




Genome-wide characterization of mammalian promoters with distal enhancer functions

Lan T M Dao^{1,6,7}, Ariel O Galindo-Albarrán^{1,7} , Jaime A Castro-Mondragon¹, Charlotte Andrieu-Soler²⁻⁴, Alejandra Medina-Rivera⁵, Charbel Souaid¹ , Guillaume Charbonnier¹, Aurélien Griffon¹, Laurent Vanhille¹, Tharshana Stephen^{2,4}, Jaafar Alomairi¹, David Martin⁴, Magali Torres¹, Nicolas Fernandez¹, Eric Soler²⁻⁴, Jacques van Helden¹ , Denis Puthier¹ & Salvatore Spicuglia¹

Gene expression in mammals is precisely regulated by the combination of promoters and gene-distal regulatory regions, known as enhancers. Several studies have suggested that some promoters might have enhancer functions. However, the extent of this type of promoters and whether they actually function to regulate the expression of distal genes have remained elusive. Here, by exploiting a high-throughput enhancer reporter assay, we unravel a set of mammalian promoters displaying enhancer activity. These promoters have distinct genomic and epigenomic features and frequently interact with other gene promoters. Extensive CRISPR–Cas9 genomic manipulation demonstrated the involvement of these promoters in the *cis* regulation of expression of distal genes in their natural loci. Our results have important implications for the understanding of complex gene regulation in normal development and disease.

Regulation of mammalian gene transcription is accomplished through the involvement of transcription start site (TSS)-proximal (promoter) and TSS-distal (enhancer) regulatory elements¹. The original definition of a promoter entails the capability to induce local gene expression, whereas the term enhancer implies the property of activating gene expression at a distance. However, this basic dichotomy of *cis*-regulatory elements has been challenged by broad similarities between promoters and enhancers, such as DNA sequence features, chromatin marks, RNA polymerase II (Pol II) recruitment and bidirectional transcription^{1–5}. Despite several findings suggesting that promoters might display enhancer activity^{6–15}, including experimental observations that enhancer elements can work as alternative promoters¹⁶, it is unclear what fraction of promoters is concerned by this property and whether their enhancer activity is involved in distal gene regulation. The advent of high-throughput reporter assays, such as STARR-seq¹³, has enabled the identification of enhancer activity solely on the basis of functionality instead of using epigenomics or location criteria¹⁷. We previously developed CapStarr-seq¹⁸, a strategy coupling capture of a region of interest with STARR-seq, allowing efficient assessment of enhancer activity in mammals. By performing CapStarr-seq in several mammalian cell lines, we found that 2–3% of coding-gene promoters display enhancer activity in a given cell line. In comparison to classical promoters and distal enhancers, these TSS-overlapping enhancers (hereafter referred to as Epromoters) displayed distinct genomic and epigenomic features and were associated with stress

response. By using comprehensive CRISPR–Cas9 genomic deletions, we demonstrated that Epromoters are involved in the *cis* regulation of the expression of distal genes in their natural context, therefore functioning as bona fide enhancers. Furthermore, human genetic variation within Epromoters was associated with a strong effect on distal gene expression. We suggest that regulatory elements with dual roles as transcriptional promoters and enhancers might ensure rapid and coordinated regulation of gene expression. These findings will enhance understanding of complex gene regulation in normal development and diseases and of how genetic variation influences the control of gene expression programs.

RESULTS

Mouse TSS-proximal DHSs display enhancer activity

To further decipher the complex relationship between proximal and distal regulatory regions for coding genes, we compared the proportions of enhancer activity for subsets of proximal and distal DNase I-hypersensitive sites (DHSs) in T cell precursors based on our previously published CapStarr-seq experiments performed in the mouse P5424 T cell and NIH-3T3 fibroblast cell lines^{13,18} (Fig. 1a,b and Supplementary Table 1). We observed that the proportions of DHSs with enhancer activity were very similar for the proximal (<1 kb from the TSS) and distal subsets in P5424 cells (Fig. 1c, left). To avoid artifactual calling of enhancer activity due to sporadic transcription from the vector¹⁹ or initiation from the promoter itself,

¹Aix-Marseille University, INSERM, TAGC, UMR 1090, Marseille, France. ²INSERM, UMR 967, CEA/DRF/IRCM, Laboratory of Molecular Hematopoiesis, Université Paris–Diderot, Université Paris–Saclay, Fontenay-aux-Roses, France. ³Labex GR-Ex, Université Sorbonne Paris Cité, Paris, France. ⁴IGMM, CNRS, Université de Montpellier, Montpellier, France. ⁵Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de México, Juriquilla, Mexico. ⁶Present address: Vinmec Research Institute of Stem Cell and Gene Technology, Hanoi, Vietnam. ⁷These authors contributed equally to this work. Correspondence should be addressed to S.S. (salvatore.spicuglia@inserm.fr).

Received 10 January; accepted 1 May; published online 5 June 2017; doi:10.1038/ng.3884

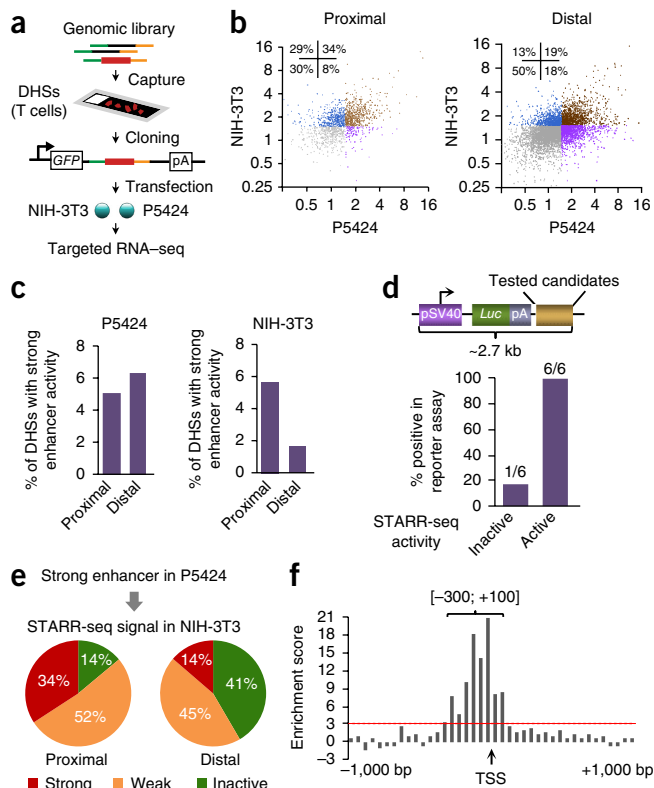


Figure 1 Comparison of proximal and distal DHSs with enhancer activity in two mouse cell lines. **(a)** Schematic of the CapStarr-seq protocol to assess the enhancer activity of promoters in NIH-3T3 and P5424 cells. **(b)** Scatterplots showing the STARR-seq signal (\log_2 scale) in P5424 and NIH-3T3 cells for proximal (left; 1,546 regions) and distal (right; 5,605 regions) DHSs. DHSs with enhancer activity in both cell lines (brown) or with activity specific to P5424 (purple) or NIH-3T3 (blue) cells are highlighted. DHSs with no enhancer activity are shown in gray. Quadrant panels show the percentage of regions in each subgroup. **(c)** Percentage of TSS-proximal and TSS-distal DHSs with strong enhancer activity (fold change >3) based on STARR-seq signal in P5424 (left) and NIH-3T3 (right) cells. **(d)** Top, reporter assay constructs. Bottom, summary of luciferase enhancer assays of proximal DHSs defined as active or inactive enhancers by STARR-seq in P5424 cells; detailed results are shown in **Supplementary Figure 1a**. Numbers correspond to the number of positive sites out of those tested. **(e)** Pie charts showing the distribution of enhancer activity in NIH-3T3 cells for the strong enhancers from TSS-proximal and TSS-distal DHSs identified in P5424 cells. **(f)** Distribution of the statistical enrichment of TSS-proximal DHSs for enhancer activity in NIH-3T3 cells. The significantly enriched region around the TSS is highlighted ($P < 0.001$, hypergeometric test).

the STARR-seq procedure was implemented to ensure that the transcripts quantified initiated from the synthetic SCP1 promoter and were polyadenylated^{9,13,18}. Reporter assays of CapStarr-seq-defined proximal enhancers confirmed their enhancer activity regardless of their orientation (**Fig. 1d** and **Supplementary Fig. 1a**). Distal enhancers identified in the P5424 T cell line were significantly enriched for lymphoid transcription factors, whereas proximal enhancers were generally depleted of these factors (**Supplementary Fig. 1b**), suggesting that the latter differ from classical distal enhancers. Consistently, the percentage of proximal T cell DHSs with enhancer activity in NIH-3T3 cells was higher than that for distal DHSs (**Fig. 1c**, right). Moreover, proximal enhancers in P5424 cells were found to be active more often in NIH-3T3 cells than distal enhancers (**Fig. 1e**)

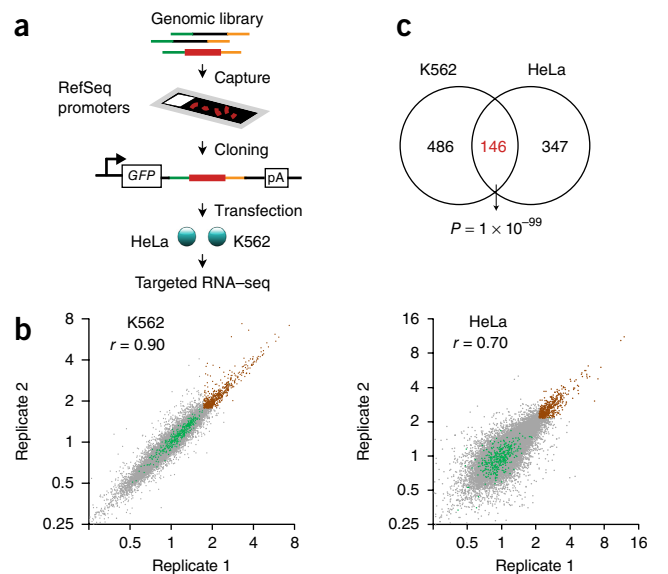


Figure 2 CapStarr-seq with human promoters. **(a)** Schematic of the CapStarr-seq strategy to assess the enhancer activity of promoters in K562 and HeLa cells. **(b)** Scatterplots showing the correlation of two STARR-seq replicates in K562 (left) and HeLa (right) cells. The data plotted are the fold change in STARR-seq signal over the input signal (\log_2 scale). Promoters with enhancer activity in both replicates are shown in brown. Random genomic regions (green) did not display enhancer activity in these assays. **(c)** Venn diagram showing the number of Epromoters found in K562 and HeLa cells. The hypergeometric test P value for the overlap between the two sets is shown.

and the proportion of proximal enhancers active in both cell lines was highly significant ($P = 1.8 \times 10^{-106}$, hypergeometric test; **Fig. 1b**), suggesting that proximal enhancers are less specific to tissue type.

