

Directionality of transcriptional regulatory elements

Received: 15 January 2025

Accepted: 24 July 2026

Published online: 10 September 2026



You Chen^{1,2,7}, Sagar R. Shah^{1,2,7}, Alden K. Leung^{2,3}, Mauricio I. Paramo^{1,2},
Caralyn Gonzales⁴, Kelly Cochran⁵, Austin Wang^{5,6}, Anshul Kundaje^{5,6},
Andrew G. Clark^{1,3}✉, John T. Lis¹✉ & Haiyuan Yu^{1,2,3}✉

Divergent transcription is a critical marker of active transcriptional regulatory elements (TREs), including enhancers and promoters, in mammals. However, distal elements with unidirectional transcriptional patterns are often overlooked, leaving their identity and function poorly understood. Here, we perform a systematic comparison between divergent and unidirectional elements, revealing their distinct architectural and functional features. Our analysis also shows that unidirectional elements have younger sequence ages and are under weaker evolutionary constraints than divergent elements, suggesting that they may represent a unique category of genomic regulatory function with more recent origins. Furthermore, we investigate factors involved in transcriptional directionality at TREs, with CTCF and ZNF143 showing dual, position-dependent roles in transcriptional modulation. Overall, the elucidation of directionality enhances our understanding of the diverse architectural models, functional features, evolutionary dynamics, and regulatory logic of TREs.

Promoters and enhancers are two essential classes of transcriptional regulatory elements (TREs) that play crucial roles in modulating the genome's regulatory programs in normal and disease states. Both elements can generate extensive transcriptional activity, and understanding their molecular architecture is crucial to uncovering their biological roles.

Active mammalian promoters often exhibit divergent transcription, resulting in both a stable sense transcript that represents the annotated gene, and an upstream antisense transcript that cannot elongate efficiently beyond the promoter region^{1,2}. The directionality of promoters can differ across various human tissues and conditions. A study of 376 samples from nine different tissue types, including cancer and benign conditions, revealed lineage-specific and cancer-specific antisense transcription³. Moreover, well-positioned CTCF and cohesin have been demonstrated as a barrier against antisense transcription initiation at divergent promoters^{4,5}. An explainable machine-learning model of transcription initiation, Puffin⁶, also identified a few motifs with orientation-specific effects at promoters, including TATA-box and YY1.

In contrast to promoters, enhancers produce short-lived RNAs called enhancer RNAs (eRNAs)⁷. The presence of divergent transcription has been shown to be a more robust predictor of active enhancers than classical epigenomic signatures (e.g., high chromatin accessibility, enrichment of H3K4me1 and H3K27ac)^{8,9}.

A study of eRNA transcription in human lymphoblastoid cell lines derived from the Yoruba population identified ~10,000 genetic variants associated with the directionality of divergent enhancers, termed directional initiation quantitative trait loci (diQTLs)¹⁰. These diQTLs are significantly enriched in the vicinity of enhancer transcription start sites (TSSs), potentially affecting the core promoter motifs such as Initiator (Inr) and TATA-like elements. Moreover, the study showed that enhancer diQTLs are significantly associated with changes in gene expression.

While divergent transcription is a well-documented phenomenon in bulk transcriptome data, a recent study using C1 CAGE profiling has challenged this paradigm by suggesting that eRNA transcripts may arise from either strand in a mutually exclusive manner within

¹Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA. ²Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, NY, USA. ³Department of Computational Biology, Cornell University, Ithaca, NY, USA. ⁴Department of Chemistry and Chemical Biology, Cornell University, Ithaca, NY, USA. ⁵Department of Computer Science, Stanford University, Stanford, CA, USA. ⁶Department of Genetics, Stanford University, Stanford, CA, USA.

⁷These authors contributed equally: You Chen, Sagar R. Shah. ✉e-mail: ac347@cornell.edu; johnlis@cornell.edu; haiyuan.yu@cornell.edu

individual cells¹¹. However, the low sequencing coverage limited the detection of eRNAs, as exemplified by the identification of only 32 divergent enhancer loci with at least 10 reads in a minimum of 5 cells in pooled single-cell data. Despite these technical limitations, the occurrence of unidirectional transcription at enhancers remains an intriguing phenomenon, warranting deeper investigation to fully understand its underlying mechanisms and biological significance.

Recently, a comprehensive evaluation of RNA sequencing approaches demonstrated that nuclear run-on assay with cap-selection^{12,13} achieves the highest sensitivity and specificity for detecting eRNAs¹⁴. In this study, we leverage the PRO-cap¹² assay to identify a substantial number of unidirectional elements with sufficient sequencing depth. This enables a systematic comparison of divergent and unidirectional elements in terms of their architectural features, functional characteristics, and evolutionary dynamics. Furthermore, we investigate genetic determinants of transcriptional directionality at TREs and identified CTCF and ZNF143 as factors with dual, position-dependent roles in transcriptional modulation.

Results

Unidirectional elements have distinct architectural features

To examine transcriptional directionality at human enhancers and promoters, we performed PRO-cap, which captures RNA polymerase II (Pol II) initiation sites during nascent transcription at base-pair resolution, on two replicates of the chronic myelogenous leukemia cell line K562 (Supplementary Data 1). Using our recently developed tool, Peak Identifier for Nascent Transcript Starts (PINTS)¹⁴, we identified a comprehensive set of active TREs genome-wide and categorized them into divergent elements (pairs of peaks on opposite strands within 300 bp of each other; $n = 34,031$) and unidirectional elements (peaks on only one strand; $n = 19,346$) (Fig. 1a, b). To ensure a fair comparison between divergent and unidirectional categories, we applied a stringent set of criteria (see Methods for details) to derive elements for the analysis presented in this manuscript (Supplementary Fig. 1a, b).

Beyond a binary classification of transcription patterns at TREs, we assessed the transcriptional directionality of divergent elements using a continuous metric, the directionality index (DI). This index is derived from the difference between PRO-cap read counts on the plus and minus strands relative to genomic coordinates, normalized by the total read counts for each element, revealing a spectrum of directionality. While some divergent elements exhibit similar transcription levels on both strands, others show pronounced, unbalanced transcription on one strand, skewing their transcriptional directionality (Fig. 1c). Although the distal category has a higher fraction of unidirectional elements than the proximal category (Fig. 1b and Supplementary Fig. 1b), among divergent elements, the proximal category tends to exhibit higher absolute DI values than the distal category (Fig. 1c).

To explore the architectural features of these divergent elements, we positioned the TREs at the midpoint between the two prominent TSSs on the plus and minus strands (Supplementary Fig. 1a). Both distal and proximal divergent elements were centered on DNase peaks and marked by the histone modification H3K27ac at their boundaries (Fig. 1d and Supplementary Fig. 1c). In the absence of distinct peaks detected on the opposite strand, unidirectional elements were anchored at the center of overlapping DNase peaks as a proxy (Supplementary Fig. 1a). We observed that transcription initiation at divergent and unidirectional elements occurs at variable distances from the center of open chromatin regions, with histone modification marks predominantly enriched on the same side that exhibits higher PRO-cap signals. Moreover, unidirectional elements have significantly lower signals for nascent transcription (PRO-cap) and epigenomic features (DNase-seq and H3K27ac ChIP-seq) compared to their divergent counterparts (Fig. 1d and Supplementary Fig. 1c).

To investigate protein binding patterns that may distinguish unidirectional from divergent elements, we analyzed ENCODE ChIP-seq data from 310 target proteins and observed significant differences. Divergent elements show greater enrichment for most proteins surveyed (Fig. 1e and Supplementary Fig. 1d), including RNA Pol II and general transcription factor components, consistent with their elevated transcription levels measured by the PRO-cap assay. In contrast, CTCF and cohesin components (RAD21 and SMC3) exhibit stronger binding signals in unidirectional elements, suggesting their potential involvement in establishing transcriptional directionality. Beyond binding strength, we examined position-specific profiles and identified four distinct clusters at divergent proximal elements (Supplementary Fig. 1d, e and Supplementary Fig. 2). Cluster 1 is characterized by a single strong tract in the upstream region and is mainly composed of sequence-specific TF families. In cluster 2, proteins are bound on both sides of divergent elements, and the side with higher PRO-cap signals also exhibits stronger ChIP-seq enrichment. This cluster includes transcriptional machinery (e.g., Pol II subunits), epigenetic regulators (e.g., PHF8), and SMAD signaling proteins (e.g., SMAD5). Clusters 3 and 4 show increased binding on the opposite strand (i.e., the side with lower transcriptional level), likely reflecting the presence of repressors. In contrast, unidirectional proximal elements occupy a smaller functional region, and proteins in cluster 2 primarily bind to the side aligned with PRO-cap peaks. Distal elements display similar but less clearly defined patterns. Collectively, these results demonstrate that unidirectional elements have architectures distinct from divergent counterparts.

Beyond the overall architecture, we examined the sequence content of individual transcriptional units within each category. Examination of single-nucleotide sequence preferences reveals two symmetric arcs of enriched AT content in divergent elements, whereas a similar pattern was detected on one side of unidirectional elements where PRO-cap peaks were observed (Fig. 2a). Closer inspection shows that each arc has a fainter parallel inner arc. Position-specific 3-mer frequencies indicated that the inner and outer arcs represent TATA-box-like and Inr-like sequences, respectively (Fig. 2b). Direct motif scans of TATA-box, Inr, and downstream promoter region (DPR) relative to prominent TSSs revealed that unidirectional TSSs and maximum TSSs (prominent TSS on the side with more read counts) of divergent elements have a slightly higher proportion of these motifs compared to minimum TSSs (prominent TSS on the side with less read counts) of divergent elements (Fig. 2c). Therefore, despite differences in their overall architectures, divergent and unidirectional elements have similar sequence content in individual core promoter regions.

Potential confounders for binary classification

To rule out potential confounders that may impact the binary classification of TREs, we first examined whether the absence of peaks on the opposite side of unidirectional elements resulted from poor sequence mappability. We evaluated the mappability using k-mers of varying lengths (24, 36, and 50 bp)¹⁵ based on the range of read lengths detected (Supplementary Fig. 3a). Both divergent and unidirectional elements have higher mappability than the random genomic background. Importantly, no apparent bias was observed in unidirectional elements on the side lacking PRO-cap peaks (Supplementary Fig. 3b).

