





Context transcription factors establish cooperative environments and mediate enhancer communication

Received: 5 May 2023

Accepted: 1 August 2024

Published online: 3 October 2024

 Check for updates

Judith F. Kribelbauer-Swietek ^{1,2}✉, Olga Pushkarev^{1,2}, Vincent Gardeux ^{1,2}, Katerina Faltejskova^{3,4}, Julie Russeil¹, Guido van Mierlo ^{1,2} & Bart Deplancke ^{1,2}✉

Many enhancers control gene expression by assembling regulatory factor clusters, also referred to as condensates. This process is vital for facilitating enhancer communication and establishing cellular identity. However, how DNA sequence and transcription factor (TF) binding instruct the formation of high regulatory factor environments remains poorly understood. Here we developed a new approach leveraging enhancer-centric chromatin accessibility quantitative trait loci (caQTLs) to nominate regulatory factor clusters genome-wide. By analyzing TF-binding signatures within the context of caQTLs and comparing episomal versus endogenous enhancer activities, we discovered a class of regulators, ‘context-only’ TFs, that amplify the activity of cell type-specific caQTL-binding TFs, that is, ‘context-initiator’ TFs. Similar to super-enhancers, enhancers enriched for context-only TF-binding sites display high coactivator binding and sensitivity to bromodomain-inhibiting molecules. We further show that binding sites for context-only and context-initiator TFs underlie enhancer coordination, providing a mechanistic rationale for how a loose TF syntax confers regulatory specificity.

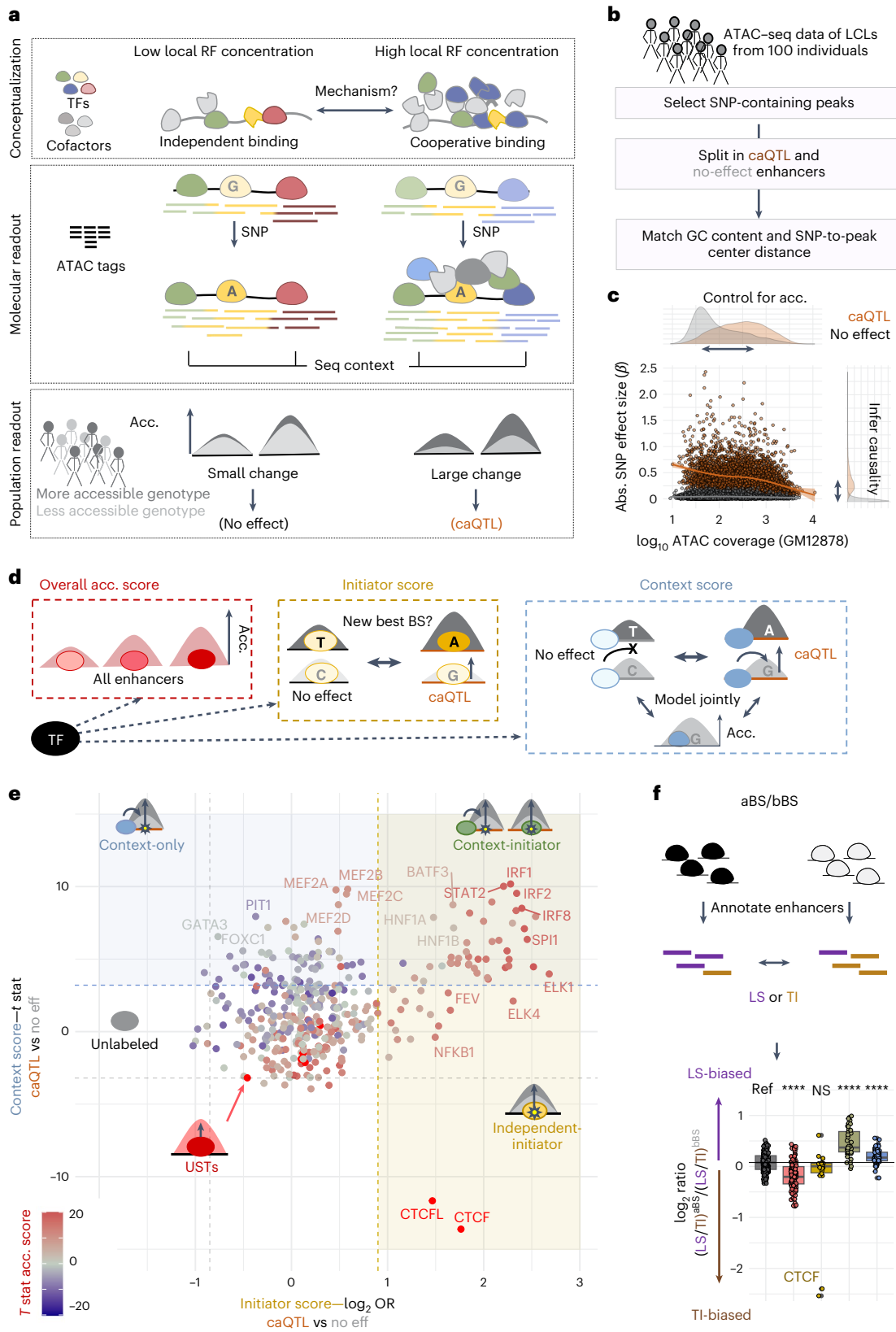
An important step in gene regulation is the recruitment of sequence-specific transcription factors (TFs) to enhancers. While binding specificities for the majority of human TFs¹ are readily available^{2–5}, motif matches alone are poor predictors for *in vivo* binding⁶. Context-dependent factors, including sensitivity to DNA modifications⁷, recruitment of transcriptional coregulators⁸ and TF complex formation⁹, among others, all contribute to TF occupancy¹⁰.

To account for genomic context, deep neural network models^{11,12} have been designed to detect dependencies between motif instances and regulatory elements that are dozens of kilobases apart^{13–15}. Although such models accurately predict TF binding,

chromatin states or DNA contact frequencies, even the most advanced ones do not generalize well to unseen contexts¹³, do not capture enhancer-based gene regulation particularly well^{16,17} or they require additional experimental inputs on top of DNA sequence^{14,15}. Progress is hampered by the difficulty of distilling human interpretable mechanisms from existing models. The emerging consensus is that context specificity is driven by ‘soft’ motif syntax rules^{11,12} with unclear molecular underpinnings. Soft binding site (BS) syntax may be reflective of the quantitative nature of TF binding¹⁸, where occupancy is determined by the sequence specificity of a TF and its local concentration^{19–22}.

¹Laboratory of Systems Biology and Genetics, Institute of Bioengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. ²Swiss Institute of Bioinformatics, Lausanne, Switzerland. ³Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, Prague, Czech Republic. ⁴Computer Science Institute, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic.

✉e-mail: judith.kribelbauer@epfl.ch; bart.deplancke@epfl.ch



Imaging experiments have indeed shown that TFs segregate into nuclear hubs with high localized concentration^{19,23–26} (see ref. 27 for a recent review). These assemblies are referred to as condensates or regulatory factor clusters, among others, and incorporate various coactivators^{28,29}, for example, bromodomain-containing proteins (BRDs). Regulatory factor clusters have been shown to form at certain

enhancers^{23–25}, in particular super-enhancers²⁸ that are defined by high levels of clustered activity^{30,31}. Yet, the exact assembly mechanism remains poorly understood³². Contributing factors include the communication with the epigenome, for example, binding of BRD4 to H3K27 acetyl lysine residues³³, and weak protein–protein interactions among intrinsically disordered domains present in both cofactors and

Fig. 1 | Enhancers rely on distinct classes of TFs to create caQTL environments. **a**, Schematic representation describing TF binding under different regulatory factor (RF) concentrations, with low concentrations as a result of independent binding and high concentrations of cooperative binding. Context-dependent differences in molecular readouts when SNPs impact TF binding are highlighted both at the level of individual ATAC-seq tags (middle) and enhancer-wide accessibility (acc.) changes at the population level (bottom). **b**, Process to prioritize SNPs based on their impact on enhancer-wide accessibility in *cis*. **c**, Comparison of absolute (abs.) SNP effect sizes against baseline enhancer accessibility in GM12878 reveals a bias of caQTL enhancers toward higher baseline accessibility. **d**, Schematic representation describing three ways of assessing TF function based on caQTL and DNA accessibility information. First, relating top BS strength to general peak accessibility (Methods; accessibility score, log-linear model, left). Second, comparing caQTLs to no-effect SNPs in terms of how frequently SNPs create a new best BS (Methods; initiator score, Fisher's exact test, middle) and third, relating BS strength in the sequence context of SNPs to the SNP effect-size status while controlling for

baseline peak accessibility (Methods; context score, log-linear model, right). **e**, TF context score (*y* axis, *t*-statistic) versus initiator score (*x* axis, log₂ odds ratio). Each point represents a specific TF motif. Accessibility scores are overlaid with a blue–red color gradient. Colored boxes indicate the following three of the four TF classes: context-only (blue), context-initiator (green) and independent-initiator (yellow). High accessibility scores indicate USTs. Dotted lines represent thresholds used to define classes. **f**, TFBS specificity toward lymphoid-specific (LS) and tissue invariant (TI) enhancers using the DHS index. LS/TI ratios are compared across enhancers (*n* = 21,752) split by above- or below-average top BS strength (aBS and bBS) for a given TF. The *y* axis shows the log₂-transformed ratio of ratios (LI/TI in aBS versus bBS; Mann–Whitney test, *P* = 2.2×10^{-15} , *P* = 0.11, *P* = 1.1×10^{-14} and *P* = 1.2×10^{-5} for the comparisons of unlabeled motifs (*n* = 151, gray) to USTs (*n* = 120, red), independent-initiator (*n* = 18, yellow), context-initiator (*n* = 44, green) and context-only (*n* = 68, blue) motifs. Box plots denote the 25th percentile, median and 75th percentile. NS, not significant (*P* > 0.05). *****P* < 0.0001.

TFs^{29,34,35}. Finally, TF binding to DNA itself can catalyze condensate formation^{36,37} and localization³⁸. Although there are several well-studied examples underscoring the importance of regulatory factor cluster formation and size control for gene regulation^{27,33,39,40}, it has proven difficult to predict *ab initio* which enhancers give rise to regulatory factor clusters. The main reason for this is a lack of data linking condensates to enhancers genome-wide.

To address this limitation, we considered the biophysical properties underlying condensate formation—condensates rely on cooperativity among regulatory factors, which leads to a nonlinear behavior when their concentration reaches a critical level^{37,40,41}. The transition to a high-concentration state occurs abruptly, meaning that a relatively small event (for example, the addition of an extra molecule) can trigger large changes in subsequent regulatory factor recruitment^{39,40}. In the context of enhancers, such a ‘seed’ event may occur when the binding of one additional TF leads to a large change in overall regulatory factor recruitment. Such changes can be observed at enhancer-centric chromatin accessibility quantitative trait loci (caQTLs)^{42–44} that are driven by single-nucleotide polymorphisms (SNPs), as SNPs create at most one new TFBS. Identifying such caQTL enhancers may thus provide a means to nominate enhancer-linked regulatory factor clusters genome-wide.

caQTL mapping also allows uncovering regulatory element communication^{44,45}, which may reflect the merging of two or more compatible regulatory factor clusters⁴¹. Regulatory element communication is present whenever the epigenomic signatures of two or more regulatory elements covary in the function of the genotype of one regulatory element-centric variant. An example thereof is the *AXIN2* locus, where the creation of a single TFBS within a transcription start site (TSS)-proximal enhancer leads to a coordinated activation of multiple regulatory elements⁴⁶. We refer to such coordinated regulatory elements as ‘chromatin modules’ (CMs), yet the underlying mechanism of their formation remains poorly understood⁴⁷.

Here, by leveraging available data on caQTLs and CMs in lymphoblastoid cells⁴⁵, we identify a class of regulators, ‘context-only’ TFs, whose dedicated function appears linked to the establishment of cooperative environments. Although context-only TFs are not associated with DNA accessibility directly, their BSs occur alongside those of cell type-specific initiator TFs, with whom they appear to engage using a loose motif syntax. In a series of computational and experimental validation analyses, we provide support for the notion that the combined binding and function of both TF classes is associated with the formation of regulatory factor clusters at a majority of caQTL enhancers, as well as with enhancer coordination.

Results

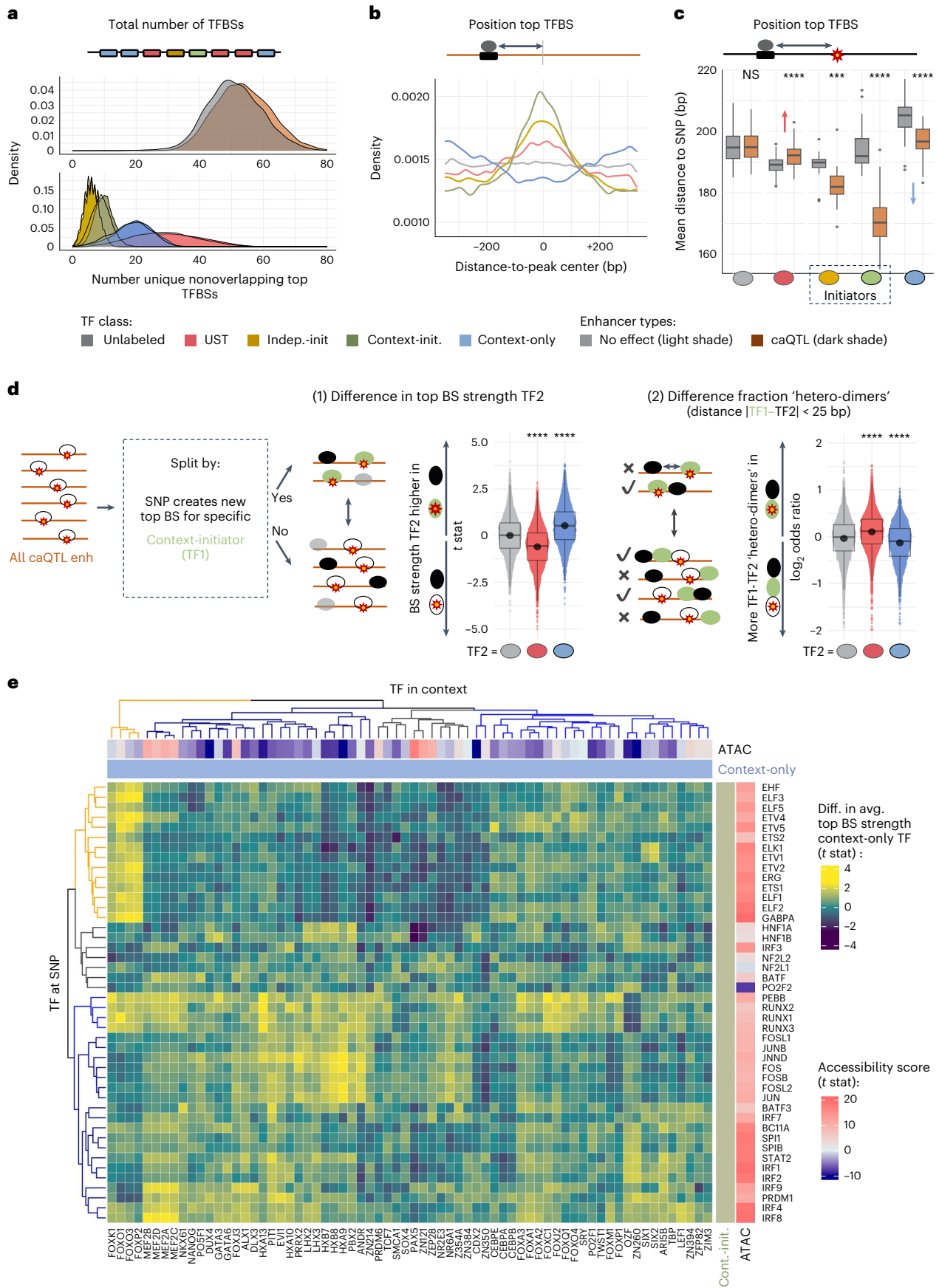
Uncovering TF signatures that create caQTL environments

