nature portfolio

https://doi.org/10.1038/s41586-021-03577-1

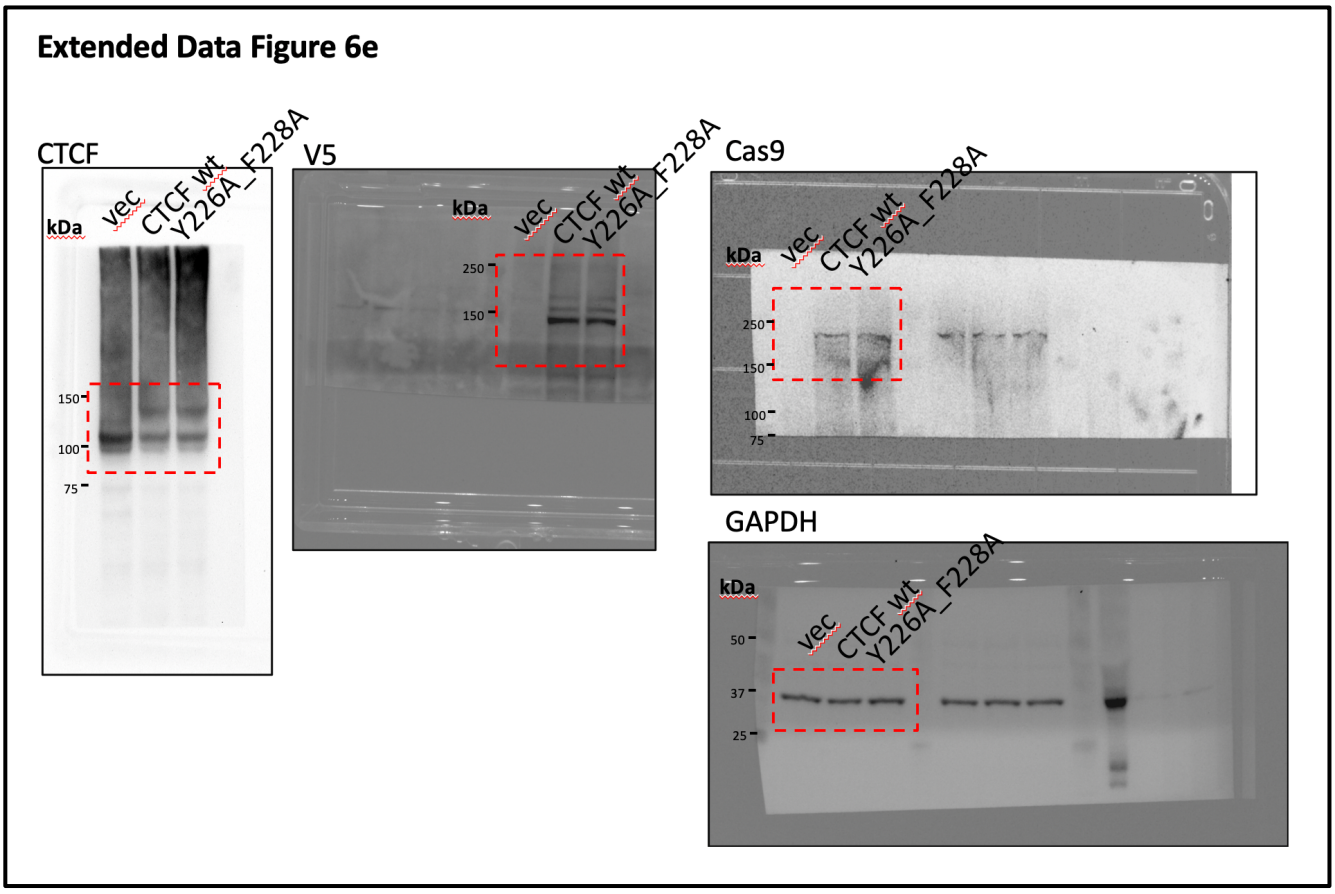
Supplementary information

Enhancer release and retargeting activates disease-susceptibility genes

In the format provided by the authors and unedited

Supplementary Figure 1. Uncropped images of Western Blots.

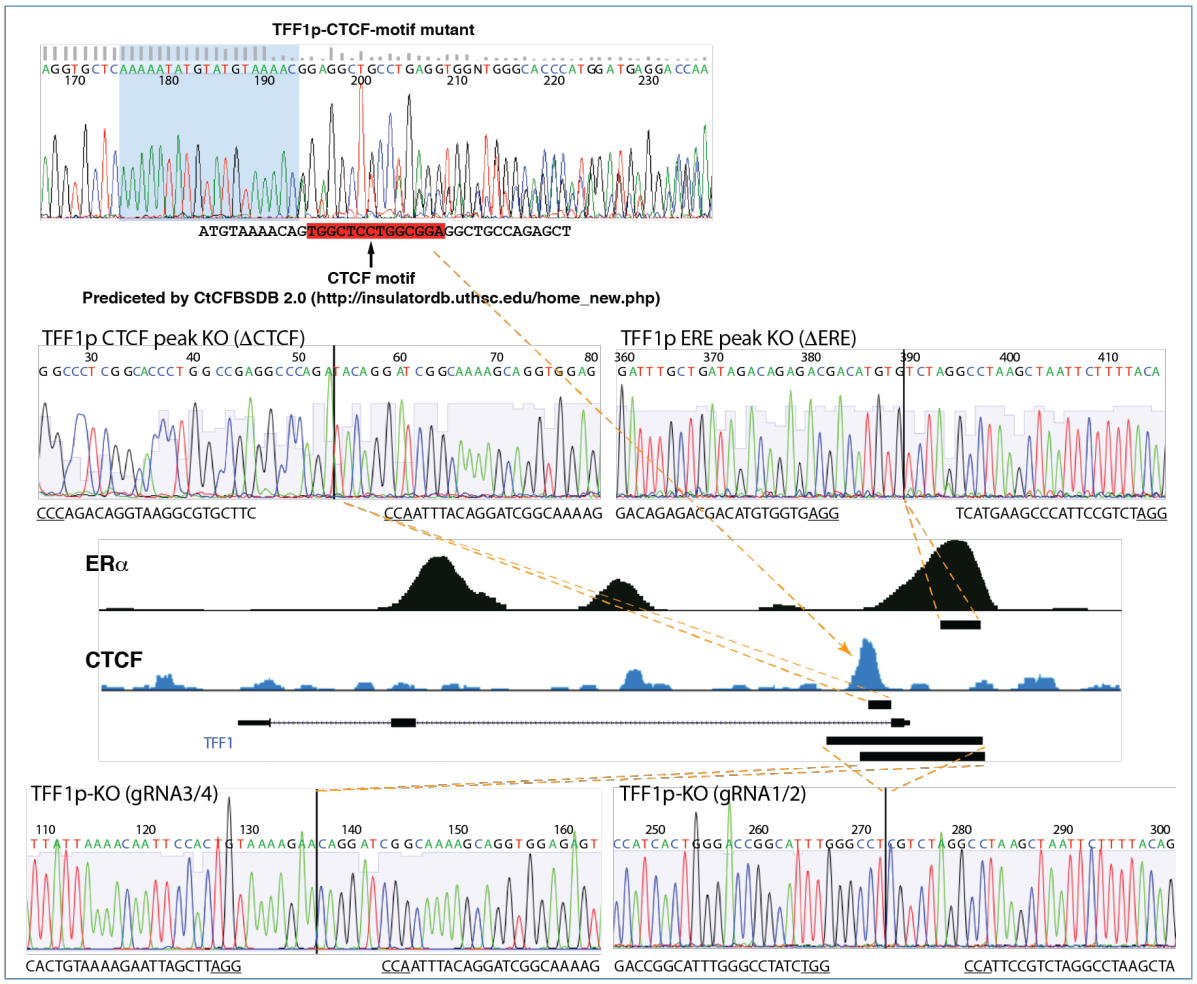
Raw Western blot images with protein molecular weight ladder and black box outlining the image being cropped to generate **Extended Data Fig.6e**. There are three proteins being examined here: dCas9-10xGCN4 (estimated MW ~200kDa, can be seen by Cas9 antibody); scFV-CTCF-V5 (WT and Y226A/F228A) (estimated MW ~140kDa), which can be seen by either CTCF or V5 antibodies.



