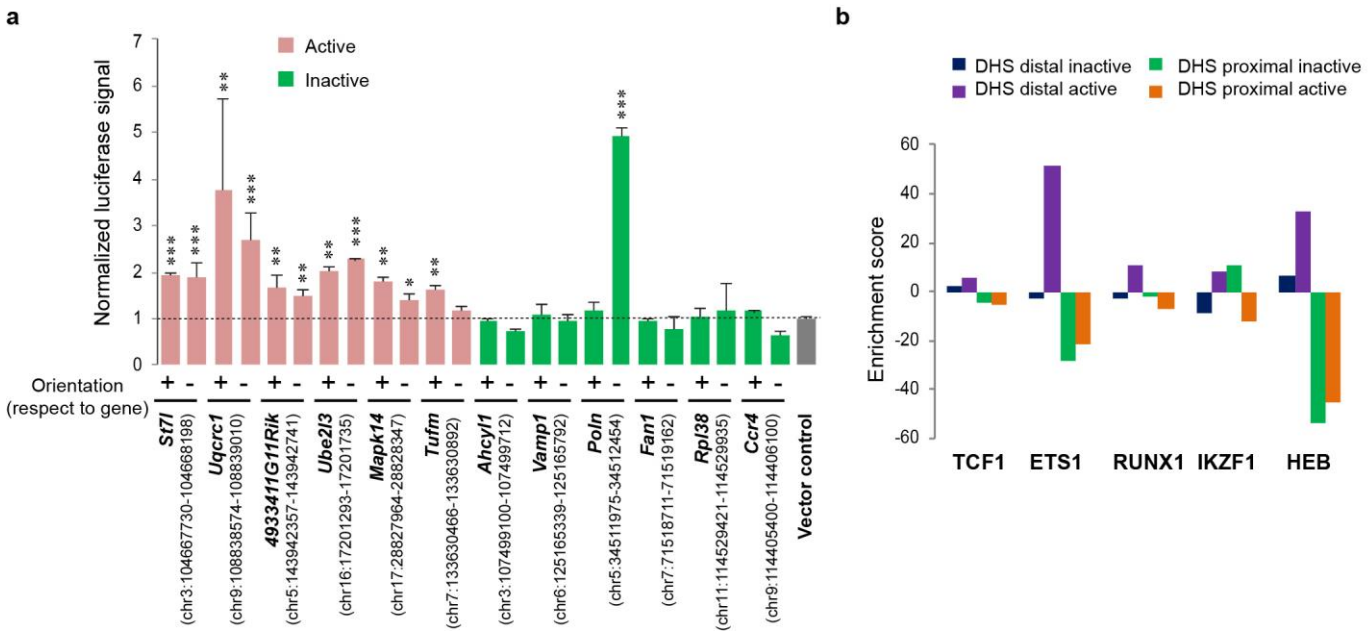
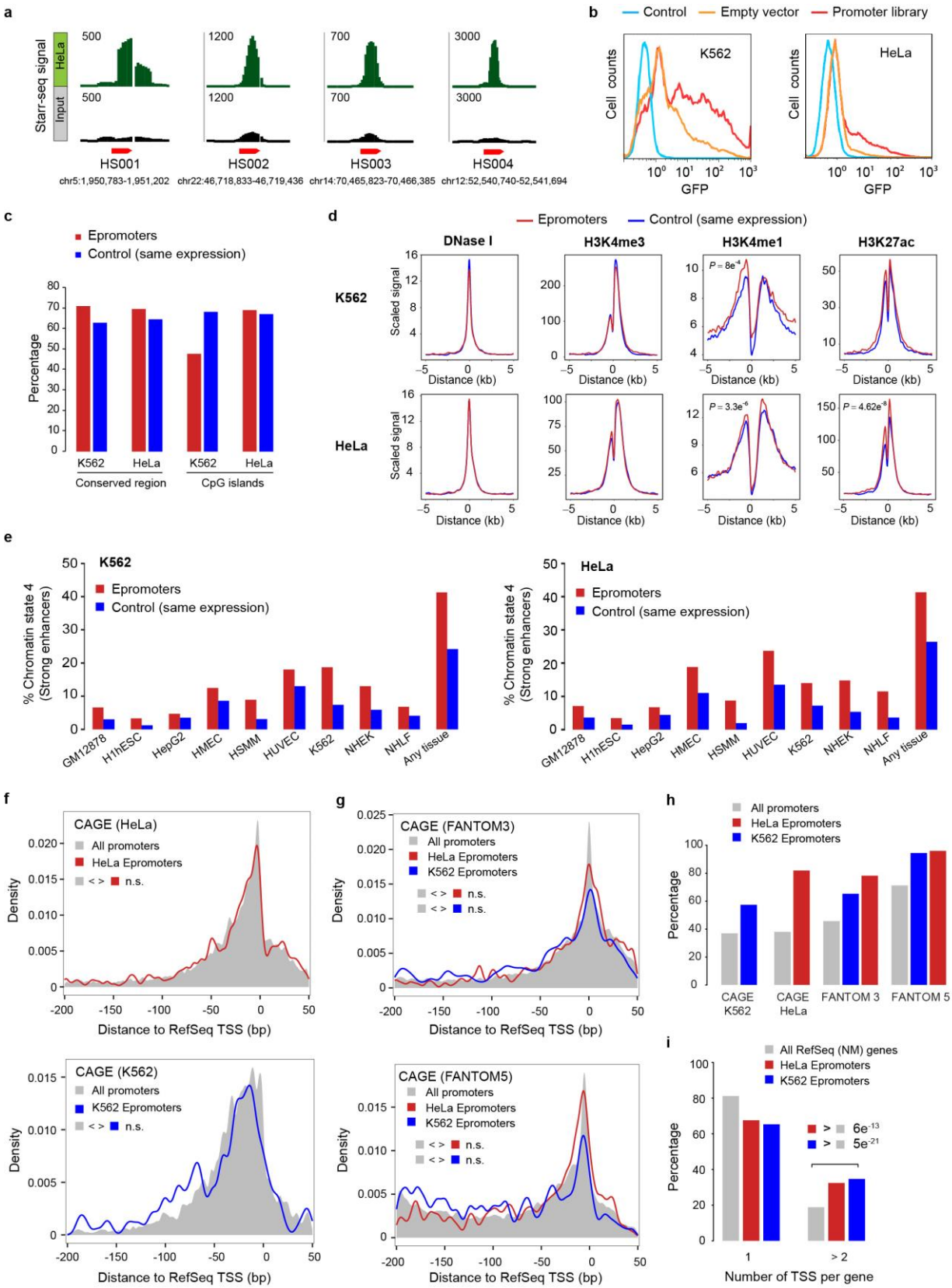


Supplementary Figure 1

Analysis of DHS STARR-seq in the P5424 cell line.

(**a**) Luciferase enhancer assays of proximal DHSs defined as active or inactive enhancers by STARR-seq in P5424 cells. For each candidate, both orientations were tested. Data represent the normalized fold change over the vector control. Error bars show s.d. from three independent transfections (***P < 0.001, **P < 0.01, *P < 0.1; two-sided Student's t test). (**b**) Enrichment score of lymphoid transcription factors at distal and proximal DHSs based on ChIP–seq data from developing thymocytes. The enrichment score was calculated as the $-\log_{10}$ (P value) obtained with a hypergeometric test (depletion is represented by negative values).