Notably, proximal enhancers were over-represented from -300 bp to $+100$ bp with respect to the TSS (**Fig. 1f**), roughly overlapping the core promoter regions where sense and antisense transcription initiation occurs and transcription factors usually bind^{10,20,21}. Collectively, these results suggest that TSS-overlapping regions displaying enhancer activity, here defined as Epromoters, might represent regulatory elements with dual promoter and enhancer functions.

Assessment of the enhancer activity of coding-gene promoters

To characterize Epromoters in an unbiased manner, we performed CapStarr-seq with all promoters of RefSeq-defined human coding genes (-200 to $+50$ bp with respect to the TSS) in the two ENCODE cell lines K562 and HeLa (**Fig. 2a** and **Supplementary Fig. 2a,b**). The enhancer activity of each captured region was calculated as the fold change of the STARR-seq signal over the input signal. We observed high correlation between replicates in both cell lines (**Fig. 2b**). Epromoters were defined as promoters for which the fold change in signal for both replicates was beyond the inflexion point of ranked promoters (Online Methods). Using these stringent criteria, we found 632 (3%) and 493 (2.37%) Epromoters among 20,719 promoters analyzed in K562 and HeLa cells, respectively (**Fig. 2b,c** and **Supplementary Table 2**). Remarkably, a highly significant proportion of Epromoters were found in both cell types, suggesting a rather ubiquitous activity. No difference in the percentage of these promoters overlapping CpG islands or in the phylogenetic conservation of these promoters among mammalian species was observed as compared to non-Epromoters (**Supplementary Fig. 2c**).

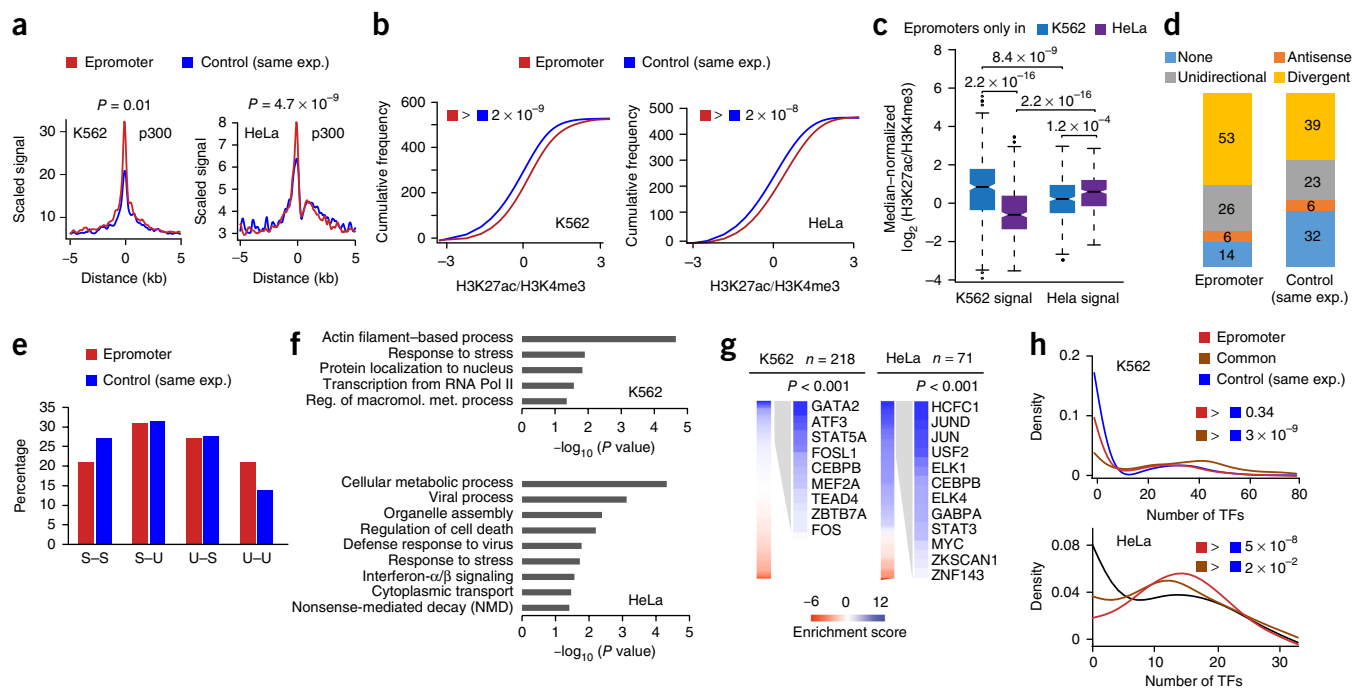


Figure 3 Genomic and epigenomic properties of Epromoters. **(a)** Average profiles of p300 in K562 (left) and HeLa (right) cells centered on the TSSs of Epromoters or control promoters with the same expression pattern for associated genes. Statistically significant differences were calculated for a region centered on the TSS (± 250 bp) using two-sided Mann–Whitney U tests. **(b)** Cumulative plots showing the H3K27ac/H3K4me3 ratio at Epromoter and control sets in K562 (left) and HeLa (right) cells (Kolmogorov test). **(c)** H3K27ac/H3K4me3 ratios at Epromoters as a function of cell type (Mann–Whitney U test). Central values represent the median of the signal, the interquartile range (IQR) corresponds to the 75th to 25th percentile, and whiskers extend to the maximum and minimum values excluding outliers. **(d)** Percentage of the Epromoter and control sets corresponding to the TSS clusters from 5' GRO–seq data defined by transcript overlap and orientation in HeLa cells. **(e)** Proportion of TSS pairs from each stability class (U, unstable; S, stable) associated with Epromoters and control sequences with the same expression) in K562 cells. **(f)** Significantly enriched biological processes for Epromoter-associated genes in K562 (top) and HeLa (bottom) cells identified using g:Profiler. **(g)** Transcription factor enrichment (ENCODE data sets) at Epromoters in K562 and HeLa cells (two-sided Mann–Whitney U test). **(h)** Density plots showing the number of transcription factors (TFs) per promoter type in K562 (left) and HeLa (right) cells. ‘Common’ refers to the set of Epromoters active in both cell lines (Kolmogorov test).

Epromoters display specific genomic and epigenomic features

We next compared the epigenomic features of Epromoters with those of a set of matched control promoters chosen from a list of common promoters lacking enhancer activity in all replicates of both cell lines (non-Epromoters) but associated with genes with similar expression levels (Supplementary Table 2). Although Epromoters displayed similar levels of DNase I hypersensitivity and histone H3 trimethylation at lysine 4 (H3K4me3) signal as the control promoters, they were generally enriched for the enhancer-associated features monomethylation of histone H3 at lysine 4 (H3K4me1), acetylation of histone H3 at lysine 27 (H3K27ac) and p300 binding (Fig. 3a and Supplementary Fig. 2d). Consistent with these findings, Epromoters displayed a higher H3K27ac/H3K4me3 ratio (Fig. 3b) and were preferentially associated with a strong enhancer state in different ENCODE cell lines (Supplementary Fig. 2e). Moreover, Epromoters had a higher H3K27ac/H3K4me3 ratio in the cell type where they were found to be active (Fig. 3c). There was no significant bias of RefSeq-defined TSSs at Epromoters, as assessed by cap analysis of gene expression (CAGE) (Supplementary Fig. 2f,g), and 94.2% and 95.7% of K562 and HeLa Epromoters, respectively, overlapped with a TSS defined by the FANTOM consortium²² (Supplementary Fig. 2h and Supplementary Table 2). However, 42.7% and 18.2% of the Epromoters active in HeLa and K562 cells lacked a TSS in the respective cell line. This might suggest that not all Epromoters are transcriptionally active (see below), although we cannot formally exclude the possibility that some individual cases could actually be promoter-proximal enhancers owing

to sites being incorrectly annotated as TSSs. While the majority of Epromoters were found in genes with only one TSS, a substantial proportion were located in genes with two or more TSSs (Supplementary Fig. 2i), reminiscent of previous findings suggesting that alternative promoters might work as enhancers¹⁶. By analyzing 5' global run-on with sequencing (5' GRO–seq) data from HeLa cells²⁰, we found that the proportion of Epromoters with divergent transcripts was higher than that for control promoters ($P = 3.1 \times 10^{-5}$, hypergeometric test; Fig. 3d). Moreover, unstable divergent transcripts, which have been shown to be a hallmark of active enhancers³, were over-represented among K562 Epromoters ($P = 5.8 \times 10^{-8}$, hypergeometric test; Fig. 3e). Altogether, the Epromoters defined by STARR-seq activity showed clear chromatin-associated enhancer features.

