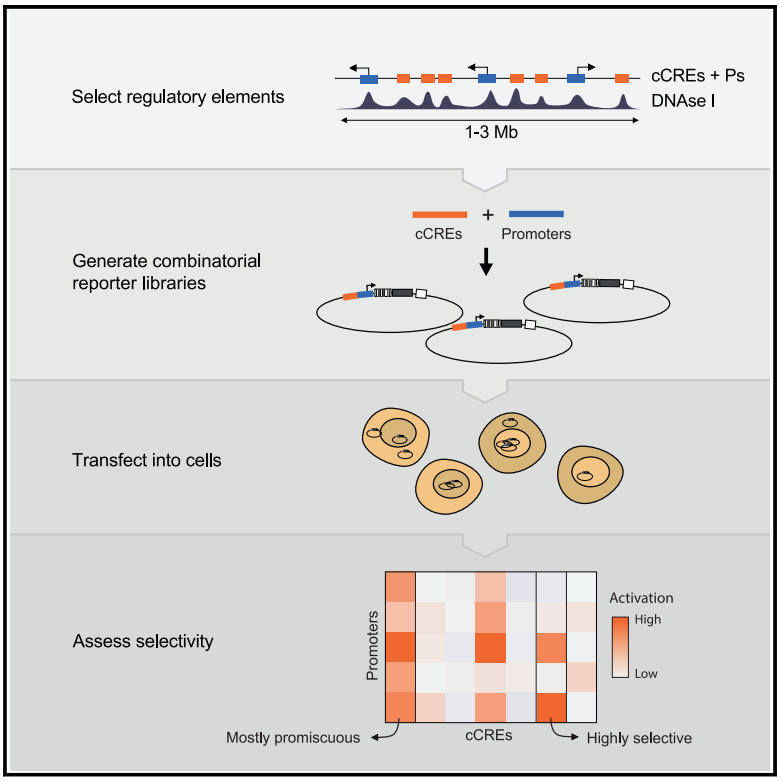
Molecular Cell

Article

Systematic analysis of intrinsic enhancer-promoter compatibility in the mouse genome

Graphical abstract



Authors

Miguel Martinez-Ara,
Federico Comoglio,
Joris van Arensbergen,
Bas van Steensel

Correspondence

b.v.steensel@nki.nl

In brief

A major question in genome biology is how enhancers "choose" their target promoter. Using high-throughput reporter assays, Martinez-Ara and colleagues surveyed the intrinsic compatibilities of thousands of enhancer-promoter pairs in mouse cells. They found a wide diversity of specificities, mostly likely driven by a complex grammar of sequence motifs.

Highlights

- Intrinsic compatibility of thousands of enhancer-promoter combinations was tested
- Compatibilities exhibit a broad spectrum, from promiscuous to highly specific
- Enhancer-promoter compatibility and chromatin looping appear to be independent

Martinez-Ara et al., 2022, Molecular Cell *82*, 2519–2531
July 7, 2022
https://doi.org/10.1016/j.molcel.2022.04.009

CellPress

Molecular Cell

CellPress
OPEN ACCESS

Article

Systematic analysis of intrinsic enhancer-promoter compatibility in the mouse genome

Miguel Martinez-Ara,[1] Federico Comoglio,[1,2] Joris van Arensbergen,[1,3] and Bas van Steensel[1,4,*]
[1]Division of Gene Regulation and Oncode Institute, Netherlands Cancer Institute, 1066 CX Amsterdam, the Netherlands
[2]Present address: enGene Statistics GmbH, Basel, Switzerland
[3]Present address: Annogen B.V., Science Park 406, Amsterdam, the Netherlands
[4]Lead contact
*Correspondence: b.v.steensel@nki.nl
https://doi.org/10.1016/j.molcel.2022.04.009

SUMMARY

Gene expression is in part controlled by *cis*-regulatory elements (CREs) such as enhancers and repressive elements. Anecdotal evidence has indicated that a CRE and a promoter need to be biochemically compatible for promoter regulation to occur, but this compatibility has remained poorly characterized in mammalian cells. We used high-throughput combinatorial reporter assays to test thousands of CRE-promoter pairs from three Mb-sized genomic regions in mouse cells. This revealed that CREs vary substantially in their promoter compatibility, ranging from striking specificity to broad promiscuity. More than half of the tested CREs exhibit significant promoter selectivity. Housekeeping promoters tend to have similar CRE preferences, but other promoters exhibit a wide diversity of compatibilities. Higher-order transcription factors (TF) motif combinations may account for compatibility. CRE-promoter selectivity does not correlate with looping interactions in the native genomic context, suggesting that chromatin folding and compatibility are two orthogonal mechanisms that confer specificity to gene regulation.