Supplementary Figure 1. Western blots show the expression of proteins in three lanes: the vec indicates cells infected with virus expressing empty vector; the 'CTCF wt' indicates cells expressing scFV-CTCF-V5 WT and dCas9-GCN4; the 'Y226A_F228A' indicates cells expressing scFV-CTCF(Y226A/F228A)-V5 and dCas9-GCN4. The antibodies used for blotting are labeled on top of the blots. Blots were probed 1:2000 with primary antibody anti-CTCF, Millipore #07-729, anti-V5 Invitrogen #R960-25, anti-Cas9 Protein-tech #26758-I-AP, anti-GAPDH; and secondary anti-rabbit-IgG-HRP conjugate (CTCF and Cas9) or anti-mouse-IgG-HRP (V5 and GAPDH) conjugate.

Supplementary Figure 2. Sanger Sequencing of genetic deletion/mutation in TFF1p.

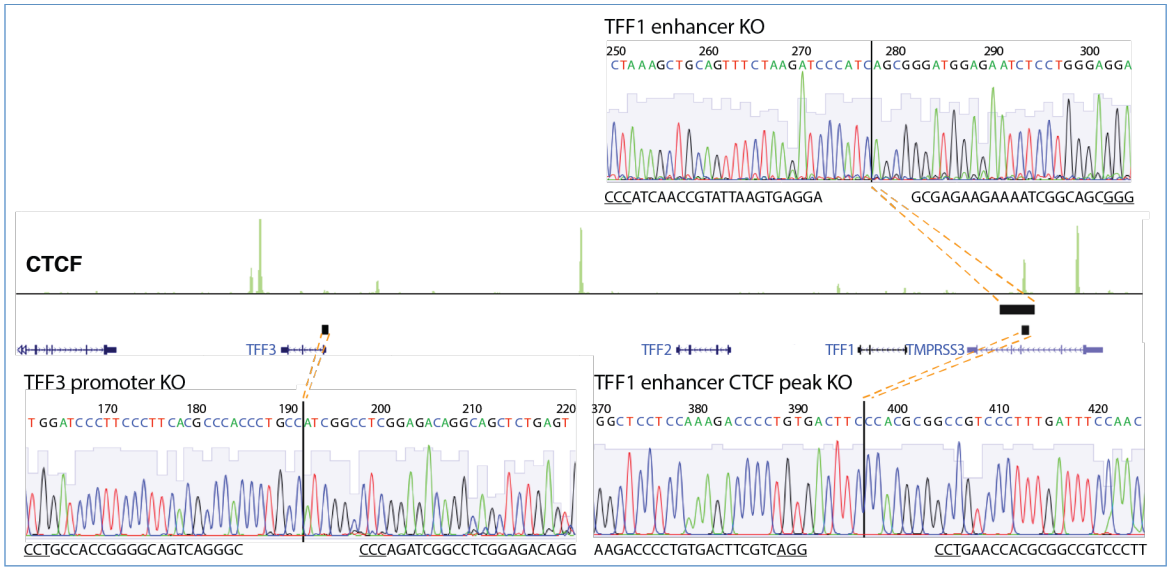
We used Sanger sequencing to validate each and every case of the genetic editing shown in **Supplementary Table 1**. The **Supplementary Figure 2** below shows the sanger validation of five cases of successful genome editing, which include the case of CTCF motif disruption in *TFF1* promoter (*TFF1p CTCF-motif mutant*, top), the deletion of the CTCF binding peak in *TFF1* promoter (ΔCTCF), the deletion of ERα binding peak (ΔERE) in the *TFF1* promoter, as well as two independent *TFF1* promoter deletion by two separate sets of gRNAs (gRNA1/2 and gRNA3/4).



Supplementary Figure 2. In the middle of the panel, we show a UCSC genome browser screenshot for ChIP-seq of CTCF, ERα, p300 and PolII binding in the *TFF1* promoter. Sanger sequencing results show the five cases of successful genome editing, which include the case of CTCF motif disruption in *TFF1* promoter (*TFF1p CTCF-motif mutant*, top), the deletion of the CTCF binding peak in *TFF1* promoter (ΔCTCF), the deletion of ERα binding peak (ΔERE) in the *TFF1* promoter, as well as two independent *TFF1* promoter deletion by two separate sets of gRNAs (gRNA1/2 and gRNA3/4). The DNA sequences below Sanger sequencing data indicate CTCF motif (only for the top panel) or gRNA sequences (for all other panels).

Supplementary Figure 3. Sanger Sequencing of genetic deletion for TFF1e and TFF3p.

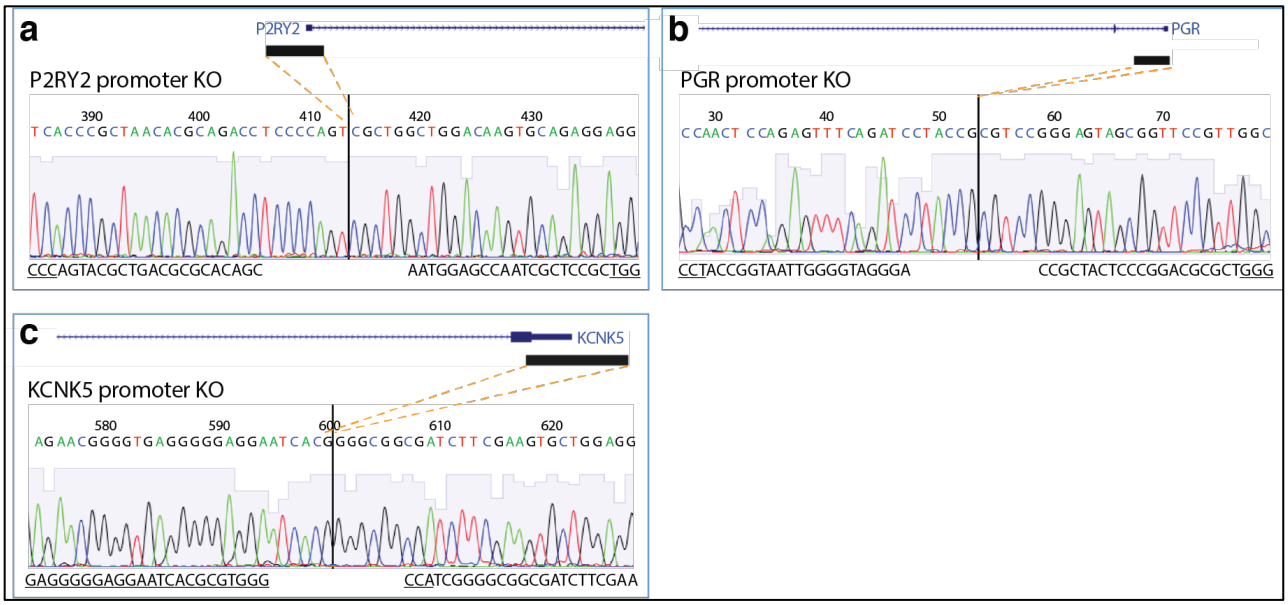
The *TFF1* enhancer is about 10kb away from the *TFF1* promoter, which bears a strong CTCF binding peak. As shown in the **Extended Data Fig.8d**, the *TFF1* enhancer is the only one of the four cases that has an enhancer CTCF peak among four estrogen-induced enhancer-gene pairs (TFF1e-p, P2RY2e-p, PGRe-p, KCNK5e-p). We deleted the entire *TFF1* enhancer, or only the CTCF peak in it using CRISPR/Cas9.