We used existing, fine-mapped⁴³ caQTL data from 100 human lymphoblastoid cell lines (LCLs) profiled with an assay for transposase-accessible chromatin with sequencing (ATAC-seq)⁴⁵ to nominate putative regulatory factor clusters at enhancers that significantly change peak accessibility due to an underlying SNP (Fig. 1a, Supplementary Fig. 1a,b and Methods). We refer to these regulatory elements as ‘caQTL enhancers’. As a control, we chose enhancers that contain SNPs without significant accessibility changes (‘no-effect’ enhancers) (Fig. 1a). To minimize confounders, we defined peaks of equal length with matched GC content and SNP-to-peak-center distances, creating a total of 50 matched control groups (Methods, Fig. 1b and Supplementary Fig. 1c). This resulted in sets of roughly 9,500 enhancers equally split into caQTL and no-effect enhancers. Because count statistics are used to infer caQTLs, caQTL enhancers tend to have higher accessibility (Supplementary Methods and Fig. 1c). To infer TFs associated with caQTL environments while accounting for this accessibility difference, we assessed the following three separate aspects of TF function derived from the top TFBS scores within enhancers (motifs from HOCOMOCO³; Methods and Fig. 1d): (1) the

Fig. 2 | Assessment of combinatorial TFBS syntax in caQTL enhancers.

a, Distribution of the number of nonoverlapping, above-average top TFBSs across caQTL (*n* = 4,747, orange or dark shade) and no-effect (*n* = 4,747, gray or light shade) enhancers. Top, the total number of sites per enhancer. Bottom, the number of sites split by TF class. **b**, Density of top TFBS positions split by TF class with respect to the ATAC-seq peak center. **c**, Distribution of average distances of top TFBSs to caQTL (orange) or no-effect (gray) SNPs split by TF class. Initiator TF classes are labeled with a dashed box. Context-only TFs move closer (blue arrow) in caQTL enhancers, while USTs move farther away (red arrow). Differences in distances are assessed using a Mann–Whitney test; *P* = 0.76, *P* = 1.1×10^{-10} , *P* = 3.3×10^{-4} , *P* < 2.2×10^{-16} and *P* = 5.4×10^{-13} . **d**, Assessing combinatorial behavior across TF classes with respect to context-initiator TFs. For each context-initiator TF (TF1, green color), caQTL enhancers are split by whether the SNP creates

a new best BS or not. (1) Middle, comparing the top TFBS strength of either unlabeled, UST or context-only TFs (TF2) between the two outlined enhancer groups (Student's *t*-test). Shown is the distribution of resulting *t*-statistics across all possible pairs grouped by the TF2 class. (2) Right, test for the formation of canonical TF heterodimers (TF1–TF2 top TFBS distance <25 bp). Shown are the odds ratios of forming a heterodimer when the caQTL SNP creates a new best BS for the context-initiator TF1 across all possible pairs grouped by the TF2 class. Significance between classes for (1) and (2) is assessed with a Student's *t*-test; all *P* values < 2.2×10^{-16} . **e**, Hierarchical clustering of the *t*-statistics for the TFBS affinity enrichment test as indicated in **d**, (1) for individual context-initiator and context-only motif combinations. In **c** and **d**, box plots denote the 25th percentile, median and 75th percentile, with whiskers extending to 1.5× the interquartile range. ****P* < 0.001, *****P* < 0.0001.



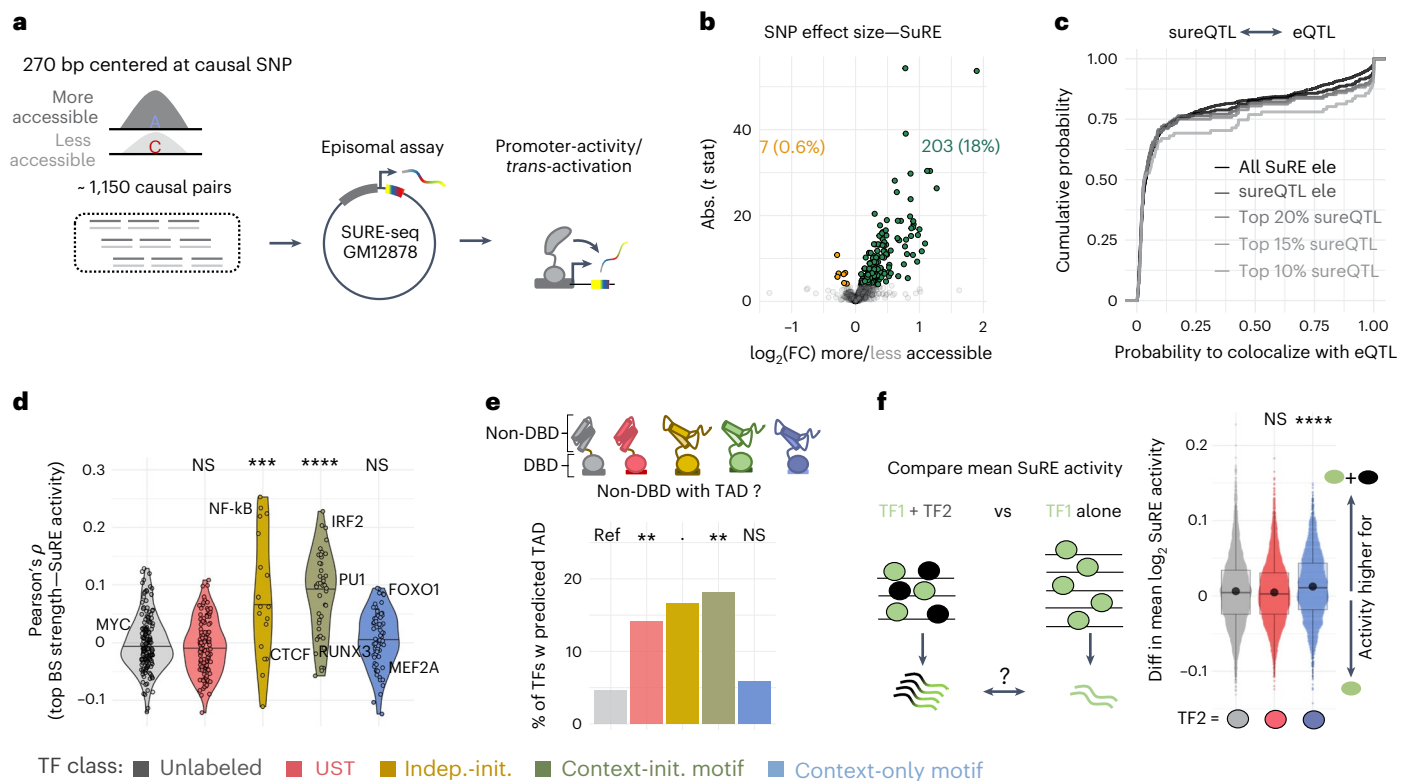


Fig. 3 | Context-only TFs boost transcriptional activity but do not initiate it. **a**, In total, more than 1,000 caQTL fragments in both genotypes were tested upstream of a TSS using the SuRE technology (autonomous promoter activity). **b**, Volcano plot of SNP effect sizes for SuRE activities (Student's t -test comparing less and more accessible genotypes). The y axis reflects the absolute t -statistic for each SNP, and the x axis reflects the $\log_2(\text{FC})$ in SuRE activity. SNPs with significant changes (sureQTLs, adjusted $P < 0.05$) are shown in different colors (green and orange for effect-size directions concordant or opposite to that of the caQTL). **c**, Empirical cumulative distribution of the probability of caQTL–eQTL colocalization. Gray lines represent different subsets of SNPs stratified by their sureQTL effect size. The black line reflects the baseline of all tested elements. **d**, Correlation of fragment-level top TFBS strength with SuRE activity (Methods) grouped by TF class. Significance is computed with respect to unlabeled TFs (Mann–Whitney test, $P = 0.65$, $P = 5.1 \times 10^{-3}$, $P = 2.2 \times 10^{-11}$ and $P = 0.17$). Violin plots

denote the medians and encompass the full range of data points. **e**, Percentage of TFs with a predicted TAD. TFs split by TF class. Differences are assessed relative to unlabeled TFs (Fisher's exact test, $P = 8.8 \times 10^{-3}$, $P = 0.08$, $P = 6.8 \times 10^{-3}$ and $P = 0.74$). **f**, SuRE activity of caQTL fragments in their less accessible genotype is compared between fragments with above-average top BSs for only a context-initiator TF (TF1), or for both a context-initiator and a second TF from either unlabeled, UST or context-only TF classes (TF2; Student's t -test). Boxplot shows the difference in mean \log_2 -transformed SuRE activity (y axis) between the two splits across all possible pairs of TF1 (context initiator) and TF2 grouped by the class of TF2 (x axis). Significance is computed with respect to unlabeled TFs (Mann–Whitney test, $P = 0.13$ and $P = 2.4 \times 10^{-12}$). Values are shown as means (dots) with box plots denoting the median and the 25th and 75th percentiles, with whiskers extending to 1.5× the interquartile range. ** $P < 0.01$, *** $P < 0.001$ and **** $P < 0.0001$.

overall association of the with DNA accessibility (accessibility score), (2) the ability of the TF to initiate an increase in accessibility at the caQTL SNP itself (initiator score, TFBS enrichment at the focal SNP) and (3) the importance of the TF in the sequence context of the caQTL SNP (context score, TFBS enrichment in caQTL enhancer contexts) (Supplementary Fig. 1d). The context score interrogates whether a group of TFs distinct from initiators contributes to the putative cooperative environment at caQTL enhancers. We also assessed the ability of TFs to act as repressors; however, we found little evidence for repressive mechanisms (Supplementary Fig. 1e).

Leveraging caQTL environments to classify TFs

Comparing TFs based on the three scores (Fig. 1e and Supplementary Fig. 1f), we defined distinct groupings with different associations with caQTL environments. Focusing on TFs with high initiator scores, we found the majority were also identified in the original caQTL study⁴⁵, including PU.1 (encoded by *SPI1*), most interferon-regulatory factors (IRFs), NF- κ B and CTCF. As expected, most initiators were predicted to drive overall accessibility (92%) with the majority of them ($n = 44$; 71%) also having high context scores, meaning that their BSs are generally enriched in caQTL enhancers. We refer to this subset of SNP-linked

TFs as 'context initiators'. The remaining 18 initiators (29%) likely reflect context-independent SNP binding of particularly potent TFs, hence the label 'independent-initiator'. A good example is CTCF, whose BS is not generally enriched in caQTL enhancers, but when an SNP creates a new CTCF BS, large accessibility changes are observed. Interestingly, we found numerous TFs that relate to overall enhancer accessibility, yet their BSs are not enriched at caQTL SNPs nor in their sequence context (Fig. 1e). These TFs overlapped significantly ($P = 1.43 \times 10^{-10}$, Fisher's exact test) with the previously identified universal stripe TFs (USTs)⁴⁸ that provide accessibility in a cell type-agnostic manner (Supplementary Fig 1g). For simplicity, we collectively refer to this class of TFs as USTs. Finally, we identified a class of TFs, including the MADS-box MEF2 and several FOX motifs, that appeared to lack the ability to initiate, but whose BSs are nonetheless enriched in the context of caQTL enhancers. The vast majority of them ($n = 58$; 85%) were not associated with enhancer accessibility, suggesting that they require context-initiator TFs for BS access and contribute to caQTL contexts through independent, currently unknown mechanisms. We thus labeled them context-only TFs. Finally, we refer to neutral TFs (that is, not significant for any of the three scores) as 'unlabeled'.