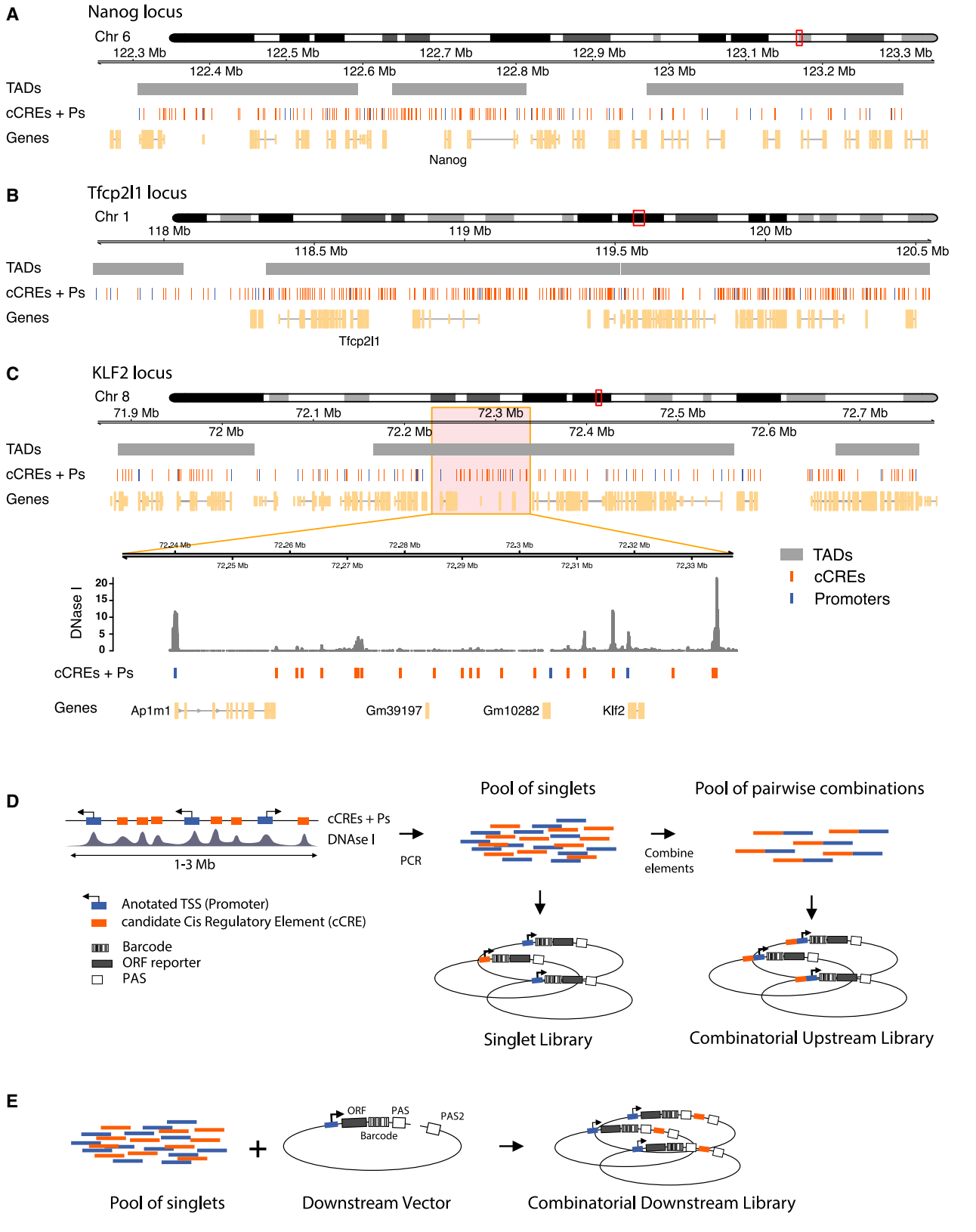
INTRODUCTION

How genes are regulated by *cis*-regulatory elements (CREs) such as enhancers and repressor elements is a long-standing topic in molecular biology (Banerji et al., 1981; Tuan et al., 1985; Fiering et al., 1995; Lettice et al., 2003; ENCODE Project Consortium, 2012; van Arensbergen et al., 2014; Zabidi and Stark, 2016; Farley et al., 2015; Robson et al., 2019; Segert et al., 2021). One conundrum is how CREs "choose" their target promoters. Some enhancers can activate multiple promoters in *cis* over short and long genomic distances (Shlyueva et al., 2014; Schoenfelder and Fraser, 2019; Furlong and Levine, 2018), while others show remarkable specificity, regulating only one of its neighboring promoters or even skipping one or more promoters to activate more distal ones. In part, 3D folding and compartmentalization of the chromatin fiber help to establish this specificity, by facilitating certain enhancer-promoter contacts and curbing others (Lupiáñez et al., 2015; Schoenfelder and Fraser, 2019; Furlong and Levine, 2018).

However, there is also substantial evidence that biochemical (in)compatibility between CREs and promoters contributes to the specificity of their regulatory interactions. This is akin to a lock-and-key mechanism: proteins bound to the CRE and the promoter must be compatible to form a productive complex. Examples of such intrinsic selectivity have been documented particularly in *Drosophila* and in some instances could be attributed to a specific sequence motif in the promoter (Li and Noll, 1994; Merli et al., 1996; Butler and Kadonaga, 2001; Juven-Gershon et al., 2008; Kwon et al., 2009). Data obtained with massively parallel reporter assays (MPRAs) in *Drosophila* cells have suggested a general separation of enhancer-promoter compatibility into housekeeping and tissue-specific classes (Arnold et al., 2017). Some of this specificity may be determined by the recruitment of co-factors (Haberle et al., 2019). However, a thorough understanding of the underlying mechanisms is still lacking.

While several studies of individual enhancer-promoter combinations indicate that biochemical compatibility also plays a role in mammals (e.g., Bertolino and Singh, 2002; Vakoc et al., 2005; Jing et al., 2008; Deng et al., 2012; Chang et al., 2004), systematic studies of this mechanism have so far been lacking in mouse or human cells. Thus, it is still unknown how widespread such intrinsic compatibility is in mammalian cells, and what drives this compatibility.

In order to address this issue, we systematically tested the compatibility of thousands of combinations of candidate CREs (cCREs) and promoters using MPRAs. We used plasmid-based MPRAs because they are highly scalable (Inoue and Ahituv,



(legend on next page)

2015; van Arensbergen et al., 2019; Sahu et al., 2021) and because episomal plasmids provide an isolated context that minimizes confounding effects of variable chromatin environments and differences in 3D folding. However, so far, MPRAs have mostly been used to assess the activity of single elements, either as enhancers or as promoters (Inoue and Ahituv, 2015; van Arensbergen et al., 2017; Arnold et al., 2013; Klein et al., 2020; Davis et al., 2020; King et al., 2020), except for one recent study that tested combinations of synthetic elements (Sahu et al., 2021). To be able to dissect compatibility between enhancers and promoters systematically, we designed cloning strategies that allowed us to test thousands of pairwise cCRE-promoter combinations in different positions and orientations in a reporter plasmid.

As models, we chose three genomic loci of 1–3 Mb in mouse embryonic stem cells (mESCs). From these loci, which each encompass ~20 genes, we tested a large fraction of all possible pairwise cCRE-promoter (cCRE-P) combinations. We found that more than half of the active cCREs exhibit significant selectivity for specific subsets of promoters. We dissected some of the underlying sequence determinants. Furthermore, we provide evidence suggesting that 3D folding and intrinsic compatibility are independent mechanisms. Our experimental strategy and datasets provide insights into the logic and mechanisms of cCRE-promoter specificity.

RESULTS

Experimental design

To maximize the probability of testing biologically relevant enhancer-promoter pairs, we combined cCREs and promoters coming from the same region in the genome. We selected three loci of 1–3 Mb in size, each roughly centered around a gene (*Nanog*, *Tfcp2l1*, or *Klf2*) that is key to the control of pluripotency of mESCs. The regulation of these genes is still incompletely understood. In addition, each locus contains about 20 other genes (Figures 1A–1C).

For promoters in the regions of interest we included approximately the −350- to +50-bp segments around all GENCODE-annotated (Frankish et al., 2019) transcription start sites (TSSs). The choice to focus on the range from −350 to +50 bp was motivated by our previous study of human promoters, which indicated that most of the relevant information for promoter function is generally contained within this range (van Arensbergen et al., 2017). This definition of promoters is longer than that of core promoters (which are usually only ~100-bp long) as was used in most previous enhancer reporter assays (Ohler et al., 2002; Haberle et al., 2019; Inoue and Ahituv, 2015; Klein et al., 2020; Davis et al., 2020; King et al., 2020; Sahu et al., 2021). We considered this to be important because the extra regulatory information contained in those additional sequences may be relevant for interactions of the promoters with CREs.

Compared with promoters, the annotation of cCREs is much less accurate. However, most cCREs are centered around DNase I hypersensitive sites (DHSs) (Groudine et al., 1983; Joshi et al., 2015; ENCODE Project Consortium, 2012). We therefore selected fragments of ~400 bp centered around all detected DHS peaks in each locus (Figures 1A–1C). This definition of cCREs within the range of typical enhancer definitions (Long et al., 2016). Some authors consider enhancers combinations of multiple DHSs or longer stretches of DNA sequences. However, other studies have shown that the activity of these long enhancers can be reproduced by shorter versions of ~500 bp (Barakat et al., 2018; Agrawal et al., 2021). Coordinates of all tested genomic fragments are provided in Data S1.

We designed two MPRA variants to test many cCRE-P combinations (Figures 1D and 1E). In the first variant, which we will refer to as Upstream assay, we obtained 82–192 individual cCREs and 18–25 Ps per locus by PCR amplification (Table S1). We pooled all of these fragments and randomly ligated them to form dimer fragments, which we then cloned *en masse* into a reporter vector, "upstream" of a randomly barcoded transcription unit that lacked a promoter itself. This resulted into highly complex libraries of cCRE-P, cCRE-cCRE, P-P, and P-cCRE pairs, with each individual element in two possible orientations. We then sequenced the libraries to identify the paired fragments, their orientations in the reporter vector, and their linked barcodes. Owing to the simple random ligation step, libraries with tens of thousands of cCRE-P combinations can be obtained with this approach (Tables S1 and S2). Here, we focus on the analysis of cCRE-P pairs, but data from all other configurations are also provided as Data S2.