Gene Ontology (GO) analysis for Epromoter-associated genes primarily showed enrichment for basic processes (Fig. 3f and Supplementary Table 3), consistent with a previous STARR-seq study in *Drosophila melanogaster* reporting that many promoters of housekeeping genes can function as enhancers⁹. We also observed a significant enrichment ($P < 0.05$) for the cellular stress response in both cell lines. K562 Epromoters were particularly associated with genes encoding actin-binding cytoskeleton proteins, which have been shown to be rapidly and transiently upregulated upon heat shock response²³, whereas HeLa Epromoters were specifically associated with genes involved in type I and II interferon responses. Indeed, the main interferon-related genes were associated with Epromoters in HeLa cells, including *MX1*, *IRF9*, *JUND*, *ISG15*, *OAS* and the IFIT cluster of genes. Epromoter-associated

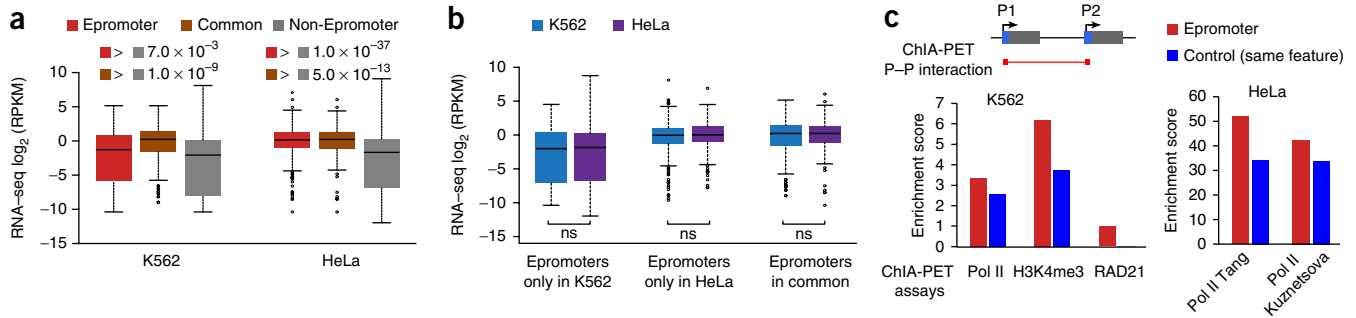


Figure 4 Expression of neighboring genes and promoter–promoter interactions. **(a,b)** Box plots comparing the expression levels of Epromoter- and non-Epromoter-associated genes in K562 and HeLa cells **(a)** and the expression of Epromoter-associated genes as a function of cell-line-specific Epromoter activity **(b)**. The expression of genes associated with Epromoters active in both cell lines (Common) is also shown. Central values represent the median of the signal, the IQR corresponds to the 75th to 25th percentile, and whiskers extend to the maximum and minimum values excluding outliers (two-sided Mann–Whitney *U* test). ns, not significant. **(c)** Top, schematic of the strategy to identify promoter–promoter (P–P) interactions. Bottom, ChIA-PET enrichment of promoter–promoter interactions for a list of promoters associated with at least one Epromoter or a non-Epromoter from a control set with the same enriched features. The sources^{37,38} of the published ChIA-PET data from HeLa cells are indicated by the name of the first author.

genes from HeLa cells were also enriched for transcriptional signatures including interferon- and tumor necrosis factor (TNF)-induced genes (**Supplementary Fig. 3c**). Differences in functional enrichment between K562 and HeLa cells might rely on cell-line-specific contexts. Indeed, interferon response genes are highly expressed in HeLa cells but not in K562 cells (**Supplementary Fig. 3a,b**), consistent with the fact that HeLa cells originated from a papillomavirus-infected tumor. We next assessed transcription factor enrichment at Epromoters using ENCODE data (**Fig. 3g**, **Supplementary Fig. 4a,b** and **Supplementary Table 4**). Consistent with the GO term enrichments, transcription factors involved in stress/interferon responses such as, JUN, FOS, IRF, ATF/CREB and STAT were enriched at Epromoters. We also found enrichment of specific transcription factor binding sites in general agreement with the transcription factor binding profiles, including strong enrichment for FOS/JUN motifs (**Supplementary Fig. 5a–d**). Moreover, Epromoters harbored a higher density of distinct bound transcription factors (**Fig. 3h**) and motifs (**Supplementary Fig. 5e**), consistent with their enhancer properties²⁴. Thus, Epromoters display genomic and epigenomic features associated with enhancer activity. While Epromoters are located close to housekeeping genes, a subset of them might be involved in stress response. In this context, some Epromoters could be required to ensure strong and rapid transcriptional output in response to environmental or intrinsic cellular stimuli.

We next asked whether enhancer and promoter (transcription of the associated gene) activities are correlated for Epromoters. We first observed that Epromoter-associated gene expression was significantly higher than that associated with non-Epromoters (**Fig. 4a**). However, enhancer activity at Epromoters did not strictly correlate with the expression levels of associated genes (**Supplementary Fig. 6a**), and differences in the enhancer activity of Epromoters between the K562 and HeLa cell lines did not correlate with significant differences in gene expression (**Fig. 4b**). This suggests that the promoter and enhancer functions of Epromoters might be partially independent, indicating potential long-range regulation of nearby genes. Chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) studies have shown that promoter–promoter interactions are a frequent phenomenon⁶. To test whether Epromoters are preferentially involved in promoter–promoter interactions, we analyzed available ChIA-PET data from the K562 and HeLa cell lines (**Supplementary Table 5**). In both cell lines, promoter–promoter interactions were found more frequently at Epromoters than at control promoters with similar levels of the corresponding histone modifications or transcription factors

(**Fig. 4c** and **Supplementary Fig. 6b**). HeLa Epromoters were enriched for HCFC1 and ZNF143 (**Fig. 3g**), two associated factors suggested to be involved in looping^{25,26}.

Epromoters function as bona fide enhancers

To experimentally address the role of Epromoters in the long-distance regulation of gene expression, we performed CRISPR–Cas9-mediated genomic deletion of the *FAF2* Epromoter, for which clear interactions with the promoters of the *NOP16*, *CLTB* and *RNF44* genes were observed by ChIA-PET in both cell lines (**Fig. 5a** and **Supplementary Fig. 7**). Deletion of the *FAF2* Epromoter (Δ Ep.*FAF2*) resulted in significant reduction of *RNF44* expression in both cell lines, while *NOP16* expression was reduced only in HeLa cells (**Fig. 5b**). A decrease in H3K27ac at the *RNF44* promoter in Δ Ep.*FAF2* K562 cells was also observed (**Fig. 5c**). We confirmed the interaction between the *FAF2* and *RNF44* promoters by circularized chromosome conformation capture and sequencing (4C–seq) in K562 cells, using either the *FAF2* or *RNF44* promoter region as the viewpoint, and observed almost complete loss of this interaction in the two Δ Ep.*FAF2* clones (**Fig. 5d** and **Supplementary Fig. 8a,b**). Consistent with these findings, the *FAF2* Epromoter was able to activate the *RNF44* promoter, as demonstrated by luciferase assay (**Fig. 5e**). Note that no luciferase activity was detected for the *RNF44* promoter vector without the *FAF2* Epromoter, ensuring that the observed enhancer activity is not due to spurious transcription¹⁹. Deletion of the endogenous *RNF44* promoter did not affect *FAF2* expression (**Fig. 5f**), indicating that distal regulation is directional. Moreover, epigenetic marks were correlated between the *FAF2* and *RNF44* loci across different cell lines (**Supplementary Fig. 8c**). To test *in vivo* whether Epromoters might function independently of their orientation, we inverted the *FAF2* Epromoter (including exon 1 of the gene) within its endogenous context in K562 cells (**Supplementary Fig. 7i–k**). Inversion of the *FAF2* Epromoter completely abolished *FAF2* expression and slightly but significantly reduced *RNF44* expression (**Fig. 5g**). However, *FAF2*–*RNF44* interaction was maintained in the inversion clones (**Supplementary Fig. 8b**) and *RNF44* expression was significantly higher than in the deletion clones (**Fig. 5h**), suggesting that *in vivo* enhancer activity is partially retained with the inverted configuration of the *FAF2* Epromoter. Finally, rescue of *FAF2* expression in either Δ Ep.*FAF2* or Inv.Ep.*FAF2* clones did not affect *RNF44* expression levels (**Fig. 5h**), indicating direct regulation of neighboring gene expression by the *FAF2* Epromoter.

To generalize our finding, we targeted three additional Epromoters with promoter–promoter interactions found either in both cell lines