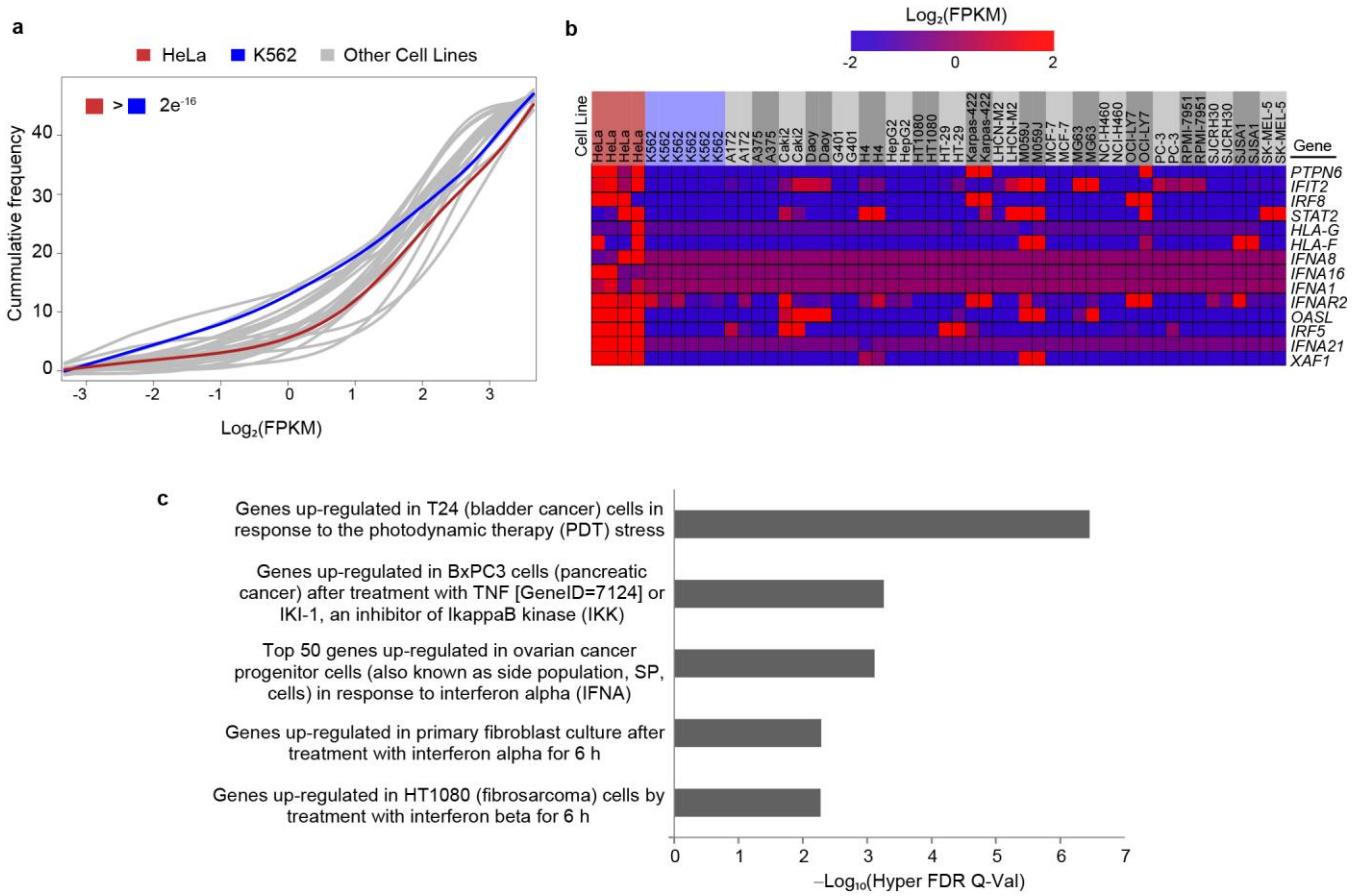


Nature Genetics: doi:10.1038/ng.3884

Supplementary Figure 2

CapStarr-seq experimental control and epigenomic profiles of Epromoters in K562 and HeLa cells.

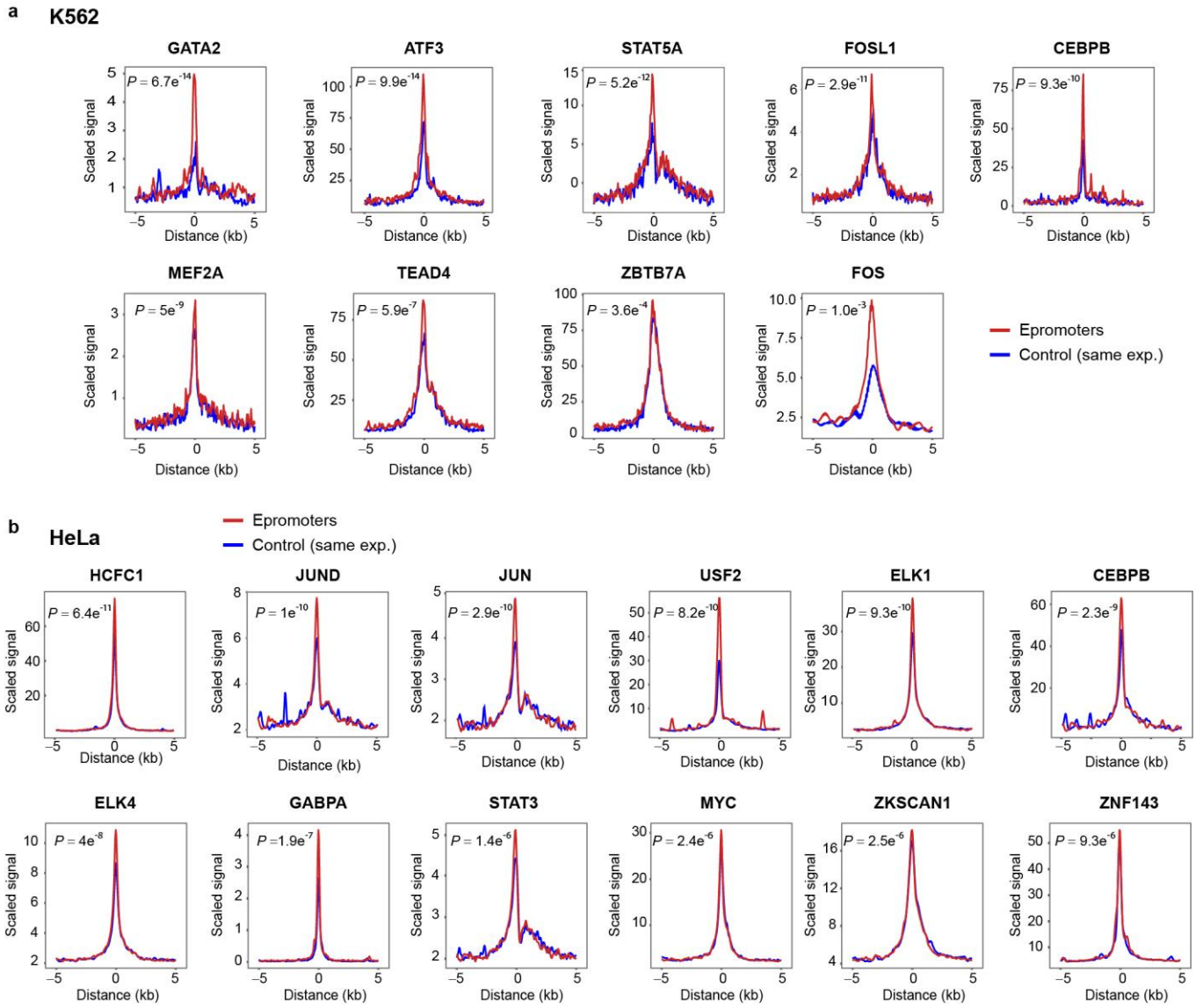
(**a**) IGV screenshots of STARR-seq signals for four STARR-seq-positive controls in HeLa cells. (**b**) FACS analysis of GFP expression in K562 (left) and HeLa (right) cells transfected with a human promoter library or empty vector. Controls were untransfected cells. The increase in GFP expression in transfected cells with the promoter library indicates potential enhancer activity in the pooled library. (**c**) Overlap with CpG islands (50%) and regions conserved in placental mammals (10%) using the EpiExplorer tool. The control is non-Epromoters with equal levels of gene expression as Epromoters in the same cell type. (**d**) Average profiles of epigenomic features for Epromoters and control promoters with the same expression pattern of associated genes. Statistical significance was calculated in a region centered on the TSS (±1 kb) using two-sided Mann–Whitney *U* tests. Only significant differences ($P < 0.001$) are shown. (**e**) Percentage of chromatin state 4 (strong enhancers) found in K562 Epromoters (left) and HeLa Epromoters (right) across ENCODE cell lines using the EpiExplorer tool. (**f**,**g**) Density plots of TSS positions corresponding to the selected promoter regions using CAGE peaks from ENCODE data in HeLa (**f**, top) and K562 (**f**, bottom) cells and data from FANTOM3 (**g**, top) and FANTOM5 (**g**, bottom) (Kolmogorov test). (**h**) Percentage of TSSs assigned to RefSeq-defined TSSs using different CAGE databases (from data in **Supplementary Table 2b**). (**i**) Comparison of the number of different RefSeq-defined TSSs per coding gene (one-sided Mann–Whitney *U* test).