In a second and complementary approach, we constructed a library in which the cCREs are placed "downstream" of the reporter gene, i.e., separated ~1 kb from the promoter (Figure 1E). This was done in two steps: we first cloned a selection of 10 promoters upstream of the barcoded transcription unit, resulting in a set of reporters with different promoters. Next, we inserted a pool of cCREs into this set, downstream of the barcoded reporter unit and in both possible orientations. We will refer to the assays done with the resulting library as Downstream assay. Due to the two-step cloning protocol, the Downstream assay is less scalable than the Upstream assay but nevertheless allows for testing of hundreds of cCRE-P combinations (Table S1).

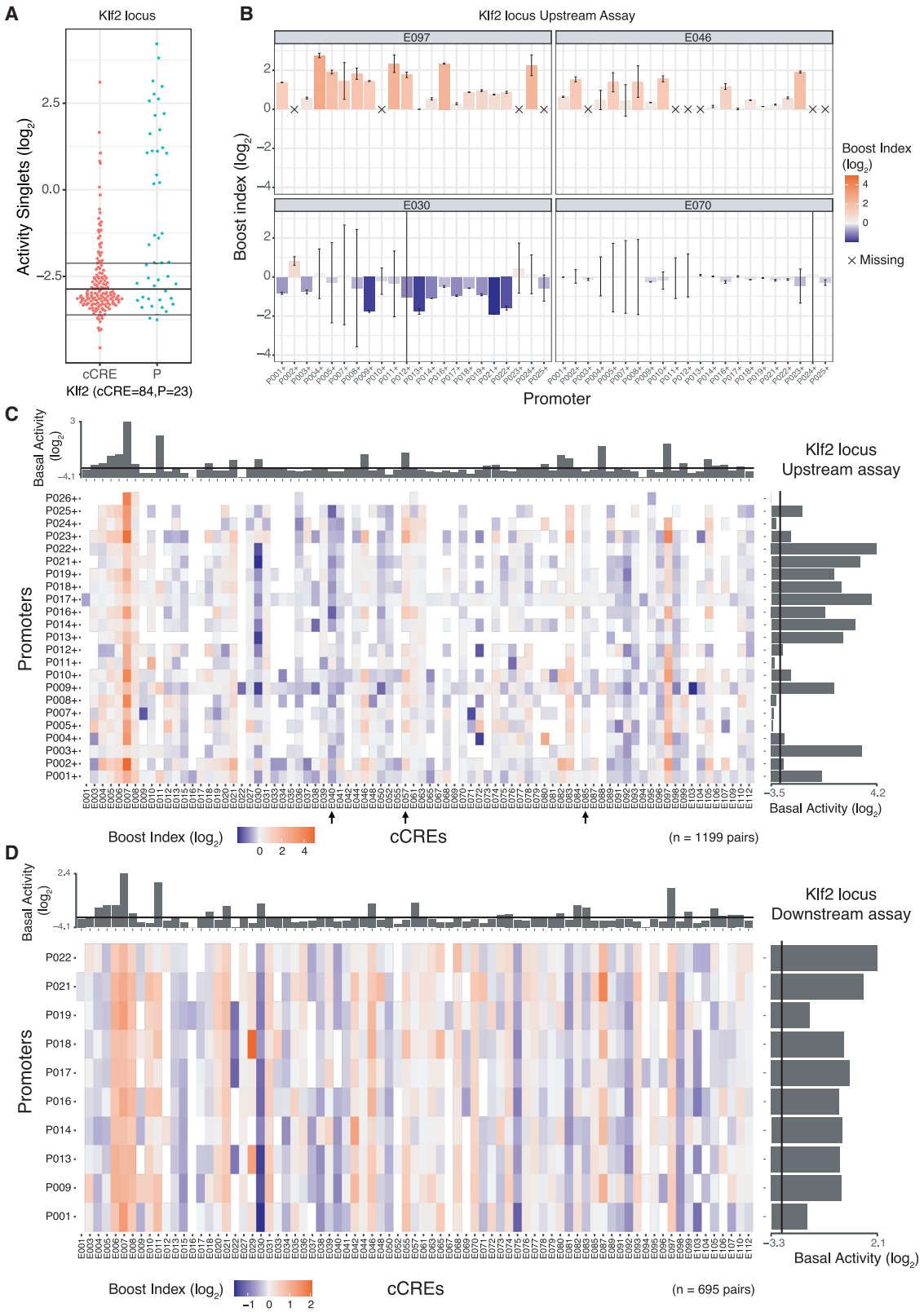
We used all P and cCRE DNA fragments from each of the three loci in separate Upstream assays, whereas we focused on ten promoters and all cCREs from the *Klf2* locus in the Downstream

Figure 1. Regulatory element selection and library construction

(A–C) Representations of *Nanog*, *Tfcp2l1*, and *Klf2* loci, respectively. In (C), the zoom-in displays a DNase I sensitivity track (Joshi et al., 2015) where peaks overlap with cCREs.

(D) Cloning strategy for the Upstream assay. cCREs and promoters were amplified by PCR from genomic DNA and pooled. Fragments in this pool were then randomly ligated to generate duplets. Singlets and duplets were cloned into the same barcoded vector to generate two libraries per locus, a singlet library and a combinatorial library.

(E) Cloning strategy for the Downstream assay. The singlet pool from the *Klf2* locus was cloned into ten vectors, each of them carrying a different promoter. The resulting ten sub-libraries were combined into one Downstream assay library.



(legend on next page)

assay. Table S1 provides summary statistics of the individual library compositions. Due to the random nature of the combinatorial cloning, we did not recover all possible pairs. Nevertheless, in the three Upstream assays combined we tested a total of 10,678 cCRE-P pairs, or 3,747 pairs if we do not take orientations into account. For the Downstream assay these numbers were 1,364 and 752, respectively. From the *Klf2* locus 847 and 676 pairs, respectively, overlapped between the Upstream and Downstream assay. As references, we also inserted each P and cCRE individually (i.e., unpaired) in the upstream position.

Boost indices estimate promoter-specific activity of cCREs

We then transiently transfected each of these libraries into mESCs. 24 h after transfection, we collected mRNA from the cells and counted the transcribed barcodes by reverse transcription followed by PCR amplification and high-throughput sequencing. In parallel, barcodes were counted in the plasmid libraries. For each barcode, we then normalized the counts in cDNA over the counts detected in the plasmid DNA. Further data processing is described in the STAR Methods. We performed 3 biological replicates per library, which correlated with an average Pearson r = 0.87 (0.83–0.90) for the Upstream assay and r = 0.98 (0.98–0.99) for the Downstream assay. (Figures S1A–S1C)

We first analyzed the transcriptional activities of all singlet (unpaired) P and cCREs in the upstream position. For promoters, these basal activities varied over a ~100-fold dynamic range (Figures 2A and S2A). Of all cCREs, 40.4% showed detectable transcriptional activity in the upstream position without any P (Figures 2A and S2A). Such autonomous transcriptional activity is a frequently observed property of enhancers (Djebali et al., 2012; Andersson et al., 2014; van Arensbergen et al., 2017), and hence, these elements are likely to be enhancers. For a few cCREs this activity was as high as some of the strongest promoters, suggesting that they may in fact be unannotated promoters or very strong enhancers.

We then determined the ability of each cCRE to alter the activity of each linked P. For this, we calculated a "boost index" for each cCRE-P pair, defined as the $\log_2$-fold change in activity of the cCRE-P pair compared with the promoter element alone. Unexpectedly, 20 negative controls that we included in the *Klf2* libraries, consisting of randomly generated DNA sequences of similar size and G/C content as the cCREs, showed a modestly negative boost index (median value −0.45 when inserted upstream) (Figure S1D). This is possibly because lengthening of the reporter constructs alters the topology, supercoiling, transfection efficiency, or a combination of these parameters. We therefore corrected all cCRE-P boost indices for this non-specific negative bias (see STAR Methods). After this correction the negative controls had a marginal residual bias (median $\log_2$ value −0.19), which we deemed acceptable (Figure S1D).