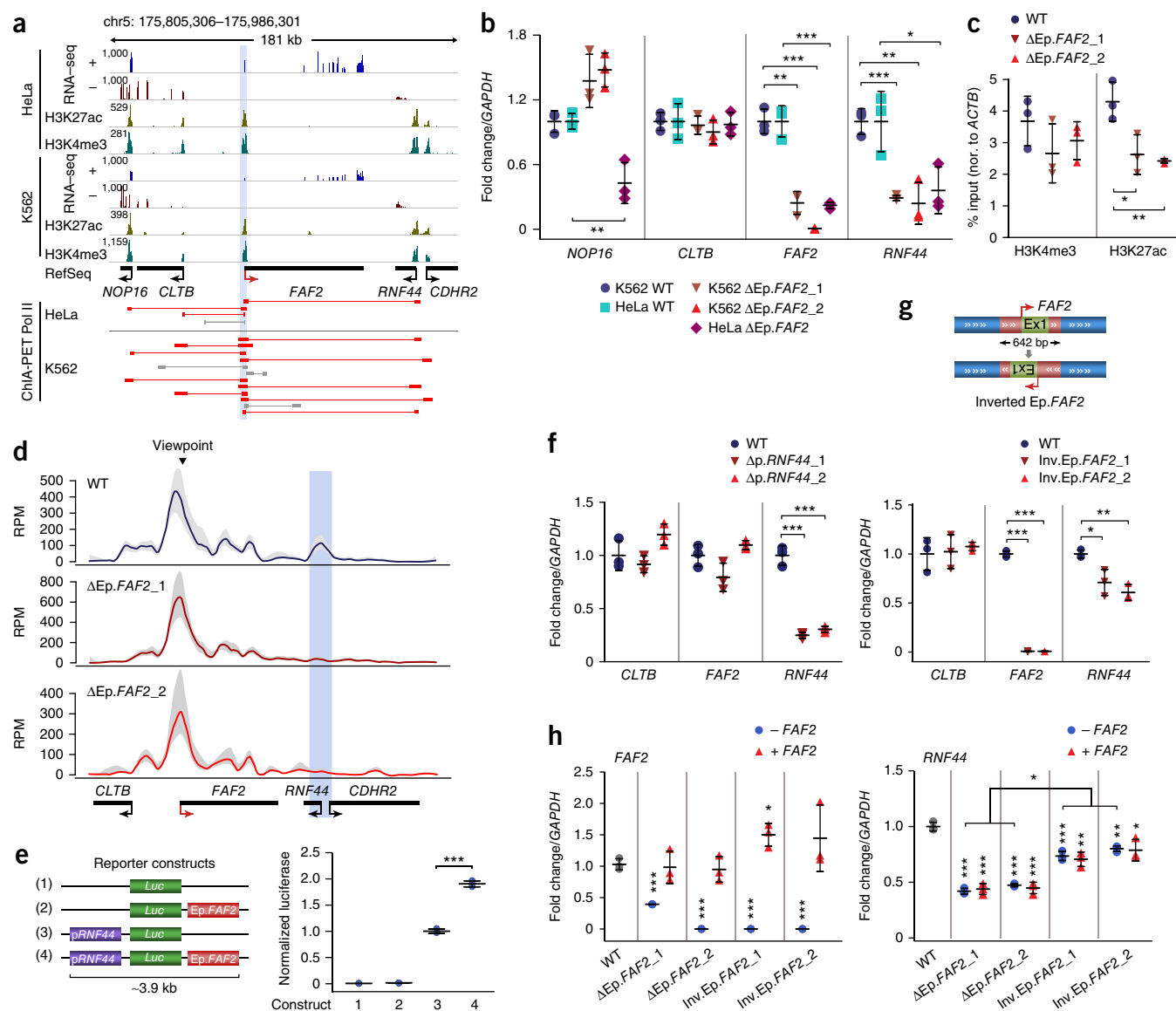


Figure 5 Epromoters function as bona fide enhancers and regulate distal gene expression. **(a)** Genomic tracks for RNA-seq, ChIP-seq and ChIA-PET Pol II around the *FAF2* locus. **(b)** qPCR analysis of gene expression in wild-type (WT) and Δ Ep.*FAF2* cell clones (the last numbers indicate the number of the independent clone). **(c)** ChIP-qPCR analysis of H3K4me3 and H3K27ac marks at the *RNF44* promoter in K562 cells. **(d)** 4C-seq analysis of *FAF2* Epromoter interactions in wild-type K562 cells and two Δ Ep.*FAF2* clones. The genomic tracks show the LOESS-normalized merge of two technical replicates (see **Supplementary Fig. 8a** for the raw data). RPM, reads per million; gray shading, 40% and 60% quantiles. The *FAF2*–*RNF44* interaction was significant in wild-type cells ($P < 1 \times 10^{-4}$) but not in Δ Ep.*FAF2* clones. **(e)** Luciferase reporter assays testing the enhancer activity of the *FAF2* Epromoter coupled with the *RNF44* promoter. **(f)** qPCR analysis of gene expression in wild-type and Δ p.*RNF44* K562 clones. **(g)** Top, schematic of knock-in of the inverted *FAF2* Epromoter. Bottom, qPCR analysis of wild-type and Inv.Ep.*FAF2* clones. Note that the intrinsic promoter activity is conserved as increased upstream antisense expression in the Inv.Ep.*FAF2* clones (**Supplementary Fig. 7k**). **(h)** qPCR analysis of the relative gene expression of *FAF2* (left) and *RNF44* (right) in wild-type, Δ Ep.*FAF2* and Inv.Ep.*FAF2* clones, in the presence or absence of *FAF2* cDNA. For the graphs in **b**, **c** and **e–h**, each point represents one of three independent RNA/cDNA preparations. Error bars, s.d.: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.1$, two-sided Student's *t* test.

(*CSDE1* and *TAGLN2*) or only in K562 cells (*BAZ2B*). Deletion of the *CSDE1* Epromoter resulted in significant reduction of *BCAS2* and *SIKE1* expression in both cell lines, while *NRAS* expression was reduced only in HeLa cells (**Supplementary Fig. 9a,b**). Deletion of the *TAGLN2* Epromoter led to significant reduction of *PIGM* and *PEA15* expression, while *DUSP23* was upregulated (**Supplementary Fig. 9c,d**). These results show specific regulation, as no effect was observed on *DCAF8*, another neighboring gene not interacting with the *TAGLN2* Epromoter. Although deletion of the *BAZ2B* Epromoter did not result in loss of *BAZ2B* expression, likely owing to alternative

promoter usage, *MARCH7* expression was significantly reduced (**Supplementary Fig. 9f–i**). Finally, the presence of CAGE-defined TSSs and spliced transcripts originating from the Epromoter regions (**Supplementary Fig. 9j**) confirmed that these loci are bona fide promoters and not incorrectly annotated distal enhancers.

Epromoters regulate distal interferon response genes

Expression of interacting gene pairs was highly correlated regardless of whether the association involved an Epromoter (**Supplementary Fig. 10a**). We therefore explored the possibility of a coordinated

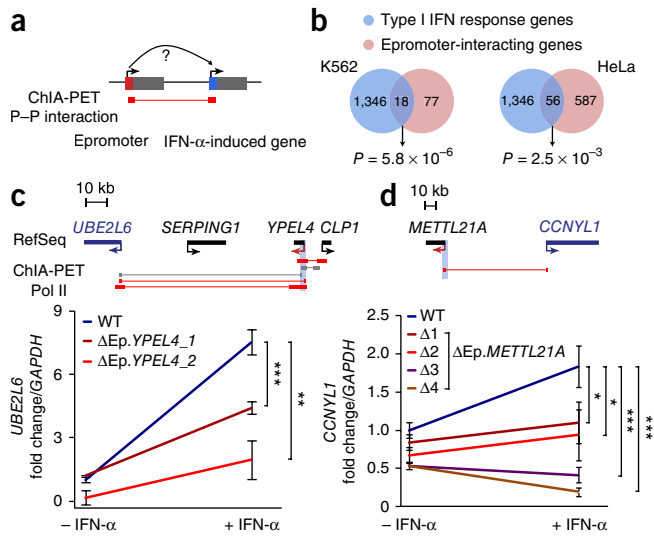


Figure 6 Epromoters are involved in a long-range response to IFN- α signaling. **(a)** Schematic of the strategy to identify IFN- α -induced genes associated with Epromoters combining ChIA-PET and STARR-seq data. **(b)** Venn diagrams showing the overlap between Epromoter-interacting genes and interferon response genes in K562 and HeLa cells (hypergeometric test). **(c,d)** qPCR analysis of the expression levels of the Epromoter-interacting genes *UBE2L6* **(c)** and *CCNYL1* **(d)** in wild-type and knockout clones with and without IFN- α stimulation. Error bars, s.d. ($n = 3$ independent RNA/cDNA preparations): **** $P < 0.001$, *** $P < 0.01$, * $P < 0.1$, two-sided Student's t test. The relative locations of genes and ChIA-PET interactions with Epromoters are shown above; Epromoters are highlighted as red arrows.

response to external stimuli mediated by Epromoters. We initially observed that key interferon response genes were found in interacting clusters associated with HeLa Epromoters (IFIT gene cluster, *ISG15–HES4* and *IRF9–PSME2–RNF31*; **Supplementary Fig. 10b**), suggesting that Epromoters are involved in the coordinated response to interferon signaling and consistent with an active interferon response in these cells (**Supplementary Fig. 3a–c**). To address whether Epromoters are involved in the activation of distal interferon-induced genes, we looked for interferon (IFN)- α -induced genes in promoter-promoter interactions with Epromoters (**Fig. 6a**). We found a significant proportion of Epromoters interacting with interferon response genes in both cell lines (**Fig. 6b** and **Supplementary Table 6**). We reasoned that in K562 cells some Epromoters might be required for proper activation of distally associated interferon response genes. To test this hypothesis, we selected two IFN- α response genes, *UBE2L6* (interacting with the *YPEL4* Epromoter) and *CCNYL1* (interacting with the *METTL21A* Epromoter) that were induced ~ 7.5 - and ~ 2 -fold after IFN- α treatment, respectively (**Fig. 6c,d**). Deletion of the interacting Epromoters did not result in consistent changes in *UBE2L6* or *CCNYL1* expression in non-stimulated cells; however, induction of these genes upon IFN- α treatment was severely reduced (**Fig. 6c,d** and **Supplementary Fig. 10c–e**). We also noted that *CLP1*, a non-interferon-responsive gene located close to *YPEL4*, displayed significant upregulation in clones in which the *YPEL4* Epromoter was deleted both before and after interferon treatment, suggesting that enhancer-promoter contacts may have been rewired in the Epromoter-knockout clones (**Supplementary Fig. 10d**). Overall, these results show that some Epromoters are involved in the rapid activation of distal genes upon external stress stimuli, supporting a model in which preformed loops between Epromoters and target genes precede gene induction²⁷.

To further rule out a plausible indirect effect mediated by Epromoter-associated genes, we analyzed allelic expression of wild-type cells and those homozygous and heterozygous for Epromoter deletion for cases where distally regulated genes harbored a SNP within the transcribed region in the K562 cell line. These genes included *PIGM* and *UBE2L6*, which are regulated by the *TAGLN2* and *YPEL4* Epromoters, respectively. In both cases, we found that allelic expression was significantly biased in the heterozygous clones (**Supplementary Figs. 9e** and **10f**), thus suggesting *cis*-specific regulation by the Epromoters.

Genetic variants within Epromoters influence distal genes

Genetic variants lying within Epromoters might influence the expression of distal genes. To address this possibility, we isolated all interacting promoter pairs (using ChIA-PET data) and those that were associated with expression quantitative trait loci (eQTLs) (**Fig. 7a,b** and **Supplementary Table 7**). We found that Epromoters more frequently overlapped eQTLs affecting the expression of distal interacting genes and that the eQTLs associated with Epromoters had a significantly stronger effect on distal gene expression than the eQTLs associated with non-Epromoters (**Fig. 7c**). We found eQTLs within three experimentally validated Epromoters (*METTL21A*, *BAZ2B* and *CSDE1*). K562 cells harbor a heterozygous eQTL variant within the *CSDE1* Epromoter (**Fig. 7d–f**) that results in DNase I accessibility and binding of transcription factors with a bias toward the reference allele (**Fig. 7g**). Allelic replacement of the reference allele resulted in decreased expression of *CSDE1* and *SIKE1* (**Fig. 7h**), as predicted by the eQTL association study. Similarly, deletion of the eQTL variant within *BAZ2B* resulted in reduced expression of the distal associated gene *MARCH7* (**Fig. 7i**). To further explore the implications of Epromoter-associated eQTLs, we analyzed *in silico* the probability of affecting transcription factor binding. We observed that SNPs potentially affecting transcription factor binding within Epromoters were biased toward having a positive effect (β) on distal gene expression, whereas this bias was not observed with non-Epromoters (**Fig. 7j,k**). Collectively, these results corroborate the functional relevance of eQTL-overlapping Epromoters, raising the intriguing possibility that disease-associated variants lying within Epromoters might directly influence distal gene expression.