Supplementary Figure 3

Assessment of the IFN-αβ signaling pathway.

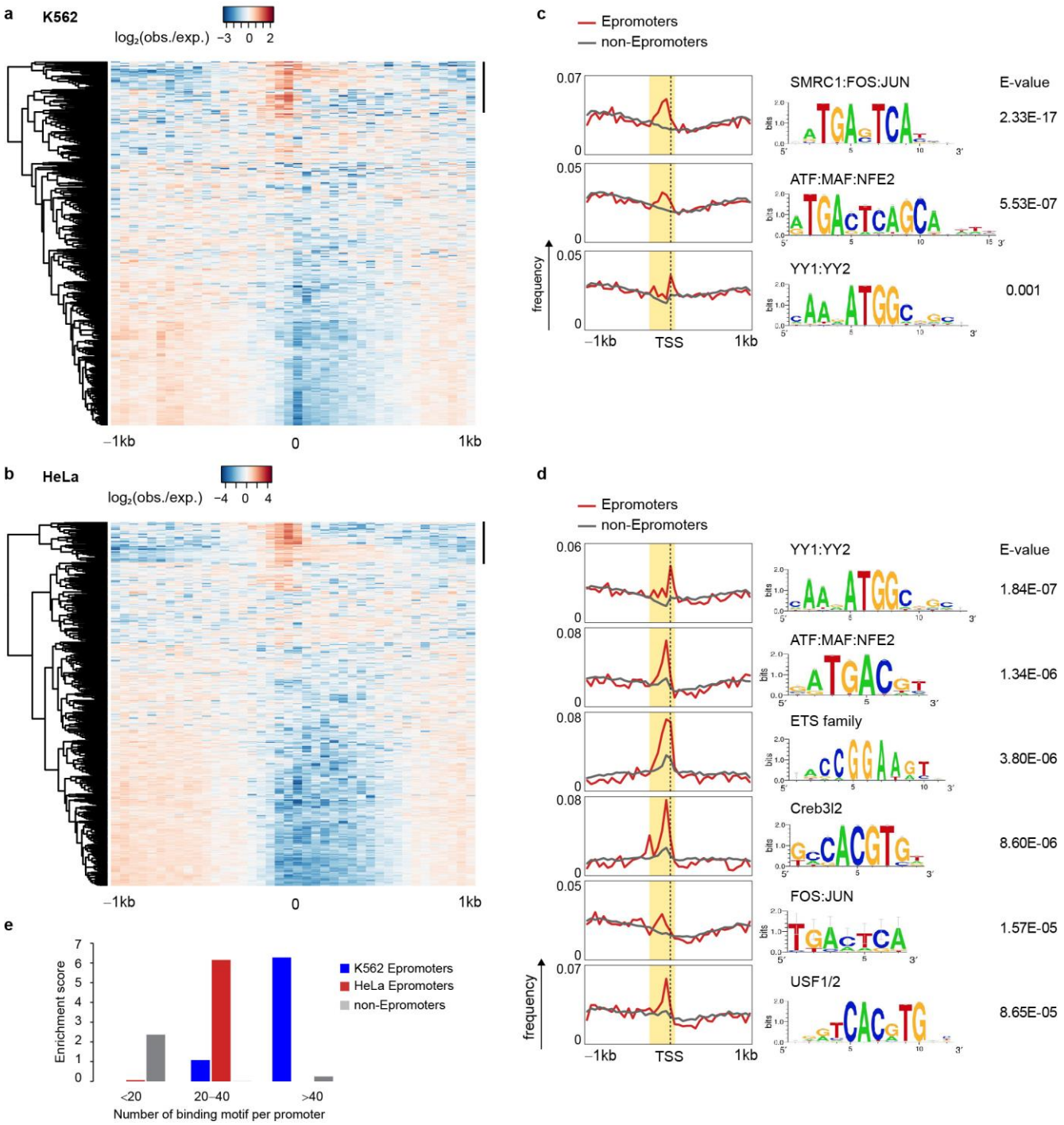
(**a**) Cumulative plot of normalized RNA levels (FPKM) for genes from the IFN-αβ signaling pathway (Reactome), based on RNA–seq data from 23 cell lines. The HeLa and K562 cell lines are highlighted (Kolmogorov test). (**b**) Heat map showing RNA–seq relative expression (FPKM) for genes from the IFN-αβ signaling pathway (Reactome) expressed at significantly higher levels in HeLa cells as compared to the 22 remaining cell lines (SAM analysis; α = 0.5). (**c**) Transcription signatures related to stress/interferon response significantly enriched in the set of Epromoter-associated genes in HeLa cells (GREAT tool).