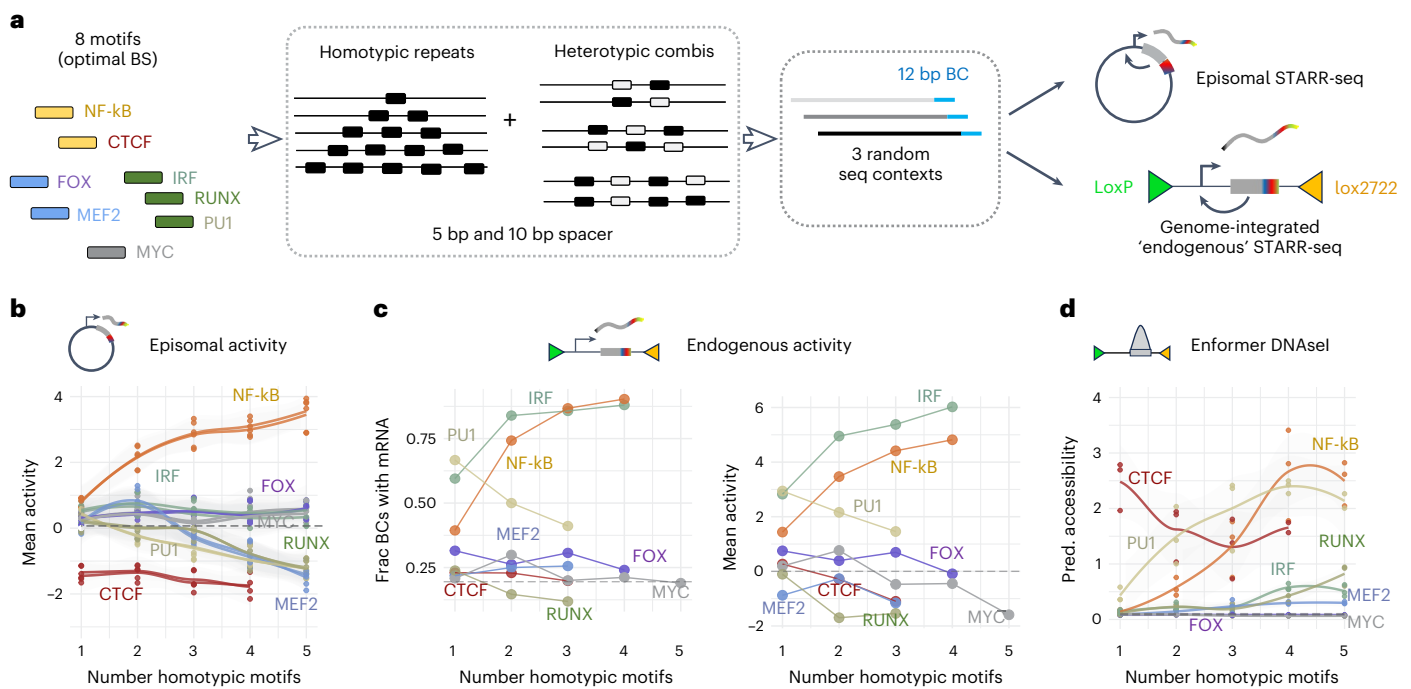


Fig. 4 | Probing activity of TF motifs in episomal and endogenous contexts.

a, Episomal and genome-integrated STARR-seq design. **b**, Average \log_2 episomal STARR-seq activity normalized by the 'no-TFBS' fragment (y axis) for homotypic motif repeats (5 bp spacer). Dots represent the three sequence contexts, and lines represent fits for each replicate. Colors reflect TFs and motif classes (compare to **a** and Fig. 1e). Gray dashed line indicates the no-TFBS reference. **c**, Two measures of

endogenous STARR-seq activity for homotypic motif combinations. Left, fraction of expressed BCs per fragment. Replicates and three sequence contexts are combined. The dashed line indicates the no-TFBS reference. Right, average \log_2 -transformed endogenous STARR-seq activity normalized by the no-TFBS activity. **d**, Enformer-predicted DNA accessibility for GM12878 for homotypic motif combinations. Individual dots represent the three different sequence contexts.

Because condensate formation has been shown to regulate cell identity genes^{31,49}, we first assessed whether BSs of a particular TF class are enriched in cell type-specific regulatory elements as annotated by the regulatory index of DNase I hypersensitive sites⁵⁰ (Methods). Comparing the ratio of above-average top TFBSs in lymphoid-specific versus tissue-invariant elements, we found that, although USTs and independent-initiator BSs favor tissue-invariant enhancers, BSs for context-initiator, and to a lesser degree context-only TFs, are enriched in lymphoid-specific enhancers (Fig. 1f).

Context-only TFs promiscuously pair with context-initiator TFs

Context-only TFs may contribute to the creation of caQTLs by enhancing the function of cell type-specific context-initiator TFs either through canonical cooperativity (an SNP-centered initiator and a context-only TF bind as a heterodimer) or by contributing to the creation of a high-concentration environment independent of a fixed BS syntax (for example, by providing stickiness for other regulatory factors). To address these possibilities, we first validated that the top TFBSs represent bona fide motifs and that there is no systematic bias in motif similarity across TF classes (Methods and Supplementary Fig. 2a,b). Next, we assessed intra-enhancer regulatory complexity by computing the number of nonoverlapping, above-average top TFBSs, finding an average of ~50 sites, of which about 20 are reserved for context-only TFs, and with an average one to two BS difference between caQTL and no-effect enhancers (Fig. 2a). Honing in on TF syntax, we analyzed the position of the top TFBSs with respect to either the peak center or the SNP itself. As expected for accessibility drivers, BSs for USTs and initiator TFs concentrate around the peak center. In contrast, context-only TFs were enriched in peak shoulders (Fig. 2b). Moreover, while the median distance to the SNP was largest for context-only TFBSs (Fig. 2c), they moved significantly closer to the SNP in caQTL versus no-effect enhancers (Fig. 2c), with UST BSs showing the opposite

trend. To investigate whether the narrowing distance might reflect the formation of heterodimers between context-only and context-initiator TFs, we split caQTL enhancers into two groups based on whether or not an SNP created a new best BS for the context-initiator TF. Next, we compared (1) the difference in BS affinity for TFs other than the context initiator and (2) the distance between context-initiator and noncontext-initiator motifs (Methods). In contrast to unlabeled TFs and USTs, we found an overall enrichment for higher affinity BSs for context-only TFs when specific context-initiator BSs are impacted by the SNP (Fig. 2d). However, context-only TFBSs were less likely to be within close proximity to the context-initiator TFBS compared to both unlabeled and UST TFBSs (Fig. 2d), thus arguing against the formation of TF heterodimers.

Plotting the BS affinity enrichment for individual pairs of context-initiator and context-only TFs (Fig. 2e), we observed a compatibility structure, albeit weak, that is not detected when pairing context-initiator or context-only TFs with USTs (Supplementary Fig. 3a,b). However, most context initiators paired with multiple context-only TFs without requiring close TFBS proximity (Supplementary Fig. 4), providing further support that the co-enrichment of the two context-TF classes is not due to canonical binding cooperativity. Even the combination of ETS factors with high-affinity FOXO TFBSs lacked enrichment for a specific syntax, despite a previously identified preferred BS spacing⁵¹ (Supplementary Fig. 4).

Context-only TFs are not associated with transcriptional activity

To investigate how context-only TFs may contribute to the creation of caQTL environments in the absence of a strict BS syntax, we assessed their ability to drive transcriptional activity. For this, we tested over a thousand caQTL enhancers (both more and less accessible genotypes; Methods and Supplementary Data 1) in a SuRE-seq (survey of regulatory

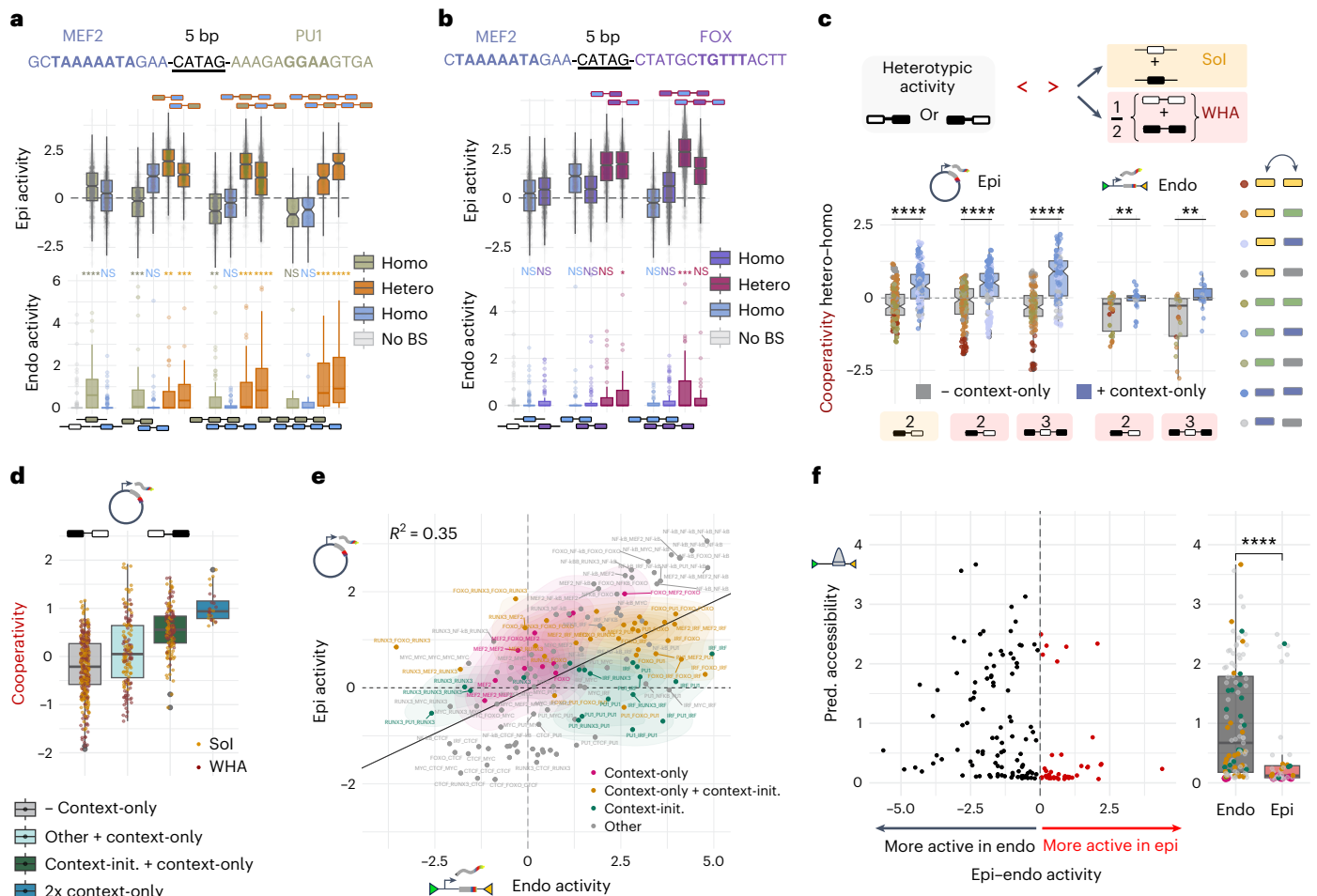


Fig. 5 | Context-only TFs induce cooperativity in enhancer activity assays. **a**, Episomal (top) and endogenous (bottom) STARR-seq activity for individual barcodes (for exact numbers, see Supplementary Data 4) of homotypic or heterotypic motif combinations of PU.1 and MEF2. Episomal activities are normalized by the no-TFBS sequence (gray dashed line). Endogenous activities are displayed without normalization with no-TFBS activities shown separately (gray boxplot). Differences in endogenous activity are given with respect to the ‘no-TFBS’ control, Mann–Whitney test; for P values, see Supplementary Table 2. Homotypic and heterotypic repeats are indicated below or above (blue, MEF2; green, PU.1; orange, MEF2 and PU.1). **b**, As in **a** but for MEF2 and FOX motifs (blue, MEF2; purple, FOX; pink, MEF2 and FOX). For exact P values, see Supplementary Table 2. **c**, Cooperativity assessment (top; Methods), sum of individual motifs (Sol, orange) or weighted average of homotypic motifs (WHA, red). Bottom, cooperativity (y axis) for episomal and endogenous assays split by motif number and method (x axis). Boxplot color indicates the presence/absence of context-only TFBSs (blue/gray). Individual points represent different sequence contexts and replicates in the episomal assay (Mann–Whitney test; $P \leq 2.2 \times 10^{-16}$,

$P = 3.7 \times 10^{-15}$, $P \leq 2.2 \times 10^{-16}$, $P = 1.8 \times 10^{-3}$ and $P = 1.3 \times 10^{-3}$). **d**, As in **c** (episomal only) for two motif combinations but split by distinct TF class combinations (x axis). Color of points represents the method described in **c**. **e**, The \log_2 -average episomal (y axis) or endogenous (x axis) STARR-seq activities normalized by the ‘no-TFBS’ contexts. The straight line represents a linear fit. Colors represent TF class combinations (pink, context-only TFBSs; orange, context-only and context-initiator BSs; green, context-initiator TFBSs). **f**, Left, volcano plot of Enformer-predicted DNA accessibility (y axis) and the difference between episomal and endogenous average \log_2 -transformed STARR-seq activities (x axis) for all TFBS combinations. The color indicates the favored assay type (episomal, red and endogenous, black). Right, difference in predicted accessibilities for episomally and endogenously favored fragments (Mann–Whitney test; $P = 2.6 \times 10^{-6}$). Colors as in **e**. In **a–d** and **f**, box plots denote the median and the 25th and 75th percentiles, with whiskers extending to 1.5× the interquartile range. Data are shown for the 5-bp spacer library. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ and **** $P < 0.0001$.

elements with sequencing) assay⁵² (Fig. 3a). We found that transcriptional activity varies about eightfold across tested fragments, with good overall replicate agreement (Pearson’s $\rho = 0.93$; Supplementary Fig. 5a). We identified 210 SNPs (~20% of caQTLs) with differential SuRE activity (sureQTLs; Fig. 3b). sureQTLs were more likely to colocalize with expression QTLs (eQTLs; Fig. 3c), and their effect-size correlated with the fold change in allele-specific accessibility (ASA) in GM12878 cells (Supplementary Fig. 5b), suggesting that a portion of caQTLs relate to changes in transcriptional activity.

Computing the correlation between the top BS strength in a fragment (Methods) and SuRE activity, we found that both initiator TF classes, but not the context-only one, were significantly correlated with

activity (Fig. 3d). Supporting this observation, the fraction of TFs with a predicted transcriptional activation domain (TAD)⁵³ (Methods) was larger for the two initiator classes, while no enrichment was seen for context-only TFs (Fig. 3e). Finally, we assessed whether context-only TFs, albeit not driving activity autonomously, may enhance the activity of context-initiator TFs. To do so, we compared the SuRE activity of fragments with above-average top TFBSs for either a context-initiator and a context-only TF, or the context-initiator TF alone. Although none of the pairs met statistical significance individually, we still observed a global trend toward higher SuRE activity when context-initiator and context-only TFBSs were paired. This was not the case for pairings involving USTs or unlabeled TFs (Fig. 3f).