Supplementary Figure 3. The plot in the middle shows the genomic landscape of the *TFF2* and *TFF3* genes relative to the *TFF1* gene and the *TFF1* enhancer. The Sanger Sequencing results above or below show full deletion of *TFF1* enhancer (i.e. TFF1 enhancer KO or TFF1e-KO in **Supplementary Table 1**), deletion of a CTCF binding peak inside the *TFF1* enhancer (i.e. TFF1e CTCF-KO), as well as the deletion of the *TFF3* promoter (i.e. TFF3p-KO) in MCF-7 cells. The DNA sequences below Sanger sequencing data indicate gRNA sequences.

Supplementary Figure 4. Sanger Sequencing of genetic deletion of P2RY2p, PGRp and KCNK5p.

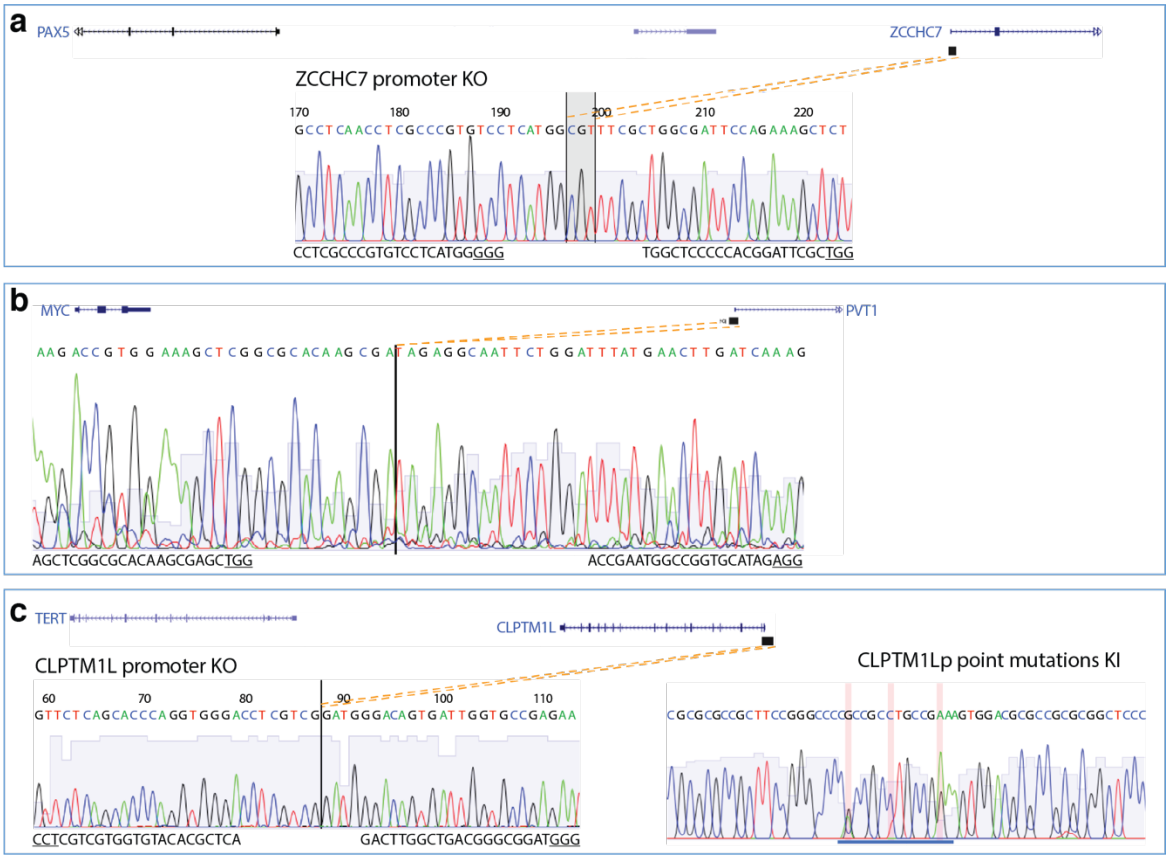
We have additionally conducted CRISPR/Cas9 deletion of three other estrogen target gene promoters (PGRp, P2RY2p and KCNK5p, also see Extended Data. Figs.1,4 and 8).



Supplementary Figure 4. Diagrams and Sanger Sequencing results show the promoter deletion of P2RY2 (panel a), PGR (panel b), and KCNK5 (panel c). The DNA sequences below Sanger sequencing data indicate gRNA sequences.

Supplementary Figure 5. Sanger Sequencing of genetic deletion or mutation of ZCCHC7p, PVT1p and CLPTM1Lp.

We have analyzed pan cancer genomic data (ICGC and PCAWG datasets, http://icgc.org/ and https://dcc.icgc.org/pcawg) to find cancer mutations/deletions that locate to oncogene-neighboring promoters (ONPs). Three prominent cases of ONP promoters next to oncogenes were selected for genetic deletion or for generating point mutations to model cancer genetic changes. These include *ZCCHC7p* that locates next to *PAX5* oncogene (panel a), *PVT1p* that locates next to *MYC* oncogene (panel b), and *CLPTM1Lp* that locates next to the *TERT* oncogene (panel c).



Supplementary Figure 5. Sanger sequencing results below are showing successful homozygous deletion of the three ONP promoters (panels a,b, and the left side of panel c). The DNA sequences below Sanger sequencing data indicate gRNA sequences. In addition to promoter deletion, for the *CLPTM1L* promoter, we used CRISRP/Cas9-mediated genome cutting followed by homologous recombination using oligonucleotides as donor to knock-in cancer point-mutations (right side of panel c, also see main manuscript **Fig.3e,f,g**). The Sanger sequencing result of the knock-in region is shown in panel c (right side). For this, a heterozygous genotype was achieved (pink highlights show the mutations).

Supplementary Figure 6 and notes.

Additional discussion that oncogene-neighboring promoters (ONPs) that are significantly mutated in human cancers.

We aim to examine the cancer mutations (or small deletions) homed in gene promoters neighboring oncogenes, which we dubbed oncogene-neighboring promoters (ONPs). The list of oncogenes was based on COSMIC database Cancer Gene Census (CGC), and promoters located in -/+200 kb genomic distance from the oncogene promoters (OPs, n=315) were referred to as oncogene-neighboring promoters (ONPs, n=1,693). We analyzed ICGC (International Cancer Genome Consortium) and the Pan Cancer Analyses of Whole Genome (PCAWG) datasets (http://icgc.org/ and https://dcc.icgc.org/pcawg).

ICGC Release 28 simple mutation and structure mutation data were obtained from ICGC data portal. Only mutations and deletions identified by whole genome sequencing (WGS) were used for further analysis. In total, 76,005,698 unique WGS mutations from 6,285 donors among 68 cancer cohorts were searched against 37,552 TSSs of 27,502 RefSeq genes annotated based on genome build hg19. The distances of mutations to TSSs were calculated by bedtools. The mutations with distance shorter than 1kb from TSSs were considered as promoter-homed mutations. Similarly, for deletions, a total of 64,113 deletions from 2,229 donors in 18 cancer cohorts were determined to be associated with promoters by distance selection.

We applied a Binomial probability model [1] to determine the statistical significance of mutational hotspots at promoter level, where the expected probability of a promoter mutation due to chance in each cancer cohort was calculated by the total mutations in all gene promoters divided by the total donor and promoter counts separately. The one-sided P-values for each gene promoter in every cancer type were determined by binomial test based on the affected donors and total tested donors, to test whether the number of donors observed was higher than the expected number by chance given the expected pattern of promoter mutation in that cancer type/cohort. The FDR values were adjusted by Benjamini-Hochberg method. In the plot below, we showed a list of significantly mutated or deleted ONPs (FDR < 0.001), and with a high number of donor counts (affected total donors > 10). (**Supplementary Figure 6**).