DISCUSSION

Here, by implementing a high-throughput reporter assay, we shed light on and characterize a set of mammalian coding-gene promoters carrying both an intrinsic ability to drive local transcription (act as a promoter) and to activate distal gene expression (act as an enhancer). These elements have distinct genomic and epigenomic features, which distinguish them from other promoters and from classical distal enhancers (**Figs. 1, 3** and **4**). For six of these loci, we demonstrated that they act as bona fide enhancers regulating distal gene expression *in vivo*. Remarkably, some Epromoters were found to regulate the expression of several distal genes (*FAF2*, *CSDE1* and *TAGLN2* Epromoters) over large genomic distances (up to 300 kb in the case of the *TAGLN2* Epromoter), implying that they might function as regulatory hubs. Our results extend and support the increasing amount of evidence pointing to a unified model of transcriptional regulation, highlighting broad similarities between enhancers and promoters^{1–5}. Furthermore, previous studies based on the frequency of promoter-promoter interactions^{6,12,14} or epigenetic features^{7,10} suggested that some promoters might display enhancer function. Consistent with our findings, previous reporter assays also showed enhancer activity from TSS-proximal regions^{9,11,13}. It is also worth noting that several

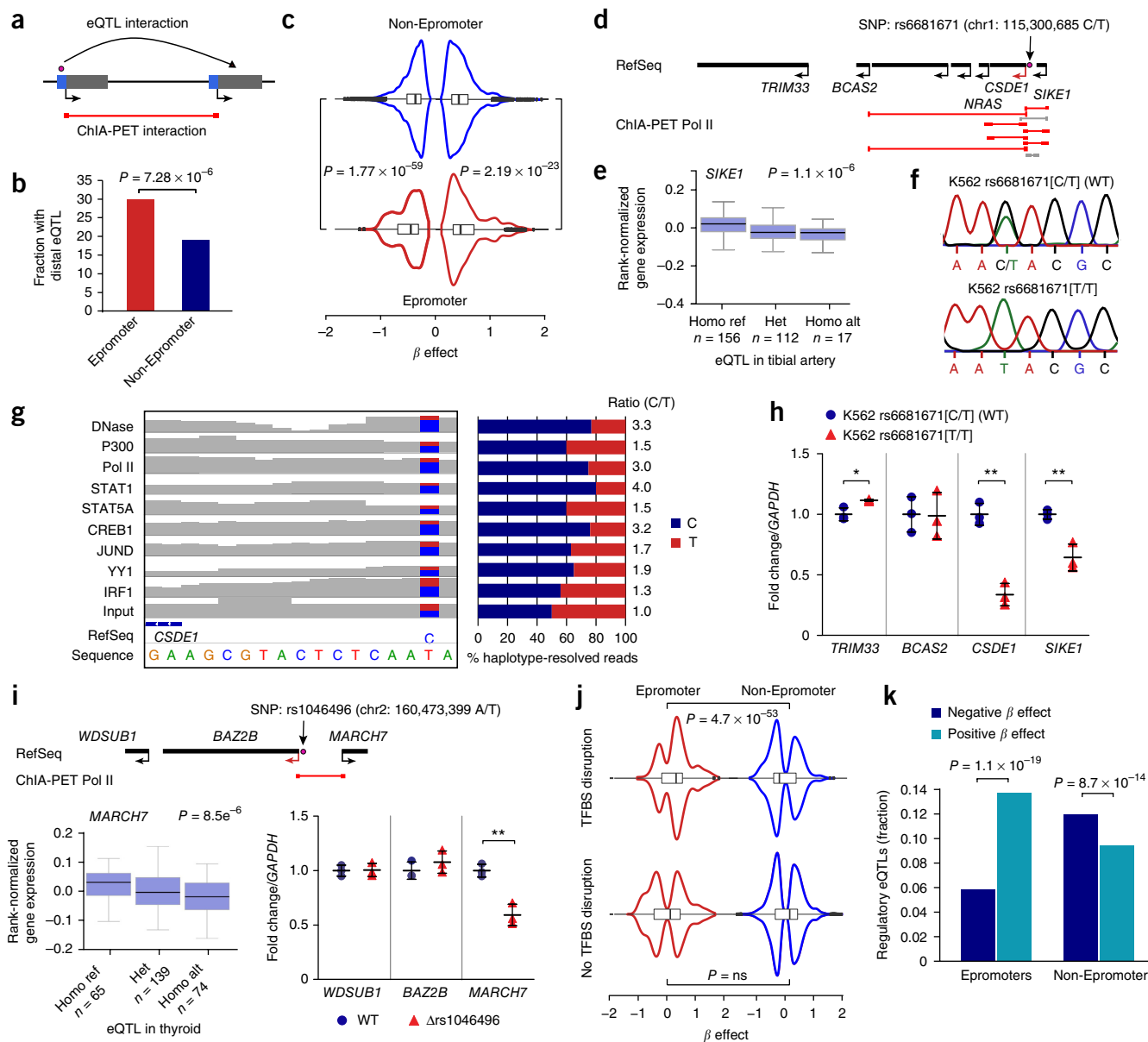


Figure 7 eQTL association within Epromoters. **(a)** Schematic of the eQTLs assessed. **(b)** Frequency of promoters having eQTLs associated with distal gene expression. Statistical significance was assessed by testing for equality of proportions. **(c)** Effects associated with eQTLs lying within promoter pairs with ChIA-PET interactions. Statistical significance was assessed independently for negative and positive scores using two-sided Mann-Whitney *U* tests. **(d)** Schematic of the eQTL SNP (rs6681671) within the *CSDE1* Epromoter associated with *SIKE1* expression. **(e)** eQTL data retrieved from the GTEx Portal. Ref, reference; Alt, alternate. Central values represent the median of the signal, the IQR corresponds to the 75th to 25th percentile, and whiskers extend to the maximum and minimum values excluding outliers. **(f)** Sequence chromatograms of wild-type and mutant K562 clones. **(g)** Coverage tracks from IGV (left) and histograms (right) showing the frequency of haplotype-resolved reads at SNP rs6681671 from the indicated ENCODE data in K562 cells. **(h)** qPCR analyses of gene expression in wild-type cells and eQTL mutants. **(i)** The eQTL SNP within the *BAZ2B* Epromoter associated with *MARCH7* expression is shown as in **d**, **e** and **h**. Δ rs1046496 denotes deletion of SNP rs1046496 in K562 cells. For **h** and **i**, error bars show s.d. ($n = 3$ independent RNA/cDNA preparations): *** $P < 0.001$, ** $P < 0.01$, * $P < 0.1$, two-sided Student's *t* test. **(j)** Effects depending on whether the eQTL disrupts a transcription factor binding site (TFBS). Statistical significance was assessed by a one-sided Mann-Whitney *U* test and corrected for multiple testing using the Benjamini-Hochberg method. **(k)** Fraction of regulatory eQTLs (affecting transcription factor binding) with positive and negative β values. Statistical significance was assessed by Fisher's exact test.

well-characterized enhancers of rapidly induced genes, including metalloproteins, histones of early cleavage stages, viral immediate-early genes (from SV40 and some cytomegaloviruses and retroviruses), heat-shock genes and the antiviral interferon genes, are located very close to the TSS⁸. Our study clearly shows that reporter-assay-based approaches can lead to the identification of TSS-overlapping promoters with bona fide enhancer activity *in vivo*.

It is possible that previous studies deleting large genomic regions overlapping promoters have underestimated the potential enhancer function of these regulatory elements (for example, see ref. 28). To our knowledge, only two studies have reported dual promoter and enhancer functions for the same regulatory elements in their endogenous context in mammals. Kowalczyk *et al.* showed that intragenic enhancers frequently act as alternative, tissue-specific promoters,

although these promoters produced a class of noncoding transcript¹⁶. Another study, published while this manuscript was under review, reported frequent distal *cis* regulation by loci associated with long noncoding RNAs (lncRNAs) and, to a lesser extent, coding genes¹⁵. Interestingly, using genetic manipulations in mouse embryonic stem cells, the authors demonstrated that these effects did not require the specific transcripts themselves, but instead involved general processes associated with their production, including enhancer-like activity of the gene promoters, the process of transcription and splicing of the transcript. On the basis of these findings, it is plausible that some of the experimentally validated Epromoters might function by other processes than enhancer-like activity. Further studies based on our catalog of Epromoters will be needed to precisely characterize the mechanisms by which these elements regulate distal gene expression.

Could it be that some of the Epromoters identified in this study are actually incorrectly annotated as promoter-proximal enhancers? The selection of captured TSS-encompassing regions was based on the annotation of coding-gene transcripts by RefSeq. Despite this conservative approach, we cannot completely rule out the possibility of erroneous TSS calls, leading to incorrectly annotated promoter-proximal enhancers. The vast majority of the tested regions overlapped with a CAGE-defined TSS. Moreover, the experimentally validated Epromoters (with the exception of *YPEL4*) did overlap with CAGE TSSs identified in the corresponding cell lines and were associated with spliced and polyadenylated transcripts of the nearest gene, confirming that these particular loci are bona fide promoters (Supplementary Fig. 9j). The analyses of CAGE-based TSSs also found that a substantial number of Epromoters did not display CAGE signal in the cell line where they were active (Supplementary Fig. 2h), in line with the poor correlation between Epromoter activity and expression of the closest gene (Supplementary Fig. 6a). However, we also found good correlation between gene pairs of interacting promoters involving at least one Epromoter (Supplementary Fig. 10a). This apparent contradiction might be explained by the existence of two types of Epromoters. One type might coordinately regulate the expression of several genes, including the closest one, therefore displaying simultaneous promoter and enhancer activities. For example, in the case of the *FAF2* Epromoter, expression of the *FAF2* and *RNF44* genes is positively correlated across different cell types (Supplementary Fig. 8c). Another type might display independent promoter and enhancer activities; in these cases, an active Epromoter could be associated with a silent or weakly expressed gene. For example, in the case of the *YPEL4* Epromoter, the *YPEL4* gene is not expressed in K562 cells, but the Epromoter regulates the expression of *UBE2L6* (Fig. 6c,d). This is reminiscent of a previous work showing that the same genomic region can have the epigenetic features of an enhancer or a promoter in different tissues⁷.

