture. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/426908v2 (2018).
29. Mikhaylichenko, O. et al. The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. *Genes Dev.* **32**, 42–57 (2018).
30. Andersson, R., Sandelin, A. & Danko, C. G. A unified architecture of transcriptional regulatory elements. *Trends Genet.* **31**, 426–433 (2015).
31. Scruggs, B. S. et al. Bidirectional transcription arises from two distinct hubs of transcription factor binding and active chromatin. *Mol. Cell* **58**, 1101–1112 (2015).
32. Paulson, M., Press, C., Smith, E., Tanese, N. & Levy, D. E. IFN-stimulated transcription through a TBP-free acetyltransferase complex escapes viral shutoff. *Nat. Cell Biol.* **4**, 140–147 (2002).
33. Zabidi, M. A. et al. Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**, 556–559 (2015).
34. Haberle, V. et al. Transcriptional cofactors display specificity for distinct types of core promoters. *Nature* **570**, 122–126 (2019).
35. Grossman, S. R. et al. Positional specificity of different transcription factor classes within enhancers. *Proc. Natl Acad. Sci. USA* **115**, E7222–E7230 (2018).
36. Yang, X. & Vingron, M. Classifying human promoters by occupancy patterns identifies recurring sequence elements, combinatorial binding, and spatial interactions. *BMC Biol.* **16**, 138 (2018).
37. Wang, X. et al. High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat. Commun.* **9**, 5380 (2018).
38. Gasperini, M. et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* **176**, 377–390.e19 (2019).
39. Dukler, N., Gulko, B., Huang, Y.-F. & Siepel, A. Is a super-enhancer greater than the sum of its parts? *Nat. Genet.* **49**, 2–3 (2016).
40. Shin, H. Y. et al. Hierarchy within the mammary STAT5-driven *Wap* super-enhancer. *Nat. Genet.* **48**, 904–911 (2016).
41. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**, 950–953 (2013).
42. Smith, R. P. et al. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat. Genet.* **45**, 1021–1028 (2013).
43. Vierstra, J. et al. Mouse regulatory DNA landscapes reveal global principles of *cis*-regulatory evolution. *Science* **346**, 1007–1012 (2014).
44. Boehning, M. et al. RNA polymerase II clustering through carboxy-terminal domain phase separation. *Nat. Struct. Mol. Biol.* **25**, 833–840 (2018).
45. Shao, W., Alcantara, S. G. & Zeitlinger, J. Reporter-*ChIP*-nexus reveals strong contribution of the *Drosophila* initiator sequence to RNA polymerase pausing. *Elife* **8**, e41461 (2019).
46. Larsson, A. J. M. et al. Genomic encoding of transcriptional burst kinetics. *Nature* **565**, 251–254 (2019).
47. Fukaya, T., Lim, B. & Levine, M. Enhancer control of transcriptional bursting. *Cell* **166**, 358–368 (2016).
48. Hay, D. et al. Genetic dissection of the α-globin super-enhancer in vivo. *Nat. Genet.* **48**, 895–903 (2016).
49. Huang, J. et al. Dynamic control of enhancer repertoires drives lineage and stage-specific transcription during hematopoiesis. *Dev. Cell* **36**, 9–23 (2016).
50. Kim, H. S. et al. Pluripotency factors functionally premark cell-type-restricted enhancers in ES cells. *Nature* **556**, 510–514 (2018).

Methods

Dual luciferase assays. The selected transcriptional response elements were individually cloned into eSTARR-seq assay vectors via LR reactions and the resulting plasmids were extracted with the E.Z.N.A. Endo-Free Plasmid Mini Kit II (catalog no. D6950; Omega Bio-tek). The plasmids were electroporated into K562 cells with the Ingenio Electroporation Kit (MIR 50115; Mirus Bio). For each electroporation, 0.5 million cells were mixed with 1–2 µg of plasmids and 50 µl of Ingenio Electroporation Solution and electroporated with a Nucleofector II *device* (Lonza) using program T-016. The pGL4.75 vector (catalog no. E6931; Promega Corporation) was co-electroporated (10 ng per electroporation) as the internal control. The electroporated K562 cells were recovered in 2 ml of culture medium at 37 °C with 5% CO_2 until collection.

The electroporated cells were collected after 24 h of recovery for the dual luciferase assay. The assay was carried out with the Dual-Glo Luciferase Assay System (catalog no. E2920; Promega Corporation) according to the manufacturer's instructions. An Infinite M1000 Microplate Reader (catalog no. 30034301; Tecan) was used to quantify the luminescence signals. Cells electroporated with only the pGL4.75 vector or with only the pDEST-hSTARR-luc-Pmyc vector were used as the background controls for firefly or *Renilla* luciferase activities, respectively.

Candidate element selection and definition. To systematically compare transcribed and untranscribed candidates within each ChromHMM class, we focused on high-confidence active TSS, upstream TSS and active enhancer predictions (posterior $P > 0.99$). This set of regions was then filtered by requiring overlap with Encyclopedia of DNA Elements (ENCODE) DHS peaks from K562 cells. Finally, ChromHMM regions were classified as either transcribed or untranscribed by overlapping with GRO-cap divergent peaks (from the supplementary files of Core et al.[13]). In total, 251 untranscribed regions were cloned using DHS peak coordinates as boundaries. Similarly, 305 transcribed regions were cloned using boundaries 60 bp downstream of each divergent TSS (TSS + 60 bp), where the TSS position is the maximum GRO-cap signal within the peak (see Extended Data Fig. 1a for element sizes within each class). TSS + 200 bp elements were cloned using boundaries 200 bp downstream of each divergent GRO-cap TSS.

As negative controls, we selected 250 sequence-verified human ORFs ranging from 600 to 2,000 bp in size. These coding sequences were screened for any exonic DHS and/or GRO-cap TSS. As positive controls, we included HS001, HS002, HS005, HS006, HS018 (ref. [2]), *MYC* E1-7 (ref. [6]) and a collection of viral promoters/enhancers (cytomegalovirus, Rous sarcoma virus and simian vacuolating virus 40).

Element cloning and input plasmid library preparation. The primers for the cloning elements were designed in batch with a webtool[51] and synthesized by Eurofins. Each primer contained a 5′-overhang, attB1′ for the forward primers and attB2′ for the reverse primers. All primer sequences used in this work can be found in Supplementary Table 1. Human genomic DNA was used as template for the PCR reactions. The amplicons were cloned into the pDONR223 vector via Gateway BP reactions. The resulting single-colony-derived entry clones were verified by Illumina sequencing as described previously[51].