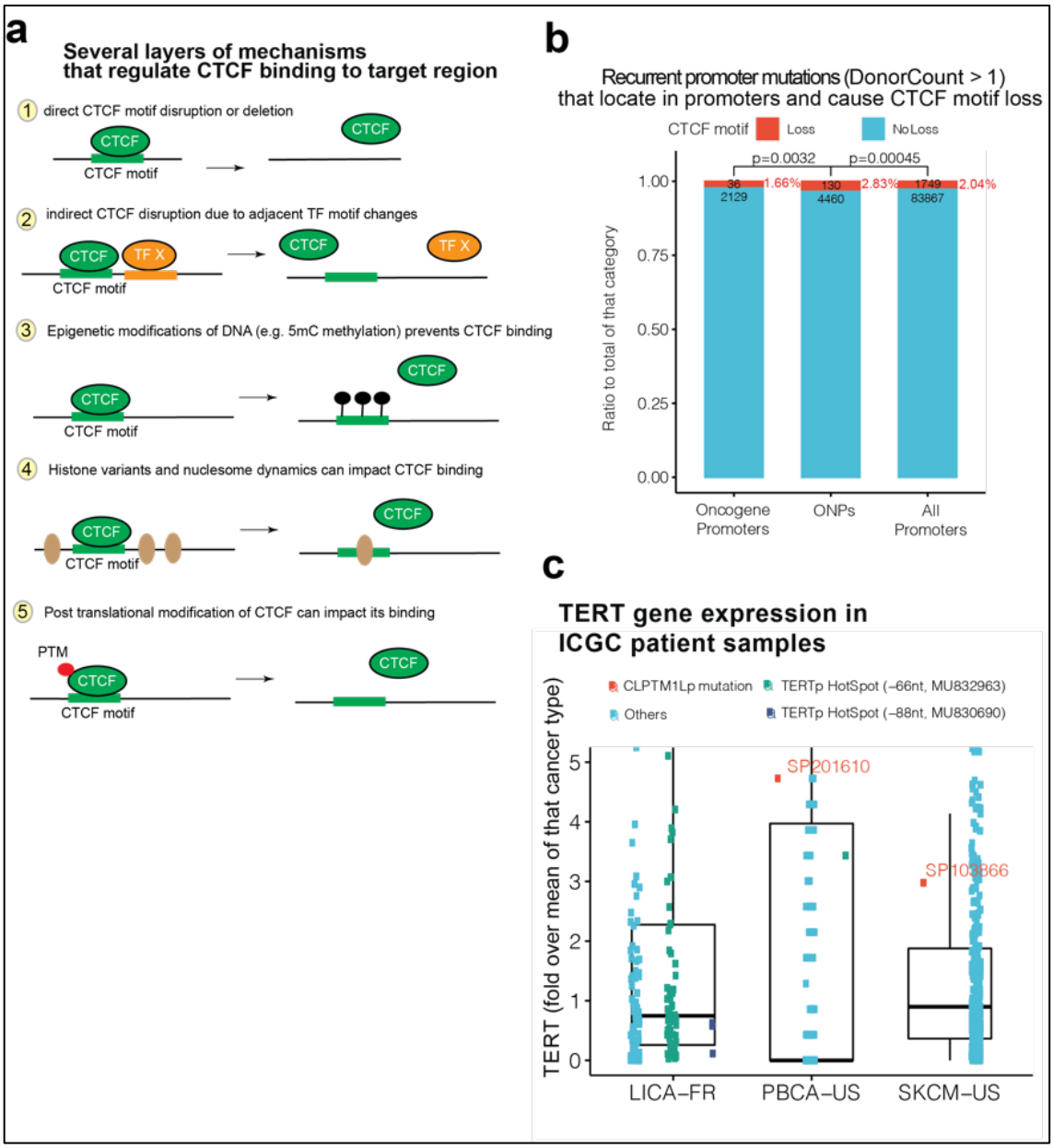
Supplementary Figure 6 legend:

In the plot below, the cancer cohorts (x axis) were ranked by the original cancer sites, and the ONP-OP gene pair (y axis) was reverse-alphabetically ranked by the neighboring oncogenes. Gene names before "->" are those of ONPs, while those after "->" are oncogenes. For example, "mutation TNFSF8-> TNC (row 3 below)" indicates that *TNC* is an oncogene listed in COSMIC, and the gene promoter of *TNFSF8* was identified as an ONP, which contains a significant number of somatic mutations in several cancer types. The dot size was scaled by the percentage of affected donors in each of that cancer type/cohort (SamplePerc). Only pairs of ONPs-OPs with mutations affecting more than 10 total donors in at least 1 cancer cohort and FDR < 0.001 by binomial test were shown here as dots with red color. Short deletion (<10 Kb) covering an ONP was shown as dots with blue color. The Cancer type abbreviations can be found in ICGC data portal (http://icgc.org/), for example: BLCA-US: Bladder Urothelial Cancer - TCGA, US.

Supplementary Figure 6

Supplementary Figure 7 and notes.
ONP cancer mutations and CTCF motif disruption.

Our results suggested that ERR is an overlooked biological/pathological process in which "a promoter defect prevents its engagement with its enhancer, which is released to scan the chromatin neighborhood, finding a new preferred target". Underlying ERR, the low-affinity CTCF binding at gene promoters appears to act as a quite common, but not universal, mechanism that dictates enhancer-promoter engagement. Direct motif disruption represents one of the several mechanisms that can cause CTCF binding loss at promoters and can trigger ERR.



Supplementary Figure 7:

a) A model figure to show several mechanisms that can impact CTCF binding at chromatin sites including promoters, which can potentially modulate/induce ERR. (TF X in the orange objects: other transcription factors).

b) Recurrent cancer mutations located in oncogene promoters, oncogene-neighboring promoters or all promoters that are predicted to disrupt CTCF motif. P values: Two-sided Fisher's exact tests. The numbers indicate the numbers of cancer mutations derived from ICGC release 28.
c) Box-and-whisker plots showing *TERT* expression levels in patients carrying mutations in ONP (i.e. *CLPTM1L* promoter) or in the OP (*TERT* promoter) itself. Well-reported two hotspot mutations in the *TERT* promoter were plotted as comparisons (-66nt and -88nt, which can also be referred to as C228T and C250T [2-4]). "Others" indicate all other cancer samples in that specific cancer cohort. Only samples with both genotype and RNA-seq data were included in this analysis. Three cancer cohorts are shown here (LICA-FR, PBCA-US, SKCM-US). PBCA-US and SKCM-US are shown because they are the only two cohorts that contain samples carrying mutations of the *CLPTM1L* promoter and paired genotype/RNA-seq datasets are available (red dots). LICA-FR does not contain any sample with *CLPTM1L* promoter mutation, but is shown here as a comparison because it is a cancer cohort with the most prevalent mutations in the *TERT* promoter itself. The boxes in the plots show 25-75% quantile of RNA-seq expression values, and whisker in the middle shows the median.

We consider that the "promoter defect" that can cause ERR can be multi-fold. Obviously, complete deletion of a promoter is the most obvious case, and we showed multiple examples of oncogene neighboring promoters that were deleted in human cancers (**Fig.3** and **Extended Data Fig.9c**, and **Supplementary Figure 6**), for which we provided experimental support that ERR operates therein.