Insufficient sequencing depth could also potentially account for the missing peak at unidirectional elements. To address this possibility, we downsampled our PRO-cap dataset of ~45 million (M) uniquely mapped PCR-free reads to lower read counts (30, 20, 10, 5, and 1 M) and performed peak calling on each downsampled dataset. We found that at extremely low depth (<10 M reads), a significant number of divergent elements are misclassified as unidirectional elements. These misclassified elements exhibit substantial transcription detected on the opposite strand at the original, higher sequencing depth (Supplementary Fig. 4a,c), indicating that a sufficient sequencing depth is

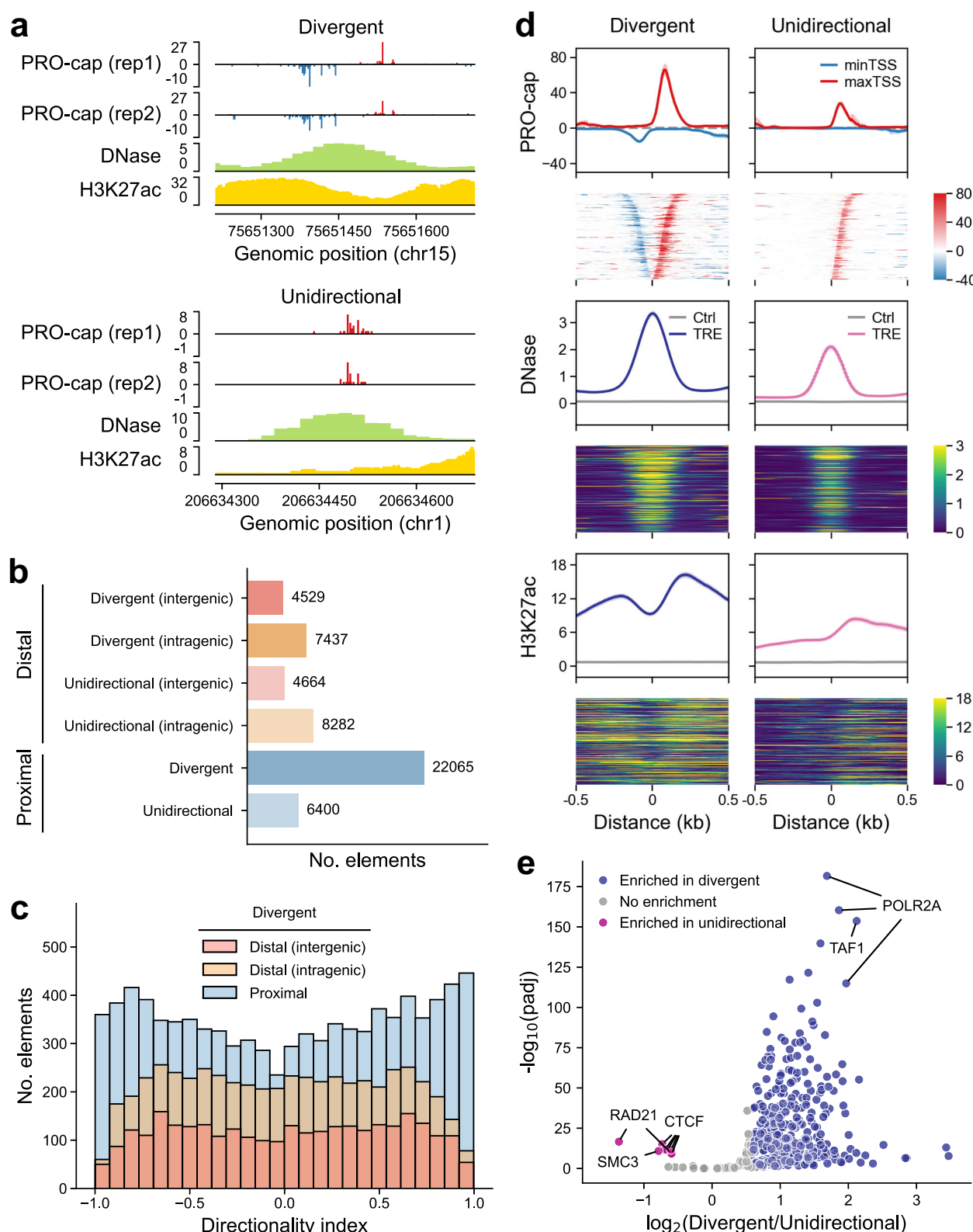


Fig. 1 | Architectural features of divergent and unidirectional distal elements.

a Genome browser tracks of PRO-cap, DNase-seq, and H3K27ac ChIP-seq for a representative divergent (top) and unidirectional (bottom) distal element. PRO-cap tracks from two replicates are shown, with red and blue signals indicating forward and reverse strands, respectively. **b** Genome-wide distribution of divergent and unidirectional elements identified by PRO-cap across proximal and distal intra- and intergenic regions. **c** Distribution of directionality index for divergent elements located in proximal and distal intra- and intergenic regions. **d** Metaplots and

heatmaps of PRO-cap, DNase-seq, and H3K27ac ChIP-seq signals at divergent and unidirectional distal elements. The metaplots display average signals with 95% confidence intervals for each assay, while the heatmaps show signals across the element set, sorted by distance from the prominent TSS to the center. **e** Volcano plot displaying ChIP-seq signal enrichment of DNA-binding proteins between divergent and unidirectional distal elements. The divergent and unidirectional groups were compared using two-sided Welch's t tests. P values were BH-adjusted (padj) for multiple testing. Source data are provided as Source Data files.

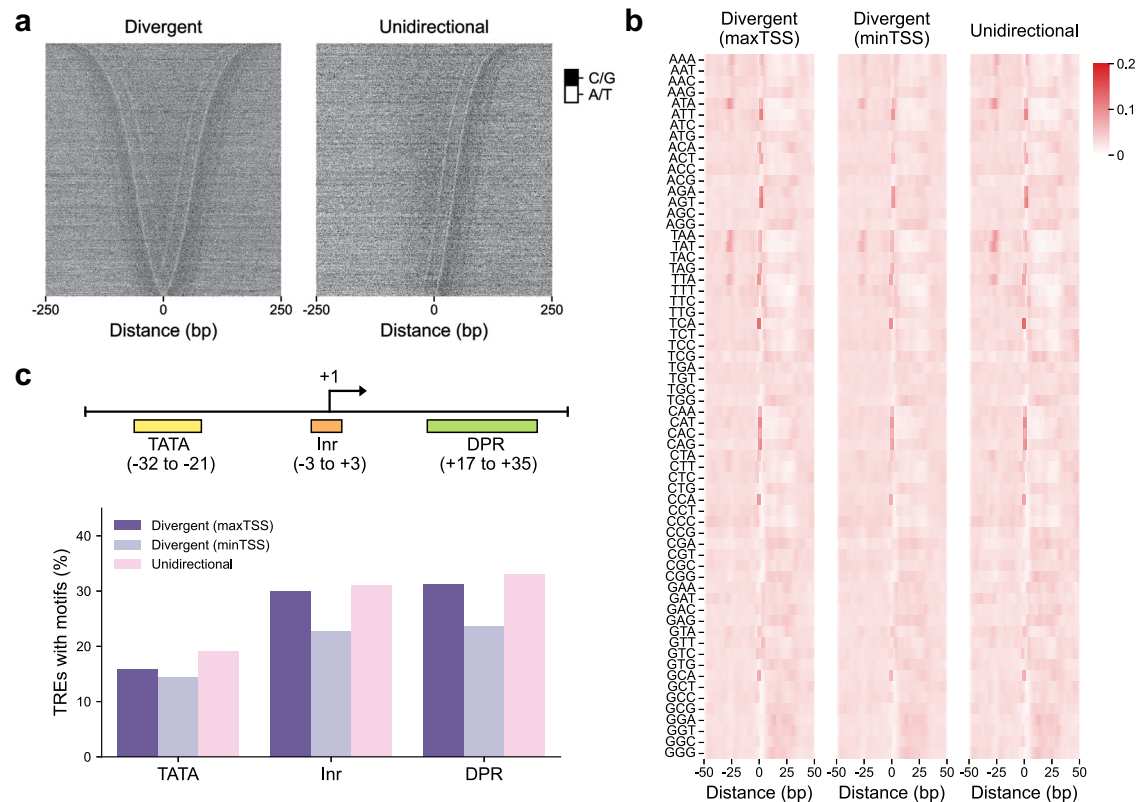


Fig. 2 | Core promoter sequences at divergent and unidirectional distal elements. **a** Genomic DNA sequence content of divergent and unidirectional distal elements, with bases A and T colored in white, and bases C and G colored in black. The heatmaps show signals across the element set, sorted by distance from the prominent TSS to the center. **b** The frequency of 64 three-mers at each position

relative to the TSS, normalized to their total frequency within the region spanning -50 to +50 bp around prominent TSSs of divergent and unidirectional distal elements. **c** Percentage of elements displaying core promoter motifs matching the TATA box, Inr, or DPR in divergent and unidirectional distal categories. Source data are provided as Source Data files.

critical for accurate discrimination of unidirectional from divergent elements. Therefore, to further explore the optimal sequencing depth required, we utilized a deep learning model, ProCapNet¹⁶, to assess transcription potential based on DNA sequences. The model was trained using peaks called at the original sequencing depth, with the exclusion of those overlapping with unidirectional elements identified at downsampled levels (profile task correlation: 0.520 ± 0.006 ; count task correlation: 0.774 ± 0.017). The predicted PRO-cap signals from unidirectional elements across varying read depths show that sequencing depths of ~20M reads improve the accuracy of binary classification (i.e., divergent versus unidirectional elements) (Supplementary Fig. 4b, d).

Building on this, we directly assessed the stability of element classification across downsampled PRO-cap datasets by comparing classifications between consecutive depth pairs in both directions. First, as coverage increased (e.g., 1M → 5M reads), we quantified the fraction of distal elements classified as unidirectional at lower depth that (i) remained unidirectional, (ii) were reclassified as divergent, or (iii) were not detected at higher depth (Supplementary Fig. 5a). Second, as coverage decreased (e.g., 5M → 1M reads), we quantified the fraction of distal elements classified as divergent at higher depth that (i) remained divergent, (ii) were reclassified as unidirectional, or (iii) were not detected at lower depth (Supplementary Fig. 5b). Classification consistency between depth pairs increased with sequencing depth, and only a subset of divergent elements at higher coverage were reclassified as unidirectional upon downsampling. To further investigate the sources of classification discrepancies, we focused on elements classified as divergent at higher coverage (Supplementary

Fig. 5b) and examined features associated with changes in classification upon downsampling. We found that elements that changed classification are, on average, more weakly transcribed (Supplementary Fig. 5c) and show more skewed strand signals (i.e., increased DI; Supplementary Fig. 5d) at higher coverage. This suggests that while reduced read counts can contribute to misclassification, such effects predominantly affect elements with pre-existing directional bias, where reduced coverage further diminishes detectable signal on the weaker strand. Similar patterns were observed for proximal elements (Supplementary Fig. 6).