All verified element clones were propagated in a lysogeny broth (LB) medium supplemented with spectinomycin. The culture was then pooled together for plasmid extraction with the E.Z.N.A. Plasmid Midi Kit (catalog no. D6904; Omega Bio-tek). The elements were cloned into the eSTARR-seq assay vector via en masse Gateway LR reactions to generate the input plasmid library. The input library was propagated in LB medium supplemented with ampicillin and the plasmids were extracted with the E.Z.N.A. Endo-Free Plasmid Maxi Kit (catalog no. D6926; Omega Bio-tek).

eSTARR-seq assay vector. The eSTARR-seq assay vectors were generated by modifying the original STARR-seq vector[2]. To engineer the pDEST-hSTARR -luc-Pmyc vector, the SCP1 promoter in the STARR-seq vector was replaced with the *MYC* promoter[6] and the truncated SuperGlo green fluorescent protein was replaced with a luciferase reporter gene (*luc2*). Additionally, the two cloning sites and the DNA fragment between them in the STARR-seq vector were replaced with an attR1-attR2 Gateway cassette. To engineer the pDEST-hSTARR-luc-Pmyc-ccw vector, the attR1-attR2 Gateway cassette in the pDEST-hSTARR-luc-Pmyc vector was removed and then recloned back to its original position in the reverse orientation. Additionally, we generated pDEST-hSTARR-luc and pDEST-hSTARR-luc-ccw vectors that are almost identical to pDEST-hSTARR-luc-Pmyc and pDEST-hSTARR-luc-Pmyc-ccw, respectively, except that the SCP1 promoter[2] was used instead of the *MYC* promoter.

Cell culture. The K562 cells (CCL-243) were purchased from ATCC. The cells were maintained in a culture medium composed of IMDM (catalog no. 30-2005; ATCC) supplemented with 10% FBS (catalog no. 30-2020; ATCC) at 37 °C with 5% CO_2. Cells used for different biological replicates were cultured separately.

eSTARR-seq library preparation. The input library plasmids were electroporated into the K562 cells with the Cell Line Nucleofector Kit V (catalog no. VVCA-1003; Lonza). For each electroporation, 1 million cells were mixed with 20 µg of plasmids and 100 µl of supplemented Nucleofector Solution V (Lonza) and electroporated with a Nucleofector II *device* using program T-016. The electroporated K562 cells were recovered in 2 ml of culture medium at 37 °C with 5% CO_2 until collection.

The electroporated K562 cells were collected after 6 h of recovery. Total RNA was extracted from the cells with TRIzol Reagent (catalog no. 15596026; Thermo Fisher Scientific) according to the manufacturer's instructions. Reverse transcription was performed with the total RNA as the template using SuperScript III Reverse Transcriptase (catalog no. 18080044; Thermo Fisher Scientific). The electroporated plasmids were extracted from the cells as described previously[52]. The first primer extension was performed with the extracted plasmids as the template. In parallel, another primer extension reaction was carried out with the input plasmid library used for transfection as the template. Reactions were treated with exonuclease I (catalog no. M0293S; New England BioLabs) to remove excess single-stranded primer, followed by purification with DNA Clean & Concentrator-5 (catalog no. D4013; Zymo Research).

The second primer extension was performed with the products of the reverse transcription and the first primer extension as the templates, respectively. In the library preparation for fusion TREs, a low-cycle PCR was performed with the products of the second primer extension as templates to add the Illumina sequencing adapters and indexing barcodes, followed by the acquisition of 240 + 360 bp pair-end reads on a MiSeq Illumina sequencer. In all the other library preparations, the products of the second primer extension went through a low-cycle pre-tagmentation PCR amplification before being tagmented with Tn5 transposomes[53]. Another round of low-cycle post-tagmentation PCR was performed to add the sequencing adapters and indexing barcodes, followed by the acquisition of 1 × 75 bp reads on a NextSeq 500 Illumina sequencer. All primer sequences can be found in Supplementary Table 1.

eSTARR-seq data analysis. Cutadapt 2.1 was used to identify attB1′ or attB2′ sequences within each read. Next, a custom Python script was used to extract element sequences and remove PCR duplicates (identical PCR barcode + first 15 bp of element). Processed reads were then aligned to candidate elements with Bowtie 2.3.4.1 (--end-to-end -a). A custom R script was used to extract alignments within 3 bp of the expected cloning boundaries, ensure complete removal of PCR duplicates and generate orientation-specific read counts for each candidate.

To identify elements with significant enhancer activity, raw read counts were processed using voom from the R Bioconductor limma package version 3.42.2. RNA and DNA counts were treated as distinct experimental conditions within each replicate. Active enhancers were defined as having a significantly elevated ratio of RNA to DNA counts with an FDR-adjusted $P < 0.1$ in both cloning orientations. Additionally, we required a $\log_2$ fold change ≥ 1 in both cloning orientations to ensure significantly higher activity than negative controls (Fig. 2c). These heuristics were validated with a linear model explicitly comparing each element to the negative control distribution. De-duplicated read counts and associated statistics are available through the public ENCODE repository.

HiDRA data analysis. Raw sequencing files were obtained from the Sequence Read Archive (accession no. SRP118092) and aligned to the hg19 genome as described by Wang et al.[37] (bowtie2 -p 6, -q and --phred33). BAM files were merged within replicates using SAMtools 1.10, then processed with a custom R script to remove multi-mappers (MAPQ < 30) and apply size selection (100–600 bp). Differential RNA versus DNA read counts were detected using voom from the R Bioconductor limma package. To minimize size bias, voom was applied separately to fragments from 100 to 150 bp, 150–200 bp, and so on. After applying voom, we only considered fragments with ≥5 DNA counts (summed from all replicates) to minimize the artifacts of low-coverage sites. Alignments with mutual overlap ≥90% and mapping to opposite strands were considered as a 'forward' and 'reverse' alignment pair. We required an FDR-adjusted $P < 0.1$ in both forward and reverse cloning orientations to call active enhancer fragments. HiDRA enhancer fragments were then analyzed relative to published GM12878 GRO-cap peaks[13]. GRO-cap peaks were collapsed to the single most-used transcription start nucleotide with a custom R script.

For dissection of unpaired GRO-cap TSS, 'upstream and TSS' fragments were defined as containing at least 200 bp upstream and 30 bp downstream of a GRO-cap TSS (size > 230 bp). 'Upstream region' fragments were taken from between 330 and 35 bp upstream of a GRO-cap TSS (size < 295 bp). 'Core promoter region' fragments were defined to contain at least 40 bp upstream and 190 bp downstream of a GRO-cap TSS (size > 235 bp).

Motif density analysis. K562 and GM12878 GRO-cap divergent pairs and processed GRO-cap data were obtained from published work[13]. Peaks were refined to a single nucleotide according to the maximum GRO-cap signal within each TSS. Divergent pairs were required to be at most 300 bp apart for visualization. Genomic sequences from −400 to +100 bp of the max TSS of each divergent pair were scanned for motifs using RTFBSDB with default match settings[54]. This scan produces a $N \times 500$ count matrix, where N is the number of sites scanned and 500 bp is the number of scanned positions. Each entry in the matrix is 0 (motif absent) or 1 (motif present). After removing divergent pairs without any matching

motifs, loci were sorted by distance between their divergent TSS and whether they were proximal (within 500 bp) or distal to a GENCODE gene annotation start coordinate. Finally, neighboring rows in the count matrix were averaged into 100 groups to compute motif density at each position for each strand and normalized to the maximum density observed in the matrix. This matrix was plotted at 4 bp resolution for visualization; most motifs are 4–12 bp. All motif density profiles shown in Fig. 3 are from K562 GRO-cap TSS, except for STAT2, which was derived from GM12878 GRO-cap TSS.