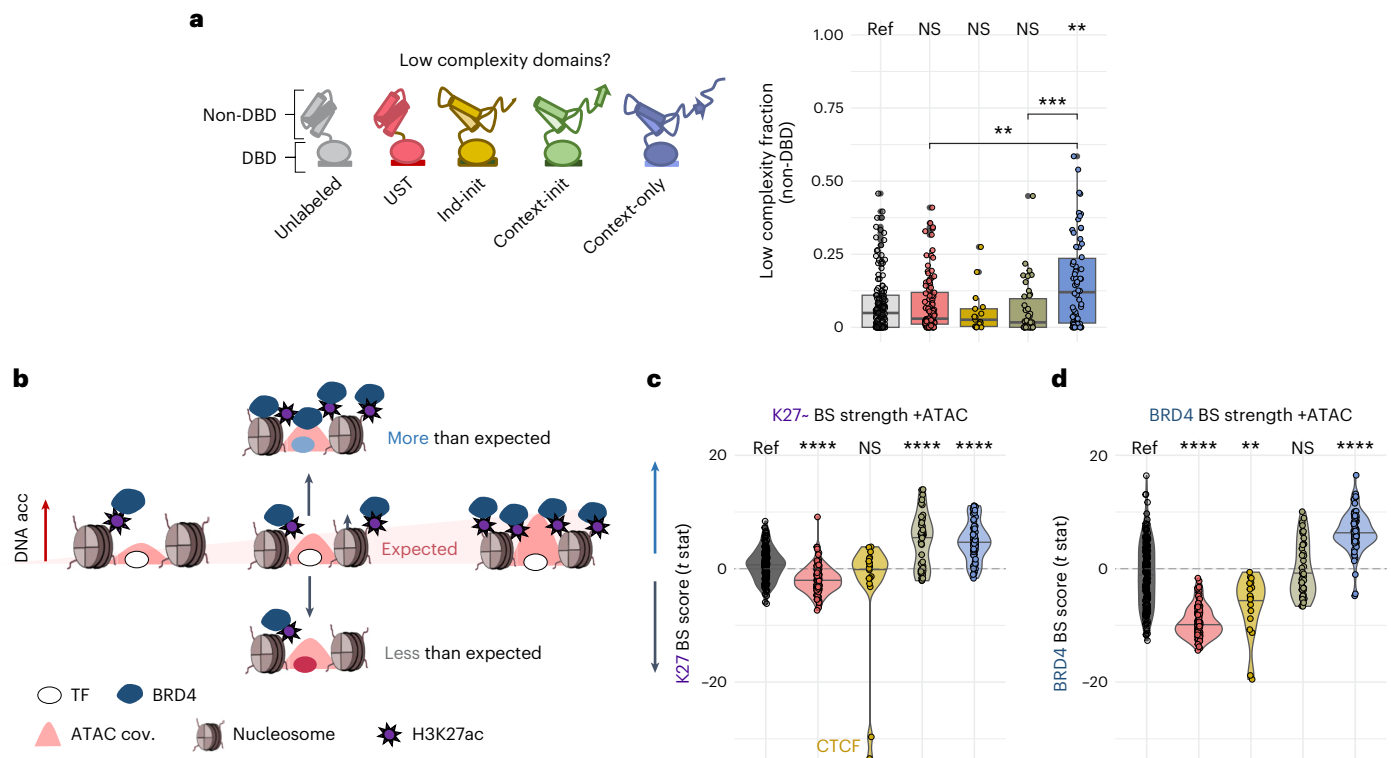


Fig. 6 | Context-only TFBSs are linked to high coactivator binding.

a, Fraction of 20-amino-acid windows with low complexity (y axis) within the non-DBD of TFs (schematics) across TF classes. Significance is computed using unlabeled TFs as a reference or by comparing across classes (Mann–Whitney test, comparison to unlabeled class: $P = 0.97$, $P = 0.36$, $P = 0.06$ and $P = 2.5 \times 10^{-3}$, respectively, P_{UST} versus $P_{\text{context-only}} = 2.9 \times 10^{-3}$, $P_{\text{context-initiator}}$ versus $P_{\text{context-only}} = 2.2 \times 10^{-4}$). Box plots denote the 25th percentile, median and 75th percentile, with whiskers extending to $1.5 \times$ the interquartile range. **b**, Schematic representation explaining the model used in **c** and **d**. Accessibility drives baseline levels of H3K27ac and BRD4 immunoprecipitation coverage (red shaded area represents the expected enrichment). TFs are evaluated on whether their top

BS strength is associated with more or less immunoprecipitation coverage after controlling for accessibility. **c,d**, Results of a log-linear model associating immunoprecipitation coverage (**c**, H3K27ac and **d**, BRD4) with top TFBS strength within an enhancer while using ATAC-seq coverage as a covariate. The t -statistics of the corresponding BS strength model coefficients are plotted on the y axis. TFs are split by TF class (**a**), and significance is assessed with respect to unlabeled TFs (Mann–Whitney test; $P = 1.2 \times 10^{-14}$, $P = 0.52$, 6×10^{-7} and $P = 3.1 \times 10^{-14}$ for **c** and $P < 2.2 \times 10^{-16}$, $P = 1.1 \times 10^{-3}$, $P = 0.38$ and $P < 2.2 \times 10^{-16}$ for **d**). Violin plots encompass the entire data range and represent the medians. ** $P < 0.01$, *** $P < 0.001$ and **** $P < 0.0001$.

Context-only TFs induce cooperativity in enhancer activity assays

Given the dozens of BSs (Fig. 2a) present in endogenous fragments, we resorted to synthetic enhancer constructs to determine if specific TFBS combinations amplify each other's activity. Specifically, we designed homotypic and heterotypic enhancers containing distinct motif combinations for eight representative motifs and tested 'enhancer' activity using STARR-seq (self-transcribing active regulatory region sequencing)⁵⁴ (Fig. 4a; motifs span different TF classes and are associated with varying SuRE activities; compare Fig. 3d). To account for motif position and flanking sequence, we used two different spacers (5 bp and 10 bp) and three randomly created and barcoded sequence contexts (Fig. 4a, Supplementary Table 1 and Supplementary Data 2 and 3). We also created a new, genome-integrated STARR-seq version (endogenous STARR-seq; Fig. 4a and Supplementary Fig. 6a–c) to assess initiator activity.

Compared to the episomal setup, genome-integrated synthetic BS combinations are only weakly expressed. On average, only ~40% of all unique barcodes detected in the genomic DNA (gDNA) library were captured in the respective mRNA samples (Supplementary Fig. 6c). Nonetheless, we observed a decent replicate agreement (Pearson's ρ between 0.51 and 0.55; Supplementary Fig. 6b). We found that motif identity, spacer length, motif number and assay type (episomal versus endogenous) all influence enhancer activity (Fig. 4b,c and Supplementary Fig. 7c). The two TFs with the strongest activity in SuRE-seq

(NF- κ B and IRF; Fig. 3d) also yielded the highest mean activity in both STARR-seq assays (Fig. 4b,c and Supplementary Fig. 7c). Interestingly, among 'one-BS' constructs, PU.1, although not driving activity from a plasmid, performed similarly well as IRF endogenously (Figs. 3d and 4b,c). This likely reflects its known pioneering ability⁵⁵, which we confirmed by predicting fragment accessibility using the deep-learning model Enformer¹³ (Fig. 4d). Endogenous activity thus likely reflects the pioneering and activation potential of TFs.

Focusing on cooperativity, we assessed whether adding a context-only to a context-initiator motif would increase activity. As an example, we selected MEF2 and PU.1 and compared enhancer activity across all tested motif pairs. Notably, the combination of both motifs resulted in enhanced episomal activity compared to each motif on its own (Fig. 5a). In the endogenous context, we found a similar trend; however, it required three or more BSs (Fig. 5a). We found similar enhancing effects for other context-TF pairs in episomal assays such as RUNX and FOX motifs (Supplementary Fig. 7a). In the episomal context, combining two context-only motifs (FOX and MEF2) resulted in even stronger enhancer activity increases, suggesting that their combined binding leads to cooperativity on nucleosome-free DNA (Fig. 5b). When integrated into the genome, the two TFs were still able to enhance one another, however, with dampened effect sizes, as neither TF on its own could drive endogenous activity (Fig. 5b) nor predicted DNA accessibility (Fig. 4d). Notably, we observed no enhancing effect when combining different context-initiator TFs (episomal context; Supplementary

Fig. 7b), supporting their primary role in providing DNA access or mediating transcriptional activity but not cooperativity per se.

To systematically assess context-only motif-driven cooperativity, we derived two measures, comparing heterotypic enhancer activity either to the sum of individual motifs or the weighted sum of homotypic motif combinations of equal length (Fig. 5c and Supplementary Methods). Episomally, heterotypic combinations that included context-only motifs were associated with positive cooperativity and scored significantly higher than those without (Fig. 5c). This was true independent of motif number and spacing (Fig. 5c and Supplementary Fig. 7d). Within the endogenous context, context-only TFs still did better than combinations of noncontext-only TFs; however, positive cooperativity required at least three BSs (Fig. 5c). Splitting heterotypic motif pairs that contained a context-only motif based on the class of the second motif, we found that, episomally, the two context-only motifs do particularly well when paired with context-initiator motifs or each other (Fig. 5d and Supplementary Fig. 7e).

Finally, to hone in on why context-only TFs strongly enhance each other on a plasmid, while, in the context of caQTL enhancers, we find them enriched with context initiators, we systematically compared activities for all BS combinations across episomal and endogenous STARR-seq experiments (Fig. 5e). Despite an overall correspondence between assays ($R^2 = 0.35$, $P = 2.8 \times 10^{-15}$), activities for enhancers made up exclusively of either context-only or context-initiator TFBSs were skewed toward episomal and endogenous assays, respectively (Fig. 5e). Combinations including both TF classes performed similar in both assays. Comparing Enformer-predicted DNA accessibility between episomally versus endogenously favored fragments revealed DNA access as a main driver of differential activity (Fig. 5f). Together, these findings indicate that to fulfill their ‘cooperativity-driving’ function, context-only TFs require context initiators to gain access to nucleosome-wrapped DNA.

Context-only TFs are linked to coactivator recruitment

Given their ability to induce cooperativity without distinct TFBS syntax preferences, we wondered whether context-TFs may contribute to caQTL environments by tethering other TFs and/or coactivators to enhancers targeted by context initiators via weak protein interactions. Comparing the fraction of non-DNA-binding domain (non-DBD) amino acids with low complexity⁵⁶, we indeed found a higher fraction of disorder in context-only TFs compared to all other TF classes (Fig. 6a).

Next, we computed how context-only TF-binding strength relates to coactivator recruitment and chromatin activity (Methods and Fig. 6b). While increases in UST binding strength resulted in lower-than-expected H3K27ac levels, higher-than-expected levels were found for context-TFBSs (Fig. 6c). Notably, only the context-only class was significantly associated with enhanced BRD4 recruitment (Fig. 6d). These findings are in line with our enhancer activity assays (Figs. 3d

and 4), with context-initiator TFs not only providing DNA access but also contributing to enhancer activation, while context-only TFs appear to amplify effect sizes by providing specificity for the coactivator BRD4, possibly mediated through disordered domains.

Communicating enhancers have stronger context-TFBSs

If the cooperativity-enhancing function of context-only TFs is indeed linked to coactivators such as BRD4, their binding to two or more enhancers may result in the formation of hubs^{41,57}. To test this hypothesis, we used information on CMs provided in the original caQTL study⁴⁵ (Methods and Fig. 7a). In most CMs (~96.5%), accessibility changes of enhancer pairs, comprising a caQTL-containing lead enhancer (CM lead) and a dependent one (CM dependent), are concordant, that is, coordinated. However, in a small subset of CMs (3.5% or ~150 pairs), anticorrelation is observed (Fig. 7a), potentially indicating enhancer incompatibility. To infer which TFs contribute to CM coordination, we created 50 control sets of independent enhancer pairs, consisting of caQTL enhancers with matched GC and caQTL potencies that only act locally (LOCAL lead) and their independent neighbors (Methods, Fig. 7a and Supplementary Fig. 8a–c).

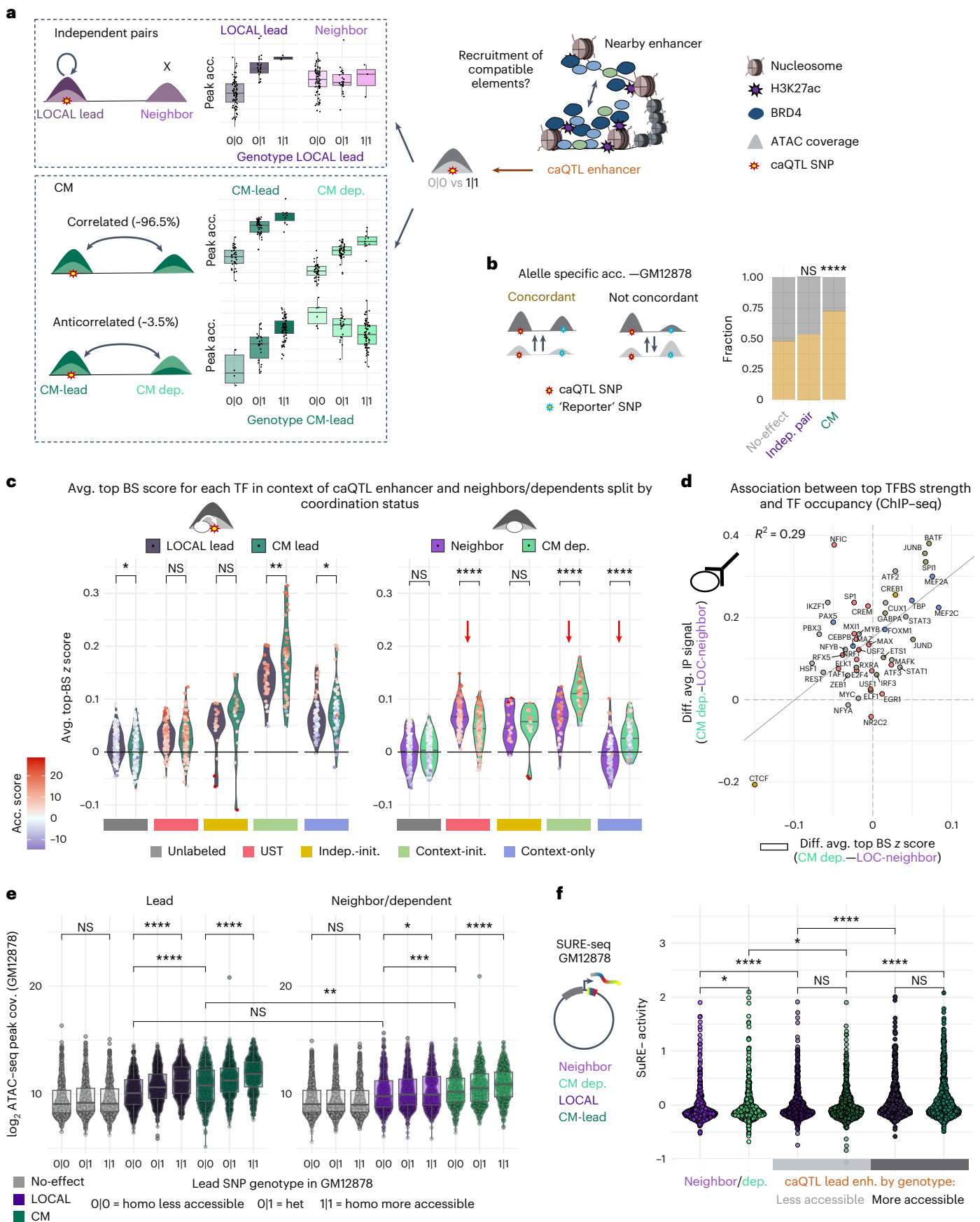
Using phased genome and ASA information in GM12878, we first validated that CMs exist within a homogenous cell population—only enhancer-centric SNPs linked through a CM showed significant concordance in ASA (Fig. 7b) with a roughly linear relationship for effect size (Supplementary Fig. 8d). The latter was true, even when accounting for several covariates (Supplementary Fig. 8e). Next, we ruled out that CM formation is explained by differences in peak-pair distance or chromatin topology as measured through (1) topologically associated domain containment, (2) the presence of a CTCF boundary between individual enhancers, and (3) CTCF motif orientation (looping; Supplementary Fig. 8f,g).