Given that stochastic sampling can increase the likelihood of observing reads predominantly on one strand at low counts, we applied the beta-binomial model to the original (non-downsampled) dataset to assess whether the observed strand imbalance exceeds that expected from stochastic sampling. We found that most loci classified as unidirectional at full sequencing depth exhibit strand imbalance beyond that expected from stochastic sampling, a pattern consistent across read-count bins and between distal and proximal elements (Supplementary Figs. 7, 8). These results indicate that unidirectional elements identified at full sequencing depth are unlikely to be driven primarily by low-count stochastic effects. We also observed that the percentage of divergent elements exhibiting strand imbalance increases with total read count. This trend is consistent with increased statistical power at higher coverage, enabling detection of more modest deviations from equal strand representation, and may also reflect greater strand asymmetry among highly expressed divergent elements (Supplementary Fig. 7e and Supplementary Fig. 8e).

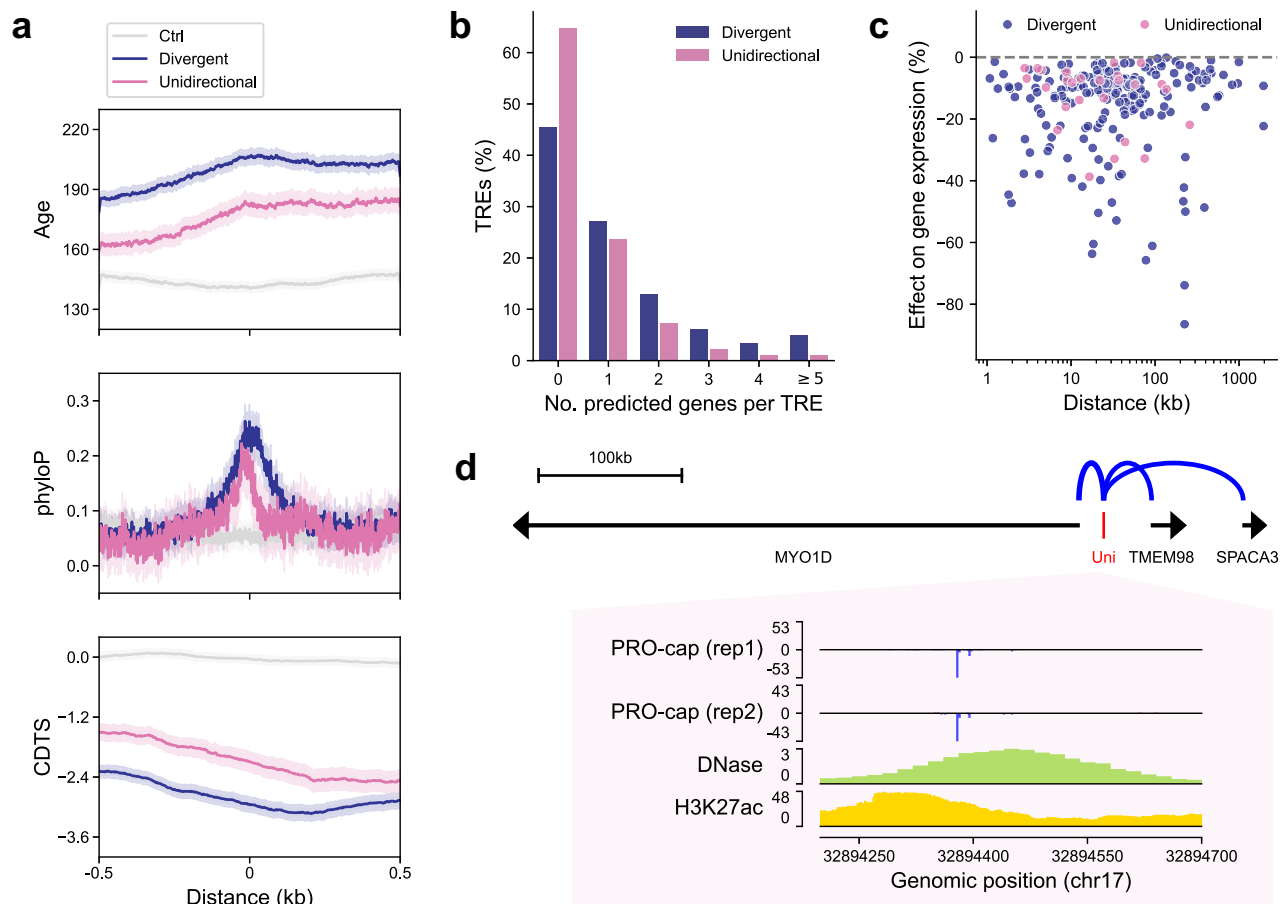


Fig. 3 | Functional features of divergent and unidirectional distal elements.

a Metaplots of sequence age (million years ago), phyloP, and CDTs across divergent and unidirectional distal elements. Distances are represented as ± 0.5 kb from the center. Metaplots display mean signals with 95% confidence intervals. **b** Histogram showing the number of genes predicted to be regulated by each divergent or unidirectional distal element as determined by the ABC model. **c** Percent change in

target gene expression following CRISPR perturbations to divergent and unidirectional distal elements. **d** Schematic (top) illustrating the regulatory effects of a unidirectional distal element on three target genes (*MYO1D*, *TMEM98*, and *SPACA3*) validated by CRISPR experiments. Below are PRO-cap, DNase-seq, and H3K27ac ChIP-seq signal tracks for this element. Source data are provided as Source Data files.

Taken together, our findings demonstrate that unidirectional elements are not solely a product of experimental artifacts, poor mappability, or insufficient sequencing depth.

A potential evolving process underlying directionality

Given the differences in overall architecture (i.e., one versus two transcription units) and the similarity in sequence content of individual transcription units, we hypothesized an evolving process from unidirectional to divergent distal elements. We first examined the sequence age of TREs, which, although not necessarily reflecting when a sequence first gained regulatory activity, provides an upper bound on the timing of activation. Sequence ages were assigned to TREs using a syntenic block aging strategy¹⁷, based on the evolutionary histories of overlapping syntenic blocks from the UCSC 100-way alignment across diverse species. We found that, while both divergent and unidirectional distal elements are significantly older than random genomic background, unidirectional elements are slightly younger than divergent counterparts (Fig. 3a). Within both TRE groups, the side with more PRO-cap reads tend to be older. To further assess evolutionary constraint, we analyzed two complementary metrics: phyloP¹⁸, which captures cross-species conservation, and context-dependent tolerance scores (CDTS)¹⁹, which reflect recent and ongoing purifying selection in the human lineage. We observed that phyloP scores delineate open chromatin regions, with the highest conservation at the

center for both TRE categories. Unidirectional elements exhibit a narrower phyloP peak, in line with the smaller functional regions observed in ChIP-seq profiles (Supplementary Fig. 2). The side with higher PRO-cap signal, which is generally older, is also less tolerant to genetic variation in the human population, as indicated by lower CDTs. The observed evolutionary patterns are consistent regardless of the anchoring strategy used, whether aligned at the element center (Fig. 3a) or at prominent TSSs (Supplementary Fig. 9). Together, these results suggest that unidirectional elements may represent functional genomic units with more recent evolutionary origins.

To explore potential mechanisms underlying the evolutionary transition from unidirectional to divergent transcription, we first analyzed motif enrichment in each TRE category using HOMER²⁰ (Supplementary Fig. 10a, Supplementary Fig. 11a, and Supplementary Data 2). Divergent and unidirectional distal elements display largely similar enriched motif types relative to background sequences. Notably, individual and composite motifs of GATA and TAL1, a core TF complex orchestrating erythroid differentiation and leukemia-related transcription programs²¹, are strongly enriched in both TRE categories. In contrast, CTCF/CTCF and ZNF143/THAP11 motifs show stronger enrichment in unidirectional distal and proximal elements.

To gain deeper insight into motif-level differences, we next focused on motif types with available ChIP-seq data in K562 and examined several properties, including motif occurrence, match score,

evolutionary conservation, and binding evidence. For most TFs, elements containing at least one instance of a given motif type show broadly similar numbers of motif occurrences per element across the two TRE categories (Supplementary Fig. 12a). Motif scores based on the best-matching instance are also comparable, although CTCF/CTCF (proximal) displays elevated scores in divergent ones (Supplementary Fig. 12b). PhyloP profiles further reveal increased conservation around ETS and AP-1 (distal) motifs in divergent elements, with CTCF/CTCF showing slightly elevated conservation in unidirectional ones (Supplementary Fig. 10b, and Supplementary Fig. 13, 14). When restricting to ChIP-seq-supported instances, most motif types are present in a larger fraction of divergent elements (Supplementary Fig. 10c and Supplementary Fig. 11b), suggesting coordinated gains in motif occupancy during the shift toward divergent transcription, whereas CTCF/CTCF motifs appear more frequently maintained in unidirectional elements. Overall, these results highlight the potential association between TF motif changes and transcriptional directionality over evolutionary time.

Transposable elements (TEs) are well-documented as a prolific source of regulatory elements, playing a crucial role in shaping the human regulatory landscape^{22,23}. To further explore the evolution of TREs, we identified TE-derived distal elements by requiring that overlapping TE sequences be at least half the size of the TREs, as a conservative estimate. Approximately 25% of unidirectional elements, slightly more than divergent ones, overlap with TEs (Supplementary Fig. 15a). The proportion and enrichment of TE classes are similar between unidirectional and divergent elements (Supplementary Fig. 15b, c). Notably, long terminal repeat (LTR) elements are significantly enriched relative to their genomic proportion, likely because they often retain their original cis-regulatory sequences post-insertion in the genome²².