Supplementary Figure 4

Enrichment of transcription factors at Epromoters.

(**a**,**b**) Average profiles of ChIP–seq signals for ENCODE transcription factors enriched at Epromoters in K562 (**a**) and HeLa (**b**) cells. Statistical significances were calculated in a region centered on the TSS (±250 bp) using two-sided Mann–Whitney *U* tests.

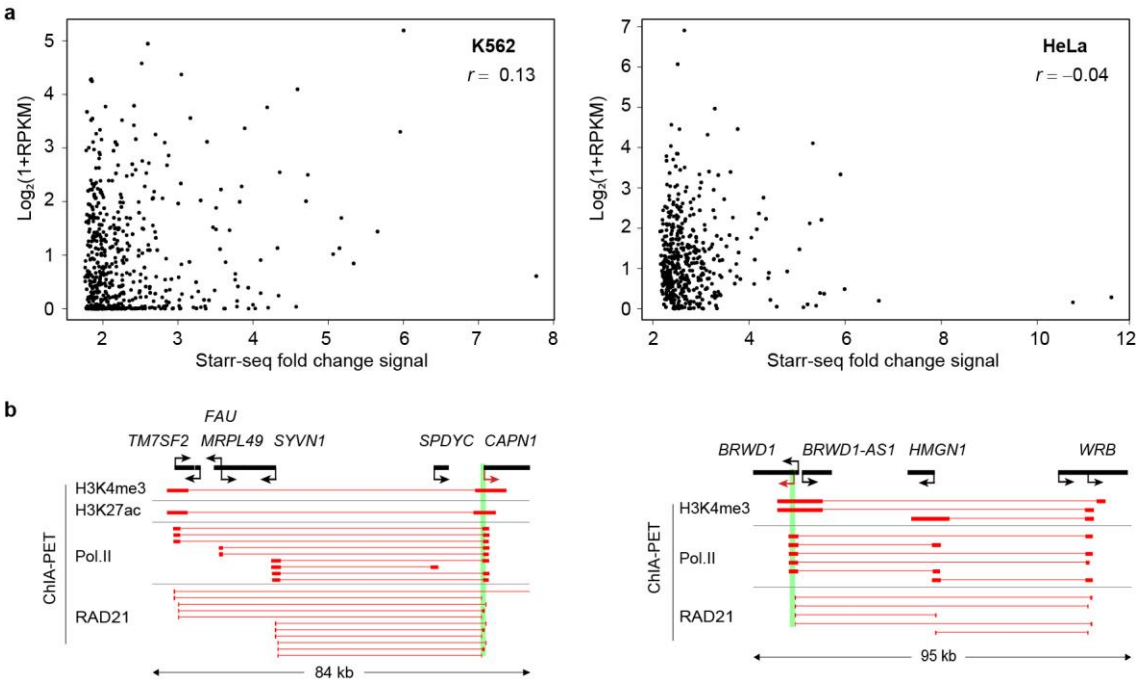


Supplementary Figure 5

Motif enrichment at Epromoters.

(**a**,**b**) Heat maps showing the enrichment distribution (log_2 (observed/expected)) of the non-redundant collection of motifs obtained by combining transcription factor binding motif (TFBM) databases (Jaspar vertebrates and Hocomoco Human). TFBMs were used to scan the extended Epromoter-associated TSS from −1 kb to +1 kb and clustering was performed based on the binding profiles in K562 (**a**) and HeLa (**b**) cells. Motifs enriched around the TSS (black line) were selected. (**c**,**d**) Significantly enriched motifs in K562 (**c**) and HeLa (**d**) cells were identified by comparing the binding enrichment within the promoter region (–200 bp to +50 bp with respect to the TSS; highlighted as orange boxes) between Epromoters and the non-Epromoters. Binding site distribution (left), motif logos (middle) and *E*

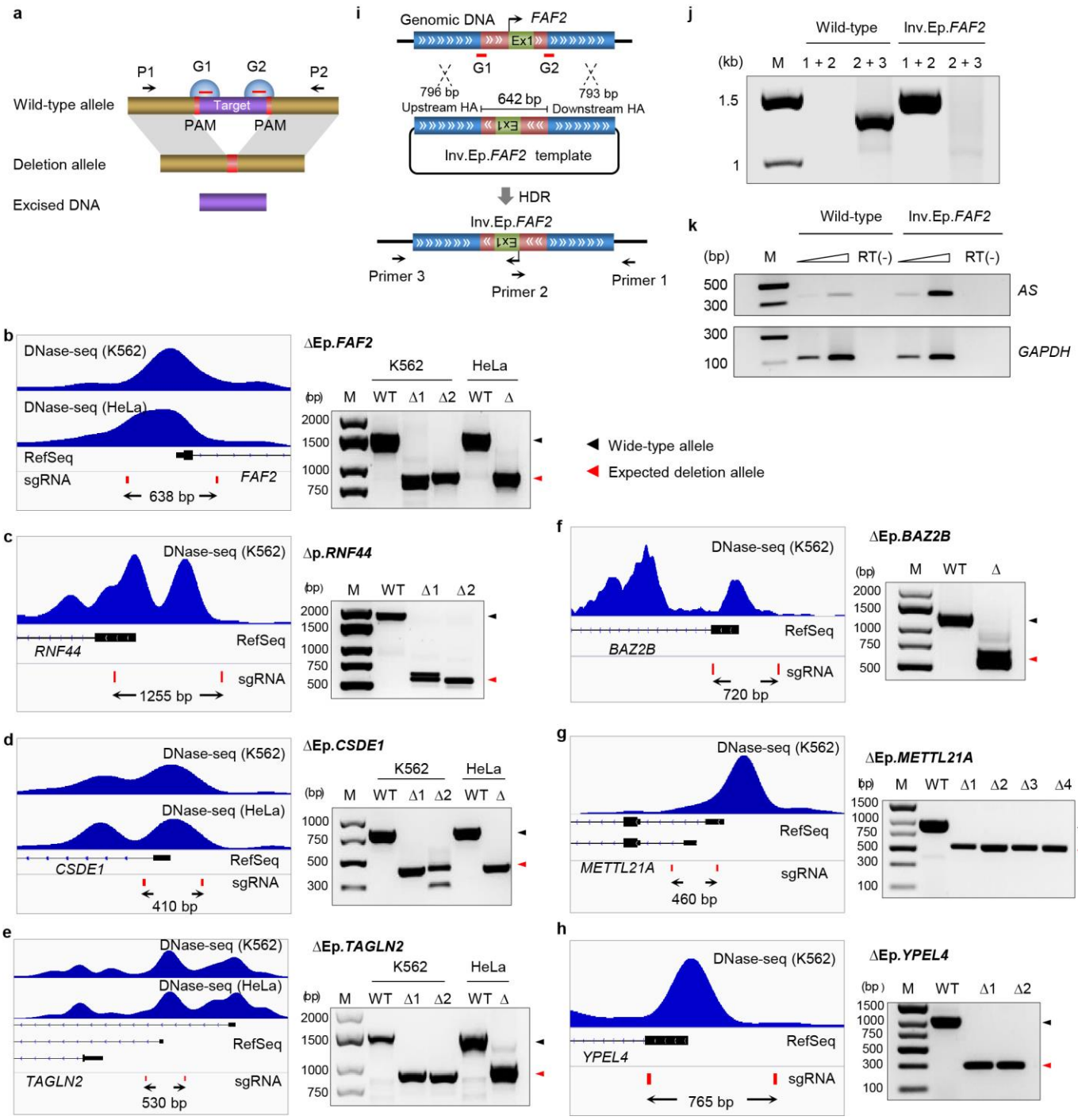
values (right) are shown only for significantly enriched motifs ($E < 0.001$; χ^2 test). (**e**) Enrichment of Epromoters and non-Epromoters as a function of the number of different TFBMs found. The enrichment score was calculated as the $-\log_{10}$ (*P* value) obtained by hypergeometric test.