After controlling for confounders, we assessed the TFBS composition in CMs and independent enhancer pairs (Methods). While no differences were found in the frequency with which caQTL SNPs create new best BSs (Supplementary Fig. 9a), independent and CM enhancer pairs differed in their overall top BS scores across TF classes. On average, both context-initiator and context-only TFs have stronger TFBSs in CMs (Fig. 7c), with more significant differences identified for distance-matched CM-dependent and LOCAL-neighbor enhancers. We also found a reverse selectivity for USTs, whose BSs were stronger in independent neighbors (Fig. 7c). Context-only TFs stood out, as their average top BS strength surpassed that of random enhancers only in CM-dependent enhancers but not in LOCAL neighbor ones (Fig. 7c). Focusing on the small set of anticorrelated CMs, we found the reverse—context-only TFBSs were depleted, whereas UST ones were enriched (Supplementary Fig. 9b).

To confirm that stronger TFBSs translate to higher occupancy, we computed the IP coverage of TFs with available chromatin

Fig. 7 | Coordinated enhancers are enriched for context-TFBSs. **a**, CM discovery based on coordinated, interindividual changes in DNA peak accessibility (y axis), with individuals split by the lead caQTL genotype (0|1 indicates more or less accessible genotypes). Independent pairs of LOCAL leads and independent neighbors are indicated in purple, and correlated or anticorrelated CM lead and CM dependents are indicated in sea green. Color shades reflect accessibility levels. **b**, Validation of CM presence in GM12878 using phased genome data. Effect-size directions of ASA at heterozygous, caQTL (red star) and ‘reporter’ SNPs (blue star, non-caQTL enhancer) are compared (yellow, concordant and gray, nonconcordant). Independent pairs ($n = 517$) and CMs ($n = 270$) are compared to no-effect ($n = 245$) enhancer pairs (Fisher’s exact test, $P = 0.16$, 1.8×10^{-8}). **c**, Average top TFBS z scores split by enhancer subcategory and TF class. Left, LOCAL- and CM-lead enhancers. Right, distance-matched neighbor/dependent enhancers (purple, LOCAL leads/LOCAL neighbors and sea green, CM leads/CM dependents). Interenhancer differences are assessed with a Mann–Whitney test (from left to right, $P = 0.02$, $P = 0.3$, $P = 0.07$, $P = 0.01$, $P = 0.05$ and 0.77 , $P = 2.0 \times 10^{-6}$, $P = 0.7$, $P = 2.6 \times 10^{-10}$ and $P = 8.5 \times 10^{-8}$). Red

arrows indicate relevant differences. **d**, Average \log_2 difference in TF occupancy (y axis) versus average top TFBS z score difference (x axis) between distance-matched LOCAL-neighbor and CM-dependent enhancers across TFs. The line represents a fitted linear model. **e**, Comparison of \log_2 -transformed peak-level DNA accessibility (y axis) in GM12878 split by the caQTL genotype (0|1 indicates more or less accessible genotypes) for no-effect ($n = 1,981$), LOCAL ($n = 1,586$) and CM ($n = 1,498$) enhancer pairs (Mann–Whitney test; see Supplementary Table 2 for exact P values). **f**, Fragment-level SuRE-seq activities (y axis) for independent and CM enhancer pairs (for n , see Supplementary Data 1). Colors differentiate ‘independent’ from ‘CM pairs’, and shades ‘leads’ from ‘neighbor/CM-dependent’ enhancers. Different caQTL genotypes are represented by gray boxes (x axis). Bars indicate means. Interenhancer differences were assessed with a Mann–Whitney test; see Supplementary Table 2 for exact P values. In **a** and **e**, box plots denote the median and the 25th and 75th percentiles, with whiskers extending to $1.5 \times$ the interquartile range. Violin plots represent the median and encompass the entire data range. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ and **** $P < 0.0001$.



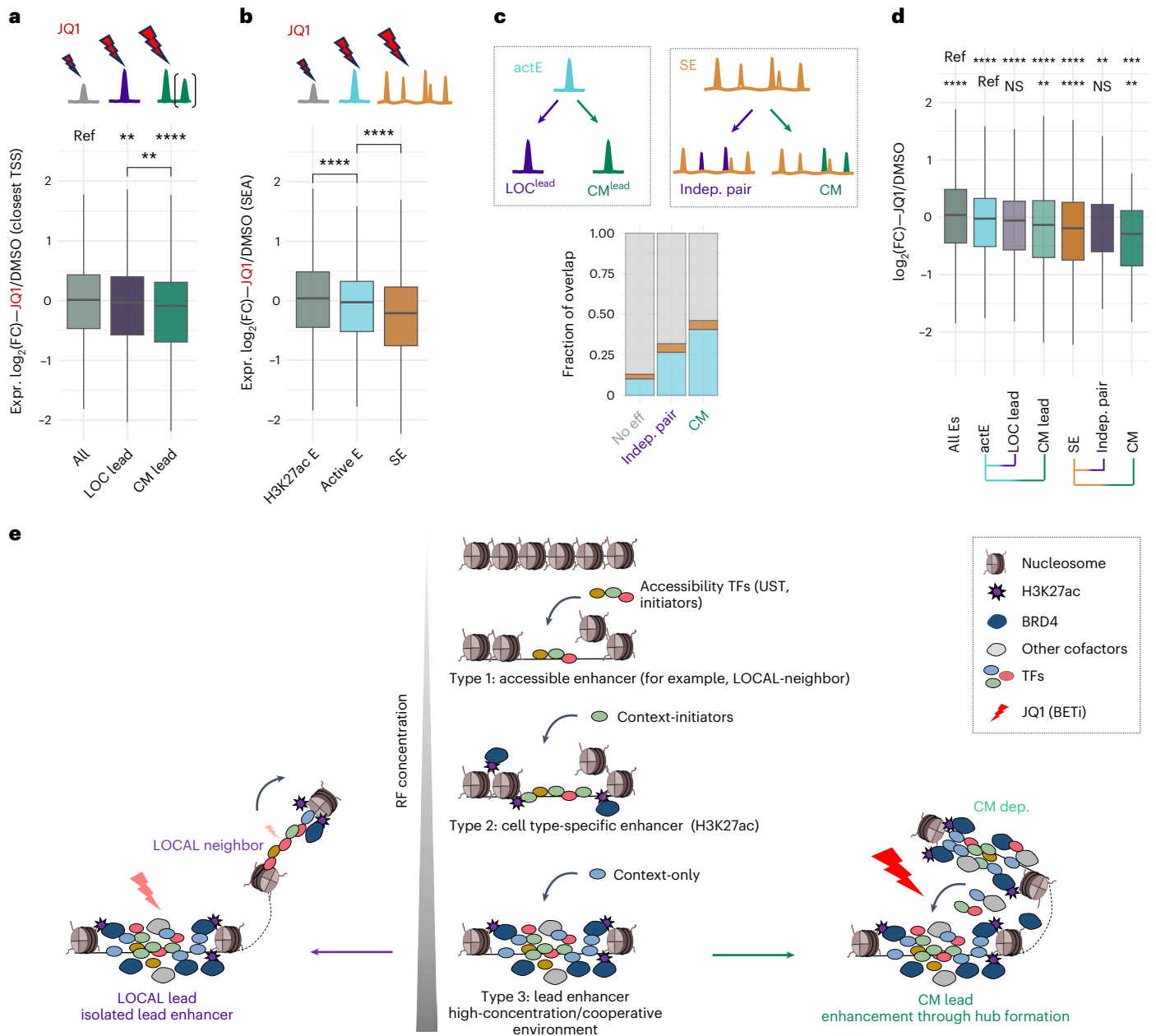


Fig. 8 | CMs are sensitive to BETis and resemble super-enhancers. a, The $\log_2(\text{FC})$ in expression after JQ1 treatment of MEC1 cells, comparing genes controlled by LOCAL-lead ($n = 1,549$) and CM-lead ($n = 1,249$) enhancers to all other genes ($n = 19,459$). Enhancer–gene mapping based on closest TSS. Schematics indicate the expected sensitivity to JQ1 (lightning symbol size). Mann–Whitney test (top, $P = 2.4 \times 10^{-3}$, $P = 1.9 \times 10^{-10}$ and bottom, $P = 6.1 \times 10^{-3}$). **b**, As in **a** but using enhancer and gene annotations from the SEA⁶² (H3K27ac enhancer, $n = 14,511$; active enhancers (active E), $n = 7,713$ and super-enhancers (SE), $n = 1087$). Mann–Whitney test; $P \leq 2.2 \times 10^{-16}$, $P = 8.0 \times 10^{-11}$. **c**, Overlap between different caQTL and no-effect enhancers with active B cell enhancers (actEs; cyan) and independent/CM pairs and no-effect–neighbor pairs with SEs (orange). Nonoverlapping enhancers are indicated in gray. **d**, As in **a** and **b** but splitting active E based on their LOCAL-lead/CM-lead annotation and SEs based on their independent-pair or CM-pair label (all enhancers, $n = 14,511$; actE, $n = 6,935$; LOCAL_actE, $n = 378$; CM_actE, $n = 360$; SE, $n = 818$; LOCAL_SE, $n = 82$ and

CM_SE, $n = 55$). Significance assessed with respect to all or active E genes using a Mann–Whitney test (top $P \leq 2.2 \times 10^{-16}$, $P = 2.8 \times 10^{-5}$, $P = 2.3 \times 10^{-8}$, $P < 2.2 \times 10^{-16}$, $P = 6.3 \times 10^{-3}$ and $P = 3.4 \times 10^{-4}$; bottom $P \leq 2.2 \times 10^{-16}$, $P = 0.15$, $P = 2.2 \times 10^{-3}$, $P = 2.3 \times 10^{-7}$, $P = 0.13$ and $P = 7.6 \times 10^{-3}$). **e**, Summary figure. Different types of enhancers form as a function of the TF class (middle). Accessibility-associated TF classes establish accessible yet independently bound enhancers with low local TF concentration (type 1, for example, LOCAL neighbor). The addition of context-initiator TFBSs can lead to cell type-specific activation (link to H3K27ac; type 2). Context-only TFBSs boost cooperativity (BRD4 retention), leading to high-concentration regulatory factor clusters (type 3, for example, caQTL leads). caQTL lead enhancers split based on the identity of nearby elements into CM-forming (right-hand side) elements or those that only act locally (left-hand side). Lightning symbol size and shading indicate the sensitivity of enhancers to JQ1. In **a**, **b** and **d**, box plots denote the 25th percentile, median and 75th percentile, with whiskers extending to 1.5 \times the interquartile range. ** $P < 0.01$, *** $P < 0.001$, and **** $P < 0.0001$.

immunoprecipitation followed by sequencing (ChIP–seq) data in GM12878 for CM-dependent enhancers and LOCAL neighbor ones. Differences in average BS strength between the two enhancer types were indeed predictive of differential binding ($R^2 = 0.29$ and $P = 4.02 \times 10^{-5}$).

However, average TF occupancy levels for CM-dependent enhancers were higher across TF classes, including USTs (Fig. 7d), whose top BS scores were higher in LOCAL-neighbor enhancers. This decoupling of BS strength and TF occupancy supports the notion of regulatory

factor cluster formation where an increase in local concentration lowers apparent dissociation constant (K_d) values²² in CM-dependent enhancers.

Interestingly, we found that the hierarchy within CMs based on Bayesian modeling⁴⁵ is recapitulated in the average BS scores for context-associated TFs, with caQTL lead enhancers having better BSs (Fig. 7c). Splitting enhancers based on the caQTL status of the lead in GM12878, we found that this difference translates to differences in accessibility (ATAC-seq) and intrinsic activity (SuRE-seq; Methods). CM-dependent ATAC-seq coverage varied as a function of CM-lead enhancer accessibility, yet starting at lower levels (Fig. 7e). While both CM enhancers were more accessible, only CM-dependent but not CM-lead enhancers were intrinsically more active than their LOCAL counterparts when isolated on a plasmid (SuRE-seq; Fig. 7f). This suggests that hierarchies within CMs are innate, and differences among caQTL leads, which are ‘stronger’ enhancers on average, only emerge in endogenous contexts where enhancer coordination can be observed.

Genes controlled by CMs are sensitive to BET inhibition

If CM enhancers indeed form hubs with high regulatory factor concentration, we would expect them to be more sensitive to bromodomain and extraterminal domain inhibitors (BETis)^{58–60}, that is, JQ1, which disrupts hubs⁶¹, especially those formed at super-enhancers⁵⁸. Using existing gene expression data of LCLs treated with either DMSO or JQ1 (ref. 58) and mapping lead enhancers to their closest TSS, we found that both LOCAL- and CM-lead genes were sensitive to JQ1. However, CM genes were impacted more (Fig. 8a). Notably, there was no difference in the distance to the TSS and in baseline expression (Supplementary Fig. 10a,b). The same gradient of JQ1 sensitivity is observed when splitting genes based on their active enhancer and super-enhancer labels using annotations from the Super Enhancer Archive (SEA)⁶² (Fig. 8b), suggesting that LOCAL leads and CMs may be synonymous with active enhancers and super-enhancers, respectively. However, when intersecting the different enhancer labels, only active enhancers were enriched for CM leads, while no differences in overlap were found for super-enhancers (Fig. 8c). Moreover, in contrast to caQTL leads, both active enhancers and super-enhancers were skewed toward highly expressed genes (Supplementary Fig. 10c). Because high activity alone does not imply communication, we next compared JQ1 sensitivity across the different intersections. When comparing to active enhancers, only the CM label but not the LOCAL label was associated with increased JQ1 sensitivity (Fig. 8d). Moreover, super-enhancers containing independent enhancer pairs were no different than active enhancers, while CM-super-enhancers had larger median reductions in gene expression than the entirety of super-enhancers (Fig. 8d and Supplementary Fig. 10d), suggesting that the super-enhancer label may lack sensitivity to detect molecular coordination.

Discussion

In this study, we present a combined experimental and analytical approach that leverages enhancer-centric caQTL SNPs to investigate how TFs initiate the formation of cooperative environments and enhancer communication. Our analyses identified a group of TFs enriched in caQTL enhancers (context-TFs), which we further stratify into context-initiator and context-only TFs based on the enrichment and positioning of their BSs relative to the caQTL SNP. Context initiators are linked to accessibility changes, the deposition of active histone marks (H3K27ac) and B cell-specific enhancer identity. Context-only TFs, on the other hand, appear to amplify rather than initiate enhancer activity, as the presence of their BSs results in increased enhancer activity in both episomal and genome-integrated assays. The functional complementarity and the resulting cooperativity between the two classes of context-TFs thus provide a rationale for why SNPs become causal in some contexts but not others.