We consider CTCF defect to be one, rather than an exclusive, mechanism underlying the loss of engagement of a promoter from its enhancer. For the subset of promoters that CTCF binding is determining, we are aware that CTCF binding can be affected due to direct motif changes, or changes in adjacent DNA sequences not directly overlapping the CTCF motif [5], or epigenetic changes [6] (e.g. DNA methylation [7] or nucleosome remodeling [8], **Supplementary Figure 7a**). Indeed, the CRISPRi method (i.e. dCas9-KRAB) that we used in **Fig.3b,c** and **Extended Data Fig.9d** suppressed promoters by creating heterochromatinization of promoters (e.g. H3K9me3), which can disrupt CTCF binding [9]. Furthermore, post-translational modifications of CTCF such as phosphorylation or poly ADP-ribosylation are known to alter CTCF binding to chromatin [10-12] (**Supplementary Figure 7a**, condition 5). There mechanisms may act at specific locus or under specific cell status to modulate CTCF binding at promoters to impact ERR and enhancer-promoter engagement. Indeed, in the *NUCKS1-RAB7L1* locus, the three SNPs in the *NUCKS1* promoter did not directly overlap any CTCF motifs, but they can still affect the binding of CTCF (**Fig.4f**).

Despite these additional layers of mechanisms, we tested how often genetic changes in human cancer can directly disrupt CTCF motifs in gene promoters. We analyzed recently released ICGC Release 28 simple mutations that are located in gene promoters, and found that ~5% (n=4,103) of all promoter-located cancer somatic mutations (n=85,616) overlap a CTCF consensus Motif (JASPAR motif MA0139.1). This employed a very stringent criterion as the mutation nucleotide has to directly overlap a FIMO-called CTCF motif. Of these, 2.04% (n=1,749) are predicted to disrupt CTCF motifs by FIMO (**Supplementary Figure 7b**). Our analysis showed that the likelihood of having CTCF motif disrupted is significantly higher for mutations located in oncogene-neighboring promoters (ONPs) than for those located in oncogene promoters (OPs) or in random promoters (**Supplementary Figure 7b**). Experimentally, we selected one case of highly mutated CTCF motif for validation. There are 35 patients that carry mutations in the promoter of *CLPTM1L*, with three of them overlapping a single CTCF motif (**Fig.3e** and **Extended Data Fig.9e**). A knock-in cell line that carries patient mutations showed disrupted CTCF binding in the *CLPTM1L* promoter, and it displayed deregulation of *CLPTM1L* and *TERT* gene expression (**Fig.3f**). These results functionally proved that some clinical mutations can activate *TERT* oncogene expression via ERR (**Fig.3e,f,g** and **Supplementary Fig. 7c**).

We also examined the expression of *TERT* gene in cancer patients' samples in which the *CLPTM1L* promoter harbored mutations. There are very few samples from cancer patients that carry ONP mutations available in the ICGC, for which paired genotype and RNA-seq datasets were conducted (common for low frequency noncoding mutations). For one of the three patients that carry CTCF disruption mutation in the *CLPTM1L* promoter (**Fig.3e** and **Supplementary Figure 7c**, PBCA-US, Specimen ID: SP201610; mutation ID: MU126588416), paired genotype and RNA-seq are available. Analysis of the RNA-seq data in this patient's sample showed that this sample had ~5-fold higher expression of *TERT* mRNA when compared to the mean expression level of *TERT* in that tumor type (**Supplementary Figure 7c**). There is another patient sample with a mutation in the *CLPTM1L* promoter but this was not predicted to disrupt CTCF motif (**Supplementary Figure 7c**, SKCM-US, Specimen ID: SP103866). This patient sample also has paired genotype and RNA-seq available, which had a ~3-fold higher level of *TERT* expression than the mean level seen in this cancer cohort (**Supplementary Figure 7c**). For comparison, we also analyzed patient samples with well-known hotspot mutations located in the *TERT* promoter (i.e., the -66nt and -88nt mutations, also referred to as C228T and C250T [2-4]). The *TERT* promoter is the most highly mutated noncoding region in human cancers [2-4]. In some cancer types, these hotspot mutations correlate with higher expression of *TERT* gene (see the Extended Data Fig.4a of a recent PCAWG paper [4]). However, in some other cancer types, including the one having the highest frequency of *TERT* promoter hotspot mutations (i.e., Liver cancer from France, or LICA-FR), the mutation-carrying tumors did not have significantly higher expression of *TERT* (**Supplementary Figure 7c**). The mechanism underlying such cancer type specificity for *TERT* promoter hotspot mutations is not clear at this stage, although tumor heterogeneity, mutation calling process and tumor sample numbers may be contributing factors. For the PBCA-US cohort, there was one case of patient carrying the -66nt hotspot mutation in the *TERT* promoter, and one case of patient carrying mutation in the *CLPTM1L* promoter (**Supplementary Figure 7c**). The gene expression of *TERT* in the ONP-mutated patient (*CLPTM1Lp* mutation, red dot) is <u>higher</u> than that in the patient with -66nt hotspot mutation in the *TERT* promoter (green dot, **Supplementary Figure 7c**).

Overall, due to the rarity of ONP-homed mutations and the limited samples for both genotype and RNA-seq, it is currently difficult to conduct statistical analysis of gene expression in cancer patient samples carrying rare noncoding mutations. But these results above support that ERR-like events take place in oncogene activation in clinical samples, and CTCF disruption in ONPs can directly activate *TERT* expression (together with **Fig.3e,f,g**). With increasing numbers of cancer samples with paired genotype (e.g., WGSs) and gene expression (i.e., RNA-Seq), future work will aim to delineate the correlation of ONP-mutated samples with cancer gene activation, and will distinguish those mutations that are potential "drivers" of ERR events versus those representing background mutations. Experimentally, it is critical to directly test the potential driver functions of these rare ONP mutations in cancer development using animal models. In this regard, it is noteworthy that even alleles mutated once in 5,338 tumors can still be tumorigenic [13], emphasizing that systematic experimental validation of rare cancer mutations can be important in addition to statistical analysis. Some of these rare noncoding mutations may provide insights to understand ~5-8% of human cancers for which no driver mutations can be identified [14].

<u>Other possible regulators of ERR at gene promoters</u>**<u>:</u>**
Our experimental data suggests that CTCF acts as a key determinant at promoters for enhancer-promoter engagement in multiple loci (using CRISPR KO, motif disruption/mutation, or mutation knock-in as well as dCas9 based chromatin tethering), but CTCF is unlikely a universal regulator for every single promoter for its engagement with enhancer. Our results showed that more than half of promoters for cell-type-specific highly expressed genes in MCF7 cells (**Fig.2b**) have a CTCF peak at their promoters, and these CTCF sites are generally-speaking weak sites. Consistently, a recent work found that a large portion of CTCF sites in human genome are tissue- or cell-type specific, with only a small

portion (~20%) of CTCF sites being cell-type-constitutive [7]. These suggest that CTCF may be important for a portion of cell type specific gene expression via E-P looping. While such numbers of promoters with CTCF binding can be affected by the peak calling algorithm and cutoff, it is clear from our results that the CTCF binding is often weak at promoters as compared to those at TAD boundaries. There are other candidate proteins that may act via promoter binding to mediate enhancer-promoter choice. For example, ZNF143 has been shown to be a DNA binding factor at promoters to modulate E-P looping [15]. We speculate that additional factors, particularly those of the zinc finger family (to which CTCF and ZNF143 belong), may be candidates for future investigation that can play some similar or redundant roles at gene promoters for their functional engagement with enhancers.

Supplementary Figures 8, 9, 10 and notes.
Additional information about GTEx analysis.