Supplementary Figure 6

Proximal and distal correlations of Epromoters with gene expression.

(**a**) Scatterplots showing the Pearson correlation between the STARR-seq signal of Epromoters and the expression of associated genes. (**b**) Examples of consistent promoter–promoter interactions observed with different ChIA-PET data sets in K562 cells.

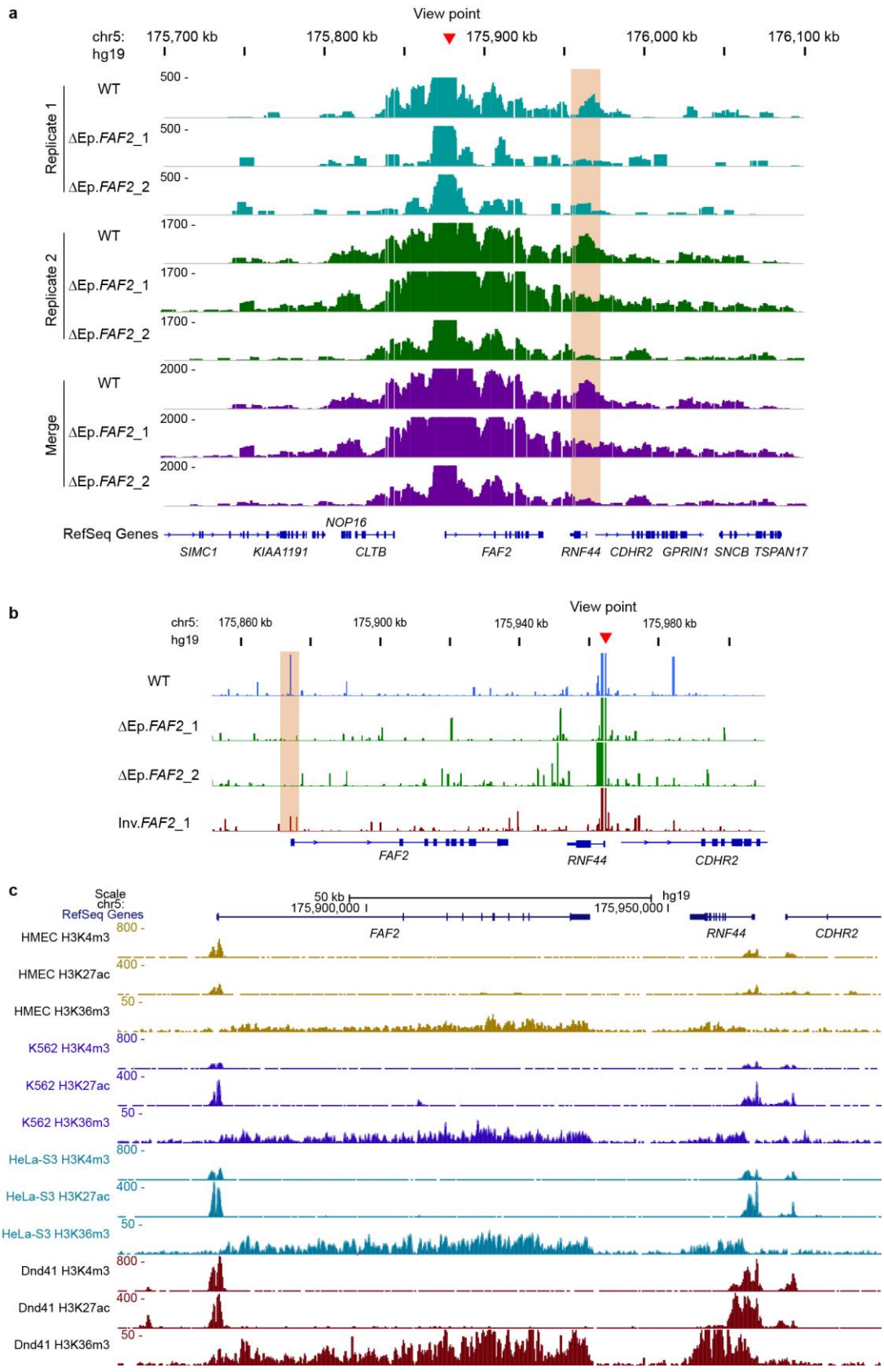


Supplementary Figure 7

Generation of knockout and knock-in cell clones via CRISPR–Cas9.

(**a**) General strategy for the generation of (E)promoter knockouts. Two gRNAs, G1 and G2, were designed flanking the genomic target to delete the intervening DNA segment. The CRISPR–Cas9 system creates two double-strand breaks (DSBs) at 3 – 4 nt upstream of the PAM sequences (red) and releases the excised DNA (purple). The resulting DSB is repaired by the NHEJ pathway. The genomic deletion is detected by PCR using primers P1 and P2. (**b**–**h**) Assessment of (E)promoter knockout. Left, IGV screenshots showing the DNase-seq (ENCODE) and RefSeq tracks for targeted regions. The locations of gRNAs (red boxes) and the expected sizes of deleted regions are indicated. Right, PCR validation of biallelic deletion in corresponding cell clones. Details on the gRNA sequences, PCR

primers and expected PCR fragments are provided in **Supplementary Table 9**. (**i**) Strategy for the generation of the inverted *FAF2* Epromoter knock-in. The two gRNAs, G1 and G2, used to generate DSBs are as in the knockout experiment. The repair template contains upstream and downstream homologous arms (HAs) flanking the inverted *FAF2* Epromoter. The HDR-mediated repair pathway generates the inverted *FAF2* Epromoter knock-in, which is detected by PCR with the combination of two primer pairs (1 + 2) and (2 + 3). (**j**) PCR validation of a successful inverted *FAF2* Epromoter knock-in cell clone using the combination of primers shown in **i**. (**k**) RT–PCR detection of antisense (AS) transcription in an Inv.Ep.*FAF2* clone. *GAPDH* was used as a cDNA loading control.