In the current model of transcription factories, the regulatory regions of neighboring genes are clustered together and each contributes to the expression of multiple genes by increasing the local concentration of regulatory factors and RNA polymerases²⁹. In this context, multigene interaction complexes have provided a structural framework for the postulated transcription factories⁶. Our results showing that Epromoters interact more frequently with other distal promoters (Fig. 4) and that eQTLs associated with Epromoters have a significantly stronger effect on distal gene expression (Fig. 7) support a model in which Epromoters have a key role within transcription factories. Whether Epromoter–promoter interactions rely on mechanisms similar to those previously shown for enhancer–promoter interactions³⁰ and what the specific contribution of Epromoters to the functioning of transcription factories is will need to be investigated in the future.

We found that a significant proportion of Epromoters interacted with interferon response genes in both cell lines analyzed (Fig. 6). Interferon response genes are not induced at baseline in K562 cells, suggesting the existence of preformed chromatin loops preceding gene induction of interferon response genes, in line with previous findings showing that TNF- α -responsive enhancers are already in contact with their target promoters before signaling²⁷. This is illustrated by the examples of the *YPEL4* and *METTL21A* Epromoters, which were found to interact with the promoters of distal IFN- α -responsive genes in unstimulated K562 cells, thus preceding gene activation. Further studies will be required to identify the transcription factors and (epigenetic) mechanisms involved in these interactions.

URLs. ENCODE, <https://www.encodeproject.org/>; R Core Team, <https://www.R-project.org/>; Reactome: interferon $\alpha\beta$ signaling, http://www.broadinstitute.org/gsea/msigdb/cards/REACTOME_INTERFERON_ALPHA_BETA_SIGNALING.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the [online version of the paper](#).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank J.-C. Andrau and J. Imbert for critical reading of the manuscript. We thank the IBiSA ‘Transcriptomics and Genomics Marseille-Luminy’ (TGML) platform for sequencing of CapStarr-seq samples and the cell biology platform for management of cell culture. Work in the laboratory of S.S. was supported by recurrent funding from INSERM and Aix-Marseille University and by specific grants from the European Union’s FP7 Programme (282510-BLUEPRINT), ARC (PJA 20151203149) and A*MIDEX (ANR-11-IDEX-0001-02). L.T.M.D., A.G. and G.C. were supported, respectively, by Vietnam International Education Development (911), CONACYT and FRM.

AUTHOR CONTRIBUTIONS

L.T.M.D. and S.S. conceptualized and designed the experiments. L.T.M.D. performed most experimental work. A.O.G.A. performed most bioinformatics analyses. J.A.C.-M. and J.v.H. performed motif analysis. C.A.-S., T.S., D.M. and E.S. performed 4C-seq experiments and analyses. C.S., A.G. and L.V. performed and analyzed data from mouse CapStarr-seq. J.A., M.T. and N.F. contributed to CRISPR screening and analyses of allelic expression. G.C. and D.P. performed ChIA-PET analyses. A.M.R. performed eQTL analysis. All authors contributed to reading, discussion and commenting on the manuscript. L.T.M.D. and S.S. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>. Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

- Kim, T.K. & Shiekhhattar, R. Architectural and Functional Commonalities between Enhancers and Promoters. *Cell* **162**, 948–959 (2015).
- Andersson, R. Promoter or enhancer, what’s the difference? Deconstruction of established distinctions and presentation of a unifying model. *BioEssays* **37**, 314–323 (2015).
- 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).
- Pekowska, A. *et al.* H3K4 tri-methylation provides an epigenetic signature of active enhancers. *EMBO J.* **30**, 4198–4210 (2011).
- Koch, F. *et al.* Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nat. Struct. Mol. Biol.* **18**, 956–963 (2011).
- Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84–98 (2012).
- Leung, D. *et al.* Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* **518**, 350–354 (2015).
- Schaffner, W. Enhancers, enhancers - from their discovery to today’s universe of transcription enhancers. *Biol. Chem.* **396**, 311–327 (2015).

9. Zabidi, M.A. *et al.* Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**, 556–559 (2015).
10. 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).
11. Nguyen, T.A. *et al.* High-throughput functional comparison of promoter and enhancer activities. *Genome Res.* **26**, 1023–1033 (2016).
12. Sanyal, A., Lajoie, B.R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113 (2012).
13. Arnold, C.D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
14. Schoenfelder, S. *et al.* The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* **25**, 582–597 (2015).
15. Engreitz, J.M. *et al.* Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539**, 452–455 (2016).
16. Kowalczyk, M.S. *et al.* Intragenic enhancers act as alternative promoters. *Mol. Cell* **45**, 447–458 (2012).
17. Dailey, L. High throughput technologies for the functional discovery of mammalian enhancers: new approaches for understanding transcriptional regulatory network dynamics. *Genomics* **106**, 151–158 (2015).
18. Vanhille, L. *et al.* High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. *Nat. Commun.* **6**, 6905 (2015).
19. Nejeplinska, J., Malik, R., Moravec, M. & Svoboda, P. Deep sequencing reveals complex spurious transcription from transiently transfected plasmids. *PLoS One* **7**, e43283 (2012).
20. Duttke, S.H. *et al.* Human promoters are intrinsically directional. *Mol. Cell* **57**, 674–684 (2015).
21. Roy, A.L. & Singer, D.S. Core promoters in transcription: old problem, new insights. *Trends Biochem. Sci.* **40**, 165–171 (2015).
22. Forrest, A.R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
23. Mahat, D.B., Salamanca, H.H., Duarte, F.M., Danko, C.G. & Lis, J.T. Mammalian Heat Shock Response and Mechanisms Underlying Its Genome-wide Transcriptional Regulation. *Mol. Cell* **62**, 63–78 (2016).
24. Hardison, R.C. & Taylor, J. Genomic approaches towards finding cis-regulatory modules in animals. *Nat. Rev. Genet.* **13**, 469–483 (2012).
25. Michaud, J. *et al.* HCFC1 is a common component of active human CpG-island promoters and coincides with ZNF143, THAP11, YY1, and GABP transcription factor occupancy. *Genome Res.* **23**, 907–916 (2013).
26. Whalen, S., Truty, R.M. & Pollard, K.S. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.* **48**, 488–496 (2016).
27. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290–294 (2013).
28. Li, Y. *et al.* CRISPR reveals a distal super-enhancer required for Sox2 expression in mouse embryonic stem cells. *PLoS One* **9**, e114485 (2014).
29. Feuerborn, A. & Cook, P.R. Why the activity of a gene depends on its neighbors. *TIG* **31**, 483–490 (2015).
30. Kagey, M.H. *et al.* Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430–435 (2010).

ONLINE METHODS

Cell culture. K562 (CCL-243), a chronic myelogenous leukemia cell line, and HeLa-S3 (CCL-2.2), a cervical carcinoma cell line, were obtained from the American Type Culture Collection (ATCC) and maintained in RPMI (Gibco) supplemented with 10% FBS (Gold, PAA) at 37 °C, 5% CO₂. The P5424 T cell line³¹ was cultured as described previously¹⁸. Cells were passaged every 2–3 d and routinely tested for mycoplasma contamination. For cell stimulation, 1 × 10⁶ K562 cells were incubated with IFN- α (Sigma, SRP4594) at 50 ng/ml for 6 h.

Mouse CapStarr-seq. Enhancer activity in the mouse P5424 and NIH-3T3 cell lines was retrieved from our previously published CapStarr-seq data¹⁸. DHS genomic regions were separated into TSS distal (>1 kb) and proximal (<1 kb) while keeping the previous definition of enhancer activity (Supplementary Table 1).

Luciferase reporter assays. For the reporter assays related to Figure 1c and Supplementary Figure 1a, proximal-defined DHS regions overlapping TSSs were selected on the basis of CapStarr-seq activity in the P5424 cell line. The tested candidates were amplified from mouse genomic DNA and cloned downstream of the luciferase gene in the pGL3-Promoter vector (Promega) at the BamHI site. For the reporter assays related to Figure 5e, the human *RNF44* promoter (1,294 bp, chr5:176,537,245–176,538,538) and/or *FAF2* Epromoter (661 bp, chr5:176,447,822–176,448,482) was amplified from K562 genomic DNA and cloned into the pGL3-Basic vector (Promega). The *RNF44* promoter was cloned upstream of the luciferase gene at the MluI–BglII sites, and the *FAF2* Epromoter was cloned downstream of the luciferase gene at the Sall site. For cell transfection, a total of 1 × 10⁶ P5424 or K562 cells were cotransfected with 1 μ g of the tested construct and 200 ng of *Renilla* vector using the Neon Transfection System (Thermo Fisher Scientific). Electroporation conditions for P5424 cells were described previously¹⁸, and conditions for K562 cells are described below (human CapStarr-seq). Twenty-four hours after transfection, luciferase activity was measured using the Dual-Luciferase Reporter Assay kit (Promega) on a TriStar LB-941 Reader. For all measurements, firefly luciferase values were first normalized to *Renilla* luciferase values (controlling for transfection efficiency and cell number). Data are represented as the fold increase in relative luciferase signal over the pGL3-Promoter vector (Supplementary Fig. 1a) or *RNF44*-pGL3-Basic vector (Fig. 5e) with s.d. Student's *t* tests (one-sided, unpaired) from three independent transfections were used to calculate significance.