We draw a connection between context-only TF-mediated cooperativity and regulatory factor cluster formation by linking context-only TFs to condensate-specific properties—context-only TFs are enriched for disordered domains, and their BS strength correlates with BRD4 recruitment, a coactivator previously implicated in condensate formation²⁸. Finally, context-only TFBSs are specifically enriched in coordinated enhancer pairs, which, similar to super-enhancers, are sensitive to BETis. Our findings therefore suggest that enhancer coordination, and thus compatibility, may rely on the creation (and merging) of two or more such cooperative environments (Fig. 8e). Moreover, we found that the decision to communicate lies predominantly with the ‘weaker’ (dependent) element. CM-dependent enhancers thus resemble the recently discovered ‘facilitator’ elements part of the erythroid α -globin SE⁶³, the ‘tethering’ elements found in *Drosophila*⁶⁴, or the synergistically acting *Sox2* enhancers⁶⁵. We thus propose that CM-dependent enhancers and facilitators, or tethering elements, are two sides of the same coin.

Although we provide a general framework of how enhancers may encode the formation of cooperative environments and enhancer compatibility, our study has several limitations (Supplementary Discussion)—first, we rely on a combination of correlative comparisons for inference. Therefore, we cannot fully rule out that mechanisms other than regulatory factor cluster formation contribute to caQTL cooperativity and enhancer coordination. To vigorously test the condensate-like nature, future experiments may involve the creation of a genome-integrated multi-enhancer system for which molecular activity can be assessed either indirectly in the presence or absence of condensate-disrupting molecules or directly by using imaging setups²⁷. Second, our analysis is limited to one cell type. Whether the concept of a degenerate TF code that relies on functional compatibility rather than pairwise combinations is generalizable, and to what degree TF motif classification is fixed across TF families and cell types, remains to be determined. Third, although our synthetic enhancers confirm that the functional cooperativity identified at the aggregate level is recapitulated when combining individual context-TFBSs, our approach does not allow for predictions at the level of individual endogenous enhancers. Regulatory factor cluster formation likely requires cumulative activity across several TFBSs, meaning that the uniqueness of each enhancer context may cloud the identification of exact thresholds. Finally, this study does not provide granularity on the types of cooperative environments. More refined experimental approaches will be needed to tackle the mechanistic intricacies of TF-driven cooperativity, such as TF domain composition⁶⁶ and/or TF-coregulator compatibility⁸.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41588-024-01892-7>.

References

1. Lambert, S. A. et al. The human transcription factors. *Cell* **172**, 650–665 (2018).
2. Rube, H. T. et al. Prediction of protein–ligand binding affinity from sequencing data with interpretable machine learning. *Nat. Biotechnol.* **40**, 1520–1527 (2022).
3. Kulakovskiy, I. V. et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-seq analysis. *Nucleic Acids Res.* **46**, D252–D259 (2018).
4. Castro-Mondragon, J. A. et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **50**, D165–D173 (2022).

5. Weirauch, M. T. et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
6. Slattery, M. et al. Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.* **39**, 381–399 (2014).
7. Kaluscha, S. et al. Evidence that direct inhibition of transcription factor binding is the prevailing mode of gene and repeat repression by DNA methylation. *Nat. Genet.* **54**, 1895–1906 (2022).
8. Neumayr, C. et al. Differential cofactor dependencies define distinct types of human enhancers. *Nature* **606**, 406–413 (2022).
9. Jolma, A. et al. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**, 384–388 (2015).
10. Isbel, L., Grand, R. S. & Schübeler, D. Generating specificity in genome regulation through transcription factor sensitivity to chromatin. *Nat. Rev. Genet.* **23**, 728–740 (2022).
11. Avsec, Ž. et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* **53**, 354–366 (2021).
12. de Almeida, B. P., Reiter, F., Pagani, M. & Stark, A. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nat. Genet.* **54**, 613–624 (2022).
13. Avsec, Ž. et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
14. Karbalayghareh, A., Sahin, M. & Leslie, C. S. Chromatin interaction—aware gene regulatory modeling with graph attention networks. *Genome Res.* **32**, 930–944 (2022).
15. Zhang, Z., Feng, F., Qiu, Y. & Liu, J. A generalizable framework to comprehensively predict epigenome, chromatin organization, and transcriptome. *Nucleic Acids Res.* **51**, 5931–5947 (2023).
16. Karollus, A., Mauermeier, T. & Gagneur, J. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biol.* **24**, 56 (2023).
17. Sasse, A. et al. Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings. *Nat. Genet.* **55**, 2060–2064 (2023).
18. Kim, S. & Wysocka, J. Deciphering the multi-scale, quantitative cis-regulatory code. *Mol. Cell.* **83**, 373–392 (2023).
19. Liu, Z. & Tjian, R. Visualizing transcription factor dynamics in living cells. *J. Cell Biol.* **217**, 1181–1191 (2018).
20. Neikes, H. K. et al. Quantification of absolute transcription factor binding affinities in the native chromatin context using BANC-seq. *Nat. Biotechnol.* **41**, 1801–1809 (2023).
21. Simicevic, J. & Deplancke, B. Transcription factor proteomics—tools, applications, and challenges. *Proteomics* **17**, 1600317 (2017).
22. Kribelbauer, J. F., Rastogi, C., Bussemaker, H. J. & Mann, R. S. Low-affinity binding sites and the transcription factor specificity paradox in eukaryotes. *Annu. Rev. Cell Dev. Biol.* **35**, 357–379 (2019).
23. Liu, Z. et al. 3D imaging of Sox2 enhancer clusters in embryonic stem cells. *eLife* **3**, e04236 (2014).
24. Mir, M. et al. Dynamic multifactor hubs interact transiently with sites of active transcription in *Drosophila* embryos. *eLife* **7**, e40497 (2018).
25. Tsai, A. et al. Nuclear microenvironments modulate transcription from low-affinity enhancers. *eLife* **6**, e28975 (2017).
26. Wollman, A. J. et al. Transcription factor clusters regulate genes in eukaryotic cells. *eLife* **6**, e27451 (2017).
27. Hayward-Lara, G., Fischer, M. D. & Mir, M. Dynamic microenvironments shape nuclear organization and gene expression. *Curr. Opin. Genet. Dev.* **86**, 102177 (2024).
28. Sabari, B. R. et al. Coactivator condensation at super-enhancers links phase separation and gene control. *Science* **361**, eaar3958 (2018).
29. Boija, A. et al. Transcription factors activate genes through the phase-separation capacity of their activation domains. *Cell* **175**, 1842–1855 (2018).
30. Whyte, W. A. et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).
31. Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
32. Uyehara, C. M. & Apostolou, E. 3D enhancer-promoter interactions and multi-connected hubs: organizational principles and functional roles. *Cell Rep.* **42**, 112068 (2023).
33. Cheng, L., De, C., Li, J. & Pertsinidis, A. Mechanisms of transcription control by distal enhancers from high-resolution single-gene imaging. Preprint at *bioRxiv* <https://doi.org/10.1101/2023.03.19.533190> (2023).
34. Brzovic, P. S. et al. The acidic transcription activator Gcn4 binds the mediator subunit Gal11/Med15 using a simple protein interface forming a fuzzy complex. *Mol. Cell* **44**, 942–953 (2011).
35. Chong, S. et al. Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science* **361**, eaar2555 (2018).
36. Shrinivas, K. et al. Enhancer features that drive formation of transcriptional condensates. *Mol. Cell* **75**, 549–561 (2019).
37. Morin, J. A. et al. Sequence-dependent surface condensation of a pioneer transcription factor on DNA. *Nat. Phys.* **18**, 271–276 (2022).
38. Meeussen, J. V. W. et al. Transcription factor clusters enable target search but do not contribute to target gene activation. *Nucleic Acids Res.* **51**, 5449–5468 (2023).
39. Chong, S. et al. Tuning levels of low-complexity domain interactions to modulate endogenous oncogenic transcription. *Mol. Cell* **82**, 2084–2097 (2022).
40. Trojanowski, J. et al. Transcription activation is enhanced by multivalent interactions independent of phase separation. *Mol. Cell* **82**, 1878–1893 (2022).
41. Alberti, S., Gladfelter, A. & Mittag, T. Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates. *Cell* **176**, 419–434 (2019).
42. Degner, J. F. et al. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).
43. Kumasaka, N., Knights, A. J. & Gaffney, D. J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.* **48**, 206–213 (2016).
44. Gate, R. E. et al. Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nat. Genet.* **50**, 1140–1150 (2018).
45. Kumasaka, N., Knights, A. J. & Gaffney, D. J. High-resolution genetic mapping of putative causal interactions between regions of open chromatin. *Nat. Genet.* **51**, 128–137 (2019).
46. Llimos, G. et al. A leukemia-protective germline variant mediates chromatin module formation via transcription factor nucleation. *Nat. Commun.* **13**, 2042 (2022).
47. Van Mierlo, G., Pushkarev, O., Kribelbauer, J. F. & Deplancke, B. Chromatin modules and their implication in genomic organization and gene regulation. *Trends Genet.* **39**, 140–153 (2023).
48. Zhao, Y. et al. ‘Stripe’ transcription factors provide accessibility to co-binding partners in mammalian genomes. *Mol. Cell* **82**, 3398–3411 (2022).
49. Zamudio, A. V. et al. Mediator condensates localize signaling factors to key cell identity genes. *Mol. Cell* **76**, 753–766 (2019).
50. Meuleman, W. et al. Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).
51. Ibarra, I. L. et al. Mechanistic insights into transcription factor cooperativity and its impact on protein-phenotype interactions. *Nat. Commun.* **11**, 124 (2020).

52. Van Arensbergen, J. et al. Genome-wide mapping of autonomous promoter activity in human cells. *Nat. Biotechnol.* **35**, 145–153 (2017).
53. Staller, M. V. et al. Directed mutational scanning reveals a balance between acidic and hydrophobic residues in strong human activation domains. *Cell Syst.* **13**, 334–345 (2022).
54. Arnold, C. D. et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
55. Minderjahn, J. et al. Mechanisms governing the pioneering and redistribution capabilities of the non-classical pioneer PU.1. *Nat. Commun.* **11**, 402 (2020).
56. Van Mierlo, G. et al. Predicting protein condensate formation using machine learning. *Cell Rep.* **34**, 108705 (2021).
57. Gibson, B. A. et al. Organization of chromatin by intrinsic and regulated phase separation. *Cell* **179**, 470–484.e21 (2019).
58. Ott, C. J. et al. Enhancer architecture and essential core regulatory circuitry of chronic lymphocytic leukemia. *Cancer Cell* **34**, 982–995 (2018).
59. Lovén, J. et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320–334 (2013).
60. Chapuy, B. et al. Discovery and characterization of super-enhancer-associated dependencies in diffuse large B cell lymphoma. *Cancer Cell* **24**, 777–790 (2013).
61. Crump, N. T. et al. BET inhibition disrupts transcription but retains enhancer-promoter contact. *Nat. Commun.* **12**, 223 (2021).
62. Chen, C. et al. SEA version 3.0: a comprehensive extension and update of the super-enhancer archive. *Nucleic Acids Res.* **48**, D198–D203 (2020).
63. Blayney, J. W. et al. Super-enhancers include classical enhancers and facilitators to fully activate gene expression. *Cell* **186**, 5826–5839 (2023).
64. Batut, P. J. et al. Genome organization controls transcriptional dynamics during development. *Science* **375**, 566–570 (2022).
65. Brosh, R. et al. Synthetic regulatory genomics uncovers enhancer context dependence at the Sox2 locus. *Mol. Cell* **83**, 1140–1152 (2023).
66. Lyons, H. et al. Functional partitioning of transcriptional regulators by patterned charge blocks. *Cell* **186**, 327–345 (2023).

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

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

© The Author(s) 2024

Methods

Ethical approval

This research did not require previous ethical approval. Information on interindividual genetic variability was obtained from a previous study⁴⁵ in the form of a publicly available dataset in summary format.

Defining enhancer categories

Information on caQTLs, peaks and CMs in LCLs from 100 individuals⁴⁵ was retrieved from Zenodo (DOI:10.5281/zenodo.1405945). We merged files with caQTL and peak information and isolated SNPs exactly one nucleotide in length and falling within peak boundaries.

To isolate caQTL and no-effect SNPs, we applied the following criteria using information provided in the original publication⁴⁵ (see Supplementary Methods for detail): caQTL = $P_{\text{caQTL}} \times P_{\text{Lead}} > 0.7$ and $P_{\text{caQTL}} > 0.999$ and SNP-to-peak-center-distance ≤ 350 bp; 'no-effect' – $P_{\text{caQTL}} < 0.2$ and $P_{\text{Lead}} < 0.5$ and no overlap with CMs or caQTL-neighbor peaks. In addition, we only retained enhancers using the annotatePeak function (ChIPseeker⁶⁷) and gene annotations from the TxDb.Hsapiens.UCSC.hg19.knownGene package.

No-effect enhancers were downsampled to a subset with equal GC content and 'SNP-to-peak center' distances as that of the caQTL set. Sampling was repeated 50 times to maximize the robustness of results (Supplementary Methods). A random sample of enhancers was generated by sampling 25,000 peaks from the full peak table without replacement. Only enhancers annotated as 'distal' or 'intron' were retained, resulting in ~21k total putative enhancers.

GM12878 ATAC-seq processing and peak scoring

DNA accessibility (ATAC-seq) data for GM12878 were retrieved from ENCODE (experiment: ENCSR637XSC). Read 1 and Read 2 fastq files were combined before aligning to the hg19 reference genome (GRCh37, release 75 from Ensembl) using the BWA-MEM tool⁶⁸ with standard settings. Alignment files were sorted and indexed with SAMtools (version 1.9)⁶⁹ and subsequently transformed to the 'bigwig' format using the bamCoverage (version 3.5.0) command from deepTools⁷⁰. To obtain an accessibility value for each enhancer, counts falling within ± 350 bp around the peak center were summed.

Human TF motifs were downloaded from the HOCOMOCO database³ (version.11). To generate enhancer context scores, peak sequences (± 350 bp of center) were scored in forward and reverse orientation using the 'calculate' function within Biopython⁷¹. For each enhancer, the best score was retained (top scoring). For caQTL enhancers, top scoring was done separately for each genotype. Different top scores across genotypes indicate the creation or destruction of an enhancer-wide top site. Sequences from the random enhancer sample were scrambled, and top scores were recomputed to control for enhancer base composition biases⁷².

TFBS score computation and TF class thresholding

To compute TF accessibility scores, we used a log-linear regression model, relating the cumulative ATAC-seq signal of randomly sampled enhancers to the top score of a TF in those enhancers while controlling for base composition bias (Supplementary Methods). We used the 'Holm-Bonferroni' correction to control for multiple hypothesis testing and $P < 0.01$ to define TFs with significant accessibility scores.