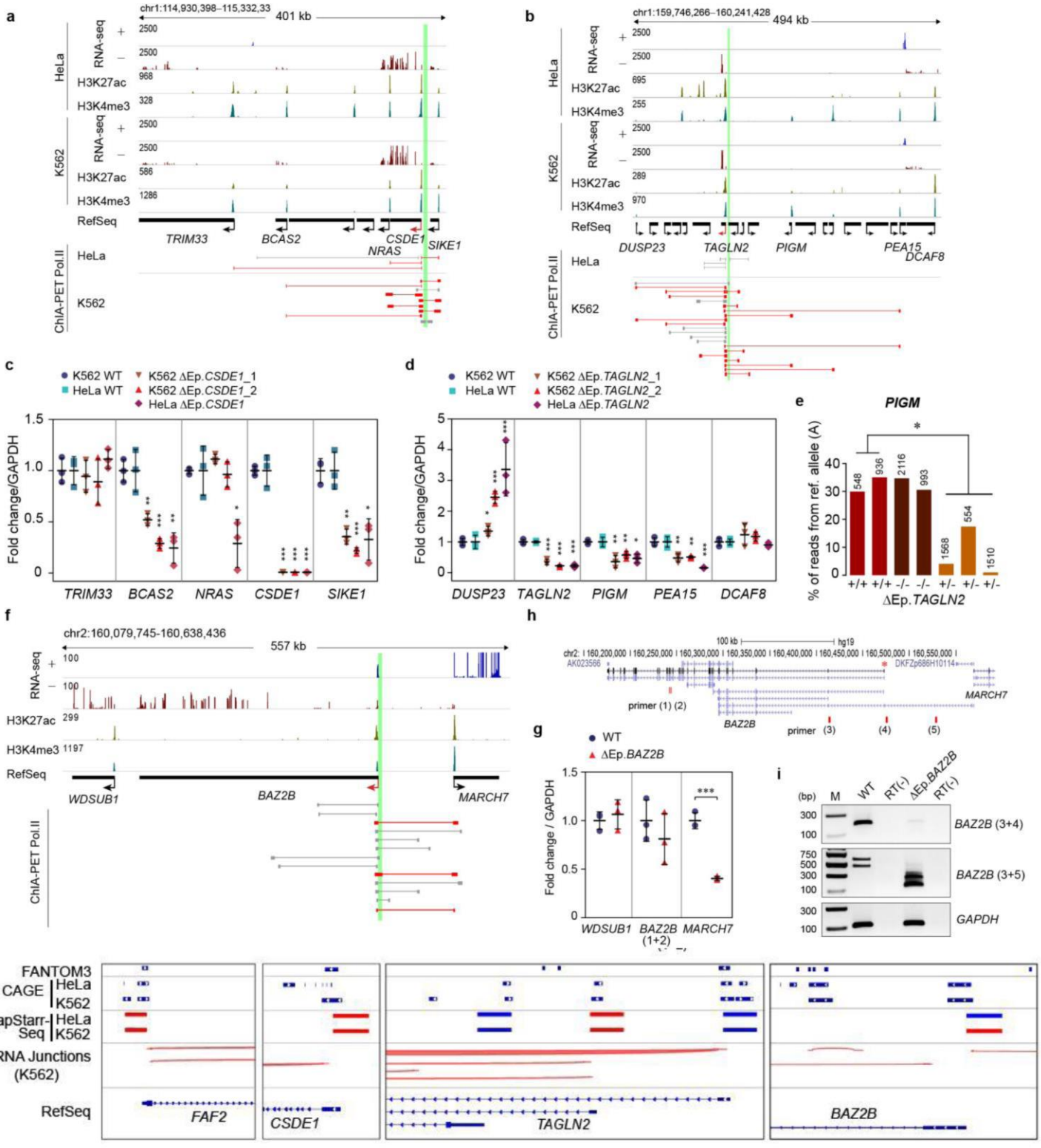


Nature Genetics: doi:10.1038/ng.3884

Supplementary Figure 8

Interaction and epigenetic co-regulation of *FAF2* and *RNF44*.

(**a**) Genomic tracks showing the 4C–seq analysis of interactions between the *FAF2* (**a**) and *RNF44* (**b**) promoters in WT and knockout K562 clones. The viewpoint from the *FAF2* Epromoter is indicated by an arrowhead. The specific interaction between the *FAF2* and *RNF44* promoters is highlighted by the orange box. (**b**) UCSC Genome Browser tracks for H3K4me3, H3K27ac and H3K36me3 at the *FAF2* locus and nearby regions across the HMEC, K562, HeLa and DND41 cell lines.