Pooled strand overlap extension PCR. Using a multichannel pipette, PCR reactions were prepared by pairing forward and reverse oligonucleotides appropriately. Then, 50 µl PCR reactions were carried out using Phusion DNA Polymerase (New England Biolabs) for 28 cycles and annealing at 58 °C. Amplicons were purified twice using AMPure XP beads (Beckman Coulter) according to the manufacturer's protocol and eluted into 40 µl of double-distilled H_2O. Each amplicon was quantified in a 96-well plate using the QuBIT dsDNA Broad Range reagents (Thermo Fisher) and a fluorometric plate reader. A pooled annealing and extension reaction was set up as follows: 10 µl of 5× HF buffer, 10 µl of 5 M betaine, 1 µl of 12.5 mM of deoxynucleoside triphosphate (dNTP) mix, 0.5 µl of Phusion DNA Polymerase, forward and reverse linker oligonucleotides to a 10-nM final concentration and double-distilled H_2O to a 50-µl final volume.

Denaturation was performed at 95 °C for 3 min. Annealing was performed by rapid cooling to 50 °C for 3 min. Extension was performed at 72 °C for 5 min. The reaction was then cooled to 4 °C for 5 min.

A final PCR reaction was performed to specifically amplify stitched products. The splicing by overlap extension (SOE)-PCR reaction mix from the previous step was used directly without any purification: 20 µl of 5× HF buffer, 20 µl of 5 M of betaine, 2 µl of 12.5 mM of dNTP mix, 1 µl of Phusion DNA Polymerase, forward and reverse primers to a 250-nM final concentration and double-distilled H_2O to a 100-µl final volume.

Amplification was performed for eight cycles to minimize bias. Denaturation was carried out at 95 °C for 3 min, annealing was carried out at 65 °C for 2 min and extension was carried out at 72 °C for 1 min. SOE-PCR amplicons were then size-selected from a nondenaturing 6% polyacrylamide gel.

Establishing homozygous deletion cell lines with CRISPR–Cas9. The gRNA sequences were designed as described previously[55]. Candidate 20-mer guides upstream of an NGG protospacer adjacent motif site and within 50 bp of the desired cutting site were identified and filtered to eliminate potential off-target effects. All candidates were reverse-complemented and aligned to the human reference genome (hg19) using Bowtie v.1.1.2, with the settings -n 2 -l 18 -p 8 -a -y --best -e 90. Guides mapped to more than one location with these settings were not used. The gRNA-coding oligonucleotides were synthesized (Eurofins) and cloned into pSpCas9(BB)-2A-Puro (pX459, plasmid no. 48139; Addgene)[56] and/or lentiCRISPRv2 neo (plasmid no. 98292; Addgene)[57] so that the gRNA-coding sequences targeting the upstream and downstream breakpoints of each desired deletion locus were cloned into different CRISPR–Cas9 vectors. Different plasmids for generating the desired pair of breakpoints were mixed (1 µg of each) and electroporated into 1 million K562 cells with Cell Line Nucleofector Kit V and recovered in 2 ml of culture medium for 24 h. The electroporated cells were then treated with 200 µg ml^{-1} G-418 (catalog no. 04727878001; Roche) and 2 µg ml^{-1} puromycin dihydrochloride (catalog no. A1113803; Gibco) for 72 h. After the *antibiotic* treatment, individual surviving cells were sorted into 96-well plates using the MA900 Multi-Application Cell Sorter (Sony) and then further propagated. Single-cell clones were confirmed with PCR and agarose gel electrophoresis. All guide sequences and genotyping primer sequences can be found in Supplementary Table 1.

Quantification of *NMU* expression. Single-cell clones with confirmed homozygous deletions in the e*NMU* locus were collected for total RNA extraction with TRIzol Reagent and Direct-zol RNA Miniprep Kit (catalog no. R2050; Zymo Research). Total RNA was reverse-transcribed into complementary DNA with Maxima H Minus Reverse Transcriptase (catalog no. EP0753; Thermo Fisher Scientific) using Oligo(dT)$_{18}$ (Integrated DNA Technologies) as primer. The qPCR reactions were carried out with the yielded cDNA as the template using SsoFast EvaGreen Supermixes (catalog no. 1725200; Bio-Rad Laboratories) according to the manufacturer's instructions in a LightCycler 480 (Roche). All qRT-PCR primer sequences can be found Supplementary Table 1.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The eSTARR-seq data are available through the ENCODE data portal (www.encodeproject.org) under accession nos. ENCSR514FNW, ENCSR729EGU and ENCSR585AGE. Processed GRO-cap data were obtained from the Gene Expression Omnibus (accession no. GSE60456). Raw sequencing files for the HiDRA study were obtained from the Sequence Read Archive (accession no. SRP118092). All candidate regulatory element clones generated in this study and used for the eSTARR-seq and luciferase assays are available upon request. Please address requests to haiyuan.yu@cornell.edu. Source data are provided with this *paper*.

Code availability

All analysis scripts are available as R Jupyter Notebooks on Github (https://github.com/hyulab/eSTARR).

References

51. Wei, X. et al. A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. *PLoS Genet.* **10**, e1004819 (2014).
52. Arad, U. Modified Hirt procedure for rapid purification of extrachromosomal DNA from mammalian cells. *Biotechniques* **24**, 760–762 (1998).
53. Picelli, S. et al. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).
54. Wang, Z., Martins, A. L. & Danko, C. G. RTFBSDB: an integrated framework for transcription factor binding site analysis. *Bioinformatics* **32**, 3024–3026 (2016).
55. Chow, R. D. et al. In vivo profiling of metastatic double knockouts through CRISPR–Cpf1 screens. *Nat. Methods* **16**, 405–408 (2019).
56. Ran, F. A. et al. Genome engineering using the CRISPR–Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).
57. Stringer, B. W. et al. A reference collection of patient-derived cell line and xenograft models of proneural, classical and mesenchymal glioblastoma. *Sci. Rep.* **9**, 4902 (2019).

Acknowledgements

The human pSTARR-seq plasmid was a gift from A. Stark (plasmid 71509; Addgene). We thank M. Gasperini, J. Tome and J. Shendure for sharing the clonal ΔeNMU K562 cells and helpful advice. We thank C. Fulco and J. Engreitz for helpful discussions and guidance. This work was supported by grants from the National Institutes of Health (HG009393 to J.T.L. and H.Y.; GM25232 to J.T.L.; DK115398 and HG008126 to H.Y.). N.D.T. was supported by a Cornell University Center for Vertebrate Genomics Scholarship and National Institutes of Health training grant T32HD057854.

Author contributions

N.D.T., J.L., A.O., J.T.L. and H.Y. conceived the project and designed the enhancer comparison screen. N.D.T. conceived the dissecting enhancer cooperativity and mechanisms. J.L. performed cloning, primer design, Cas9 deletions and all eSTARR- and Clone-seq assays. N.D.T. optimized and prepared the enhancer fusions with guidance from A.O., H.Y. and J.T.L. N.D.T. and A.K-Y.L. performed the analysis with feedback from J.G.B., J.L., A.O., J.T.L. and H.Y. S.D.W. designed the single guide RNAs. N.D.T. wrote the manuscript with feedback from all authors.