To explore how the local regulatory environment influences the fate of TEs, we examined their spatial relationships with the closest non-TE TREs. We observed that TEs overlapping with TREs (d_2 and d_3) are significantly closer to non-TE elements compared to TEs not overlapping with any TREs (d_1), suggesting that TEs located near existing regulatory elements are more likely to acquire regulatory function themselves (Supplementary Fig. 15d). TEs overlapping with unidirectional elements (d_2), particularly those from LTR and long interspersed nuclear element (LINE) groups, are farther from non-TE divergent elements than TEs overlapping with divergent ones (d_3). These observations highlight an association between TE location and TRE directionality, raising the possibility that unidirectional TREs may, in some cases, evolve into divergent TREs over evolutionary time.

Unidirectional distal elements have gene regulatory functions

Considering their potentially recent evolutionary origins, we aimed to determine whether unidirectional elements have acquired regulatory functions. We first examined their genomic distribution relative to genes and observed that unidirectional distal elements have similar genomic distributions to their divergent counterparts (Supplementary Fig. 16), implying that they may play a role in gene regulation.

To evaluate the regulatory potential of unidirectional distal elements, we employed the activity-by-contact (ABC) model²⁴ to predict target genes in the K562 cell line. While a significant number of unidirectional distal elements are predicted to regulate target genes, a higher proportion of these elements have no predicted target genes than divergent distal elements (Fig. 3b). This result can be attributed to the lower activity of unidirectional distal elements, as indicated by DNase and H3K27ac ChIP-seq signals, despite having relatively comparable average 3D genome contact frequencies across all tested promoter-enhancer pairs with those involving divergent distal elements (Supplementary Fig. 17a). We further compared predicted target genes unique to either group or shared between both groups

(Supplementary Fig. 17b). The unique and shared target gene sets are under similar selection strength as measured by loss-of-function observed/expected upper bound fraction (LOEUF) scores²⁵, and exhibit comparable gene expression levels based on RNA-seq data (Supplementary Fig. 17c,d). Bergman et al.²⁶ identified two classes of promoters where P2 promoters showed stronger intrinsic activity and less responsiveness to enhancers. Among the target genes unique to unidirectional distal elements, there is a higher proportion associated with P2 promoters, followed by those unique to divergent distal elements and shared ones (Supplementary Fig. 17e). Consistent with this finding, the promoters of these target genes unique to unidirectional distal elements display higher DNase and H3K27ac signals compared to the other two groups (Supplementary Fig. 17f).

In addition to *in silico* target gene predictions, we evaluated the regulatory effects of unidirectional distal elements using previously curated CRISPR perturbation experiments²⁷. Although a larger fraction of divergent distal elements than unidirectional elements showed a significant ability to activate expression of cognate genes, 23 unidirectional distal elements were found to activate the expression of their target genes (Fig. 3c). For instance, the deletion of one unidirectional distal element significantly reduced the expression levels of three target genes (*MYO1D*, *TMEM98*, and *SPACA3*) by 30–40% (Fig. 3d). Altogether, these results demonstrate that at least a subset of unidirectional distal elements possess gene regulatory function.

CTCF binding blocks transcription initiation at TREs

Beyond the binary classification of divergent and unidirectional elements, we further investigated contributing factors underlying transcriptional directionality along a continuous spectrum. Noting our observation that CTCF and cohesin components are enriched in unidirectional elements in the K562 cell line (Fig. 1e and Supplementary Fig. 1d), we first examined whether these factors influence transcriptional directionality at TREs. A previous study by the Blobel lab demonstrated that CTCF regulates the directionality of divergent promoters by blocking antisense transcription⁵. To determine if CTCF also represses transcription initiation at enhancers, we performed PRO-cap in the human colorectal carcinoma cell line HCT116, both before and after inducing CTCF degradation using an auxin-inducible degron system.

In the unperturbed HCT116 cell line, we observed higher levels of CTCF and RAD21 ChIP-seq signals at unidirectional distal elements compared to divergent ones, consistent with our findings in the K562 cell line (Supplementary Fig. 18a). By further classifying divergent elements based on their absolute DI (Q1 representing the most balanced elements and Q4 the most skewed), we found that highly skewed elements exhibit stronger CTCF and RAD21 binding compared to more balanced ones (Supplementary Fig. 18b).

Next, we investigated the impact of acute CTCF depletion on transcription initiation at distal divergent and unidirectional elements. For subsequent analyses, the maximum TSS side is defined as the side with more read counts for divergent and unidirectional elements before CTCF degradation, while the opposite side is termed the minimum TSS side. We observed a significant increase in transcription levels at 223 distal elements on the minimum TSS side following CTCF loss (Fig. 4a, c), whereas the maximum TSS side remains largely unchanged (Supplementary Fig. 19a). The majority (~95%) of these transcriptionally upregulated elements are initially bound by CTCF, providing further evidence for CTCF's direct role in these transcriptional changes (Supplementary Fig. 18c). Furthermore, CTCF-bound elements in the upregulated group exhibit stronger CTCF and RAD21 binding compared to those in the unchanged group (Supplementary Fig. 18d). Notably, a higher proportion of unidirectional and transcriptionally skewed divergent elements, which have higher CTCF and RAD21 binding signals, show an increase in transcription after CTCF loss (Supplementary Fig. 18e). Moreover, elements with upregulated

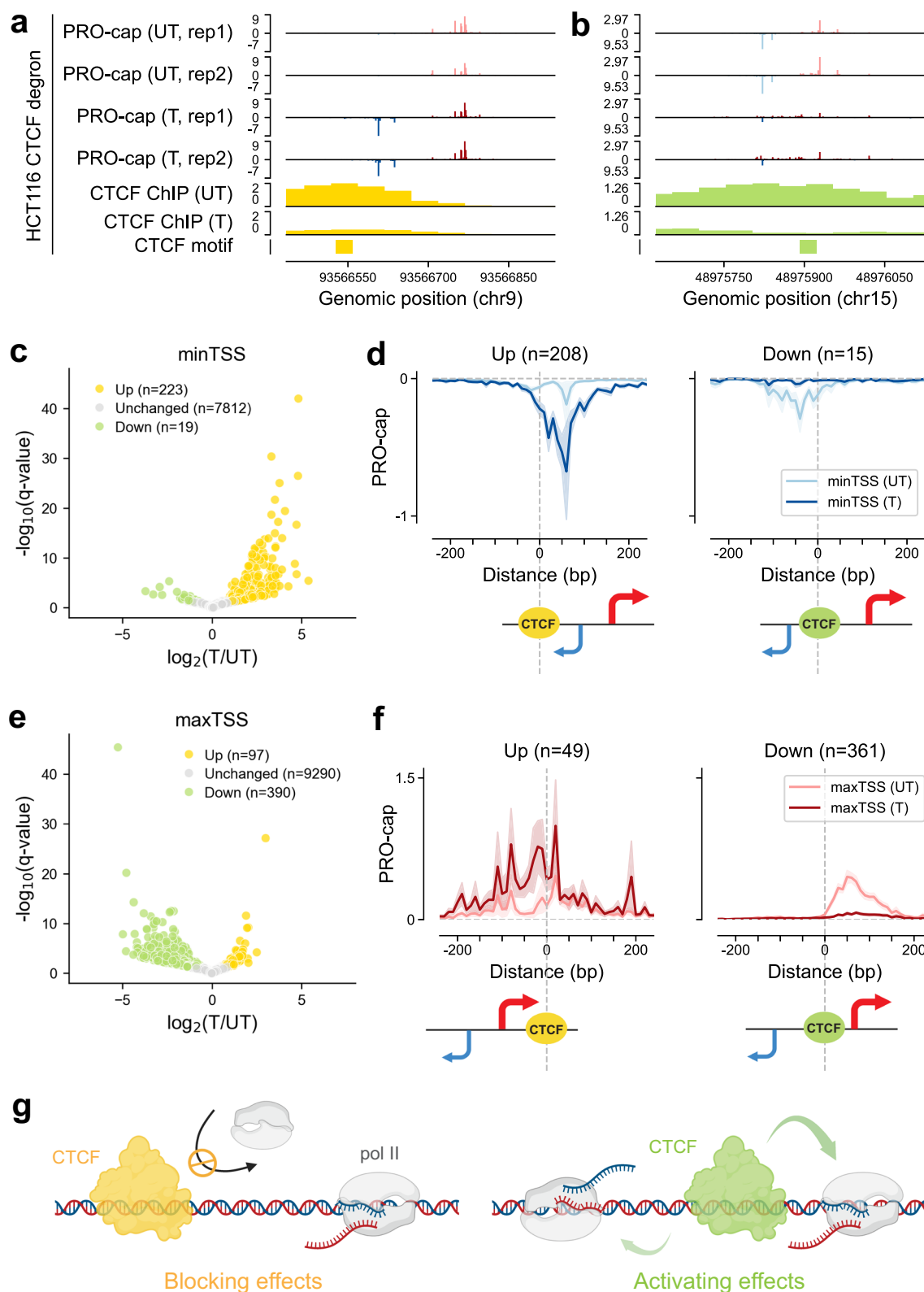


Fig. 4 | CTCF has a dual function in modulating transcription initiation.

a, b PRO-cap and CTCF ChIP-seq browser tracks of two representative distal elements, showing either upregulated (**a**) or downregulated (**b**) transcription before (untreated, UT) and after (treated, T) CTCF degradation. The positions of CTCF motifs are indicated and PRO-cap signal tracks from two replicates are shown. **c** Volcano plot of PRO-cap signals on the minimum TSS side of distal elements before and after CTCF degradation. Differentially expressed distal elements are highlighted in yellow for upregulation and green for downregulation. **d** Metaplots of PRO-cap signals (RPM normalized, 10-bp bins) on the minimum TSS side of distal

elements in the upregulated (left) and downregulated (right) groups, centered on CTCF motifs. Metaplots display mean signals with standard errors. A schematic of the CTCF motif position relative to transcription initiation sites is shown below each panel. Created in BioRender. Chen, Y. (<https://BioRender.com/i76a199>). **e** Same as (**c**), but for the maximum TSS side of distal elements. **f** Same as (**d**), but for the maximum TSS side of distal elements. Created in BioRender. Chen, Y. (<https://BioRender.com/i76a199>). **g** Illustration of CTCF's dual function in modulating transcription initiation based on its position relative to transcription

transcription are enriched at chromatin loop anchors (Supplementary Fig. 18f) and chromatin domain boundaries (Supplementary Fig. 18g). Thus, beyond the previous observation at promoters⁵, our results indicate that CTCF occupancy at both proximal and distal TREs blocks transcription initiation on one side, thereby tuning their transcription directionality.