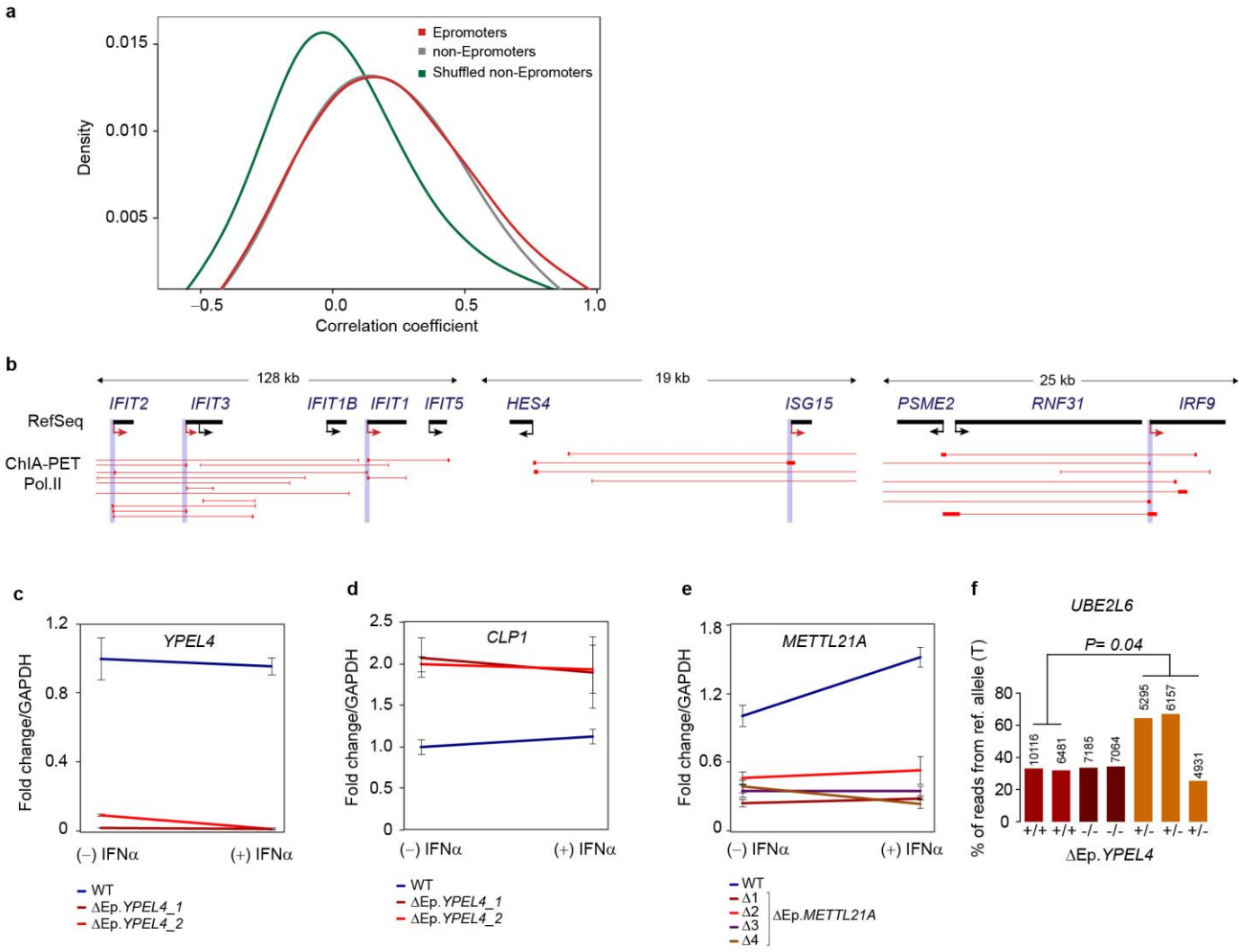


Nature Genetics: doi:10.1038/ng.3884

Supplementary Figure 9

Additional validations of distal gene regulation by Epromoters.

(**a**,**b**) IGV tracks for RNA–seq, ChIP–seq and ChIA-PET Pol II data in K562 and HeLa cells at the *CSDE1* (**a**) and *TAGLN2* (**b**) loci and nearby regions. The promoter–promoter interactions for Epromoters are highlighted in red. (**c**,**d**) qPCR analysis of gene expression in WT, ΔEp.*CSDE1* (**c**) and ΔEp.*TAGLN2* (**d**) clones. The number following the gene name is the number of independent cell clones. (**e**) Allelic frequency of the A versus T variant (chr1:160000435) in *PIGM* transcripts in WT, ΔEp.*TAGLN2* homozygous and ΔEp.*TAGLN2* heterozygous K562 clones. The total number of reads is indicated for each sample. The significant deviation of allelic frequency in heterozygous clones with respect to homozygous samples was calculated by performing a one-sided Student's *t* test. (**f**) IGV screenshot showing tracks for RNA–seq, ChIP–seq and ChIA-PET Pol II data in K562 cells at the *BAZ2B* locus and the nearby region. (**g**) qPCR analysis of gene expression in WT and ΔEp.*BAZ2B* clones. Knockout of the *BAZ2B* Epromoter resulted in significant reduction of *MARCH7* expression but had no effect on the nearby gene *WDSUB1* or *BAZ2B* (using primers 1 and 2 shown in **h**). (**h**) UCSC Genome Browser tracks showing the different *BAZ2B* transcripts and primers used in **g** and **i**. (**i**) Alternative promoter usage for the *BAZ2B* gene was assessed by RT–PCR in K562 cells. The smaller fragment size observed in ΔEp.*BAZ2B* clones corresponds to the deletion of exon 1 (asterisk in **h**). (**j**) IGV tracks for FANTOM3 and ENCODE CAGE data, CapStarr-seq regions and RNA junctions around the TSS of the indicated gene. The red color in CapStarr-seq tracks represents active Epromoters. For **c**, **d** and **g**, error bars show s.d. (n = 3 independent RNA/cDNA preparation; ***P < 0.001, **P < 0.01, *P < 0.1, two-sided Student's t test).



Supplementary Figure 10

Epromoters involved in IFN-α signaling in K562 cells.

(**a**) Distribution of expression correlation for ChIA-PET interacting gene pairs including at least one Epromoter (red) or excluding Epromoters (gray) and randomly rewired gene pairs (green) using RNA–seq data from ENCODE. Statistical significance was assessed by Kolmogorov test. (**b**) Examples of clusters of interferon response genes (green labels) associated with Epromoters (red arrows) in HeLa cells. (**c**–**e**) qPCR analysis of gene expression in WT, ΔEp.*YPEL4* (**c**,**d**) and ΔEp.*METTL21A* (**e**) cell clones. Error bars show s.d (*n* = 3 independent RNA/cDNA preparations). (**f**) Allelic frequency of the T versus C variant (chr11:57319339) in *UBE2L6* transcripts in WT, ΔEp.*YPEL4* homozygous and ΔEp.*YPEL4* heterozygous K562 clones. The total number of reads is indicated for each sample. The significant deviation of allelic frequency in heterozygous clones with respect to homozygous samples was calculated by performing a one-sided Student's *t* test.

Supplementary Note

Extended Methods

Construction of the human promoter library

Genomic library was generated from a pool of genomic DNA extracted from peripheral blood cells of healthy donors. For target enrichment, a home-designed 3 bp resolution oligonucleotide microarray covering from −200 to +50 bp relative to the TSS of 20,719 human protein-coding genes was constructed using the SureSelect technology (Agilent, 1M format) and the eArray tool default settings (https://earray.chem.agilent.com/earray/). In addition, 4 STARR-seq positive controls previously identified as enhancers in HeLa[1] and 370 random genomic regions (250 bp) without active epigenomic features in ENCODE cell lines were included (**Supplementary Table 2a**).

TSS analyses and comparison with CAGE data

To compare the number of distinct TSS from coding genes associated with Epromoters or non-Epromoters, the hg19 RefSeq annotation was retrieved from UCSC Table Browser (http://genome.ucsc.edu/cgi-bin/hgTables) and the number of transcripts from the same coding gene with different start coordinates was computed and graphed by R software in a bar plot (**Supplementary Fig. 2i**). To corroborate the TSS position of Epromoters, CAGE tags TSS data from FANTOM3 (http://gerg01.gsc.riken.jp/cage_analysis/export/hg17prmtr), FANTOM5 (http://fantom.gsc.riken.jp/5/data/) or HeLa and K562 CAGE peaks from ENCODE (https://www.encodeproject.org/) were obtained (source data in **Supplementary Table 10**). The FANTOM3 data was lifted into hg19 genome annotation (LiftOver by UCSC tools) and processed to obtain a CAGE set (tag clusters with $\geq$ 2 tags) according to Hayashizaki *et al*.[2]. Intersection between CAGE-defined TSSs and the promoter regions (from −200 to +50 bp relative to the RefSeq-defined TSS) was retrieved by BedTools (v2.25.0) (**Supplementary Table 2b)**. The percentages of intersections are shown in **Supplementary Fig. 2h**. Density plots were graphed by R software. A Kolgomorov test was performed between each pair of promoter sets (**Supplementary Fig. 2f, g**).