Identification of activating and repressive cCREs

For each of the three genomic loci, the matrix of corrected boost indices shows a wide diversity of patterns across the cCREs. We observed this both in the Upstream and Downstream assays (Figures 2B–2D and S2B–S2D). For example, in the *Klf2* locus Upstream assay, cCRE E097 activates most of the tested promoters, while E046 (Figure 2B) and E057 (arrow in Figure 2C) only activate a distinct subset of promoters. Several elements are primarily acting as repressors (e.g., E030 [Figure 2B] and E040, [arrow in Figure 2C]), and some seem neither activating nor repressive (e.g., E070 [Figure 2B] and E085 [arrow in Figure 2C]).

We broadly classified the cCREs according to their overall effects on the linked promoters (Figure S3A). In the Upstream assays, 21% of cCREs showed positive boost indices that were significantly higher than the rest of cCREs across all tested promoters, indicating that they can act as enhancer elements. About 17% of the cCREs showed negative boost indices significantly below the rest of cCREs and hence are putative repressor elements. For the remaining 62% of cCREs, the boost indices across their linked promoters were not significantly higher or lower than the rest; these "ambiguous" elements either have no regulatory effects at all, or they have a mixed repressive/activating/inactive effect that depends on the linked P (see below).

We were somewhat surprised to identify similar numbers of putative enhancers and repressors because most annotated cCREs in mammalian genomes are predicted to be enhancers rather than repressive elements (ENCODE Project Consortium et al., 2020; ENCODE Project Consortium, 2012). In some cases, this repression may be underestimated in our analysis, as the estimates of negative boost indices for lowly active promoters are less reliable due to the higher noise-to-mean ratios at low expression levels (Figure S3B).

For activating elements, the boost indices varied in part according to the basal activities of the cCRE and promoters. Strong boosting occurred primarily at promoters with low basal activities, while highly active promoters were more difficult to boost (FigureS3C). This suggests a saturation effect, or it could indicate that promoters with high basal activity are less dependent on distal enhancers. For cCREs, their basal activity is generally a strong positive predictor of their enhancer potency (Figure S3D). However, exceptions to this rule occur, as some

Figure 2. Singlet and combinatorial activities of cCREs and promoters from the *Klf2* locus

(A) Transcription activities of singlet cCREs and promoters. Each dot represents the mean activity of one singlet. Horizontal lines represent the average background activity of empty vectors (black line) plus or minus two standard deviations (gray lines). Elements with activities more than two standard deviations above the average background signal are defined as active.

(B) Examples of Upstream assay cCRE-P combinations for cCREs E097, E046, E030, and E070 of the *Klf2* locus. Bar plots represent the mean boost index of each combination, vertical lines represent the standard deviations. Crosses mark missing data.

(C and D) Boost index matrices of cCRE-P combinations from the *Klf2* locus according to Upstream (C) and Downstream (D) assays. White tiles indicate missing data. Bar plots on the right and top of each panel show basal activities of each tested P or cCRE, respectively, with the black line indicating the background activity of the empty vector. All data are averages of 3 independent biological replicates.

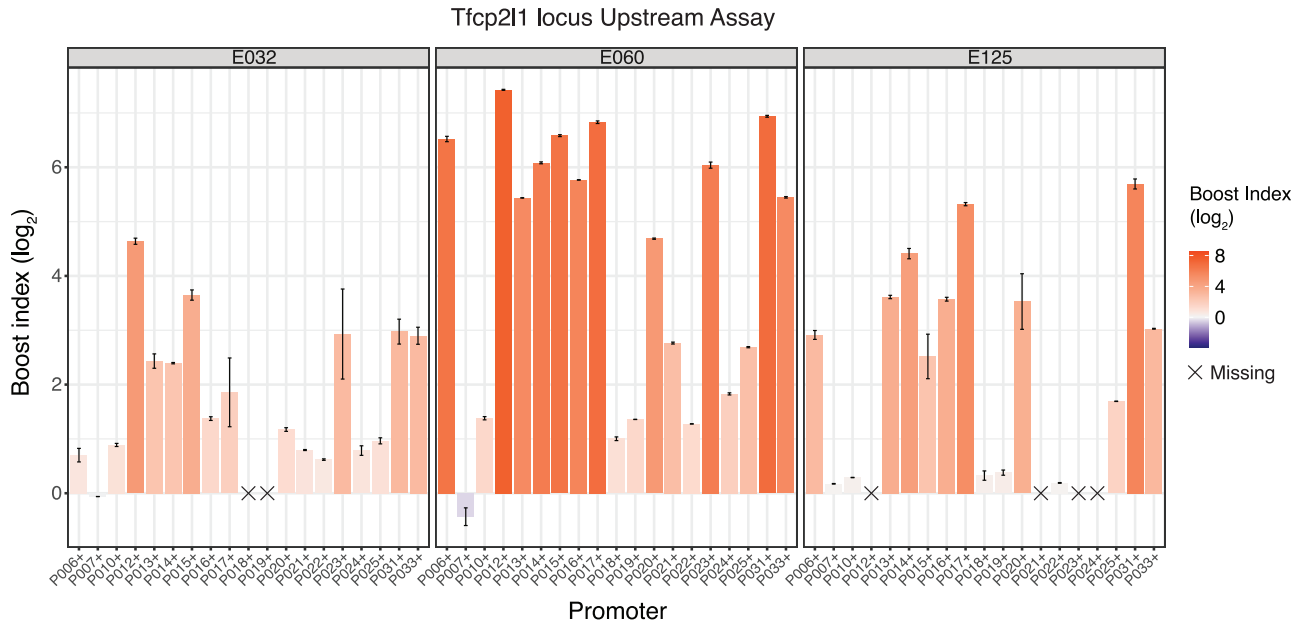


Figure 3. Examples of selective cCREs from the *Tfcp2l1* locus

Boost indices obtained in the Upstream assay are shown for cCRE-P combinations of cCREs E032, E060, E125, of the *Tfcp2l1* locus. Bar plots indicate the mean boost index of each combination, vertical lines indicate standard deviations. All data are averages of 3 independent biological replicates.

cCRE-P pairs show high boost indices even though the basal activity of the cCRE is low (Figure S3D, upper left quadrant).

cCRE effects are partially orientation and position independent

Next, we asked whether the ability of cCREs to regulate the linked promoters was generally independent of their orientation and position. This was originally posited for enhancers (Banerji et al., 1981) and in some cases also reported for repressive elements (Segert et al., 2021). Indeed, in the Upstream assays, we found a general positive correlation of the boost indices between the two orientations of the cCREs (Pearson's r = 0.68) (Figure S4A). These results are similar to those recently obtained with a minimal core promoter (Klein et al., 2020). In the Downstream assay, the correlation between orientations was somewhat lower (Pearson's r = 0.47) (Figure S4B). This may be due to the lower dynamic range of the Downstream assay data (Figure S1C). To simplify, for all other analyses we combined the boost indices of + and − orientations of the cCREs by averaging.