CTCF has a dual function in modulating transcription in a position-dependent manner

Alongside the changes in transcription levels, we noted an intriguing shift in TSS positions following the depletion of CTCF. For subsequent analyses, downstream denotes the direction in which transcription proceeds after a given TSS, while upstream refers to the opposite direction. In the upregulated group, some divergent elements exhibit an upstream repositioning of their minimum TSSs, moving closer to their respective maximum TSSs (Supplementary Fig. 20a). In one representative element, transcription initiates downstream of the CTCF motif before degradation, but shifts upstream upon CTCF loss, suggesting that CTCF normally blocks these cryptic TSSs (Supplementary Fig. 20b, c).

To further evaluate the relationship between CTCF function and its position relative to the TSS, we analyzed changes in transcription initiation focused on the minimum TSS side of TREs centered on the CTCF motif. We found that in the transcriptionally upregulated group, the CTCF motif is predominantly located downstream of the minimum TSS (Fig. 4d). Conversely, in the transcriptionally downregulated group, the CTCF motif resides in the upstream region, suggesting that CTCF has a dual function in tuning transcription initiation depending on its position relative to the original TSS. Similar trends were noted when analyzing changes in PRO-cap signals at the maximum TSS side (Fig. 4b, e, f). Moreover, our analysis of changes in the minimum and maximum TSS sides revealed more elements in the upregulated and downregulated groups, respectively (Fig. 4c, e). This pattern reflects a biological bias towards identifying CTCF as a blocker on the minimum TSS side, characterized by lower transcription levels, and as an activator on the maximum TSS side, characterized by higher transcription levels. Collectively, these findings indicate that CTCF functions as a transcriptional blocker when located downstream of the TSS and as an activator when positioned upstream (Fig. 4g).

Furthermore, we observed that elements with upregulated transcription at the minimum TSS side or downregulated transcription at the maximum TSS side show limited changes in transcription levels at their respective opposite sides (Supplementary Fig. 19a–d). Additionally, transcriptionally skewed elements become more balanced either by increasing transcription levels at the minimum TSS side or decreasing them at the maximum TSS side following CTCF depletion (Supplementary Fig. 19e, f). Similar patterns were observed in proximal elements (Supplementary Fig. 21), supporting a dual role for CTCF in regulating the transcription levels and directionality of both distal and proximal elements in a position-dependent manner.

Additional motifs critical for transcriptional directionality of TREs

Since CTCF can only account for a limited proportion of elements with skewed transcription patterns, we expanded our search for additional contributors to directionality in a more systematic manner. We trained a ProCapNet model using all elements found in the HCT116 cell line before CTCF degradation (profile task correlation: 0.600 ± 0.004 ; count task correlation: 0.768 ± 0.014) to identify motifs critical for transcription level (Fig. 5a and Supplementary Fig. 22a). We reasoned that motifs with orientation- and/or position-dependent effects on total PRO-cap read counts could be a source of unbalanced transcription between the two strands.

We first assessed whether the model has learned CTCF's dual function in transcription initiation. We predicted PRO-cap signals

before and after *in silico* deletion of CTCF motifs in elements with up- or down-regulated transcription levels in the CTCF degron experiments (using the same elements and motif positions as shown in Fig. 4d-left, f-right). The model demonstrated the ability to predict motif position-dependent changes consistent with our observations from the CTCF perturbation experiments, even though it was trained using only PRO-cap data obtained before CTCF degradation (Supplementary Fig. 23).

Next, we examined which motifs exhibited orientation-dependent effects on transcription levels. While most motifs do not show a clear orientation bias, some (TATA box, CTCF, and ZNF143) are preferentially located on the same strand as the prominent TSSs of unidirectional elements (Supplementary Fig. 22b). The lack of bias observed for certain motifs, such as CREB and NRF1, can be attributed to their palindromic sequences. On the other hand, orientation preference may arise from structural constraints. For instance, the oriented interaction of TATA box binding protein to the TATA box aligns with the assembly of a pre-initiation complex possessing a unique polarity²⁸. Additionally, CTCF acts as a structural regulator of chromatin looping, where the orientation is crucial for loop topology and gene expression²⁹.

We then tested whether other TFs might also exhibit a dual function in a position-dependent manner, similar to what was observed with CTCF. Since the ProCapNet model learned CTCF's dual function solely from PRO-cap data before CTCF degradation, we reasoned that it might have also detected the dual function of other motifs, if any, without requiring PRO-cap data obtained after knocking down the corresponding TFs. We performed *in silico* deletion of motif instances and compared the fold change in transcription levels on either the minimum or maximum TSS side. In addition to CTCF, we found that ZNF143, API, SP, and NFY also exhibit dual roles in modulating transcription at TREs (Fig. 5b). Specifically, these TFs are more likely to activate transcription when located upstream of TSSs (decreased transcription after deletion) while repressing transcription when located downstream (increased transcription after deletion).

To experimentally validate the top candidate, we performed PRO-cap assays in HEK293T dTAG lines before and after ZNF143 degradation. We observed differential PRO-cap signals at TREs, particularly pronounced in proximal regions (Supplementary Fig. 24a,b). Regions showing PRO-cap changes exhibited strong ZNF143 ChIP-seq binding before treatment, which decreased following degradation (Supplementary Fig. 24c). In contrast, regions with unchanged PRO-cap signals showed weaker baseline binding. These results demonstrate that the observed PRO-cap changes are a direct consequence of ZNF143 degradation. Position-specific transcription profiles centered on ZNF143 motifs confirmed the dual effects predicted by the ProCapNet model trained in HCT116 cells (Fig. 5c). Importantly, this experimental validation in HEK293T cells demonstrated that these predicted effects generalize across cell types. A ProCapNet model trained directly on HEK293T PRO-cap data prior to ZNF143 degradation further recapitulated motif position-dependent changes (Supplementary Fig. 25). Taken together, motifs with orientation- and/or position-dependent effects on transcription are potential contributors to transcription directionality.

Discussion

Our high-quality PRO-cap data, combined with our peak calling tool PINTS, enabled us to identify a comprehensive list of divergent and unidirectional elements. Our systematic comparison of epigenetic features, protein binding profiles, and sequence content revealed that unidirectional elements have architectures that are distinct from those of divergent elements. Furthermore, our data demonstrate that a subset of unidirectional elements has acquired regulatory activity, supported by *in silico* target gene predictions and CRISPR perturbation experiments. Unidirectional elements typically have lower

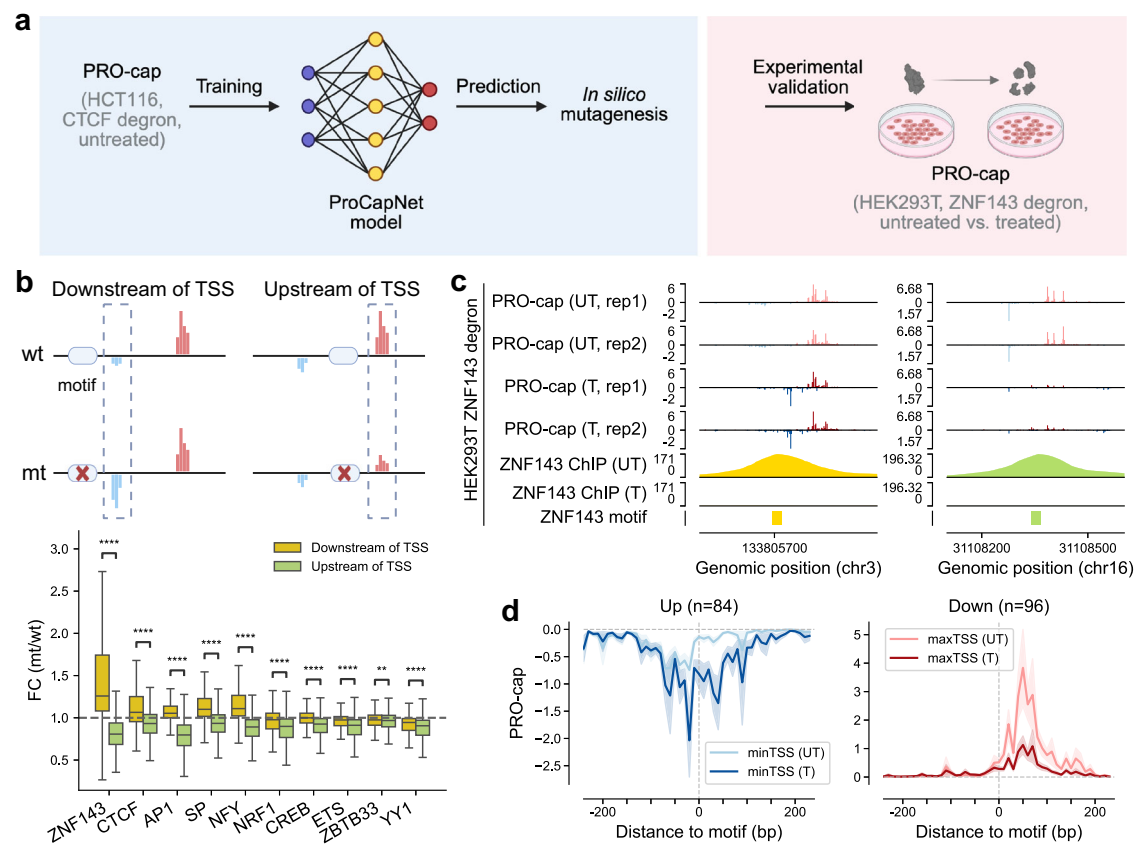


Fig. 5 | TFs with position-dependent dual functions in tuning transcription initiation. **a** Schematic of deep learning model training and prediction, followed by experimental validation of the top candidate identified by the model. Created in BioRender. Yu, H. (<https://BioRender.com/d9yuty2>). **b** (Top) Schematic of in silico motif deletion: motifs downstream of the minimum TSS were deleted to measure fold changes in transcription at the minimum TSS side, and motifs upstream of the maximum TSS were deleted to measure fold changes at the maximum TSS side. Created in BioRender. Yu, H. (<https://BioRender.com/u01b410>). (Bottom) Boxplots showing the resulting fold changes in predicted transcription levels at the respective TSS sides following these deletions. The center line indicates the median; box limits represent the upper and lower quartiles; whiskers extend to 1.5× the

interquartile range. Comparisons were performed using two-sided Mann-Whitney *U* tests. BH-adjusted *p* values (padj) are displayed as follows: **, 0.001 < padj ≤ 0.01; ****, padj ≤ 0.0001. **c** PRO-cap and ZNF143 ChIP-seq browser tracks of two representative proximal elements, showing either upregulated (left) or downregulated (right) transcription before (untreated, UT) and after (treated, T) ZNF143 degradation. The positions of ZNF143 motifs are indicated and PRO-cap signal tracks from two replicates are shown. **d** Metaplots of PRO-cap signals (RPM normalized, 10-bp bins) for proximal elements with increased transcription at the minimum side (left) or decreased transcription at the maximum side (right) upon ZNF143 perturbation, centered on the motif. Metaplots display mean signals with standard errors. Source data are provided as Source Data files.