Competing interests

The authors declare no competing interests.

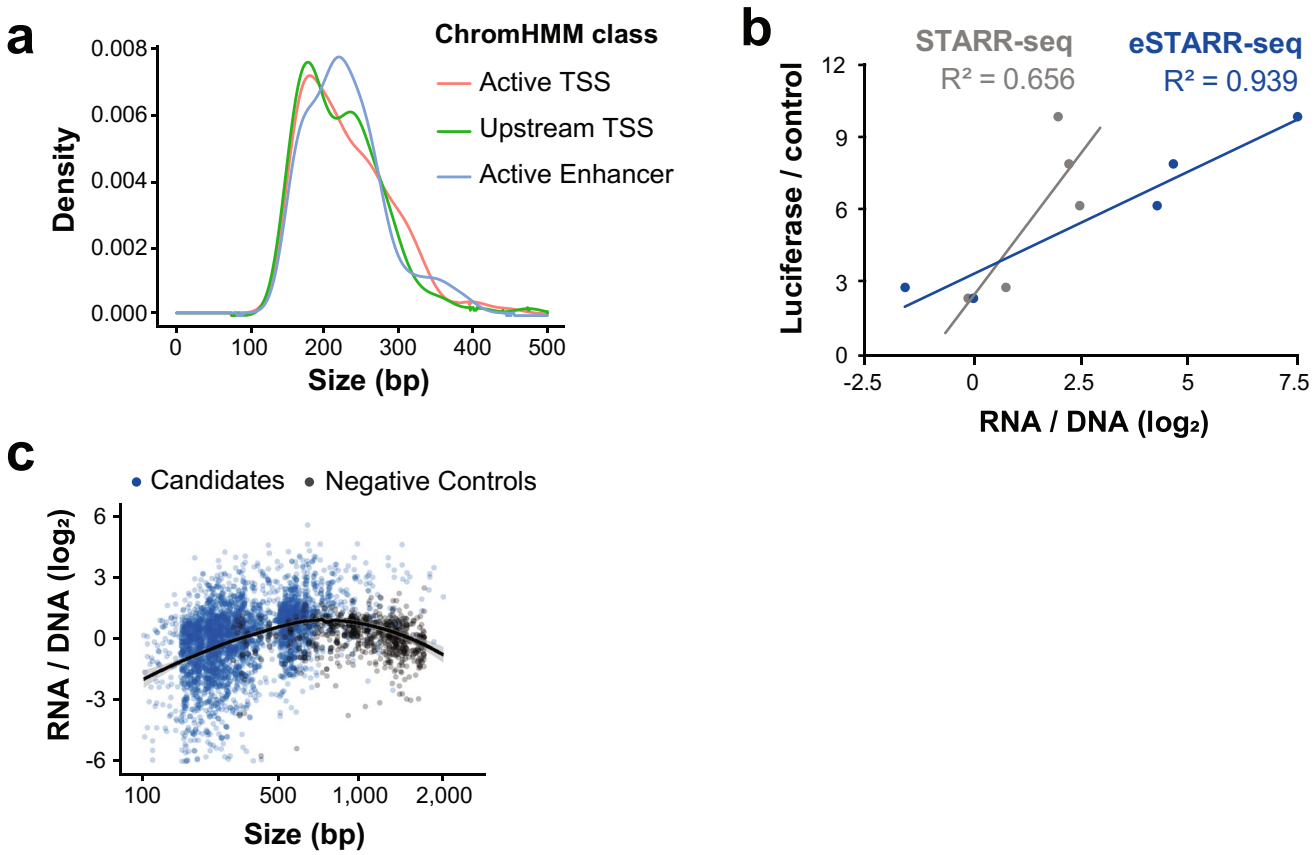
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-020-0686-2.

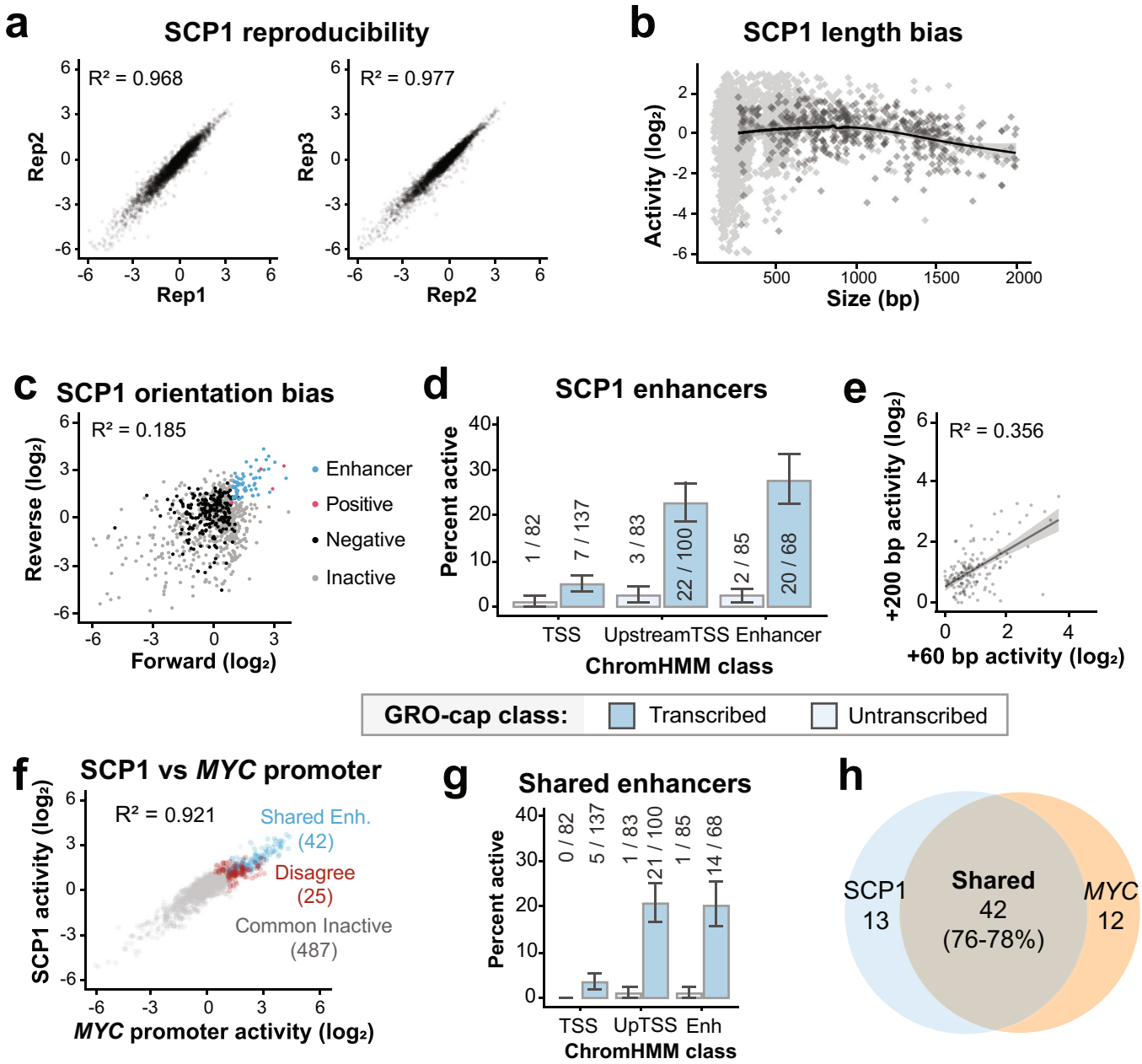
Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-0686-2.