For the initiator score, we computed the number of times a SNP created a new best site for a TF when going from the less to the more accessible genotype and compared frequencies between caQTL and no-effect enhancers using Fisher's exact test. A final \log_2 odds ratio was assigned by averaging across the 50 samples of no-effect enhancers, and a cutoff of 0.9 was set to define initiator TFs (Supplementary Methods). We repeated the analysis to look for 'repressors' by asking how often a caQTL versus no-effect SNP creates a new best site when going from the more to the less accessible genotype.

For the context score of a TF, we used a linear regression model relating the top BS score of a TF in the less accessible genotype to the binary enhancer status (caQTL or no-effect), while controlling for overall ATAC-seq coverage for each TF separately (Supplementary Methods). The procedure was repeated across the 50 samples of no-effect enhancers, and the average across all samples was used as the final context score (t -statistic of regression coefficient). Enhancers in which the SNP created a new best site for a given TF were dropped before fitting each model. A threshold of $t > 3.2$ was used to define context TFs (Supplementary Methods).

All TF motifs with a significant initiator score obtained the 'initiator' label. Next, initiators were subdivided into context or independent initiators based on their context score (t -statistic $> < 3.2$). TF motifs with high context- but low initiator scores were labeled 'context-only'. Finally, TF motifs passing the accessibility score threshold, but not those of the other two scores, were labeled 'USTs', with the name derived from ref. 48. Overlap with USTs was assessed using a Fisher's exact test.

TF cell type specificity index

To link TFs to enhancer annotations, we leveraged the DNA hypersensitive sites (DHS) index⁵⁰ and isolated enhancers annotated as either 'lymphoid-specific' (specific) or 'tissue-invariant' (universal). We derived a 'lymphoid specificity index' for TFs in the following manner: we split the random set of enhancers into those with above- or below-average top TFBS scores (mean-centered), computed the ratio of lymphoid-specific versus tissue-invariant enhancer annotations for each of the two splits, and took the ratio of ratios as the final indicator.

Analysis of TF BS syntax

Sequence logos across caQTL enhancers were generated using the R package 'ggseqlogo' (Supplementary Methods). Motif similarities were computed with the TomTom tool part of the MEME-Suite using default settings and visualized by hierarchical clustering. TF class-wide summaries were generated by dividing the number of similar motif pairs ($q < 0.1$) by the number of possible motif pairs across classes (Supplementary Methods).

To compute the number of nonoverlapping top TFBSs, we first extracted the position of top sites with above-average scores. Next, we merged overlapping motifs (less than 10 bp apart) to obtain a final, nonoverlapping count of TFBSs in an enhancer. We computed distances to the peak center or the enhancer-centric SNPs for all TFs separately. For the distribution of top TFBSs with respect to the peak center, all TF-peak center distances within a TF class were considered. For the positioning with respect to the SNP, the average top TFBS-to-SNP distance per TF was computed first, and class-wide differences were assessed across these averages.

For the analysis of combinatorial TF binding, caQTL enhancers were split based on whether a new best BS for TF1 was created or not. Between the two splits, we compared (1) the top TFBS scores for TF2 using a two-sided Student's t -test and (2) the distances between the top TFBS of TF2 and the SNP (TF1 site). For (2), we set the cutoff for a potential fixed BS syntax to distances < 25 bp and used Fisher's exact test to compute odds ratios (Supplementary Methods). To compare associated P values, we took the negative $\log_{10} P$ value first and assigned a negative sign for odds ratios < 1 (Supplementary Fig. 4).

ASA in GM12878

We performed ASA detection across the whole genome using GM12878 ATAC-seq data downloaded from ENCODE⁷³ (experiment: ENCSR637XSC). We downloaded phased genotype information from the Genome in a Bottle database (<https://www.nist.gov/programs-projects/genome-bottle>). We (1) aligned the .fastq file on the hg19 reference genome (GRCh37, release 75 from Ensembl) using BWA-MEM⁶⁸ (v0.7.17-r1188), (2) marked duplicated reads using Picard⁷⁴ (v2.17.8)

and (3) counted the reads for each allele of the variants described in the phased genotype VCF file using FreeBayes⁷⁵ (v1.3.4) with the following options (--report-monomorphic, --only-use-input-alleles, --min-alternate-fraction 0, --variant-input (vcf file); Supplementary Methods) and summarized the ASA files by counting the RO (reference) versus AO (alternate) allele for heterozygous calls.

SuRE-seq

We selected an equal number of caQTL-probability- and GC-content-matched LOCAL- and CM-lead enhancers, as well as their respective neighbor/dependent elements encompassing a total of 270 bp (for a detailed selection strategy, see Supplementary Methods). A full list is included in Supplementary Data 1. The designed sequences were submitted to Annogen (<https://www.annogen.bio/>) for SuRE screening services in GM12878. Transfections and data acquisition were handled by Annogen, following the transfection protocol described here⁵².

Count tables were processed as follows: barcodes with <2 coverage in the barcode-to-element mapping library ('Reads' in the raw count tables), that mapped to more than one fragment or that had 0 counts in the plasmid library were discarded. Elements with less than four barcodes were discarded. Different orientations of fragments were combined before computing the SuRE activity of each element-barcode combination as the \log_2 ratio of mRNA over plasmid DNA counts (Supplementary Methods). To compute replicate agreement, each element was summarized by its average SuRE activity across barcodes.

sureQTLs were defined as caQTL elements whose SuRE activity differed significantly across genotypes. Significance was assessed using the Student's *t*-test and the 'Bonferroni' method to adjust *P* values. The sureQTL-eQTL overlaps were calculated using eQTL information given in the variant table of the original LCL study⁴⁵, conditioning on either all sureQTLs or sureQTLs with increasingly large effect sizes.

We used Pearson's correlation to compute the relationship between average SuRE activity (less accessible genotype) and the top TFBS within fragments. To compare sureQTL to caQTL (ASA in GM12878) effect sizes, we split fragments by their sureQTL status first. To test for the enhancing effect of TF motif pairs, we compared averaged SuRE activities across fragments (less accessible genotype) containing above-average top TFBSs for either both TF motifs or only one of them using a two-sided Student's *t*-test. The differences in means were extracted as a summary statistic.

STARR-seq

STARR-seq libraries containing 110 bp of sequence context and fixed primer sites on each end were ordered from Twist Biosciences. For a full list of sequences in either the 5-bp or 10-bp spacer library, see Supplementary Data 2 and 3. A detailed design overview is given in the Supplementary Methods. In brief, we generated three random sequence contexts (no-TFBS) in which different homotypic and heterotypic motif repeats were embedded using eight TF motifs (not enriched but highly expressed, *MYC* = CCACGTGC; independent initiator, *NFKB* (*REL*) = GGGAAATCCC and *CTCF* = CCACCAGGGGGCGC; context initiators, *SPI* (PU.1) = AAAGAGGAAGTGA, *IRF* = GAAAGCGAACT and *RUNX* = TTTGTGGTTT; context-only, *MEF2* = GCTAAAATAGAA and *FOX* = CTGTTTACTTT).

Libraries were amplified with primers carrying a 12-bp random barcode (3' end), Nextera sequencing adapters, as well as overhangs for assembly into the STARR-seq vector (Addgene, 99296; Supplementary Table 1) according to the manufacturer's instruction (ten cycles). Amplified pools were cloned into the STARR-seq vector by Gibson Assembly (New England Biolabs, NEB) and transformed into Dh5 α high-efficiency competent cells (NEB). Transformed cells were spread on Ampicillin plates to achieve ~50,000 individual clones for both 5-bp and 10-bp spacer libraries. Colonies were scraped and transferred to a liquid growth medium (Luria Broth) and grown at 37 °C for 2 h before plasmid DNA isolation (Maxiprep Kit, Invitrogen).

MEC1 cells were cultured in Iscove's Modified Dulbecco's Medium (IMDM, Gibco) supplemented with 10% fetal bovine serum and 1% penicillin-streptomycin to obtain ~50 million cells. Transfections of 5-bp and 10-bp libraries were performed in duplicate using $\sim 6 \times 10^6$ cells and 30 μ g STARR-seq plasmid per replicate. Transfections were carried out with the Neon Transfection System (Thermo Fisher Scientific)—three pulses of 20 ms at 1,200 V in R buffer using 100 μ l tips. Cells were collected 24 h post-transfection and lysed in Trizol (Thermo Fisher Scientific). RNA was extracted by chloroform extraction, followed by isopropanol precipitation with a 1:1 ratio. RNA pellets were resuspended in RNase-free water and digested with DNase I for 20 min at room temperature (Zymo). DNase I-treated RNA was purified with an RNA Clean and Concentrator-25 Kit (Zymo). A total of 10–12 μ g RNA was reverse transcribed using the STARR-seq-specific reverse transcription primer (Supplementary Table 1). For each microliter of Maxima H minus reverse transcriptase (Thermo Fisher Scientific), 3 μ g input was used. Reverse transcription reactions were diluted 1:4 with RNase and DNase-free water before running the splice-junction PCR using the STARR-seq thio-splice-junction primer and a custom reverse primer (Supplementary Table 1). About one-third of the RT reaction was amplified for 19–22 cycles. Illumina barcodes were added by a final PCR with eight cycles following Illumina Nextera guidelines. After each amplification step, a cleanup using AMPure XP magnetic beads (Beckman Coulter) was performed. Plasmid input libraries were amplified from the plasmid pool directly with Nextera Read 1 and Read 2 indexing primers (600 ng plasmid, six separate reactions, eight cycles). Plasmid and RNA libraries were sequenced on a NEXTSeq or MiSeq (Illumina) desktop sequencer at EPFL's Gene Expression Core Facility with a 75- or 150-cycle kit.

Barcodes were assigned to fragments using custom Python (version 3.9.5) and R (version 4.1.0) scripts. Barcodes mapping to more than one fragment or having less than five counts in the plasmid library were discarded. STARR-seq activity was defined as the \log_2 RNA to DNA ratio for each barcode. The fragment-level activity was derived by averaging across barcodes linked to each fragment. Normalized activities were derived by subtracting the average \log_2 ratio of the respective 'no-TFBS' sequence contexts alone (Supplementary Methods).

To compute cooperativity for heterotypic multimotif combinations, the following two approaches were considered: expressing cooperativity as the difference between normalized, heterotypic fragment activities and (1) the sum of the underlying monomeric motif activities, or (2) the weighted average of the homotypic motif activities with equal length. For an example and more detail, see Supplementary Methods.

Genome-integrated STARR-seq

An endogenous version of the STARR-seq assay was created by adding a loxP and a mutant lox2272 site flanking the origin of replication (*ori*) and the polyadenylation site of the aforementioned STARR-seq library carrying the homotypic and heterotypic motif combinations of 8 TFs with a 5 bp spacer (for a description of the assembly strategy, see Supplementary Methods and Supplementary Table 1). The resulting plasmid pool was transformed into One Shot OmniMax 2T1 high-efficiency competent cells (Invitrogen) and spread on Ampicillin plates to achieve ~30,000 individual clones and DNA was extracted as for the episomal STARR-seq library.

To assay endogenous STARR-seq activity, we established a stable MEC1 cell line (MEC1 cells were purchased from DSMZ; ACC 497) expressing Cre recombinase under the control of a blasticidin selection marker and containing a Cre-recombinase-mediated cassette-exchange landing pad (loxP and lox2722 sites flanking a CMV-GFP). The cassette was integrated within the *AXIN2* gene locus where it replaces the TSS as well as the 3 kb regulatory sequence upstream of the TSS (*loxP* downstream *AXIN2* sequence, GCCGCCGGCGGCCCGAAATCCATCGTC and lox2272 upstream *AXIN2* sequence, CTGCGACTGTAGCAAGAGGG-GACTGGGACT; locus described in ref. 46). Genetically modified MEC1

cells were cultured to ~50 million cells in IMDM media (Gibco), with 10% fetal bovine serum, 1% penicillin/streptomycin and 5 $\mu\text{g ml}^{-1}$ blasticidin. Eight transfections were performed using the same procedure as for the episomal STARR-seq libraries. Transfected cells were cultured for another 10 days to allow for the removal of GFP in cells with successful cassette exchange. After 10 days, around 70×10^6 cells were sorted on either FACSAriaII or FACSAriaFusion flow cytometers (BD Biosciences), resulting in ~2 Mio GFP⁻ cells. Sorted cells were cultured for another 8–9 days before extracting genomic DNA (gDNA) and mRNA from a total of 5 million cells in triplicates using the mini AIIprep DNA/RNA extraction kit (Qiagen). A total of 20–24 μg of total RNA per replicate was reverse transcribed using the STARR-seq-specific RT primer (Supplementary Table 1). For each microliter of Maxima H minus RT (Thermo Fisher Scientific), 3 μg input was used. The resulting cDNA was purified with a Clean and Concentrator Kit (Zymo) before amplification. The total amount of cDNA and gDNA for each replicate was amplified using the reverse primer already used for the episomal STARR-seq assay (Supplementary Table 1) and either the STARR-seq thio-splice-junction primer (cDNA) or a primer targeting the STARR-seq intron (gDNA; Supplementary Table 1). PCRs were run at 72 °C and 62 °C for a total of 20–21 cycles. Correct sizes of spliced mRNA and nonspliced gDNA products were assessed using gel electrophoresis. Illumina barcodes were added using nine additional cycles. After each amplification step, PCR products were cleaned up using AMPure XP magnetic beads (Beckman Coulter). Plasmid input libraries were amplified as described for the episomal STARR-seq assay, and plasmid, gDNA and mRNA libraries were sequenced on a NextSeq Illumina desktop sequencer at EPFL's Gene expression Core Facility with either a 75- or a 150-cycle kit.

Barcode-to-fragment mapping was executed as described for the episomal assay. gDNA and mRNA counts were normalized by sequencing depth using a 'reads per million (RPM) sequenced' conversion. The three no-TFBS sequence contexts were considered indiscriminately in downstream analyses. To compute endogenous activities, only barcodes with at least two RPM across all three gDNA library replicates were retained. To compute the fraction of BCs with nonzero mRNA levels, the number of unique BCs with a nonzero mRNA count in at least one of the three replicates was divided by the total number of BCs in the gDNA sample. Endogenous, fragment-level activities were computed by taking the average mRNA over gDNA ratio across the three replicates first, before averaging across respective BCs. To minimize effects due to extreme outliers, the top 5% of barcode-level activities were removed. In addition, fragments with less than ten unique expressed barcodes were excluded. Normalization of fragment-level activities was done as described for the episomal assay. BC-level endogenous activities were not normalized to account for the large number of 'zero' RNA counts.

Prediction of DNA accessibility using Enformer

We used the pretrained Enformer model (<https://tfhub.dev/deepmind/enformer/1>; model head number 69) to predict DNA accessibility in GM12878 on an NVIDIA GeForce RTX 3090 GPU for each genome-integrated STARR-seq fragment, padded with the genomic sequence surrounding the landing pad (*AXIN2* locus; Supplementary Methods). A two-sided Mann–Whitney test was performed to compare predicted DNA accessibility across inserts favored in episomal versus endogenous STARR-seq assays.