The search for potential ERR events in the human population was conducted using data from the Genotype-Tissue Expression (GTEx) project V7 release [16] for 48 tissue types. In order to identify potential ERR events, we first searched for cis-eQTLs in promoter-proximal regions (defined as regions 2kb upstream and 1kb downstream of GENCODE annotated gene TSSs). We used cis-eQTL-gene interactions deemed significant by GTEx project's standards (gene-wise FDR<0.05). We refer to these genes that host cis-eQTLs in their promoters as 'Gene-CP' (i.e., cognate promoter), and these eQTLs are referred to as P-eQTLs (**Extended Data Fig.11a**). We then select the subset of Gene-CP promoters whose P-eQTLs also target another distal gene within the same chromosomal neighborhood in the same tissue type (this distal gene is referred to as 'Gene-AP' for alternative promoter). In particular, our methodology requires the same P-eQTL variant having opposite effects on the allelic expression of the two genes in that specific tissue type. We defined the chromosomal neighborhood in this context by setting the eQTL distance threshold to 200kb (based on an estimated ~180kb median size of chromatin contact domains [17]). Additionally, to avoid erroneous signals arising from Gene-CP promoter directly affecting the activity of Gene-AP, we maintained a minimum distance of 5kb between the TSSs of the two genes. At this stage, our analysis found 19,231 unique Gene-CP/Gene-AP pairs showing opposite correlation with P-eQTLs in Gene-CP promoter (**Supplementary Table 4a**). Out of these, 45.8% (8,799/19,231) are identified recurrently in multiple tissues, and 65.7% (12,638/19,231) of them harbor multiple P-eQTLs in the Gene-CP promoter. While we identified unique Gene-CP/Gene-AP pairs, we counted the events based on gene names rather than the numbers of P-eQTLs. For example, if "GeneA" and "GeneB" were regarded as a unique Gene-CP/Gene-AP pair, GeneA/GeneB was counted only once, even if "GeneA" promoter harbors multiple P-eQTLs. Furthermore, if "GeneA" and a second neighboring gene - "GeneC" - were identified as another unique Gene-CP/Gene-AP pair, then GeneA/GeneC will be counted separately from GeneA/GeneB pair, although they shared "GeneA" promoter P-eQTLs.

Next, we identified GTEx cis-eQTLs that overlapped with an exhaustive set of enhancer regions experimentally identified by the ENCODE [18], FANTOM5 [19], and Roadmap Epigenomics [20] projects. Out of the 19,231 Gene-CP/Gene-AP pairs that share P-eQTLs in the Gene-CP promoter, we searched for the subset that Gene-CP also possesses a cis-eQTL in its chromosomal neighborhood (-/+200kb) in the given tissue type that overlaps an enhancer region. The specific occurrence of P-eQTL/Gene-CP/Gene-AP/enhancer-cis-eQTL was inferred as a potential ERR event; and this cis-eQTL targeting Gene-CP that overlaps an enhancer was referred to as enhancer-eQTL, or E-eQTL. The potential ERR events combined across all the tissue types were found to be constituted of 872 unique Gene-CP/Gene-AP pairs across the genome. The steps described above are illustrated in the flowchart in **Extended Data Fig.11a**, and subsequent downstream analyses were carried out using R [21] and Bioconductor packages for genomic analyses [22]. Among the 872 unique pairs of genes undergoing ERR-like regulation, 61.4% (535/872) were recurrently observed in multiple tissues, and 77.1% (672/872) have multiple P-eQTLs in the Gene-CP promoter (**Extended Data Fig.11b**). P-eQTLs in the Gene-CPs of the potential ERRs were then considered for subsequent clinical/disease association analysis (**Supplementary Table 4c)**.

We examined potential ERR events from GTEx for clinical implications by looking for SNV-trait associations in the Gene-CP promoter regions. These promoter regions were queried for P-eQTL variants, associated with traits and diseases that were identified in genome wide association studies (GWAS) in the NHGRI-EBI GWAS Catalog [23] and GWASdb v2 [24]. The resulting list (**Supplementary Table 4c**) was further scrutinized for any clinically relevant cases, particularly if similar disease functions or associations exist for Gene-AP that may explain the disease risk SNPs in Gene-CP.

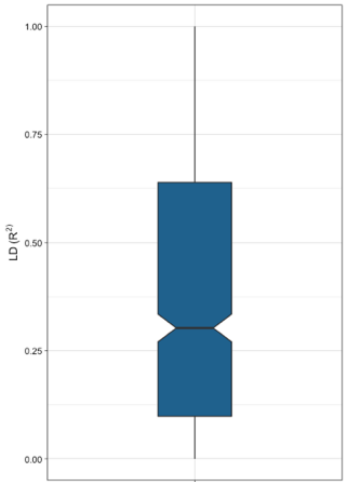
Since the sample numbers vary considerably for different tissue types in GTEx dataset, one would expect tissue types with higher numbers of samples to have higher counts of potential ERR events. We performed a linear regression analysis to analyze the relationship between the number of samples and the number of potential ERR events per tissue type. Tissues were found to display generally a proportional number of potential ERR events relative to the sample numbers. But some tissues appear to be outliers with either fewer or greater than the expected number of potential ERR events (**Fig. 4a**). In the future, as the numbers of available samples for different human tissues or diseases increase, we expect to see an increase in the number of potential ERR events.

Although we have carefully identified potential "ERR-like events" based on computational analysis using GTEx datasets as described above, we cannot exclude that other mechanisms may be involved in a certain portion of the potential ERR gene pairs that underlie their gene expression patterns. We have considered three additional potential mechanisms, 1) promoter-promoter interaction, 2) lncRNA, and 3) linkage disequilibrium.

About direct promoter-promoter interactions: Other possibilities may underlie a subset of potential ERR events that we found in our analyses of GTEx data. However, promoter-promoter interaction cannot explain the phenomenon of ERR. Instead, promoter-promoter interactions may result in genes changing in the same direction (i.e. deletion of one reduced the expression of the other, which has been observed in some loci by recent work [25, 26]). But reportedly only a small portion of human promoters (2-3% by estimation) may act to activate neighboring promoters [25]. Interestingly, our CRISPRi screening of ONPs indeed revealed that some promoters, once inhibited, will reduce the neighboring gene expression (**Fig.3c**). But this type of regulation takes place at a relatively lower prevalence (<10%, or 1-2 cases of 25 tested in **Fig.3c**) than ERR (~32%, or 8 of 25 cases tested in **Fig.3c**), broadly in agreement with the estimation by Dao et al. [25].

About SNPs in lncRNAs: There is a possibility that some promoter-homed SNPs lying in lncRNAs differentially regulate the gene adjacent by affecting lncRNA function. For the 872 identified potential ERRs, when we look at the overlap between the P-eQTLs and annotated transcripts in the genome, only a small fraction of identified ERR P-eQTLs (19/1,650) were found to overlap with lncRNA transcripts.

About Linkage Disequilibrium (LD): Our primary goal in this study is to identify potential ERR gene pairs. Therefore, the list of P-eQTLs we identified in Gene-CP does not necessarily demonstrate all of them to be causal variants, nor does it completely exclude any other nearby SNPs from contributing to the gene expression variation. Indeed, it is a unique challenge for functional genomics studies to distinguish causal variants from those that are non-

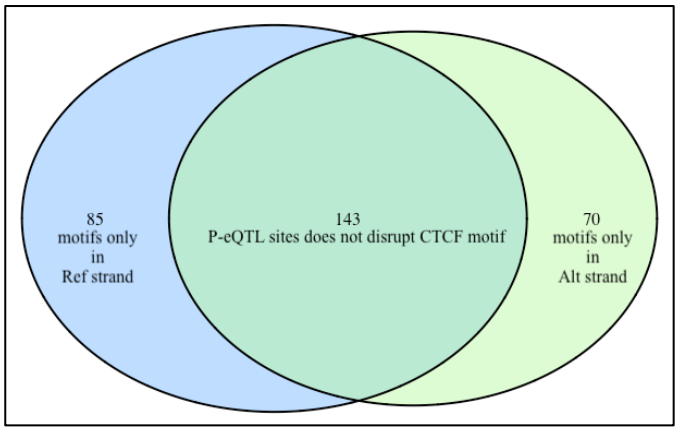


Supplementary Figure 8: Distribution of linkage disequilibrium (LD) between promoter-eQTLs and enhancer-eQTLs defined in our ERR analysis.