Correspondence and requests for materials should be addressed to J.T.L. or H.Y.

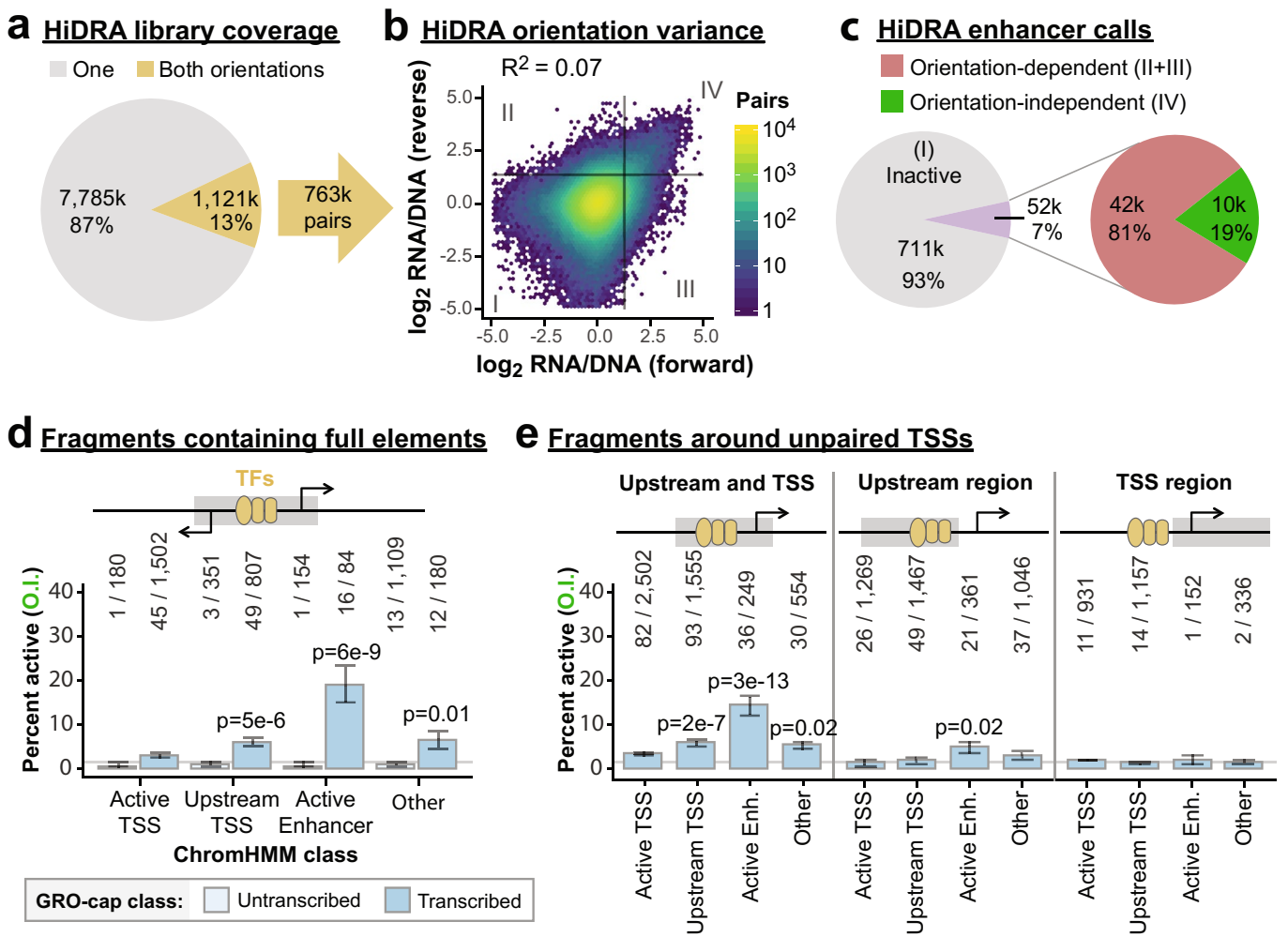
Reprints and permissions information is available at www.nature.com/reprints.



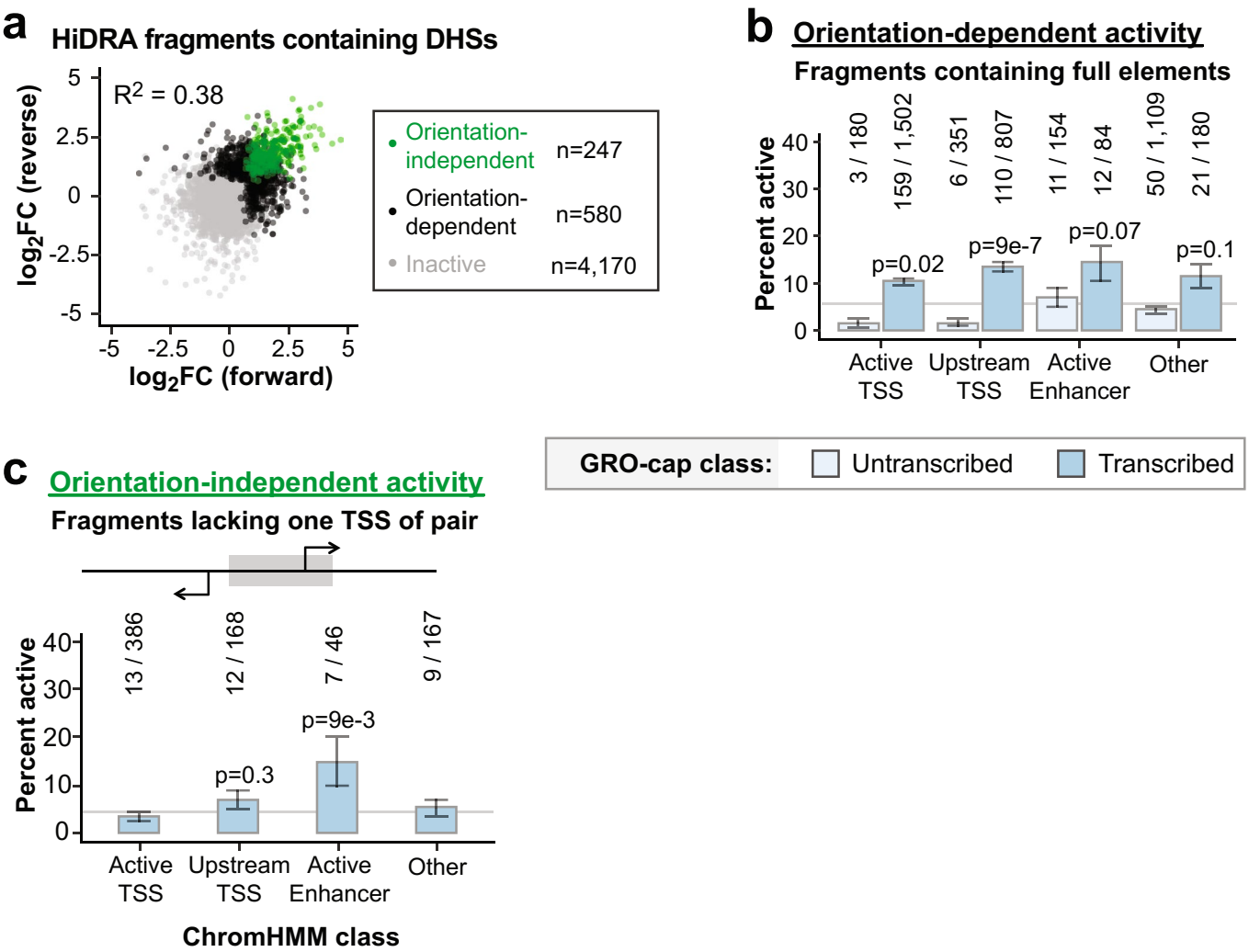
Extended Data Fig. 1 | Design and validation of eSTARR-seq and selected candidates. a, Size distribution of candidates is shown by ChromHMM class. **b**, Correlation between luciferase, STARR-seq, and eSTARR-seq reporter activity in HeLa cells. Luciferase and STARR-seq data are from (Arnold et al., 2013). **c**, eSTARR-seq activity is shown relative to each elements' size for both candidate elements (blue) and negative controls (gray). Line indicates a fitted loess curve estimate of size bias for eSTARR-seq and 95% confidence interval in gray.