transcription levels, which may partly account for their weaker enhancer potential relative to their divergent counterparts. This pattern aligns with another study from our group³⁰, where nascent transcription level was positively associated with enhancer activity across LentiMPRA, ATAC-STARR-seq, and WHG-STARR-seq assays. Our analyses also suggest a potential evolutionary process underlying the directionality of TREs. Future PRO-cap data from a diverse array of species may provide more conclusive evidence for this progression. Moreover, such data will enable the examination of how changes in directionality at TREs can affect the transcription levels of target genes, particularly those under positive selection that drive phenotypic adaptations over longer evolutionary timescales. In addition to potential evolutionary mechanisms, TF blocking may also contribute, at a more immediate regulatory level, to the establishment of unidirectional elements. Consistent with this model, we observed that 3–5% of unidirectional elements showed upregulation on the opposite strand upon CTCF degradation, suggesting that TF-mediated blocking contributes to restricting bidirectional activity at a subset of unidirectional elements.

To assess the robustness of the main findings, we compared the effects of directionality classification (divergent versus unidirectional) in simple regression models (Y - classification) and fully adjusted ones

that additionally included log(PRO-cap total counts), GC content, and mappability (Supplementary Table 1). In the simple models, directionality classification showed significant associations in the expected direction (i.e., divergent elements exhibited higher DNase-seq and H3K27ac signals, more target genes, higher phyloP, and lower CDTs). In the fully adjusted models, directionality classification remained significant for DNase-seq total counts, H3K27ac total counts, and number of target genes, with attenuated but directionally consistent effect sizes, indicating that classification captures regulatory differences beyond the included covariates. In contrast, phyloP and CDTs showed weaker unadjusted associations that were not robust to covariate adjustment. Interpretation of the adjusted associations is complicated by the role of log(PRO-cap total counts), which may function as a partial mediator if divergent architecture drives greater transcriptional output. In this case, adjustment may distort the estimated effect of directionality classification, so a non-significant coefficient after adjustment does not necessarily indicate the absence of a directionality effect, but may reflect shared biological signals among correlated features.

In addition, we explored other contributors to directionality using the CTCF and ZNF143 degron system coupled with PRO-cap, as well as ProCapNet deep learning models. We identified several TFs with dual

functions in modulating nascent transcription based on their relative positions to the TSS. Interestingly, another group recently reported this spatial grammar through independent lines of functional evidence³¹, including TF perturbation (NRF1 and YY1 knockdown csRNA-seq data in human U2OS cells; NFY knockdown Start-seq data in mouse embryonic fibroblasts), natural genetic variation (csRNA-seq data in bone-marrow-derived macrophages from two mouse strains; PRO-cap dataset of 67 human lymphoblastoid cell lines), and TSS-MPRA. Our study adds to this growing body of evidence by characterizing high-quality CTCF and ZNF143 perturbation datasets with the following advantages¹⁴: 1) PRO-cap offers better sensitivity and specificity in eRNA detection compared to csRNA-seq and Start-seq; 2) PRO-cap can identify precise TSS positions of TREs; and 3) our study utilized PINTS, which outperforms HOMER, the peak-calling tool used in their study, in identifying active TREs genome-wide. Additionally, we employed deep learning models to decode the spatial grammar of other TFs, providing complementary evidence for position-dependent functional profiles.

We also acknowledge the potential limitations in our study. To systematically delineate the features of unidirectional TREs, we performed stringent filtering of the elements to ensure the quality and fairness of the comparative analyses. While this approach enhanced the clarity of our findings, it also led to the exclusion of certain elements that may warrant further investigation in the future. For example, unidirectional elements that do not overlap with DNase peaks were excluded from our analyses. However, this exclusion does not necessarily indicate that these elements are merely noise. Instead, it is plausible that their presence could reflect the higher sensitivity of the PRO-cap assay or differences in cell populations assayed by PRO-cap and DNase-seq experiments. Additionally, elements with other TREs within a 500 bp region were excluded from this study because we were unable to confidently dissect the contribution of individual TREs, which could have confounded our analyses. Future studies on these elements could yield new insights into the interaction and coordination among TREs, potentially uncovering complex regulatory networks.

Overall, our dissection of the mechanistic underpinnings of transcriptional directionality augments our understanding of the rules and logic governing gene regulation.

Methods

Preparation of PRO-cap libraries

Library preparations for two replicates, each with 10 million K562 cells, were conducted as previously described^{12,32,33}. Likewise, we generated PRO-cap libraries from two replicates of 10 million HCT116 cells, genetically modified using CRISPR to target *H. sapiens* CTCF and *O. sativa* LOC4335696³⁴, before and after 6-h treatment with 1 μ M 5-phenyl-1H-indole-3-acetic acid. In addition, we performed PRO-cap on two replicates of 10 million HEK293T cells in which ZNF143 had been endogenously C-terminally tagged with a dTAG using CRISPR^{35–39}, before and after 6-h treatment with 500 nM dTAGV-1. Briefly, the cells were permeabilized, followed by run-on reactions. After isolating RNA, two adaptor ligations and reverse transcription were performed using custom adaptors. 5' cap selection reactions were conducted between adaptor ligations through a series of enzymatic steps. RNA washes, phenol:chloroform extractions, and ethanol precipitations were performed between reactions, all under RNase-free conditions. Libraries were sequenced on Illumina's NovaSeq platform after PCR amplification and library clean-up. Details of library information can be found in Supplementary Data 1.

Data preprocessing

This dataset was managed and analyzed using the Resource Management System (<https://github.com/aldenleung/rmsp/>). Raw reads were preprocessed with fastp³⁶ (v0.23.4) for adapter trimming and unique

molecular index (UMI) processing, retaining only reads ≥ 18 bp for downstream analyses. The processed reads were aligned to the human reference genome hg38 (GCA_000001405.15) and ribosomal DNA (UI3369.1) using STAR³⁷ (v2.7.11a). For the HCT116 PRO-cap library with *Drosophila* spike-in, the alignment also included the *Drosophila* reference genome (dm6). Uniquely mapped reads were filtered with samtools³⁸ (v1.18) and deduplicated using umi_tools³⁹ (v1.1.5). The PCR-free, uniquely mapped reads obtained from replicates of each sample were converted to bigwig format and merged with biodata-tools (v0.0.7) for peak calling.

Identification and post-processing of TREs

We performed peak calling with PINTS (v1.1.10), using the parameter `--min-lengths-opposite-peaks` set to 5. The identified peaks were then classified into two categories: (1) divergent peak pairs that consist of peaks located on opposite strands and within 300 bp of each other, and (2) unidirectional peaks that originated from only one strand. These elements were further divided into proximal (within ± 500 bp) and distal (outside ± 500 bp) categories based on their distances to TSSs of all genes annotated by GENCODE v37⁴⁰.

All elements were resized to a ± 250 bp window centered on the anchor point to minimize size-related bias in subsequent analyses. For divergent elements, the anchor point was defined as the midpoint between prominent TSSs on the plus and minus strands. For unidirectional elements, the center of overlapping DNase peaks was used as a proxy. To do so, the following inclusion criteria were specifically applied for unidirectional elements: (1) overlapping with DNase peaks; and (2) the center of the DNase peak must be appropriately positioned (i.e., on the left of the prominent TSS if the unidirectional element is on the forward strand, or on the right if on the minus strand). The resized divergent and unidirectional elements were further filtered using the following exclusion criteria: (1) elements extending beyond chromosome ends; (2) overlapping with ENCODE blacklist regions⁴¹; (3) original peak region not fully contained within the resized region; and (4) overlapping with any other elements.

Metaplots and heatmaps

To make the comparison and visualization more straightforward in the metaplots, the sides with more PRO-cap read counts were flipped to the right regardless of their relative orientation to genomic positions. The random genomic regions depicted in the metaplots were generated as controls with the following criteria: (1) matched on chromosome number and element size, and (2) not overlapping GENCODE v37 exon regions and ENCODE blacklist regions. For metaplots in 10-bp bins, signals are summed within consecutive 10-bp windows tiled from the anchor point outward. The anchor bin (position 0) spans -5 to $+4$ bp relative to the anchor point, and each flanking position $\pm n$ represents a 10-bp window whose center is $10n$ bp from the anchor. Signals are summed within each bin for each element before averaging across elements. In the heatmaps, elements were sorted based on the distance between their prominent TSSs and the anchor points. Any deviations from this sorting method were carefully noted.

Annotating elements with protein binding features

We downloaded human TF ChIP-seq narrowPeak bed files in K562 cells under unperturbed conditions from the ENCODE database⁴², comprising 430 datasets targeting 310 proteins. For datasets processed by the ENCODE4 pipeline, we used "IDR thresholded peaks," and for datasets only available from the ENCODE3 pipeline, we used "optimal IDR thresholded peaks." When both replicate-level and pooled files were available, we consistently analyzed the pooled files. To annotate elements with protein binding features, we overlapped them with all narrowPeak files. For a given dataset, each element was assigned the maximum *signalValue* of any overlapping peak, or 0 if no peak overlapped. We then computed the fold-change (FC) of average signals

between divergent and unidirectional categories. This analysis was restricted to datasets where both groups had non-zero means and at least one group contained a minimum of 30 elements with non-zero values. *P* values were calculated using two-sided Welch's *t* tests, followed by Benjamini-Hochberg (BH) correction. Enrichment was defined as $|FC| > 1.5$ and adjusted *p* values (padj) < 0.05 .