Boxplot showing distribution of R^2 values between P-eQTLs and the most significant E-eQTL for the set of all unique ERRs detected in GTEx tissues. With the median R^2 at 0.3, most of the P-eQTL and E-eQTL SNPs are not in LD, suggesting they independently regulate Gene-CP and Gene-AP. The boxplot centerline represents median; box limits indicate the 25th and 75th percentiles; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles.

functional, because many SNPs are located in strong Linkage Disequilibrium (LD). In particular, common variants with high allele frequencies may display strong LDs with neighboring SNPs [27]. Some computational methods have prioritized potentially causal variants based on overlap with regulatory elements with active histone marks or chromatin opening, e.g., [28, 29]. We examined the LD relationship between our P-eQTLs and E-eQTLs for their contribution to the expression of Gene-CP. We found that the P-eQTLs and E-eQTLs involved in our final list of potential ERRs are rarely found in the same LD (**Supplementary Figure 8**), and therefore largely contributed to the gene expression of Gene-CP independently. This analysis suggested that while we cannot completely exclude the contribution of some of the eQTLs in high LD with P-eQTLs to the gene expression variation of Gene-CP, it is likely to be infrequent. In the rare cases that P-eQTLs and E-eQTLs are in a high LD domain, they may play biologically synergistic or redundant roles.

About SNPs overlapping CTCF motifs: Most GWAS SNPs are not predicted to impact Transcription Factor (TF) binding. Indeed, even by combining DNA motifs for a collection of human TFs, only a small fraction of causal SNPs associated with immune diseases can be interpreted as motif-disrupting (e.g. ~10-20% estimated by Farh et al., [28]). It is therefore unlikely that ERR events are universally caused by altered CTCF motif/binding at Gene-CP promoters. We analyzed the SNPs defined as P-eQTLs in our analysis of potential ERRs, using a similar MEME/FIMO method as those used for calculating CTCF motif changes by cancer somatic mutations. We found that out of the 1,650 P-eQTLs, 228 SNPs (~13-14%) are located in CTCF core motifs. Based on the reference genome, about 37% of these CTCF-located P-eQTL SNPs (85/228) are predicted to disrupt the CTCF motif by FIMO/MEME in between the reference or alternative alleles (**Supplementary Figure 9**).



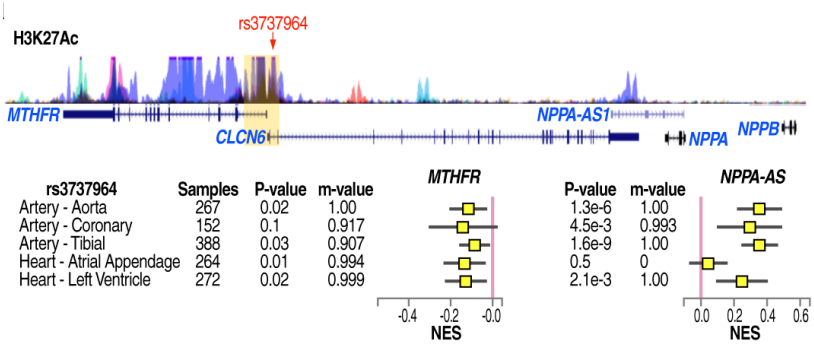
Supplementary Figure 9: A portion of ERR P-eQTLs may disrupt CTCF motifs.

Venn Diagram showing that out of 1,650 P-eQTLs we identified to locate in potential ERR Gene-CP promoters, a ~13-14% of them overlap a predicted CTCF motif at either Watson or Crick strand. Out of these, about 1/3 can in silico alter the CTCF motif as predicted by FIMO considering the reference (Ref) or alternative alleles (Alt).

Overall, we consider that defective CTCF binding due to direct motif change (also see **Supplementary Figure 7a**) may explain a portion of, but surely not all, ERR-like events. Indeed, in the example of *NUCKS1-RAB7L1* case we studied in depth in **Fig.4**, the three SNPs are not directly located in core CTCF motifs and are not predicted to disrupt CTCF motifs, but they locate close to ChIP-Seq peaks of CTCF (**Extended Data Fig.11e**). Importantly, our experimental evidence indicated that the alternative allele bears lower CTCF binding than the reference allele at the promoter *of NUCKS1*, which correlated with their defective engagement with the active enhancer (**Fig.4d,f**). These results support the role of CTCF in promoter engagement with an active enhancer for a portion of the potential ERR pairs we identified, but it is unlikely the universal regulator.

Additional ERR candidate loci associated with genetic disease risk: In addition to *NUCKS1-RAB7L1* locus shown in the main figures, additional Gene-CP/Gene-AP pair identified by our analyses include *MTHFR*/*NPPA-AS1*. ERR may provide new insights into understanding a prominent cardiovascular disease risk SNP rs3737964 [30] in the promoter of *MTHFR*. This SNP correlates with lower expression of *MTHFR* but higher expression of *NPPA-AS1* (**Supplementary Figure 10**). Notably, *NPPA-AS1*

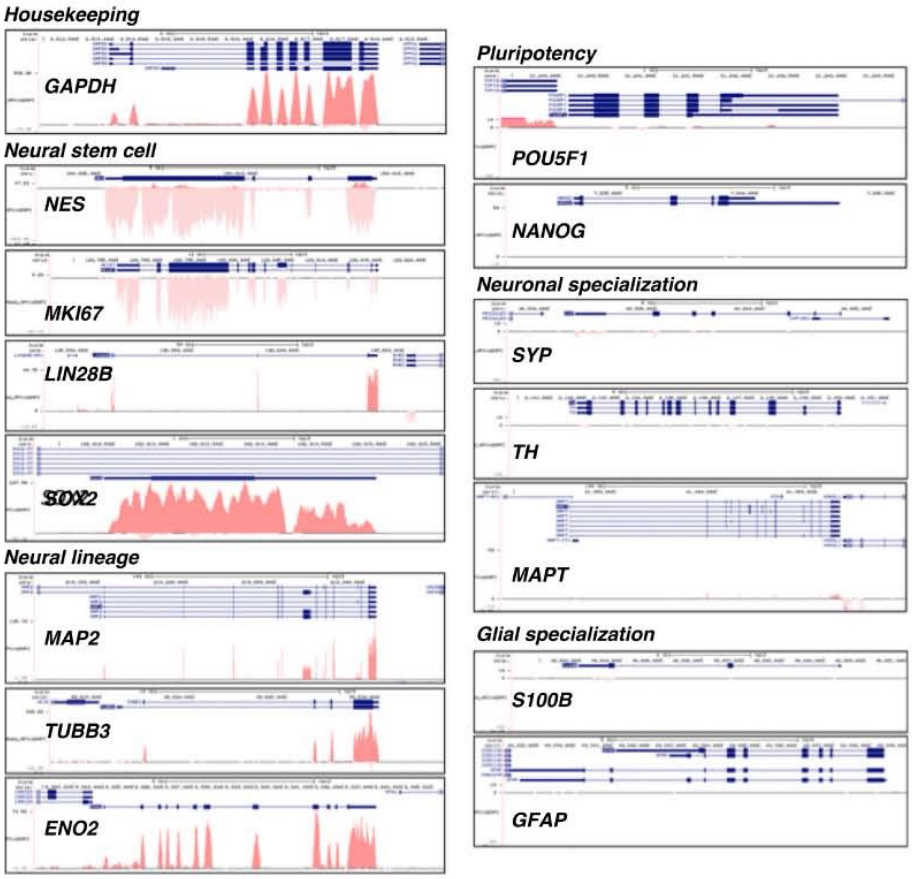
encodes a long non-coding RNA that acts to maintain cardiovascular and metabolic homeostasis by post-transcriptionally regulating the expression of *NPPA* [31]. Therefore, while locating far away from *NPPA-AS1*, rs3737964 may play a role in disease risk by increasing *NPPA-AS1* expression via ERR. We expect that additional loci that do not show genome-wide significance in GWAS studies may also work via ERR to modulate their neighboring gene expression and contribute to disease risk.