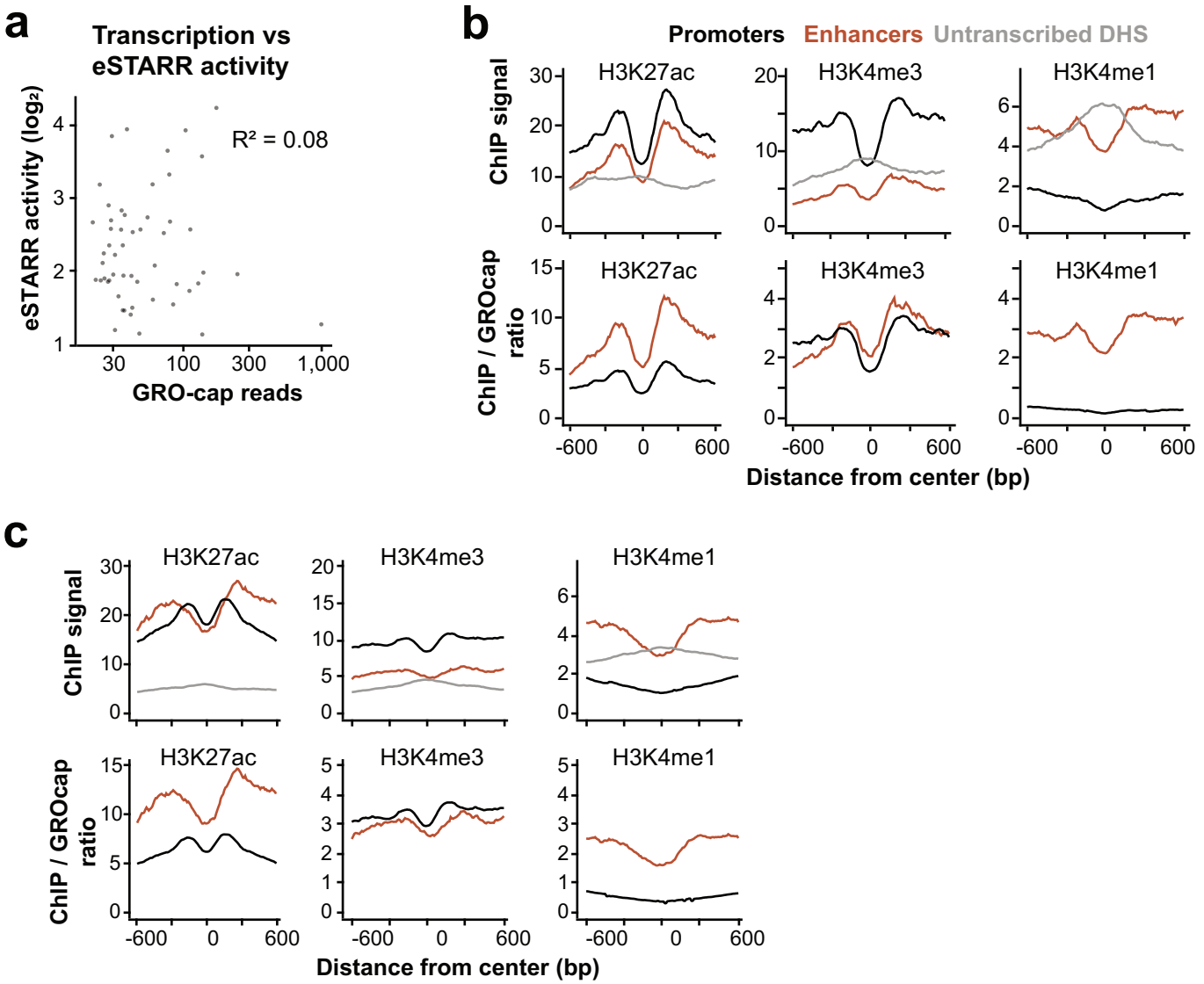
Extended Data Fig. 2 | Comparison with the SCP1 promoter. a, Correlation between replicates using SCP1. **b**, eSTARR-seq activity vs element length using SCP1, averaged from n = 3 transfection replicates. **c**, eSTARR-seq activity in forward vs reverse cloning orientations using SCP1 (averaged from n = 3). **d**, Percent of elements from each ChromHMM class with significant enhancer activity for SCP1. Error bars indicate standard error calculated for a sample of binary trials, centered on the observed success rate. **e**, SCP1 eSTARR-seq activity of elements cloned using TSS + 60 bp boundaries (x) or TSS + 200 boundaries (y). Gray area shows 95% confidence interval of linear regression from n = 93 elements. **f**, eSTARR-seq activity of MYC (x) vs SCP1 (y) as the promoter. Colors indicate enhancers shared by both promoters (blue), active with only one promoter (red), or inactive with both promoters (gray). **g**, Percent of elements from each ChromHMM class with significant enhancer activity for both MYC promoter and SCP1. Error bars indicate standard error calculated for a sample of binary trials, centered on the observed probability. **h**, Venn diagram showing overlap of the MYC promoter and SCP1 active enhancer sets.