Clustering of protein binding profiles

We downloaded the corresponding “fold change over control” bigWig files for all datasets used in the “Annotating elements with protein binding features” section. For each dataset, we focused on a 1 kb window centered on each element and computed the average ChIP-seq signal at each position across all elements. To enable fair comparison across datasets with different sequencing depths, position-specific signals were converted to Z-scores within each dataset. K-means clustering ($k = 4$, selected using the elbow method) was performed for the divergent proximal category, which showed the clearest patterns. For proteins with multiple datasets, we evaluated consistency across datasets: if more than half of the datasets assigned the protein to the same cluster, inconsistent datasets were removed; otherwise, all datasets for that protein were excluded.

Detection of core promoter elements

The presence of an initiator (Inr) was annotated for each prominent TSS when the sequence of -3 to $+3$ matched the consensus sequence BBCABW. The TATA box (-32 to -21 relative to the prominent TSS) and the downstream promoter region (DPR) elements ($+17$ to $+35$ relative to the prominent TSS) were identified using a previously published support vector regression model⁴³.

TF motif features

We performed known motif enrichment analysis for each TRE category using HOMER's default motif database, with background sequences matched for GC content and length. Significant motifs (HOMER *p* value $< 1e-10$) in any TRE category were retained as the union set. The complete enrichment results are provided in Supplementary Data 2. To provide a concise overview of motif enrichment (Supplementary Fig. 10a, and Supplementary Fig. 11a), we grouped motifs by sequence similarity. Similarity was assessed using Tomtom, and motifs were hierarchically clustered based on $-\log_{10}(q \text{ value})$. For each sequence-similarity group within a given TRE category, we displayed one representative motif showing the strongest enrichment, limited to motifs whose corresponding TF has TPM > 1 in K562.

For subsequent analyses, we shifted from TF families to individual motif types, restricting to those with available ChIP-seq data in K562 to ensure a consistent set across analyses and examine multiple aspects of these motifs. Because different analyses address distinct biological questions, we applied tailored strategies for selecting motif instances. For motif occurrence per element, motif scores and phyloP analyses, ChIP-seq support was not required, allowing assessment of intrinsic sequence features and evolutionary constraint across TREs, while providing a complementary perspective to the ChIP-seq-supported motif enrichment analysis.

Motif occurrence per element

To quantify how often each motif appears in divergent versus unidirectional elements, we considered elements with at least one instance of a given motif type. Overlapping instances of the same motif type were merged, and the number of non-overlapping occurrences per element was counted to provide a more conservative estimate of motif frequency.

Motif scores and phyloP

To examine the strength of motif matches and local evolutionary constraint, we selected the single best-matching instance of each motif

type per element. HOMER motif scores reflect the similarity between a sequence and the position-weight matrix (higher scores indicate stronger matches). Motif scores were compared between categories using Mann-Whitney *U* tests, followed by BH correction. For phyloP analysis, we plotted position-specific scores in a 30-bp window centered on each motif (Supplementary Fig. 13, 14), accounting for their orientation. In Supplementary Fig. 10b, phyloP profiles were aligned with sequence logos.

Motif enrichment

To compare divergent and unidirectional TREs directly, we performed enrichment analyses using best-matching motif instances that also overlapped a ChIP-seq peak (Supplementary Fig. 10c, and Supplementary Fig. 11b). This yields a high-confidence set supported by both sequence and binding data. We used Fisher's exact test followed by BH correction.

ProCapNet model training

We trained the ProCapNet models (<https://github.com/kundajelab/ProCapNet>) for each biosample (K562, HCT116 prior to CTCF degradation, and HEK293T prior to ZNF143 degradation) using default parameters except that we set `in_window` as 1000, `out_window` as 500, and `source_fracs` as [1,0]. To ensure sufficient training dataset, we included all bidirectional and unidirectional elements (direct output from PINTS without additional filtering). For the model discussed in the “Potential confounders for binary classification” section, we additionally excluded elements that overlapped with unidirectional elements we aimed to predict at various downsampled levels. In the “Additional motifs critical for transcriptional directionality of TREs” section, we computed base-level contribution scores using DeepSHAP⁴⁴ and applied TF-ModISco⁴⁵ to identify de novo motifs from these scores. The resulting sequence patterns were matched to known TF motifs using TOMTOM. Motif instances from the filtered set of TREs (see the “Identification and post-processing of TREs” section) were called with Fi-NeMo (<https://github.com/kundajelab/Fi-NeMo>), which leverages contribution scores for improved sensitivity and specificity.

Statistical testing of strand imbalance

We applied an empirical Bayes beta-binomial model to the original K562 PRO-cap dataset (the processed elements in the section “Identification and post-processing of TREs”) to test whether observed strand imbalance exceeds stochastic sampling expectations at a given read count. This framework explicitly accounts for sampling variability while allowing overdispersion across elements. We defined *p* as the proportion of forward reads and modeled its distribution across elements using a beta distribution with $\alpha = \beta$, such that the distribution is symmetric around $p = 0.5$. We started with empirical Bayes, where the parameter α was estimated separately for distal and proximal elements using maximum likelihood. Simulations from the fitted beta-binomial model accurately recapitulated the observed distribution of forward-read proportions (Supplementary Fig. 7a, and Supplementary Fig. 8a), indicating that the model captures the underlying variability in the data. Furthermore, consistent with empirical Bayes shrinkage, posterior mean estimates for low-count elements were pulled toward 0.5, whereas estimates for high-count elements remained close to their observed values (Supplementary Fig. 7b, and Supplementary Fig. 8b), demonstrating that the model appropriately accounts for sampling uncertainty in a count-dependent manner. The percentage of elements with posterior probability > 0.99 of being biased toward either strand (i.e., $\Pr(p > 0.5) > 0.99$ or $\Pr(p < 0.5) > 0.99$) was calculated across read-count bins. We also tested other prior distributions with Jeffreys prior ($\alpha = \beta = 0.5$, a non-informative, U-shaped prior), uniform prior ($\alpha = \beta = 1$, a flat prior that assigns equal density to all values of *p*), and $\alpha = \beta = 2$ (a bell-shaped prior that strongly favors values around

$p=0.5$), to test whether the pattern of strand imbalance beyond sampling noise is robust across these choices.

Target gene prediction

We employed the ABC model to predict enhancer-gene interactions in the K562 cell line, following the protocol described in (<https://abc-enhancer-gene-prediction.readthedocs.io/en/latest/>). A threshold of 0.025 was chosen as it achieved 70% recall in the CRISPR benchmark following the pipeline described in (https://github.com/EngreitzLab/CRISPR_comparison). We assigned our PRO-cap distal elements to ABC regions that have an overlapping length of at least half the element size.

Enhancer-gene interactions validated by CRISPR screens

To validate the regulatory effects of our distal elements, we intersected them with regions screened by CRISPR experiments curated at (https://github.com/EngreitzLab/CRISPR_comparison), requiring an overlapping length of at least half the element size.

Sequence age and evolutionary metrics

To estimate the sequence age of TREs, we employed a syntenic block aging strategy as described by Fong et al.¹⁷. We obtained the hg38 100-way vertebrate Multiz alignment data from the UCSC database⁴⁶ and assigned an age to each syntenic block based on the most recent common ancestor (MRCA) of the species in the alignment block. To simplify the analysis, we grouped MRCA nodes into 10 age categories and reported the age as the oldest ancestral branch. phyloP scores (100-way) were downloaded from the UCSC database to measure the conservation across species, while CDTs computed on 7794 unrelated individuals was obtained from the literature¹⁹ to assess recent and ongoing purifying selection in the human lineage.

Annotation of TE-derived TREs

Genomic TE annotations were obtained from RepeatMasker⁴⁷. TE-derived distal elements were defined by requiring overlapping TE sequences to be at least half the size of the TREs, which serves as a conservative estimate. The enrichment of TE families for TREs was calculated as $\log_2 \frac{\text{No. TREs derived from TEs in a given family} / \text{No. total TREs}}{\text{Total length of TEs in a given family} / \text{Genome size}}$.

CTCF-related analysis

We only included elements present in the HCT116 cell line before CTCF degradation. All elements were resized to a ± 250 bp window centered on the anchor point, as described in the “Identification and post-processing of TREs” section to quantify transcriptional changes in the maximum and minimum TSS sides. Differential expression analysis was performed using DESeq2 with thresholds set as $|FC| > 1.5$ and $\text{padj} < 0.05$. We used FIMO⁴⁸ with default parameters to identify matches for the CTCF motifs (MA0139.2, MA1929.2, and MA1930.2) in the JASPAR database⁴⁹. For motif-centered plots, instances from different motif variants were aligned based on a shared core motif sequence. Only elements with high-confidence CTCF motifs and confirmed ChIP-seq binding in unperturbed HCT116 cells (ENCSR000D0T) were included to ensure precise positioning. For visualization of representative loci, we used CTCF ChIP-seq data from Auxin-treated and untreated HCT116 CTCF degran cells (GSE179540).

We evaluated the proportion of CTCF motifs located at chromatin loop anchors using the CTCF chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) data in HCT116 (ENCF431BUN) obtained from the ENCODE database. The insulation scores were calculated using Hi-C data in HCT116 (ENCF5730PJ) with window size of 30 kb via cooltools⁵⁰. Locations with lower scores, indicating reduced contact frequencies between upstream and downstream loci, are often referred to as boundaries.

ZNF143-related analysis

We only included elements present in the HEK293T cell line before ZNF143 degradation. All elements were resized to a ± 250 bp window centered on the anchor point, as described in the “Identification and post-processing of TREs” section to quantify transcriptional changes in the maximum and minimum TSS sides. Differential expression analysis was performed using DESeq2 with thresholds set as $|FC| > 1.5$ and $\text{padj} < 0.05$. We used FIMO to identify ZNF143 motif instances using position-weight matrices from JASPAR (MA0088.2) and a motif derived from TF-MoDisco. For motif-centered plots, instances from different motif variants were aligned based on a shared core motif sequence. Only elements with high-confidence motifs and confirmed ChIP-seq binding in unperturbed HEK293T cells (GSE266489) were included to ensure precise positioning.

Regression analysis

We fitted regression models separately for distal and proximal elements. Directionality classification (coded as 0 = unidirectional, 1 = divergent) served as the primary predictor, with adjustment for $\log(-\text{PRO-cap total counts})$, GC content, and mean mappability. Variance inflation factors (VIFs) were computed for all predictors, and no substantial multicollinearity was observed (all VIFs < 5). We modeled multiple element-level features as outcomes, including DNase-seq total counts, H3K27ac ChIP-seq total counts, the number of target genes (for distal elements only), maximum phyloP (higher values indicate stronger conservation), and minimum CDTs (lower values indicate greater intolerance to variation). Negative binomial regression was used for count outcomes (DNase-seq total counts, H3K27ac total counts, and number of target genes) to account for overdispersion. Ordinary least squares regression with heteroskedasticity-robust standard errors was used for continuous outcomes (phyloP and CDTs). We evaluated the effect of directionality classification in both unadjusted models (Y - classification) and fully adjusted models including covariates. *P* values for the directionality classification predictor were adjusted using BH correction.