Supplementary Figure 10:
ERR-like regulation at the MTHFR - NPPA locus. The hypertension risk SNP in the MTHFR promoter is indicated (top). Box plots showing normalized effect size (NES) of MTHFR or NPPA-AS1 expression segregated by a hypertension risk SNP rs3737964 in relevant tissues such as artery and heart tissues (bottom). The definitions of NES, p-value and m-value could be found in GTEx portal (www.gtexportal.org). We directly quote the definition here. For m-value, it denotes the posterior probability that an eQTL effect exists in each tissue tested in the cross-tissue meta-analysis; the m value ranges in between 0 and 1. For p-value, it was generated by a t-test that compares the observed NES from single-tissue eQTL analysis to a null NES of 0. The normalized effect size (NES) was used to denote the slope of the linear regression of normalized expression data versus the three genotype categories using single-tissue eQTL analysis, representing eQTL effect size. The normalized expression values are based on quantile normalization within each tissue, followed by inverse quantile normalization for each gene across samples.

Supplementary Figure 11.
Validation of iPSC-derived neural progenitor cells (NPCs) by RNA-Seq.

We confirmed the identity of iPSC derived neural progenitor cells (NPCs) by RNA-seq (**Supplementary Figure 11**). As expected, we did not observe expression of pluripotency markers *NANOG* and *POU5F1* (Oct4); but we can see strong expression of markers of (neural) stem cell identity, *NES* (Nestin), *MKI67* (Ki67), *LIN28B*, and *SOX*2. We observed robust expression of neuronal-lineage markers *TUBB3* (Tuj1), *ENO2* (NSE), and *MAP2*; but non or relatively low expression of specialized genes of neuronal identity, such as *MAPT*, *SYP*, or *TH*; or glial differentiation, such as *GFAP* and *S100B*. Together, these results corroborate that our cells are of NPC identity.



Supplementary Figure 11. RNA-seq data shown as snapshots of the UCSC genome browser track (hg19), the pink color indicates expression signals of RNA-seq of the housekeeping gene (*GAPDH*), Neural stem cell specific genes (*NES, MKI67, SOX2*), Neural lineage specific genes (*MAP2, TUBB3, ENO2*), Pluripotency specific gene (*POU5F1, NANOG*), Neuronal specific genes (*SYP, TH, MAPT*), Glial specific genes (*S100B, GFAP*).

References for Supplementary Information:

1. Chang, M.T., et al., *Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity.* Nat Biotechnol, 2016. **34**(2): p. 155-63.
2. Horn, S., et al., *TERT promoter mutations in familial and sporadic melanoma.* Science, 2013. **339**(6122): p. 959-61.
3. Huang, F.W., et al., *Highly recurrent TERT promoter mutations in human melanoma.* Science, 2013. **339**(6122): p. 957-9.
4. Rheinbay, E., et al., *Analyses of non-coding somatic drivers in 2,658 cancer whole genomes.* Nature, 2020. **578**(7793): p. 102-111.
5. Alipanahi, B., et al., *Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning.* Nat Biotechnol, 2015. **33**(8): p. 831-8.
6. Ghirlando, R. and G. Felsenfeld, *CTCF: making the right connections.* Genes Dev, 2016. **30**(8): p. 881-91.
7. Maurano, M.T., et al., *Role of DNA Methylation in Modulating Transcription Factor Occupancy.* Cell Rep, 2015. **12**(7): p. 1184-95.
8. Wen, Z., et al., *Histone variant H2A.Z regulates nucleosome unwrapping and CTCF binding in mouse ES cells.* Nucleic Acids Res, 2020.
9. Tarjan, D.R., W.A. Flavahan, and B.E. Bernstein, *Epigenome editing strategies for the functional annotation of CTCF insulators.* Nat Commun, 2019. **10**(1): p. 4258.
10. Del Rosario, B.C., et al., *Exploration of CTCF post-translation modifications uncovers Serine-224 phosphorylation by PLK1 at pericentric regions during the G2/M transition.* Elife, 2019. **8**.
11. Yu, W., et al., *Poly(ADP-ribosyl)ation regulates CTCF-dependent chromatin insulation.* Nat Genet, 2004. **36**(10): p. 1105-10.
12. Luo, H., et al., *LATS kinase-mediated CTCF phosphorylation and selective loss of genomic binding.* Sci Adv, 2020. **6**(8): p. eaaw4651.
13. Kim, E., et al., *Systematic Functional Interrogation of Rare Cancer Variants Identifies Oncogenic Alleles.* Cancer Discov, 2016. **6**(7): p. 714-26.
14. Consortium, I.T.P.-C.A.o.W.G., et al., *Pan-cancer analysis of whole genomes.* Nature, 2020. **578**(7793): p. 82-93.
15. Bailey, S.D., et al., *ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters.* Nat Commun, 2015. **2**: p. 6186.
16. Consortium, G.T., et al., *Genetic effects on gene expression across human tissues.* Nature, 2017. **550**(7675): p. 204-213.
17. Rao, S.S.P.H., Miriam H.; Durand, Neva C. ; Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, Ido Machol, Arina D. Omer, Eric S. Lander, Erez Lieberman Aiden, *A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping.* Cell, 2014.
18. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
19. Consortium, F., et al., *A promoter-level mammalian expression atlas.* Nature, 2014. **507**(7493): p. 462-70.
20. Roadmap Epigenomics, C., et al., *Integrative analysis of 111 reference human epigenomes.* Nature, 2015. **518**(7539): p. 317-30.
21. Team, R.C., *R: A language and environment for statistical computing.* , in. *R Foundation for Statistical Computing, Vienna, Austria. URL* https://www.R-project.org/. 2017.
22. Huber, W., et al., *Orchestrating high-throughput genomic analysis with Bioconductor.* Nat Methods, 2015. **12**(2): p. 115-21.

23. MacArthur, J., et al., *The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog).* Nucleic Acids Res, 2017. **45**(D1): p. D896-D901.
24. Li, M.J., et al., *GWASdb v2: an update database for human genetic variants identified by genome-wide association studies.* Nucleic Acids Res, 2016. **44**(D1): p. D869-76.
25. Dao, L.T.M., et al., *Genome-wide characterization of mammalian promoters with distal enhancer functions.* Nat Genet, 2017. **49**(7): p. 1073-1081.
26. Jung, I., et al., *A compendium of promoter-centered long-range chromatin interactions in the human genome.* Nat Genet, 2019. **51**(10): p. 1442-1449.
27. Khurana, E., et al., *Role of non-coding sequence variants in cancer.* Nat Rev Genet, 2016. **17**(2): p. 93-108.
28. Farh, K.K., et al., *Genetic and epigenetic fine mapping of causal autoimmune disease variants.* Nature, 2015. **518**(7539): p. 337-43.
29. Gjoneska, E., et al., *Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease.* Nature, 2015. **518**(7539): p. 365-9.
30. Del Greco, M.F., et al., *Genome-wide association analysis and fine mapping of NT-proBNP level provide novel insight into the role of the MTHFR-CLCN6-NPPA-NPPB gene cluster.* Hum Mol Genet, 2011. **20**(8): p. 1660-71.
31. Annilo, T., K. Kepp, and M. Laan, *Natural antisense transcript of natriuretic peptide precursor A (NPPA): structural organization and modulation of NPPA expression.* BMC Mol Biol, 2009. **10**: p. 81.