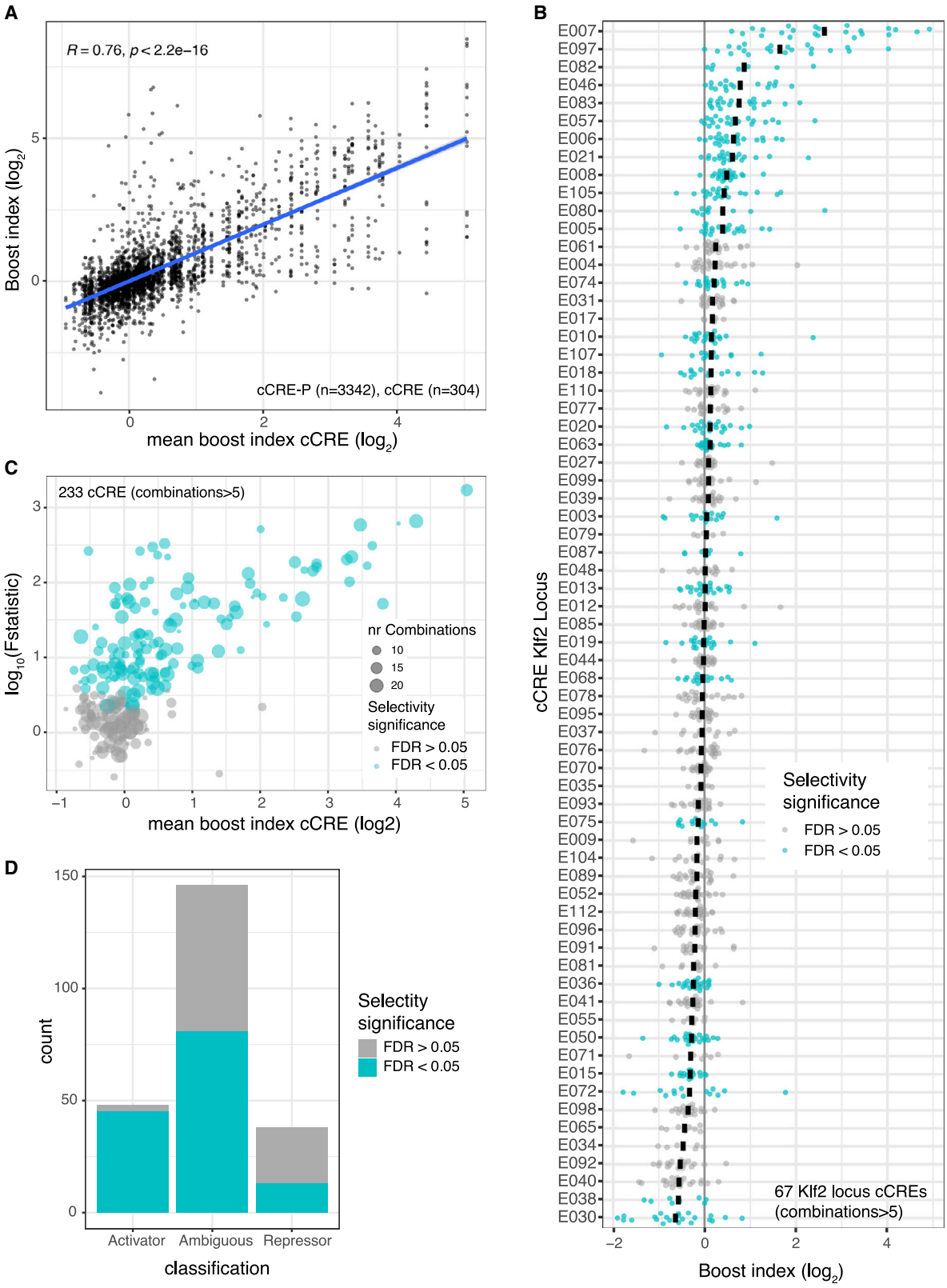
We then investigated the degree of position-independence, by comparing the overlapping P-cCRE pairs from the *Klf2* locus Downstream and Upstream assays. This showed an overall Pearson correlation of 0.64 (Figure S4C). We conclude that repressive and activating effects of cCREs are substantially but not completely position independent, at least for the ten tested promoters from the *Klf2* locus.

Extensive selectivity of cCREs for promoters

Visual inspection of the boost index matrices suggested that some cCREs alter the expression of most promoters to similar degrees, while others selectively alter the expression of a subset of promoters. In addition to the examples in Figure 2B from the *Klf2* locus, strikingly specific promoter responses to some cCREs are illustrated for the *Tfcp2l1* locus in Figure 3. For example, E060, which forms part of an annotated super-enhancer (Khan and Zhang, 2016), activates most of the tested promoters, but with boost indices that can vary >50-fold between promoters. Two other remarkable examples from the *Tfcp2l1* locus are E032 and E0125, which each show different degrees of specificity, between low or no activation of some promoters and very strong activation of others. Much broader specificity is observed for E064, E073, E074, and E090 from the *Nanog* locus, which are part of previously identified super-enhancers (Blinka et al., 2016) (Figure S2D).

We investigated the degrees of selectivity more systematically. Figures 4A and 4B depicts the distribution of the boost indices for each cCRE. Clearly, some cCREs have a much broader range of boost indices than others. We used an ANOVA approach with Welch F test to systematically identify cCREs for which the variance of boost indices was larger than could be explained by experimental noise (see STAR Methods). Strikingly, out of 233 cCREs with more than 5 tested cCRE-P combinations, a total of 139 (59.9%) (Figures 4B and 4C) showed significant unexplained variance at an estimated false discovery rate (FDR) cutoff of 5%. Thus, at least throughout the three loci that we tested, cCRE-P selectivity is widespread, ranging from strong specificity for a few promoters to low specificity as seen in quantitative differences in the regulation of a broad set of promoters.

Intersection of the ANOVA-based classification of selective/unselective cCREs with the above broad classification into enhancers and repressors indicates that almost all (94%) general enhancer elements exhibit significant P selectivity. In contrast,



(legend on next page)

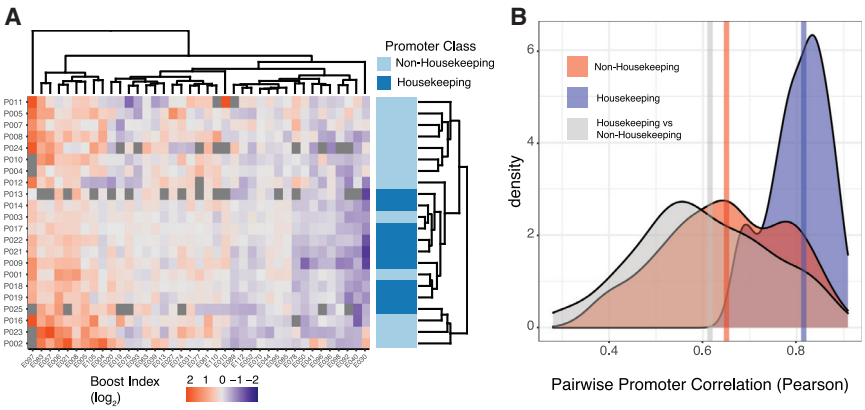


Figure 5. Housekeeping promoters show a distinct pattern of cCRE compatibility

(A) Hierarchical clustering of the Upstream assay boosting matrix of the *Klf2* locus. In order to facilitate hierarchical clustering, the matrix has been restricted to almost complete cases (cCREs > 15 combinations).

(B) Density plot of pairwise Pearson correlation coefficients of the boost indices of *Klf2* locus promoters classified as either housekeeping or non-housekeeping (Hounkpe et al., 2021). Blue: correlations between all pairs of housekeeping promoters; red: all correlations between pairs of non-housekeeping promoters; gray: all correlations between one housekeeping and one non-housekeeping promoter. Vertical lines represent the median of each group. Unlike in (A), all promoters in the Upstream assay were included in this analysis. All data are averages of 3 independent biological replicates.

only 34% of the repressors are detectably biased toward a subset of promoters (Figure 4D). However, we note that this percentage may be underestimated because at low expression levels the noise levels are higher (Figure S3B). Interestingly, among the "ambiguous" cCREs, 55% are in fact selective. Such elements mostly activate or repress only very few promoters and leave most other promoters unaffected. The remainder of the ambiguous cCREs are probably not functional (e.g., E70 from the *Klf2* locus, Figure 2B). In summary, these results indicate that more than half of all tested cCREs exhibits significant preference for specific promoters.