Human CapStarr-seq. Construction of the human promoter library is detailed in the Supplementary Note. The principle of CapStarr-seq was described previously¹⁸. The detailed step-by-step protocol is accessible on Protocol Exchange³². The human promoter library was transfected into K562 and HeLa cells using the Neon Transfection System (Thermo Fisher Scientific; pulse voltage 1,450 V and 1,005 V, pulse width 10 and 35 ms, pulse number 3 and 2 for K562 and HeLa cells, respectively). For each replicate, 30 × 10⁶ cells were transfected with 150 μ g of library; two independent transfection replicates were performed for each cell line. The transfected and non-transfected (plasmid input) libraries were single-end sequenced on the Illumina NextSeq 500 platform, and reads were mapped to the hg19 reference genome using standard procedures. Supplementary Table 8 summarizes the number of sequenced and mapped reads for each sample. The coverage of each genomic region was calculated using BEDTools (v2.17.0), and the ratio of the CapStarr-seq coverage over the input (fold change) was computed for each sample. Promoter regions with enhancer activity were defined by determining the inflexion point of the ranked fold change (Supplementary Table 2a). Epromoters were defined as promoters displaying enhancer activity in both replicates. A common set of non-Epromoters was also defined as promoters lacking enhancer activity in all replicates of both cell lines. STARR-seq-positive controls displayed enhancer activity in our assays (Supplementary Fig. 2a).

Flow cytometry. We primarily observed enhanced GFP expression from the pooled promoter library as compared to the empty vector by FACS analysis (Supplementary Fig. 2b). A total of 5 × 10⁶ K562 or HeLa cells were transfected with 25 μ g of the empty STARR-seq screening vector¹³ or the promoter library using the Neon Transfection System (Thermo Fisher Scientific) with the conditions described above. Twenty-four hours after electroporation, GFP

expression was assessed on a FACSCalibur (BD Biosciences). Data were analyzed and visualized with FlowJo software.

RNA transcription and selection of the control set. Transcript quantification by RPKM (K562 and HeLa cell lines, four samples each) was obtained from the ENCODE Consortium (Supplementary Table 9). The data were normalized using the Normalizer package³³ with the quartiles $-\log_2$ option, and the mean of the four samples was obtained. A control (with the same expression) for each cell line was obtained by comparing Epromoters to promoters without enhancer activity (using transcription values for the nearest gene), and a list was generated of the same number of observations using a tool developed in house. The expression levels of genes associated with Epromoters and control sets in each cell line were compared to each other or to CapStarr-seq fold changes in signal and graphed using R software (R Core Team).

Epigenomic analysis. ChIP-seq data for the H3K4me3, H3K4me1 and H3K27ac histone marks, as well as DNase-seq data, were obtained from the ENCODE Consortium (Supplementary Table 9). Median average profiles were generated by extracting ChIP-seq signal from wiggle files for the 5-kb regions centered on TSSs. To test whether the differences between different classes of promoters were statistically significant, we first extracted the average signal for the top 25% of the signal in 2-kb regions centered on TSSs. A two-sided Mann–Whitney *U* test was then performed for each pair of promoter sets.

TSS analyses. To define promoter classes, clusters of 5' GRO-seq transcripts from HeLa cells were obtained from Duttler *et al.*²⁰. The clusters overlapping a 500-bp region extended from the promoter coordinates were retrieved. Bidirectional coding genes (TSS closer than 1.5 kb and in the opposite direction) were omitted. Each promoter was defined as a function of the orientation of the overlapping clusters of 5' GRO-seq transcripts: unidirectional, only one transcript in the same direction as the gene; divergent, two RNA fragments in opposite directions; antisense, only one transcript in the opposite direction as the gene. Definition of TSS pairs as a function of RNA stability (UU, unstable–unstable; US, unstable–stable; SS, stable–stable) in K562 cells was obtained from Core *et al.*³. The TSS pairs overlapping a 500-bp region extended from the promoter coordinates were retrieved. Further analyses of TSSs and comparison with CAGE data are provided in the Supplementary Note.

Functional enrichment. GO enrichment in biological processes and pathways was assessed using g:Profiler³⁴ and default options (Supplementary Table 3). For the statistical background, we used the list of all genes associated with the capture promoters. Enrichment scores were calculated using the g:GOST native method. Enrichment analysis for transcriptomic signatures was performed using GREAT³⁵ with all capture promoters as the background. Only gene signatures involved in TNF and interferon responses are shown in Supplementary Figure 3c.

To analyze the expression of type I interferon response genes, transcript quantification data (FPKM) for 23 cell lines (including HeLa and K562 cells) were obtained from the ENCODE Consortium (Supplementary Table 9) and normalized as described above. The FPKM values of genes involved in the 'Reactome: interferon $\alpha\beta$ signaling' pathway were graphed using R software in a cumulative plot (Supplementary Fig. 3a). A Kolmogorov test was then performed to compare the HeLa and K562 cell lines. Genes in the 'Reactome: interferon $\alpha\beta$ signaling' pathway that were differentially expressed in HeLa cells relative to the remaining 22 cell lines were identified by performing Significance Analysis of Microarrays (SAM) with TMEV (4.9)³⁶ software using a delta value of 0.5.

Transcription factor enrichment and density. ChIP-seq data (wiggle and peak files) from 71 (56 unique) and 218 (116 unique) transcription factors for the HeLa and K562 cell lines, respectively, were obtained from the ENCODE Consortium (Supplementary Table 9). To test whether the differences between Epromoters and control promoters (with the same expression) were statistically significant, we quantified the ChIP-seq signal from –200 to +50 bp with respect to the TSS. A Mann–Whitney *U* test was then performed for each pair of promoter sets. An enrichment score was calculated using the following

formula: $-\log_{10}(P \text{ value})$ if fold change >1 or $\log_{10}(P \text{ value})$ if fold change <1 . A heat map of the scores was generated using Multiple Experiment Viewer³⁶. We considered transcription factors to be enriched if they had a fold change >1.2 and $P < 0.001$. The average profiles for significantly enriched transcription factors were generated by extracting ChIP-seq signal from wiggle files for the 5-kb regions centered on TSSs. To assess the number of transcription factors bound per promoter (transcription factor density), the overlap of transcription factor peaks with Epromoters and control promoters (same expression) was assessed using BEDTools. The presence (1) or absence (0) of overlapping transcription factors for each promoter was summed and the density of transcription factors for each promoter was graphed using R software. A Kolmogorov test was then performed for each pair of promoter sets.

Motif analysis in Epromoters. Epromoter sequences from K562 and HeLa cells were scanned with a non-redundant collection of TFBSs (Supplementary Note) to detect over-represented and positionally biased motifs relative to control sequences (non-Epromoters). We detected motifs over-represented in Epromoters relative to non-Epromoters with the program matrix-enrichment (default parameters), which computes the cumulative distributions of scores for a given motif and computes the significance of over-representation at each possible score threshold with the binomial law. In addition to assessing global over-representation, we ran position-scan, which runs a chi-squared homogeneity test to detect motifs whose positional distribution differs between two sequence sets. We tuned the position-scan parameters to detect motifs showing a specific peak of enrichment near the core promoter (from -250 to $+50$ with respect to the TSS) of Epromoters relative to non-Epromoters. For graphical representation, the positional distributions of predicted sites were drawn on an extended region (± 1 kb relative to the TSS), whereas the chi-squared test was restricted to the core promoter using a bin width of 50 bp and scanning with a threshold of $P \leq 1 \times 10^{-3}$. The background model was a first-order Markov chain trained with dinucleotide frequencies from all human core promoters.

Computations of ChIA-PET enrichment scores for promoter-promoter interactions. Pol II ChIA-PET interactions from HeLa and K562 cells were obtained from published data^{37,38} and ENCODE Consortium data (Supplementary Table 9), respectively. ChIA-PET fragments for which the two mates intersected a 1-kb region encompassing two distinct TSSs were selected to define promoter-promoter interactions (Supplementary Table 5). Control sets were subsets of promoters without enhancer activity in both cell lines, as defined above. For each mark, each Epromoter was associated with a control promoter with the closest ChIP-seq signal computed from ENCODE Consortium data (Supplementary Table 9) to create a control list matched to the Epromoter list for signal distribution. To obtain enrichment scores, the fraction of promoters with promoter-promoter interactions was computed. Next, the number of interacting promoters labeled as Epromoter or control promoter was retrieved. ChIA-PET interactions mediated by H3K27ac, H3K4me2 and H3K4me1 were not significant for any set and are not displayed in Figure 4c. The corresponding enrichment scores were computed from hypergeometric tests using the following formula: $-\log_{10}(P \text{ value})$.

Gene expression correlation for interacting gene pairs. RNA-seq quantification data (FPKM) for 23 cell lines were retrieved from the ENCODE Consortium (Supplementary Table 9) and normalized as described above. Pearson's correlation between coding-gene pairs on the same chromosome and having a ChIA-PET interaction in K562 or HeLa cells (Supplementary Table 5) was assessed using R software (R Core Team). Correlation scores for gene pairs involving at least one Epromoter or only non-Epromoters were graphed using R software. A control set containing shuffled gene pairs from the ChIA-PET interacting pairs was also plotted.

CRISPR-Cas9 genome editing. Targeted Epromoter and promoter regions were defined by CapStarr-seq and DNase-seq peaks ranging from 410 bp to 1,255 bp in length (Supplementary Fig. 7b–h, left). For the knockout experiments, the general strategy is shown in Supplementary Figure 7a. Two gRNAs were designed for each end of the targeted region using the CRISPRdirect tool³⁹. The gRNAs were cloned into a gRNA cloning vector (Addgene, 41824) as previously described⁴⁰. Two million cells were transfected with 15 μg of

the hCas9 vector (Addgene, 41815) and 7 μg of each gRNA using the Neon Transfection System (Thermo Fisher Scientific). Three days after transfection, the bulk of transfected cells were plated in 96-well plates at limiting dilution (0.5 cells per 100 μl per well) for clonal expansion. After 10–14 d, individual cell clones were screened for homologous allele deletion by direct PCR using Phire Tissue Direct PCR Master Mix (Thermo Fisher Scientific) according to the manufacturer's protocol. Forward and reverse primers were designed bracketing the targeted regions, allowing for the detection of knockout or wild-type alleles. Clones were considered to have undergone homologous allele deletion if they had at least one deletion band of the expected size and no wild-type band (Supplementary Fig. 7b–h, right). If more than two cell clones were obtained for a given locus, the most precise deletion was chosen. All gRNAs and primers are listed in Supplementary Table 10. The generation of clones in which the *FAF2* Epromoter was inverted and eQTL SNPs were mutated is described in the Supplementary Note.