Statistics and reproducibility

No statistical method was used to predetermine sample size. The experiments were not randomized. The Investigators were not blinded to allocation during experiments and outcome assessment. PRO-cap replicates were highly reproducible, with Pearson correlation coefficient of 0.90–0.94 for read counts in 50-bp genomic bins.

Schematics

All schematics in Figs. 4, 5, and Supplementary Figs. 1, 15, 20, 21, and 23 were created using BioRender, with the appropriate publication licenses.

Reporting summary

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

Data availability

The raw and processed PRO-cap datasets have been deposited in the ENCODE portal and publicly available under the following accession numbers: K562 (ENCSR220XSM); HCT116 CTCF_U (ENCSR165QKR); HCT116 CTCF_T (ENCSR261OWE); HEK293T ZNF143_U (ENCSR473KVF); HEK293T ZNF143_T (ENCSR733DAS). The raw PRO-cap sequencing data are also available in the NCBI SRA under the following accession numbers: K562 (SRX27358268 and SRX27358269); HCT116 CTCF_U (SRX18485643 and SRX18485644); HCT116 CTCF_T (SRX18485830 and SRX18485831); HEK293T ZNF143_U (SRX33762549 and SRX33762550); HEK293T ZNF143_T (SRX33762551 and

SRX33762552). Source data are provided with this paper (<https://doi.org/10.5281/zenodo.20672301> [<https://zenodo.org/records/20672301>])⁵¹. A detailed summary of the public datasets is listed at https://github.com/haiyuan-yu-lab/TRE_directionality/blob/main/data_preprocessing/2.Other_resources.ipynb.

Code availability

All code is available at the GitHub repository https://github.com/haiyuan-yu-lab/TRE_directionality under the MIT License and archived on Zenodo (DOI: 10.5281/zenodo.20672574 [<https://zenodo.org/records/20672574>])⁵².

References

- Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008).
- Seila, A. C. et al. Divergent transcription from active promoters. *Science* **322**, 1849–1851 (2008).
- Balbin, O. A. et al. The landscape of antisense gene expression in human cancers. *Genome Res* **25**, 1068–1079 (2015).
- Bornelöv, S., Komorowski, J. & Wadelius, C. Different distribution of histone modifications in genes with unidirectional and bidirectional transcription and a role of CTCF and cohesin in directing transcription. *BMC Genom.* **16**, 300 (2015).
- Luan, J. et al. CTCF blocks antisense transcription initiation at divergent promoters. *Nat. Struct. Mol. Biol.* **29**, 1136–1144 (2022).
- Dudnyk, K., Cai, D., Shi, C., Xu, J. & Zhou, J. Sequence basis of transcription initiation in the human genome. *Science* **384**, eadj0116 (2024).
- Kim, T.-K. et al. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).
- Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
- Tippens, N. D. et al. Transcription imparts architecture, function and logic to enhancer units. *Nat. Genet.* **52**, 1067–1075 (2020).
- Kristjánsdóttir, K., Dziubek, A., Kang, H. M. & Kwak, H. Population-scale study of eRNA transcription reveals bipartite functional enhancer architecture. *Nat. Commun.* **11**, 5963 (2020).
- Kouno, T. et al. C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution. *Nat. Commun.* **10**, 360 (2019).
- 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).
- 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).
- Yao, L. et al. A comparison of experimental assays and analytical methods for genome-wide identification of active enhancers. *Nat. Biotechnol.* **40**, 1056–1065 (2022).
- Karimzadeh, M., Ernst, C., Kundaje, A. & Hoffman, M. M. Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Res* **46**, e120 (2018).
- Cochran, K. et al. Dissecting the cis-regulatory syntax of transcription initiation with deep learning. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.05.28.596138>. (2024).
- Fong, S. L. & Capra, J. A. Modeling the evolutionary architectures of transcribed human enhancer sequences reveals distinct origins, functions, and associations with human trait variation. *Mol. Biol. Evol.* **38**, 3681–3696 (2021).
- Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**, 110–121 (2010).
- di Iulio, J. et al. The human noncoding genome defined by genetic diversity. *Nat. Genet.* **50**, 333–337 (2018).
- Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
- Han, G. C. et al. Genome-wide organization of GATA1 and TAL1 determined at high resolution. *Mol. Cell Biol.* **36**, 157–172 (2016).
- Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).
- Fueyo, R., Judd, J., Feschotte, C. & Wysocka, J. Roles of transposable elements in the regulation of mammalian transcription. *Nat. Rev. Mol. Cell Biol.* **23**, 481–497 (2022).
- Fulco, C. P. et al. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
- Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
- Bergman, D. T. et al. Compatibility rules of human enhancer and promoter sequences. *Nature* **607**, 176–184 (2022).
- Gschwind, A. R. et al. An encyclopedia of enhancer-gene regulatory interactions in the human genome. *bioRxiv.org* <https://doi.org/10.1101/2023.11.09.563812>. (2023).
- Kays, A. R. & Schepartz, A. Virtually unidirectional binding of TBP to the AdMLP TATA box within the quaternary complex with TFIIA and TFIIB. *Chem. Biol.* **7**, 601–610 (2000).
- Guo, Y. et al. CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. *Cell* **162**, 900–910 (2015).
- Zhang, J. et al. Comprehensive evaluation of diverse massively parallel reporter assays to functionally characterize human enhancers genome-wide. *Genome Biol.* **26**, 378 (2025).
- Duttke, S. H. et al. Position-dependent function of human sequence-specific transcription factors. *Nature* **631**, 891–898 (2024).
- Mahat, D. B. et al. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat. Protoc.* **11**, 1455–1476 (2016).
- Judd, J. et al. A rapid, sensitive, scalable method for Precision Run-On sequencing (PRO-seq). Preprint at *bioRxiv* <https://doi.org/10.1101/2020.05.18.102277> (2020).
- Yesbolatova, A. et al. The auxin-inducible degron 2 technology provides sharp degradation control in yeast, mammalian cells, and mice. *Nat. Commun.* **11**, 5701 (2020).
- Dong, J., Sathyan, K. M., Scott, T. G., Mukherjee, R. & Guertin, M. J. ZNF143 binds DNA and stimulates transcription initiation to activate and repress direct target genes. *Nucleic Acids Res.* **53**, (2025).
- Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
- Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
- Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Giga-science* **10**, (2021).
- Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res* **27**, 491–499 (2017).
- Frankish, A. et al. GENCODE 2021. *Nucleic Acids Res* **49**, D916–D923 (2021).
- Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE blacklist: identification of problematic regions of the genome. *Sci. Rep.* **9**, 9354 (2019).
- ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
- Vo Ngoc, L., Huang, C. Y., Cassidy, C. J., Medrano, C. & Kadonaga, J. T. Identification of the human DPR core promoter element using machine learning. *Nature* **585**, 459–463 (2020).

44. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. *arXiv* <https://doi.org/10.48550/arXiv.1705.07874> (2017).
45. Shrikumar, A. et al. Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5.6.5. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.1811.00416> (2018).
46. Perez, G. et al. The UCSC Genome Browser database: 2025 update. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkae974>. (2024).
47. Open, R. 4.0 <http://www.repeatmasker.org>. Preprint at (2015).
48. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
49. Rauluseviciute, I. et al. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **52**, D174–D182 (2024).
50. Open2C et al. Cooltools: enabling high-resolution Hi-C analysis in Python. *PLoS Comput. Biol.* **20**, e1012067 (2024).
51. Chen, Y. et al. Directionality of transcriptional regulatory elements. Preprint at <https://doi.org/10.5281/ZENODO.20672300> (2026).
52. Chen, Y. Haiyuan-Yu-lab/TRE_directionality: v10. *Zenodo*, <https://doi.org/10.5281/ZENODO.20672574>. (2026).

Acknowledgements

We would like to thank Jennifer Jou and Ben Hitz for assistance with our PRO-cap data uploads to the ENCODE portal. We appreciate the ENCODE consortium for generating other datasets used in this paper and providing HCT116 CTCF mAID lines. We are grateful to the Guertin lab at the University of Connecticut for kindly providing HEK293T ZNF143 dTAG cells.

Author contributions

H.Y. and J.T.L. conceived and oversaw the study. Y.C. and S.R.S. designed the study and interpreted the results with input from all authors. S.R.S. and M.I.P. performed the PRO-cap experiments. Y.C. and A.K.L. implemented the computational analyses. C.G. prepared nuclei from HEK293T degron cell lines. K.C. developed the ProCapNet model and A.W. developed Fi-NeMo, both under the supervision of A.K. Y.C. and S.R.S. wrote the manuscript, and all authors approved the final version. H.Y., J.T.L. and A.G.C. supervised the project, edited the manuscript, and provided funding and resources.

Funding

We acknowledge funding support from the National Institutes of Health (R01 DK115398 and R01 AG077899 to H.Y.; UM1 HG009393, 5R01HG012970, and RM1 5RM1GM139738 to J.L. and H.Y.; 5U24HG007234, U01HG009431, and U01HG012069 to A.K.). S.R.S. was supported in part by the Center for Vertebrate Genomics Scholar Award and the Breast Cancer Coalition's Postdoctoral Fellowship. K.C. was supported in part by the Pierre & Christine Lamond Stanford Graduate Fellowship.

Competing interests

S.R.S. is an equity holder and member of the scientific advisory board of NeuScience, Inc., a consultant at Third Bridge Group Limited and Dialectica, and a member of the Venture Scout Community at Juniper VC, which are not related to this work. A.K. is on the scientific advisory board of SerImmune, AINovo, TensorBio and OpenTargets. A.K. was a scientific co-founder of RavelBio, a paid consultant with Illumina, was on the SAB of PatchBio and owns shares in DeepGenomics, Immunai, Freenome and Illumina. K.C. is a paid consultant with ImmunoVec and owns shares in Inceptive Nucleics. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-026-76415-5>.

Correspondence and requests for materials should be addressed to Andrew G. Clark, John T. Lis or Haiyuan Yu.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.