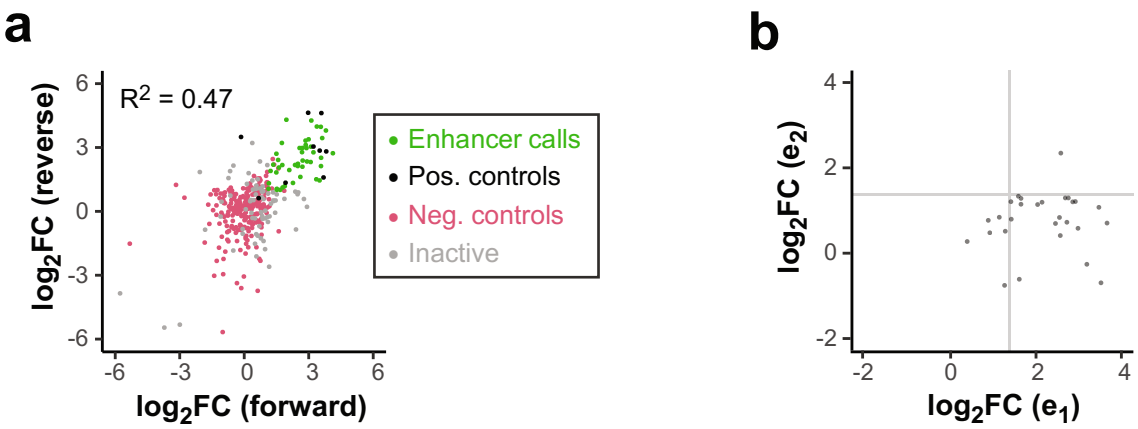
Extended Data Fig. 3 | Validation of strand bias and TSS function from HiDRA. a, Pie chart indicating the fraction of HiDRA fragments tested in one (gray) or both (gold) orientations. Some fragments have pairings with more than one fragment in the opposing orientation, providing 763,000 distinct pairs. **b**, Comparison of HiDRA enhancer activities from opposing orientations of fragment pairs. Color indicates the number of pairs. Gray lines denote approximate statistical cut-off for active enhancers. Quadrants II and III denote orientation-dependent 'enhancer' fragment pairs; quadrant IV fragments are active in both orientations. **c**, Pie chart indicating the percent of HiDRA fragment pairs classified as inactive, orientation-dependent, and orientation-independent. **d**–**e**, Bar charts indicating the percentage of orientation-independent enhancer calls from HiDRA fragments sample from DHSs within the indicated ChromHMM classes. **d**, fragments are further classified as untranscribed or transcribed (contains divergent GRO-cap TSSs). P-values are from two-sided Fisher's exact test between indicated ratio and total enhancer ratio (140/4,367). **e**, fragments are sampled from different areas around unpaired GRO-cap TSSs (see cartoon and Methods). Raw fragment counts are shown above each bar. Gray line marks the average percent activity of all fragments. P-values are from two-sided Fisher's exact test between indicated ratio and total enhancer ratio (402/11,579). All error bars indicate standard error calculated for a sample of binary trials, centered on the observed probability.



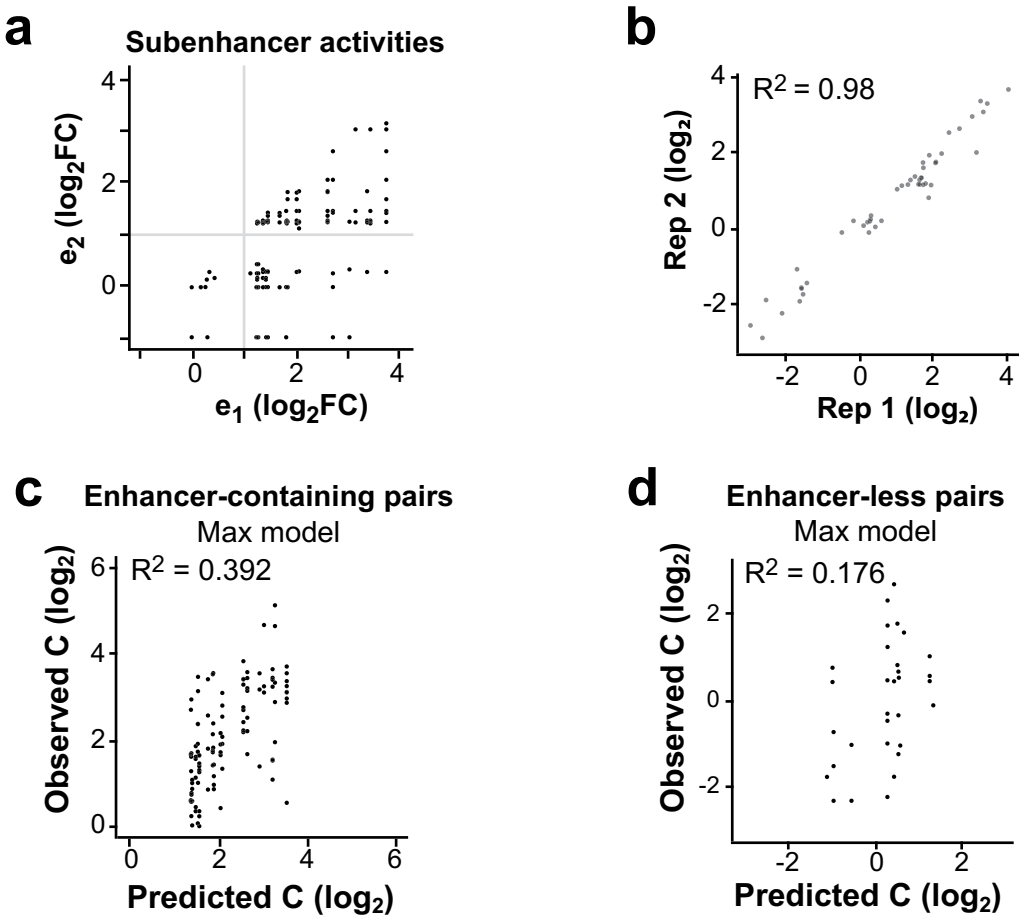
Extended Data Fig. 4 | Orientation dependence in the HiDRA dataset. a, Comparison of forward vs reverse cloning orientation for HiDRA fragments overlapping GM12878 DHS peaks. Data points are shown as log2 fold-change of RNA vs DNA read counts. Elements with significantly elevated activity in both orientations are called orientation-independent enhancers (green). Elements with significantly elevated activity in one orientation are called orientation-dependent (black). Remaining fragments are called inactive (gray). **b**-**c**, Percent of orientation-dependent (**b**) or -independent (**c**) fragments within each GRO-cap and ChromHMM class. Raw fragment counts are shown above each bar. Gray line marks the percent activity of all fragments judged by the same criteria. P-values are from two-sided Fisher's exact test between indicated ratio and total enhancer ratio (372/4,367 for b, 41/767 for c). Error bars indicate standard error calculated for a sample of binary trials, centered on the observed probability.