Promoters of housekeeping and developmental genes in *Drosophila* were reported to have distinct specificities toward cCREs (Zabidi et al., 2015). To investigate whether such a dichotomy could also be observed in our data, we focused on the *Klf2* locus, which has roughly equal numbers of housekeeping and non-housekeeping promoters (Hounkpe et al., 2021) (the *Tfcp2l1* and *Nanog* loci have only three and zero housekeeping genes, respectively). Indeed, hierarchical clustering of the boost index matrix showed a rough separation of the two classes of promoters (Figure 5A). However, this is largely due to the highly similar cCRE specificities among the housekeeping promoters, whereas the cCRE specificities of the non-housekeeping promoters are much more diverse and generally as distinct from each other as from the housekeeping promoters (Figure 5B). To test whether a housekeeping versus non-housekeeping dichotomy may largely explain our identification of cCREs with significant selectivity (Figures 4B and 4C), we repeated this analysis after removing all housekeeping promoters. This yielded highly similar results (123 of 221 cCREs are significantly selective at 5% FDR cutoff, Figure S5). We conclude that housekeeping promoters may be similarly regulated, but cCRE selectivity goes beyond a simple distinction between housekeeping and non-housekeeping promoters.

Similar results with an independent MPRA dataset from human cells

In a parallel study, Bergman et al. (2022) conducted similar large-scale cCRE-P combinatorial MPRAs in human K562 cells. Based on mathematical modeling of their data, the authors proposed that intrinsic cCRE and P activities generally combine multiplicatively to determine reporter activity. To test whether cCRE-P selectivity could also be detected in these K562 data, we subjected these data to our ANOVA with Welch F test. Indeed, for the majority of cCREs (90.46% at 0.1% FDR cutoff) the variance of boost indices was significantly larger than could be explained by experimental noise (Figures S6A and S6B). We conclude that the K562 data also point to a significant layer of cCRE-P selectivity.

Selectivity may be mediated by combinations of multiple TF motifs

Taken together, these results point to a broad spectrum of cCRE specificities for promoters, ranging from largely indiscriminate to highly selective. We searched in our dataset for sequence motifs that may account for these effects, focusing on binding motifs of transcription factors (TFs) that are expressed in mESCs.

We first searched for TF motifs in the cCREs that correlate with boost indices across all promoters. This yielded several dozens of TFs that are candidate activators or repressors (Figure S7A). Several of these, such as Sox2, Nanog, ETV4, and GABPA are

Figure 4. Promoter selectivity of cCREs

(A) Plot showing the broad diversity of boost indices of many cCREs. Data are from Upstream assays of *Klf2*, *Nanog*, and *Tfcp2l1* loci combined. Vertical axis indicates boost indices of all tested cCRE-P pairs, which are horizontally ordered by the mean boost index of each cCRE. R is pearson correlation and p its corresponding p-value.

(B) Boost index distributions for each cCRE from the *Klf2* locus (Upstream assay). Each dot represents one cCRE-P combination; black bar represents the mean. Turquoise coloring marks cCREs that have a larger variance of their boost indices than may be expected based on experimental noise, according to the Welch F test after multiple hypothesis correction (5% FDR cutoff).

(C) Summary of Welch F test selectivity analysis results for all cCREs from the three loci with more than 5 cCRE-P combinations. Each dot represents one cCRE; the size of the dots indicates the number of cCRE-P pairs. Significantly selective cCREs (5% FDR cutoff) are highlighted in turquoise.

(D) Proportion of significantly selective (turquoise) cCRE in the three categories as shown in Figure S3A. All data are averages of 3 independent biological replicates.

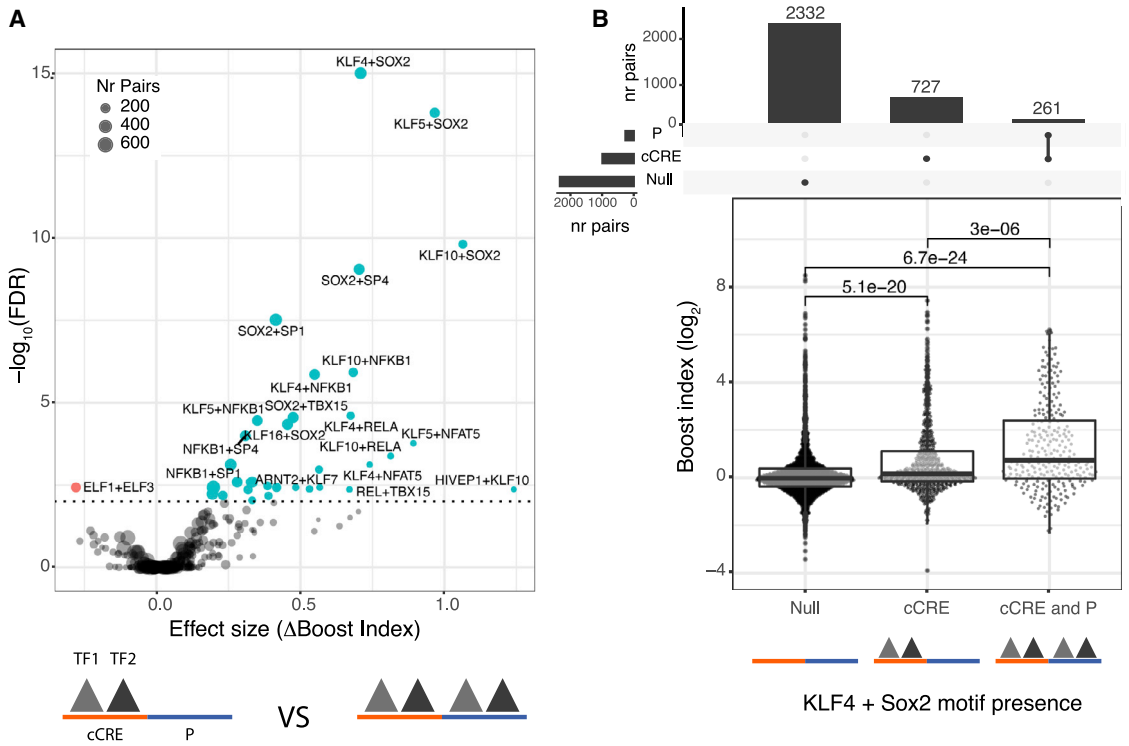


Figure 6. Association of TF motif duos with higher boost indices

(A) Results of TF survey for self-compatible TF motif duos. TF motif duos associated with higher or lower boost indices at a 1% FDR cutoff are highlighted.

(B) Association of Sox2 + Klf4 motifs at both cCRE and P with higher boost indices. cCRE-P combinations are split into 3 groups according to presence or absence of Sox2 + Klf4 motifs both at the cCRE and the promoter, or only the cCRE. Numbers at the top of horizontal brackets are the p values obtained from comparing the different groups boost index distributions using a Wilcoxon rank-sum test. Boxplots represent median and interquartile ranges. Bar plots at the top represent the number of combinations in each group.

known key regulators in mESC cells (Kim et al., 2008; Akagi et al., 2015; Kinoshita et al., 2007). These TFs may broadly contribute to enhancer activity.