Gene expression. Total RNA was extracted using TRIzol reagent (Thermo Fisher Scientific). 3 μg of RNA was then treated with DNase I (Ambion) and reverse transcribed into cDNA using Superscript VILO Master Mix (Thermo Fisher Scientific). Real-time PCR was performed using Power SYBR Master Mix (Thermo Fisher Scientific) on a Stratagene Mx3000P instrument. Primer sequences are listed in Supplementary Table 10. Gene expression was normalized to that of *GAPDH*. Relative expression was calculated by the $\Delta\Delta C_T$ method, and all data shown are reported as the fold change over the control. For each cell clone, the Student's *t* test was performed (unpaired, two-tailed, 95% confidence interval) from three independent RNA/cDNA preparations. Data are represented with s.d. For conventional RT-PCR, one-twentieth of the synthesized cDNA was used as the template for one reaction; PCRs were performed with Phusion polymerase (Thermo Fisher Scientific), $T_m = 60$ °C, 30 cycles.

FAF2 rescue experiments. Human *FAF2* cDNA was purchased from Origene (SC100662). K562 cell clones in which the *FAF2* Epromoter was knocked out or inverted were transfected with 2 μg of *FAF2* cDNA plasmid, and samples were collected 24 h after transfection for gene expression analysis as described above.

Allelic expression. Genetic variants within the transcribed regions of the *PIGM* (chr1:160,000,435) and *UBE2L6* (chr11:57,319,339) genes were identified by visual assessment of RNA-seq data from the K562 cell line using the IGV tool (version 2.3.67)⁴¹. PCR primers containing Illumina adaptors were designed flanking each variant (Supplementary Table 10). cDNAs from wild-type K562 clones and clones with homozygous and heterozygous deletion of the *TAGLN2* and *YPEL4* Epromoters were amplified using *PIGM*- and *UBE2L6*-specific primers, respectively. In the case of *UBE2L6*, the cDNA was generated from IFN- α -treated cells. A second PCR was performed using NEBNext Multiplex Oligos for Illumina (New England BioLabs), the product was subjected to single-end sequencing on the Illumina NextSeq 500 platform and reads were mapped to the hg19 reference genome using standard procedures. Allelic frequency was computed using the IGV tool.

Haplotype-resolved analysis of DNase-seq and ChIP-seq data. Transcription factors for which a ChIP-seq peak in K562 cells (ENCODE Consortium) overlapped the eQTL SNP rs6681671 in the *CSDE1* Epromoter were selected. BAM files from corresponding ChIP-seq data, along with DNase-seq data and input, were directly retrieved with the IGV tool, and the frequency of the haplotype-resolved reads was manually computed. Only samples with at least ten reads were selected.

Chromatin immunoprecipitation and qPCR. Generation of ChIP samples is described in the Supplementary Note. ChIP eluates and input were assayed by real-time PCR (Stratagene Mx3000P instrument) in a 20- μl reaction with one-thirtieth of the elution material using Power SYBR Master Mix (Thermo Fisher Scientific). The primers used in the real-time PCR assays are listed in Supplementary Table 10. Data represent the percentage of input normalized to *ACTB* with s.d. Student's *t* test (two-tailed, unpaired) was used to test for significance from three independent chromatin preparations.

4C analysis. 4C-seq experiments were carried out as described^{142–44}. 4C libraries were prepared using NlaIII–DpnII enzyme combinations for the *FAF2* and *RNF44* promoters. Primer sequences are listed in **Supplementary Table 10**. For the *FAF2* viewpoint, two technical replicates each of one wild-type K562 clone and two Δ Ep.*FAF2* clones were analyzed. For the *RNF44* viewpoint, one wild-type K562 clone, two Δ Ep.*FAF2* clones and one Inv.Ep.*FAF2* clone were analyzed. Samples were sequenced and used for downstream analysis as independent replicates and as a merged data set. 4C-seq data processing was performed as described⁴⁵ using the NCBI human assembly GRCh37 (hg19), and detailed analysis and visualization were carried out using r3Cseq and FourCseq software^{46,47}. For a visible data profile, normalized RPM data were smoothed via a running-mean approach and quantiles (40%, 50% and 60%) were further smoothed and interpolated with the R loess function using Basic4Cseq⁴⁸.

Distal association with interferon response. Human type I interferon response genes were retrieved from Interferome database v2.01 (ref. 49). We then selected the interferon response genes distally interacting with an Epromoter on the basis of ChIA-PET data (**Supplementary Table 5**). The list of Epromoters distally interacting with interferon response genes is provided in **Supplementary Table 6**.

eQTL analysis. eQTL data were obtained from GTEx project portal version 6 and lifted over to hg19 coordinates to match capture promoter data. Using GenomicRanges⁵⁰, capture promoter coordinates were extended 1.5 kb to each side to capture overlapping eQTLs that could be mechanistically related to these promoters. ChIA-PET promoter–promoter pairs were obtained as described below. Promoter–promoter pairs were annotated using capture promoters and eQTL overlaps to determine long- and close-range interaction effects between pairs. We were able to annotate 4,310 of 7,825 pairs (**Supplementary Table 7**). Customized R scripts were used to analyze the relationship between eQTL β value (effect size) and long- and close-range gene promoter interactions in the annotated promoter–promoter pairs and to determine whether eQTLs were located within the extended region of an Epromoter or a non-Epromoter. Taking only eQTLs affecting the distal gene in the pair, the β -value bimodal distributions of these eQTLs were split into negative and positive values by fitting a two-component mixture model (R mixtool package⁵¹) and looking for the cutoff where the probability of a negative value being generated by the left distribution was ≥ 0.5 . To test whether Epromoter-associated β values were stronger than the ones associated with non-Epromoters, we independently compared negative and positive β -value sets using a one-tailed non-parametric Wilcoxon rank-sum test (wilcox.test R function) and corrected for multiple testing using the Benjamini–Hochberg method (p.adjust R function). The statistical analyses to predict the impact of eQTL SNPs on transcription factor binding sites is detailed in the **Supplementary Note**.

Statistics. All experiments were performed using at least three independent samples or transfections. R/Bioconductor or GraphPad Prism 6.0 was used for statistical analysis. For comparisons in Venn diagram representations, a hypergeometric test was performed. Unless otherwise indicated in the figure legends, for comparisons between two groups of equal sample size and small n (like in qPCR dot plots), an unpaired two-tailed Student's t test was performed; for comparisons between two groups of equal sample size and large n (as in box-plot representations), a two-tailed Mann–Whitney U test was performed. For comparisons of two distributions, a Kolmogorov–Smirnov test was

performed. $P < 0.05$ was considered to be statistically significant, and error bars represent s.d. Investigators were not blinded to sample identity.

Data availability. All custom scripts have been made available at <https://github.com/arielgalindoalbarra/Epromoters>. Human CapStarr-seq and 4C data generated during the current study are available in the Gene Expression Omnibus (GEO) under accessions **GSE83296 (Supplementary Table 8)** and **GSE98194**, respectively. Mouse CapStarr-seq data analyzed during the current study were published previously¹⁸ and are available in GEO under accession **GSE60029**. All public data sets and primers used are described in **Supplementary Tables 9 and 10**, respectively.

31. Mombaerts, P., Terhorst, C., Jacks, T., Tonegawa, S. & Sancho, J. Characterization of immature thymocyte lines derived from T-cell receptor or recombination activating gene 1 and p53 double mutant mice. *Proc. Natl. Acad. Sci. USA* **92**, 7420–7424 (1995).
32. Dao, L.T.M., Vanhille, L., Griffon, A., Fernandez, N. & Spicuglia, S. CapStarr-seq protocol. *Protocol Exchange* <http://dx.doi.org/10.1038/protex.2015.096> (2015).
33. Glusman, G., Caballero, J., Robinson, M., Kutlu, B. & Hood, L. Optimal scaling of digital transcriptomes. *PLoS One* **8**, e77885 (2013).
34. Reimand, J. *et al.* g:Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* **44**, W1 W83–9 (2016).
35. McLean, C.Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
36. Saeed, A.I. *et al.* TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34**, 374–378 (2003).
37. Kuznetsova, T. *et al.* Glucocorticoid receptor and nuclear factor kappa-b affect three-dimensional chromatin organization. *Genome Biol.* **16**, 264 (2015).
38. Tang, Z. *et al.* CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* **163**, 1611–1627 (2015).
39. Naito, Y., Hino, K., Bono, H. & Ui-Tei, K. CRISPRdirect: software for designing CRISPR/Cas guide RNA with reduced off-target sites. *Bioinformatics* **31**, 1120–1123 (2015).
40. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
41. Thorvaldsdóttir, H., Robinson, J.T. & Mesirov, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
42. Stadhouders, R. *et al.* Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. *Nat. Protoc.* **8**, 509–524 (2013).
43. Stadhouders, R. *et al.* HBS1L-MYB intergenic variants modulate fetal hemoglobin via long-range MYB enhancers. *J. Clin. Invest.* **124**, 1699–1710 (2014).
44. Vieux-Rochas, M., Fabre, P.J., Leleu, M., Duboule, D. & Noordermeer, D. Clustering of mammalian Hox genes with other H3K27me3 targets within an active nuclear domain. *Proc. Natl. Acad. Sci. USA* **112**, 4672–4677 (2015).
45. Stadhouders, R. *et al.* Control of developmentally primed erythroid genes by combinatorial co-repressor actions. *Nat. Commun.* **6**, 8893 (2015).
46. Thongjuea, S., Stadhouders, R., Grosveld, F.G., Soler, E. & Lenhard, B. r3Cseq: an R/Bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data. *Nucleic Acids Res.* **41**, e132 (2013).
47. Klein, F.A. *et al.* FourCSeq: analysis of 4C sequencing data. *Bioinformatics* **31**, 3085–3091 (2015).
48. Walter, C., Schuetzmann, D., Rosenbauer, F. & Dugas, M. Basic4Cseq: an R/Bioconductor package for analyzing 4C-seq data. *Bioinformatics* **30**, 3268–3269 (2014).
49. Rusinova, I. *et al.* Interferome v2.0: an updated database of annotated interferon-regulated genes. *Nucleic Acids Res.* **41**, D1040–D1046 (2013).
50. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLOS Comput. Biol.* **9**, e1003118 (2013).
51. Benaglia, T., Chauveau, D., Hunter, D.R. & Young, D.S. mixtools: An R Package for Analyzing Finite Mixture Models. *J. Stat. Softw.* **32**, 1–29 (2009).