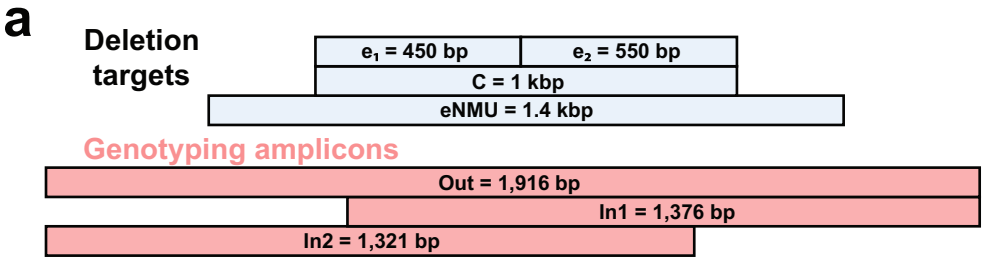
Extended Data Fig. 5 | Features of eSTARR-seq enhancers. a, Scatterplot of activity vs GRO-cap reads from eSTARR enhancers in K562 cells. **b**, Metaplots of average H3K27ac, H3K4me3, and H3K4me1 *ChIP*-seq signal from different element classes defined in K562 cells. Promoters are defined as GRO-cap divergent TSSs within 500 bp of GENCODE gene start, whereas enhancers are defined as GRO-cap divergent TSSs with significant eSTARR activity. Below, *ChIP*-seq to GRO-cap signal ratio is shown within the window. **c**, Metaplots of average H3K27ac, H3K4me3, and H3K4me1 *ChIP*-seq signal from different element classes defined in GM12878 cells. Promoters are defined as GRO-cap divergent TSSs within 500 bp of GENCODE gene start, whereas enhancers are defined as GRO-cap divergent TSSs with significant HiDRA activity. Below, *ChIP*-seq to GRO-cap signal ratio is shown within the window. n = 860 promoter DHS, 119 transcribed enhancer DHS, 1,100 untranscribed DHS.



Extended Data Fig. 6 | Functional dissection of genomic TSS clusters. a, Comparison of forward vs reverse cloning orientation for all tested TSS clusters. Data points are shown as log2 fold-change vs negative controls (magenta), averaged from three replicates. Positive controls (black) are known MYC or viral enhancers. Clusters with significantly elevated activity in both orientations are called enhancers (green). All other clusters are called inactive (gray). **b**, Comparison of sub-element activities within active enhancer clusters. The stronger sub-element is always chosen to be e1, and the weaker sub-element is e2. Gray lines indicate approximate significance cut-offs.

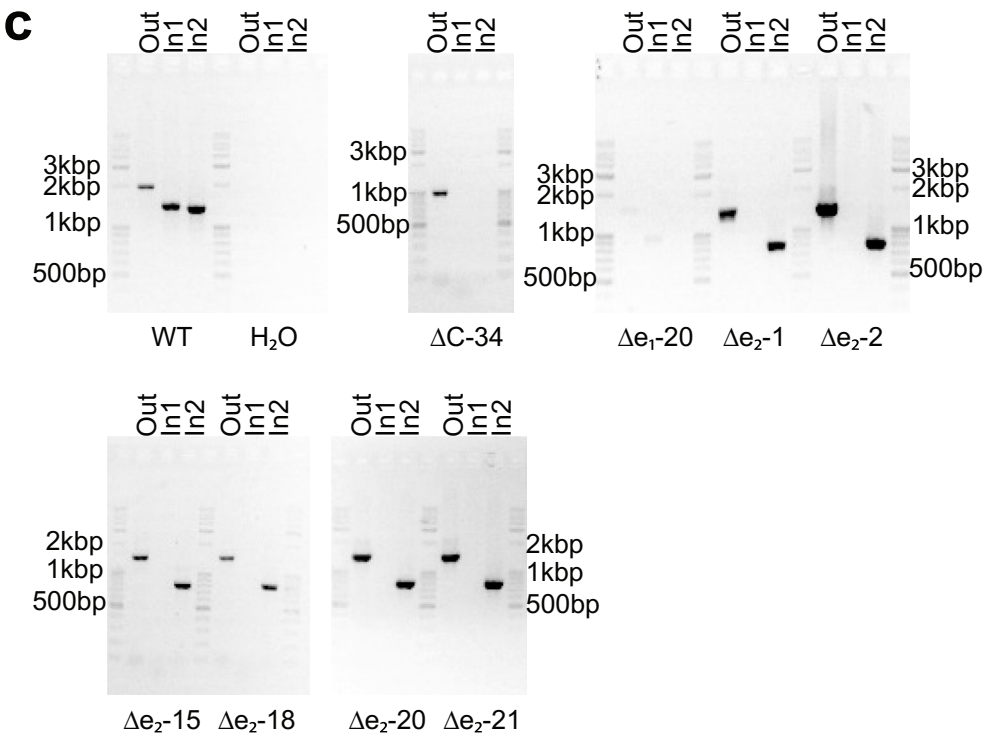


Extended Data Fig. 7 | Design and evaluation of synthetic unit pairs. a, Comparison of sub-element activities within synthetic enhancer clusters. The stronger sub-element is always chosen to be e_1, and the weaker sub-element is e_2. Gray lines indicate approximate significance cut-offs. **b**, Correlation between individual eSTARR-seq activities tested previously and re-tested as controls in the synthetic fusion screen (n = 48 elements). **c**, Agreement between predicted and observed cluster activities ('C') for enhancer-containing synthetic pairs. **d**, Agreement between predicted and observed cluster activities ('C') for enhancer-less synthetic pairs.



b

		Genotype				
	Expected size (bp)	WT	ΔeNMU	ΔC	Δe1	Δe2
Amplicon	Out	1,916	495	965	1,455	1,426
	In1	1,376	-	-	-	831
	In2	1,321	-	-	915	-



Extended Data Fig. 8 | Genotyping of Cas9 deletion clones. a, Illustration of genotyping PCR amplicon design and size relative to elements targeted for deletion. **b**, Table listing expected amplicon sizes from various genotypes. '-' indicates that no amplification is expected. **c**, Gel images from K562 clonal lines used for qRT-PCR experiments in Fig. 6. (eNMU clones were generated, genotyped and generously provided by the Shendure lab.) Genotyping PCRs were performed only once, but biological replication was achieved through independent clones.

nature research

Corresponding author(s): NG-A52545R2

Last updated by author(s): Jul 10, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	None
Data analysis	Ubuntu 18.04.3 LTS, cutadapt 2.1, bowtie2 2.3.4.1, R 3.6.1, Bioconductor 3.9, limma_3.40, rtracklayer_1.44, GenomicRanges_1.36, ggplot2_3.1.1, tidyverse_1.2.1, lattice_0.20-38, Rsamtools_2.0, BiocParallel_1.18, GenomicAlignments_1.20, rtfbsdb_0.4.5, MotifDb_1.26

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Figures 1-6 use data from GSE60456. Figure 4 uses data from GSE104001.
The rest of the figures rely on eSTARR-seq data that is publicly available through the ENCODE data portal: ENCSR514FNW, ENCSR729EGU, and ENCSR585AGE

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample sizes were computed. Sample sizes were maximized within acceptable costs of library construction, and are mostly >50 samples per group.
Data exclusions	None
Replication	Re-analysis of a recently published dataset supports key findings of our work. All experiments were replicated across three independent experimental batches.
Randomization	None
Blinding	None

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Purchased from ATCC in 2017
Authentication	Authenticated by ATCC, https://www.atcc.org/products/all/CCL-243.aspx
Mycoplasma contamination	Experimental batches used in this study tested negative by PCR, alongside positive control reactions.
Commonly misidentified lines (See ICLAC register)	None