Building of a non-redundant database of TFBS

A non-redundant motif database was built by merging 641 motifs from the Hocomoco Human motif database[3], and 519 motifs from the JASPAR core vertebrate[4], versions 2016 for both databases. Motif analysis was performed with the Regulatory Sequence Analysis Tools suite[5]. The merged collection was reduced to 486 non-redundant motifs with matrix-clustering. We used very stringent parameters (correlation $\geq$ 0.85, width-normalized correlation $\geq$ 0.7) in order to merge only motifs of high similarity and sizes. Matrices were regrouped by hierarchical clustering, using the width-normalized correlation as similarity metric

Generation of *FAF2*-Epromoter inverted clones

For the inversion of *FAF2*-Epromoter, the upstream homology arm (796 bp; chr5:176,447,045-176,447,840) and downstream homology arm (793 bp; chr5:176,448,483-176,449,275) flanking the inverted *FAF2*-Epromoter (642 bp; chr5:176,447,841-176,448,482) were PCR amplified, purified and assembled using Gibson Assembly Master Mix (NEB). The assembled product was then cloned into

Nature Genetics: doi:10.1038/ng.3884

pGEM-T Easy vector (Promega) generating the repair template for homologous directed repair pathway (HDR) (**Supplementary Fig. 7i**). K562 cells were transfected with 8 μg of hCas9 vector, 4 μg of each gRNA (same as for the knockout experiment) and 4 μg of repair template. After 3 days of transfection, cells were plated in 96-well plates for clonal expansion as described above. For inversion detection, the specific primer pairs were designed as shown in **Supplementary Fig**. **7i** and **Supplementary Table 9**. Primer 1 and 3 were designed outside of inverted region, while primer 2 was inside and has the same direction as primer 3, allowing the detection of inverted *FAF2*-Epromoter in genomic DNA. The inverted *FAF2*-Epromoter clones were defined as having PCR amplification of inversion band (with primer 1 and primer 2) and absence of wild-type band (with primer 2 and primer 3) (**Supplementary Fig. 7j**).

Generation of eQTL-SNP mutated clones

For the study of eQTL SNPs, a gRNA was design to create a break near the target (the 20 nt of gRNA overlap with the target SNP; **Supplementary Table 9**). A 100 bp single-stranded Oligo Donor (ssODN) centered on the SNP was used as HR template. High-quality ssODNs were synthesized and PAGE purified (Sigma Aldrich). K562 cells were transfected with 5μg of gRNA, 10 μg of hCas9 and 1 μl of 100 μM ssODN template. The clonal expansion was performed as above. For clonal screening, individual cell clones were subjected to PCR using Phire Tissue PCR Master Mix (ThermoFisher Scientific) followed manufacture's protocol. Forward and reverse primers were designed bracketing the target SNP. The PCR products were then purified using MinElute Purification kit (Qiagen) and sequenced (Eurofins Genomics). For *CSDE1* SNP (rs6681671; NC_000001.10:g.115300685C>T) we obtained a clone harboring a homozygous replacement of the reference allele (C) by the alternative allele (T) and selected for further analyses (rs6681671_T/T). For *BAZ2B* SNP (rs1046496; NC_000002.11:g.160473399A>T) no homozygous replacement was obtained; instead we selected a homozygous deletion of the SNP (Δrs1046496).

Chromatin immunoprecipitation (ChIP)

ChIP Total 40×10^6 K562 cells were crosslinked in 1% formaldehyde for 10 min at 20 ^{o}C, followed by quenching with glycine at a final concentration of 250 mM. Pelleted cells were washed twice with ice-cold PBS, and then re-suspended in lysis buffer (20 mM Hepes pH 7.6, 1% SDS, 1X PIC) at final cell concentration of 15×10^6 cells/ml. Chromatin was sonicated with Bioruptor (Diagenode) to an average length of 200-400 bp (5 pulses of 30 sec ON and 30 sec OFF). An aliquot of sonicated cell lysate equivalent to 0.5×10^6 cells was diluted with SDS-free dilution buffer (1% Triton X-100, 1.2 mM EDTA, 16.7 mM Tris pH 8.0, 167 mM NaCl) for single immunoprecipitation. Specific antibodies (1 μg per ChIP) and proteinase inhibitor cocktail were added to the lysate and rotated overnight at 4 ^{o}C. The antibodies used were as follows: H3K4me3 (C15410003-50) and H3K27ac (C15410196) (Diagenode). On the next day, Protein A magnetic beads (Invitrogen) were washed twice with dilution buffer (0.15% SDS, 1% Triton X-100, 1.2 mM EDTA, 16.7 mM Tris pH 8, 167 mM NaCl and 0.1% BSA) and added to the lysate and rotated 1 hour at 4 ^{o}C. Then, beads were washed with each of the following buffers: once with Wash Buffer 1 (2 mM EDTA, 20 mM Tris pH 8, 1% Triton X-100, 0.1% SDS, 150 mM NaCl), twice with Wash Buffer 2 (2 mM EDTA, 20 mM Tris pH 8, 1% Triton X-100,

Nature Genetics: doi:10.1038/ng.3884

0.1% SDS, 500 mM NaCl), twice with Wash Buffer 3 (1 mM EDTA, 10 mM Tris pH 8). Finally, beads were eluted in Elution buffer (1% SDS, 0.1 M $NaHCO_3$) and rotated at RT for 20 min. Eluted materials were then added with 0.2 M NaCl, 0.1 mg/ml of proteinase K and incubated overnight at 65 ^{o}C for reverse cross-linking, along with the untreated input (10% of the starting material). The next day, DNA was purified with QIAquick PCR Purification Kit (Qiagen) and eluted in 30 µl of water.

Prediction of eQTL impact on TF binding sites

In order to predict the effect on transcription factor binding of eQTL variants associated with distal gene regulation (**Supplementary Table 7**), we used the tool *variation-scan* from the RSAT tool suite[6]. In order to reduce false positives we set out to assesses the impact of each eQTL allele on TF binding using only motifs for biologically relevant TFs that were found to be over-represented in the Epromoters sequence set (**Supplementary Fig. 4**), as suggested previously[7]. For each eQTL within the assayed promoters the binding affinity for one motif was assessed for both alleles, if one of the alleles had a binding score with a *P* value $\leq 1x10^{-3}$ then a ratio between the *P* values for both alleles from the eQTL were compared, if the ratio was ≥ 10 then the eQTL was considered as having a putative effect on TFBSs. We compared the number of eQTLs affecting TF binding vs the not affecting between Epromoters and non-Epromoters using a fisher exact test. Using the same test we also compared the number of eQTLs affecting binding in Epromoters and non-Epromoters between eQTLs with positive and negative beta values. *P* values for fisher tests were corrected using Benjamini & Hochberg method in p.adjust R command. Distribution of beta-values for eQTLs putatively affecting and not affecting TF binding were compared between non-Epromoters and Epromoters using a one tailed non-parametric Wilcoxon Rank Sum Test (wilcox.test R function, alternative "less"), and corrected for multiple testing using Benjamini & Hochberg (p.adjust R function).

References

1 Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074-1077, (2013).

2 Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nature genetics* **38**, 626-635, (2006).

3 Kulakovskiy, I. V. *et al.* HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic acids research* **44**, D116-125, (2016).

4 Mathelier, A. *et al.* JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic acids research* **44**, D110-115, (2016).

5 Medina-Rivera, A. *et al.* RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic acids research*, (2015).

6 Medina-Rivera, A. *et al.* RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic acids research* **43**, W50-56, (2015).

7 Andersen, M. C. *et al.* In silico detection of sequence variations modifying transcriptional regulation. *PLoS Comput Biol* **4**, e5, (2008).

Nature Genetics: doi:10.1038/ng.3884