Computing TF domain properties

Information on TF domains was extracted from Ensemble (Ensembl.Hsapiens.v86). DNA-binding domains were removed from the full-length sequence before predicting *trans*-activation domains (TADs) and low complexity domain scores (approaches described previously^{53,56} and summarized in Supplementary Methods). Significance between TF classes was assessed using a Fisher's exact test (TAD presence) and a two-sided Mann–Whitney test (low complexity score), respectively.

Association between TFs and epigenomic states

Fastq files for H3K27ac and BRD4 ChIP-seq experiments in GM12878 cells were retrieved from ENCODE (ENCFF000ASP, ENCFF000ASU for H3K27ac) or the sequence read archive (SRA) (SRR1636861 for BRD4) and processed as described for the ATAC-seq data but considering ± 650 bp around the peak center. A log-linear model was used to relate the enhancer IP signal to TFBS top scores (see Supplementary Methods for a full description). The t -statistic of the TFBS score coefficient was used as the evaluating metric.

Defining independent and CM enhancer pairs

caQTL enhancers were split into LOCAL leads/CM leads based on whether they were contained within the CM file ('dag.txt' in ref. 45). CM leads were subdivided into correlating and anticorrelating, depending on the effect direction of their linked, dependent enhancers (see 'dag.txt'). For a detailed explanation, see Supplementary Methods.

LOCAL and correlated CM leads were matched with respect to GC content and the SNP 'potency' (P_{caQTL}^{45}) using 50 subsampling steps (Supplementary Methods). To create enhancer pairs (LOCAL lead and LOCAL neighbor; CM lead and CM dependent), we additionally required distance-matching (Supplementary Methods). For illustration purposes, we chose the subsample with the assessed enhancer property closest to the sample average.

CM quality control

CM presence in isogenic cell lines was confirmed by computing the concordance of effect-size direction across SNPs present in phased enhancer pairs in GM12878 (Supplementary Methods). To establish a quantitative relationship, we used a linear model to predict the effect size of SNPs in neighbor/dependent enhancers as a function of the lead caQTL SNP, controlling for peak and SNP-to-peak-center distances and enhancer type (Supplementary Methods). To rule out bias introduced by 3D genome topology, we downloaded information on topologically associating domains in GM12878 from ENCODE (ENCFF788UTU), lifted coordinates to the hg19 genome version (BSgenome.Hsapiens.UCSC.hg19, version 1.4.3) and compared domain inclusion across the different enhancer-pair types (Fisher's exact test). To compare the CTCF boundary signal between enhancer pairs, CTCF ChIP-seq data for GM12878 was downloaded from ENCODE and the maximum IP coverage between peaks was compared (Mann–Whitney test). CTCF motif orientation surrounding enhancer pairs was extracted using the searchSeq function within the R package TFBSTools. For a detailed description, see Supplementary Methods.

Evaluating TF binding and accessibility differences in CMs

To compare TFBSs, we z -transformed top scores using the random enhancer sample as a baseline. For each TF, we computed average z scores within a given enhancer type by averaging across enhancers and across the 50 samplings (Supplementary Methods). For anticorrelated CMs, comparisons were performed against scrambled enhancer sequences.

Available TF ChIP-seq data that had a linked HOCOMOCO motif (48 TFs) were downloaded from ENCODE (see Data availability section; GM12878). Processing was done in the same way as was done for enhancer accessibility scores. Differences in occupancy across LOCAL neighbors and CM dependents were computed by averaging the IP signal within enhancer types before computing the ratio between them.

Enhancers were split by type (caQTL/no-effect/CM, etc.) and the caQTL genotype in GM12878 (phased genome information was used to assign genotypes for neighbor/dependent enhancers). GM12878 ATAC-seq coverage was compared across enhancer types and genotypes using a two-sided Mann–Whitney test.

Assessing BETi sensitivity in CMs

Paired-end RNA-seq FASTQ files belonging to replicates of either JQ1 or DMSO-treated MEC1 cells were retrieved from the SRA (SRR7815327,

SRR7815329, SRR7815330, SRR7815331, SRR7815332 and SRR7815333). Paired-end reads were aligned to the hg19 (GRCh37.75) reference genome using the STAR aligner⁷⁶, sorted, indexed and deduplicated with SAMtools⁶⁹ and loaded into R using the 'Rsubread' package⁷⁷. The 'DESeq2' package⁷⁸ was used for differential gene expression analysis, keeping genes with at least ten counts across all replicates.

Enhancer-to-gene mapping was done by finding the closest TSS. Distances to the TSS were compared between LOCAL leads and CM leads using a Student's *t*-test. Annotation data for super-enhancers and active B cell enhancers were retrieved from the SEA⁶² (v3.0). Enhancers were defined as overlapping if they shared at least 200 bp. Specifics on different enhancer-type intersections and subsamplings are described in Supplementary Methods. Significance assessment between log₂(FCs) in expression across groups was done using a two-sided Mann–Whitney test.

Statistics and reproducibility

All statistical analyses were carried out in R, and figures were generated using the ggplot2 package (version 3.4.2). Raw *P* values are visualized as NS = *P* > 0.05, **P* < 0.05, ***P* < 0.01, ****P* < 0.001 and *****P* < 0.0001. *P* values < 2.2 × 10⁻¹⁶ are reported as '*P* < 2.2 × 10⁻¹⁶', the default cutoff in R. *P* values are unadjusted unless otherwise indicated. Where applicable, multiple testing correction was performed and is indicated in the respective Methods or Supplementary Methods. For all comparisons across TF motif classes, the following sample sizes were given: *n* = 151 for unlabeled TF motifs, *n* = 120 for USTs, *n* = 18 for independent-initiator, *n* = 44 for context-initiator and *n* = 68 for context-only motifs. For the Mann–Whitney and Student's *t*-test, an unpaired, two-sided test was performed unless otherwise specified.

No statistical method was used to predetermine the sample size. No data were excluded from the analyses with filtering steps specified in the respective Methods. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary

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

Data availability

Raw and processed sequencing data for STARR-seq experiments are available from the Gene Expression Omnibus, under accession GSE229646. Raw and processed count data for SuRE-seq and STARR-seq experiments, information about STARR-seq and SuRE-seq libraries, as well as processed summary statistics for all enhancer pairs, and TF classifications are available from Zenodo (<https://doi.org/10.5281/zenodo.12684117>). Motifs were downloaded from HOCOMOCO (version 11: <https://hocomoco11.autosome.org>). The following genome information was used: GRCh37, release 75 from Ensembl. Existing datasets used in this study are available at <https://zenodo.org/records/1405945> and the SRA database (<https://www.ncbi.nlm.nih.gov/sra>) under the following accessions: SRR1636861, SRR7815327, SRR7815329, SRR7815330, SRR7815331, SRR7815332, SRR7815333, or from ENCODE (<https://www.encodeproject.org/>) under the following accessions: ENCSR637XSC, ENCF000ASP, ENCF000ASU, ENCF000NSE, ENCF000NSI, ENCF000NSP, ENCF000NUB, ENCF000NUK, ENCF002BFH, ENCF000ROU, ENCF000VSS, ENCF000VUJ, ENCF000NVC, ENCF000NVM, ENCF000VUY, ENCF000NVQ, ENCF000NWE, ENCF000NWM, ENCF002EAM, ENCF000VVV, ENCF000VXD, ENCF0379ZPU, ENCF000VYD, ENCF000VXQ, ENCF000VXY, ENCF000VYM, ENCF000NXA, ENCF000NXE, ENCF000VYT, ENCF002EBS, ENCF000ROK, ENCF000NYH, ENCF000VZR, ENCF000VZZ, ENCF000WFR, ENCF000WAZ, ENCF000NZO, ENCF000OAD, ENCF000NYY, ENCF000WDK, ENCF000OCO, ENCF000OEH, ENCF000OEV,

ENCF000OBS, ENCF000WFD, ENCF000WFJ, ENCF000OGC, ENCF000WFP, ENCF000OHD, ENCF000WGB, ENCF000OIB and ENCF0788UTU. Enhancer annotation data were obtained from <https://www.meuleman.org/research/dhsindex/>. Phased genome data were downloaded from the genome in the bottle project (<https://www.nist.gov/programs-projects/genome-bottle>). The Enformer model is available at <https://tfhub.dev/deepmind/enformer/1>. Information on super-enhancers can be found in the SEA (<http://sea.edbc.org>).

Code availability

Custom code to process and subset data and compute scores can be found at <https://github.com/DeplanckeLab/Context-TFs> and <https://doi.org/10.5281/zenodo.12732162> (ref. 79).

References

- Wang, Q. et al. Exploring epigenomic datasets by ChIPseeker. *Curr. Protoc.* **2**, e585 (2022).
- Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at <https://doi.org/10.48550/arXiv.1303.3997> (2013).
- Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
- Ramírez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
- Cock, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
- Van den Berge, K. et al. Normalization benchmark of ATAC-seq datasets shows the importance of accounting for GC-content effects. *Cell Rep. Methods* **2**, 100321 (2022).
- Luo, Y. et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.* **48**, D882–D889 (2020).
- Picard toolkit. *GitHub* <https://broadinstitute.github.io/picard/> (2019).
- Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at <https://doi.org/10.48550/arXiv.1207.3907> (2012).
- Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
- Liao, Y., Smyth, G. K. & Shi, W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.* **47**, e47 (2019).
- Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
- Gardeux, V. & Jukri. DeplanckeLab/Context-TFs: initial release (v1.0). Zenodo <https://doi.org/10.5281/zenodo.12732162> (2024).

Acknowledgements

We thank H. Bussemaker and C. Canto for reviewing the manuscript and providing valuable feedback. We further thank Annogen, the EPFL Flow Cytometry and Gene Expression core facilities, as well as the EPFL's Scientific IT and Application Support (SCITAS). This work was supported by a Swiss National Science Foundation (grant 310030_197082), EPFL's Center for Imaging and Marie Skłodowska-Curie fellowships for J.F.K. (895426), O.P. (860002) and G.v.M. (101026623), as well as EMBO long-term fellowships for J.F.K. (1139-2019) and G.v.M. (2020-895).

Author contributions

J.F.K. and B.D. conceived and designed the study. J.F.K. and J.R. performed the experiments. J.F.K. analyzed the data and performed the statistical analyses, with help from O.P., V.G. and G.V.M. K.F.

implemented the Enformer model. J.F.K. and B.D. wrote the manuscript with input from V.G., G.v.M., O.P. and K.F.

Competing interests

The authors declare no competing interest.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41588-024-01892-7>.

Correspondence and requests for materials should be addressed to Judith F. Kribelbauer-Swietek or Bart Deplancke.

Peer review information *Nature Genetics* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

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

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

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

n/a | Confirmed

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

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Data analysis

samtools 1.9
 deeptools 3.5.0
 STAR 2.7.10a
 Freebayes (v1.3.4)
 MEME suite - TomTom tool <https://meme-suite.org/meme/>

Custom code is archived here <https://github.com/DeplanckeLab/Context-TFs> and here <https://doi.org/10.5281/zenodo.12732162>.

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

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Raw and processed sequencing data for STARR-seq experiments are available from the Gene Expression Omnibus, under accession number GSE229646. Raw and processed count data for SuRE-seq and STARR-seq experiments, information about STARR-seq and SuRE-seq libraries, as well as processed summary statistics for all enhancer pairs, and TF classifications are available from Zenodo (<https://zenodo.org/records/12684117>). Motifs were downloaded from HOCOMOC (version 11: <https://hocomoco11.autosome.org>). The following genome information was used: GRCh37, release 75 from Ensembl. Existing datasets used in this study are available from here (<https://zenodo.org/records/1405945>), the SRA database (<https://www.ncbi.nlm.nih.gov/sra>) under the following accession numbers: SRR1636861, SRR7815327, SRR7815329, SRR7815330, SRR7815331, SRR7815332, SRR7815333, or from ENCODE (<https://www.encodeproject.org/>) under the following accession numbers: ENCSR637XSC, ENCF000ASP, ENCF000ASU, ENCF000NSE, ENCF000NSI, ENCF000NSW, ENCF000NUB, ENCF000NUK, ENCF002BFH, ENCF000ROU, ENCF000VSS, ENCF000VUJ, ENCF000NVC, ENCF000NVM, ENCF000VUY, ENCF000NVQ, ENCF000NWE, ENCF000NWM, ENCF002EAM, ENCF000VVV, ENCF000VXD, ENCF379ZPU, ENCF000VYD, ENCF000VXQ, ENCF000VXY, ENCF000VYM, ENCF000NXA, ENCF000NXE, ENCF000VYT, ENCF002EBS, ENCF000ROK, ENCF000NYH, ENCF000VZR, ENCF000VZZ, ENCF000WFR, ENCF000WAZ, ENCF000NZO, ENCF000OAD, ENCF000NYY, ENCF000WDK, ENCF000OCO, ENCF000OEH, ENCF000OEV, ENCF000OBS, ENCF000WFD, ENCF000WFJ, ENCF000OGC, ENCF000WFP, ENCF000OHD, ENCF000WGB, ENCF000OIB, and ENCF788UTU. Enhancer annotation data was obtained here: <https://www.meuleman.org/research/dhsindex/>. Phased genome data was downloaded from the genome in the bottle project (<https://www.nist.gov/programs-projects/genome-bottle>). The Enformer model is available here (<https://tfhub.dev/deepmind/enformer/1>). Information on Super Enhancers can be found here: SEA v3.0.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

Reporting on race, ethnicity, or other socially relevant groupings

Population characteristics

Recruitment

Ethics oversight

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size

Data exclusions	No data was excluded.
Replication	Experiments included internal replications, using different sequence embeddings and randomized barcodes, and different spacings. Additionally, all experiments were done in duplicates or triplicates. We achieved replicate agreements (individual correlations between 0.5-0.93) that are in line with studies using similar approaches. Different sequence embeddings produced similar results. For computational analyses, statistics are derived by averaging across fifty samplings to ensure robustness.
Randomization	We used sequence scrambling to account for base composition biases in endogenous enhancer sequences and three different sequence embedding for enhancer activity assays to avoid biases introduced by context. Instead of randomizing enhancers, we generated matched control groups to assess statistical significance.
Blinding	Blinding does not apply as this is a study based on prior data on caQTLs and genotypes and experimental data cannot be altered based on 'observer-bias'. In principle, we performed the discovery of motif caQTL-enhancer context associations blind. Any association with the context of caQTL enhancers is surprising.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern
<input checked="" type="checkbox"/>	<input type="checkbox"/> Plants

Methods

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s)	MEC1 (ACC 497, DSMZ)
Authentication	Non of the cell lines were authenticated.
Mycoplasma contamination	Cell lines tested negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	None.