Next, we searched for motifs associated with cCRE-P selectivity. We reasoned that selectivity may be due to certain combinations of TFs bound to cCRE and P. First, we asked whether for any TF the simultaneous presence of its motif at cCRE and P correlated with boost indices (Figure S7B). This only yielded a weak association of FOXO motifs (at a 5% FDR cutoff). Possibly this is due to FOXO1, a known regulator in mESCs (Zhang et al., 2011). We then asked if selectivity may be mediated by multiple TFs rather than single TFs. For this purpose, we took the TF motifs associated with enhancer activity with effect sizes >0.1 (n = 66) and searched for combinations of motifs that would be associated with higher boost indices if present at both the cCRE and the P (Figures 6A and 6B). This yielded a few dozen stronger associations (at a 1% FDR cutoff). Some of these associations may be redundant either because of motif similarity or because of motif co-occurrence. For example, the 5 associations between Sox2 and Klf motifs may represent the Klf4-Sox2 pair (Figure 6B), which are known to cooperate in mESCs (Wei et al., 2009). These results indicate that selectivity may be mediated by combinations of multiple TF motifs. Our dataset does not provide sufficient statistical power for an exhaustive search of such combinations.

Chromatin looping is independent of compatibility

Finally, we considered that certain pairs of cCREs and promoters frequently contact each other in the nucleus, as is indicated by focal or stripe-like enrichment patterns in high-resolution Hi-C maps (Bonev et al., 2017; Hsieh et al., 2020). While long-range contacts are irrelevant in our MPRAs because the tested elements are directly linked, we asked whether such physical contacts in the native genomic context are related to the selectivity of cCREs for certain promoters according to our MPRAs. We considered two models. In one model, the biochemical interactions that underlie cCRE-P selectivity may promote or stabilize cCRE-P looping interactions. Alternatively, looping interactions and cCRE-P selectivity may be independent aspects of cCRE-P interplay that each work by different mechanisms.

To discriminate between these two models, we investigated whether the boost indices of cCRE-P pairs correlate with their contact frequencies in micro-C, a high-resolution variant of Hi-C (Hsieh et al., 2020). Remarkably, we found no correlation between these two quantities (Figure 7A). We also found an extremely weak, although statistically significant, correlation between higher boost indices and longer linear distances of cCRE-P pairs along the genome (Figure 7B).

We conclude that cCRE-P contacts in the nucleus may be independent of their functional compatibility as detected in our

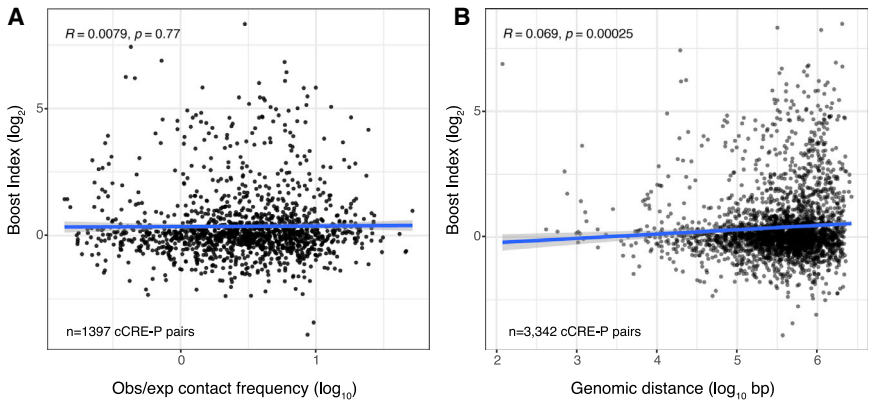


Figure 7. Relationship between 3D organization and boost indices

Absent or very weak correlation between boost indices and (A) contact frequencies according to micro-C (Hsieh et al., 2020) or (B) linear genomic distance, for all cCRE-P pairs from the three loci combined. All boost index data are averages of 3 independent biological replicates. R is pearson correlation and p its corresponding p-value.

reporter assays, raising the interesting possibility that chromatin looping and compatibility are two orthogonal mechanisms of gene regulation.

DISCUSSION

Only a few other studies have so far attempted to analyze cCRE-P compatibility systematically. An early survey of 27 cCRE-P combinations in human cells did not find evidence for specificity (Kermekchiev et al., 1991), but the assay employed may have been insufficiently quantitative, and the choice of tested elements may have been biased. In contrast, testing of ~200 cCRE-P pairs in zebrafish pointed to extensive specificity (Gehrig et al., 2009). An MPRA study in *Drosophila* cells using seven different promoters and genome-wide cCREs suggested that cCRE-P specificity broadly separates between housekeeping and tissue-specific promoters (Zabidi et al., 2015). Parallel to our work, a MPRA study was reported of thousands of cCRE-P combinations in a human cell line (Bergman et al., 2022).

Our results reveal a broad spectrum of specificities: some cCREs are promiscuous, others are highly specific for certain promoters, and in many instances the specificity is quantitative rather than qualitative. By statistical analysis, we found that more than half of the cCREs exhibit a degree of specificity that cannot be explained by experimental noise. We also found evidence for such specificity in the MPRA data of Bergman et al. (2022). Although this study focused on the observation that a substantial fraction of the variance in reporter activity can be explained by a multiplicative combination of enhancer and promoter activities, both datasets indicate that CRE-P selectivity as well as intrinsic enhancer and promoter activities contribute to reporter expression (see also Bergman et al. [2022]). We suggest that such a general multiplicative rule and a more complex grammar of enhancer-promoter specificities are two sides of the same coin of gene regulation.

It is likely that cCRE-P compatibility is governed by a complex grammar of TF combinations. Underlying this grammar may be a diversity of molecular mechanisms, including direct and indirect TF-TF interactions (e.g., Wei et al. [2009], local concentration of activating factors [Davis et al., 2020; Tak et al., 2021], or functional bridging by co-factors [El Khattabi et al., 2019; Haberle et al., 2019]). Due to the complexity of this grammar, its elucidation may require much larger cCRE-P combinatorial datasets than generated here, as well as systematic mutational analysis (Fuqua et al., 2020; Kircher et al., 2019) of individual cCRE-P combinations. Nevertheless, our statistical analysis highlights several candidate combinations of TF motifs that may contribute to the compatibility of some cCRE-P pairs.

Our data indicate that some of the cCREs tested may be repressive elements rather than enhancers even though they were selected from DHSs. This is similar to a recent screen of cCREs in human cells, which identified a large set of candidate repressive elements (Pang and Snyder, 2020) and to another screen in *Drosophila* (Gisselbrecht et al., 2020). It will be interesting to further explore the physiological regulatory role of these elements.

Surprisingly, we found that the boost indices of cCRE-P pairs generally do not correlate with their contact frequencies in the native chromatin context. This suggests that 3D genome organization and compatibility are regulated by different mechanisms. We envision that compatibility and 3D organization may be two independent layers necessary for correct selective gene regulation: 3D organization such as the formation of chromatin loops and compartments may determine whether CREs and promoters are able to interact, while compatibility may determine whether such an interaction is functional, i.e., gives rise to a change in P activity.

Limitations of the study

Our current data were generated with transiently transfected plasmids. Advantages of this "reductionist" approach are that it largely eliminates possible confounding effects of chromatin packaging and 3D folding and that thousands of cCRE-P combinations could be tested. However, MPRAs are intrinsically artificial, as the tested elements are taken out of their natural sequence context, may be too small or incorrectly spaced, or require a natural chromatin context to function properly. Thus, further studies are needed to verify and analyze the impact of the observed specificities in the native genomic context.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY

- Lead contact
- Materials availability
- Data and code availability

- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Cell culture
- METHOD DETAILS
  - Selection of cCREs and promoters
  - Upstream assay library generation
  - Downstream assay library generation
  - Inverse PCR and sequencing to link inserted elements to barcodes
  - Libraries transfection
  - RNA extraction and cDNA sequencing
  - Plasmid DNA (pDNA) barcode sequencing
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Linking barcodes to element singlets or duplets
  - Pre-Processing of cDNA and pDNA reads
  - Post processing of cDNA and pDNA counts
  - Calculation of boost indices
  - Identification of activating and repressive cCREs
  - Analysis of selectivity
  - Analysis of Houskeeping and non-housekeeping promoter selectivity
  - Analysis of selectivity on K562 ExP data
  - TF motif Survey
  - Micro-C data correlation

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.molcel.2022.04.009.

ACKNOWLEDGMENTS

We thank Tao Chen for initial input on the study design, the NKI Genomics, Robotics, and Research High Performing Computing facilities for technical support, Barak Cohen and his lab for insightful discussions, and Gioacchino Natoli and his lab for providing the TF motif database. We thank the J. Engreitz lab for constructive discussions and for sharing their data prior to publication. Supported by ERC advanced grant 694466 (B.v.S.) and Swiss National Science Foundation postdoctoral fellowship P2EZP3_165206 (F.C.). Oncode is partly funded by the Dutch Cancer Society (KWF).

AUTHOR CONTRIBUTIONS

M.M.-A., F.C., and B.v.S. designed the study. M.M.A. and F.C. developed computational methods and performed analyses. M.M.A. and J.v.A. developed experimental methods. M.M.A. performed experiments. B.v.S. and M.M.A. wrote the manuscript, with input from F.C. and J.v.A. B.v.S. supervised the study.

DECLARATION OF INTERESTS

J.v.A. is founder of Gen-X B.V. and Annogen B.V. F.C. is a co-founder of enGene Statistics GmbH. B.v.S. is member of the advisory board of Molecular Cell.

Received: October 29, 2021
Revised: February 17, 2022
Accepted: April 5, 2022
Published: May 19, 2022

REFERENCES

Agrawal, P., Blinka, S., Pulakanti, K., Reimer, M.H., Jr., Stelloh, C., Meyer, A.E., and Rao, S. (2021). Genome editing demonstrates that the -5 kb Nanog enhancer regulates Nanog expression by modulating RNAPII initiation and/or recruitment. J. Biol. Chem. *296*, 100189.

Akagi, T., Kuure, S., Uranishi, K., Koide, H., Costantini, F., and Yokota, T. (2015). ETS-related transcription factors ETV4 and ETV5 are involved in proliferation and induction of differentiation-associated genes in embryonic stem (ES) cells. J. Biol. Chem. *290*, 22460–22473.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461.

Arnold, C.D., Gerlach, D., Stelzer, C., Boryń, Ł.M., Rath, M., and Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science *339*, 1074–1077.

Arnold, C.D., Zabidi, M.A., Pagani, M., Rath, M., Schernhuber, K., Kazmar, T., and Stark, A. (2017). Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution. Nat. Biotechnol. *35*, 136–144.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. *37*, W202–W208.

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell *27*, 299–308.

Barakat, T.S., Halbritter, F., Zhang, M., Rendeiro, A.F., Perenthaler, E., Bock, C., and Chambers, I. (2018). Functional dissection of the enhancer repertoire in human embryonic stem cells. Cell Stem Cell *23*, 276. e8–288.e8.

Bergman, D.T., Jones, T.R., Liu, V., Siraj, L., Kang, H.Y., Nasser, J., Kane, M., Nguyen, T.H., Grossman, S.R., Fulco, C.P., et al. (2022). Compatibility rules of human enhancer and promoter sequences. Nature. Published online May 19, 2022. https://doi.org/10.1038/s41586-022-04877-w.

Bertolino, E., and Singh, H. (2002). POU/TBP cooperativity: a mechanism for enhancer action from a distance. Mol. Cell *10*, 397–407.

Blinka, S., Reimer, M.H., Jr., Pulakanti, K., and Rao, S. (2016). Super-enhancers at the Nanog locus differentially regulate neighboring pluripotency-associated genes. Cell Rep. *17*, 19–28.

Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G.L., Lubling, Y., Xu, X., Lv, X., Hugnot, J.P., Tanay, A., and Cavalli, G. (2017). Multiscale 3D genome rewiring during mouse neural development. Cell *171*, 557. e24–572.e24.

Butler, J.E., and Kadonaga, J.T. (2001). Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. Genes Dev. *15*, 2515–2519.

Chang, T.H., Primig, M., Hadchouel, J., Tajbakhsh, S., Rocancourt, D., Fernandez, A., Kappler, R., Scherthan, H., and Buckingham, M. (2004). An enhancer directs differential expression of the linked Mrf4 and Myf5 myogenic regulatory genes in the mouse. Dev. Biol. *269*, 595–608.

Davis, J.E., Insigne, K.D., Jones, E.M., Hastings, Q.A., Boldridge, W.C., and Kosuri, S. (2020). Dissection of c-AMP response element architecture by using genomic and episomal massively parallel reporter assays. Cell Syst. *11*, 75. e7–85.e7.

Deng, W., Lee, J., Wang, H., Miller, J., Reik, A., Gregory, P.D., Dean, A., and Blobel, G.A. (2012). Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. Cell *149*, 1233–1244.

Diaferia, G.R., Balestrieri, C., Prosperini, E., Nicoli, P., Spaggiari, P., Zerbi, A., and Natoli, G. (2016). Dissection of transcriptional and *cis*-regulatory control of differentiation in human pancreatic cancer. EMBO J. *35*, 595–617.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. Nature *489*, 101–108.

El Khattabi, L., Zhao, H., Kalchschmidt, J., Young, N., Jung, S., Van Blerkom, P., Kieffer-Kwon, P., Kieffer-Kwon, K.R., Park, S., Wang, X., et al. (2019). A

pliable mediator acts as a functional rather than an architectural bridge between promoters and enhancers. Cell *178*, 1145. e20–1158.e20.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

ENCODE Project Consortium, Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699–710.

Farley, E.K., Olson, K.M., and Levine, M.S. (2015). Regulatory principles governing tissue specificity of developmental enhancers. Cold Spring Harb. Symp. Quant. Biol. *80*, 27–32.

Fiering, S., Epner, E., Robinson, K., Zhuang, Y., Telling, A., Hu, M., Martin, D.I., Enver, T., Ley, T.J., and Groudine, M. (1995). Targeted deletion of 5′HS2 of the murine beta-globin LCR reveals that it is not essential for proper regulation of the beta-globin locus. Genes Dev. *9*, 2203–2213.

Frankish, A., Diekhans, M., Ferreira, A.M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. *47*, D766–D773.

Fuqua, T., Jordan, J., Van Breugel, M.E., Halavatyi, A., Tischer, C., Polidoro, P., Abe, N., Tsai, A., Mann, R.S., Stern, D.L., and Crocker, J. (2020). Dense and pleiotropic regulatory information in a developmental enhancer. Nature *587*, 235–239.

Furlong, E.E.M., and Levine, M. (2018). Developmental enhancers and chromosome topology. Science *361*, 1341–1345.

Gehrig, J., Reischl, M., Kalmár, E., Ferg, M., Hadzhiev, Y., Zaucker, A., Song, C., Schindler, S., Liebel, U., and Müller, F. (2009). Automated high-throughput mapping of promoter-enhancer interactions in zebrafish embryos. Nat. Methods *6*, 911–916.

Gisselbrecht, S.S., Palagi, A., Kurland, J.V., Rogers, J.M., Ozadam, H., Zhan, Y., Dekker, J., and Bulyk, M.L. (2020). Transcriptional silencers in *Drosophila* serve a dual role as transcriptional enhancers in alternate cellular contexts. Mol. Cell *77*, 324. e8–337.e8.

Groudine, M., Kohwi-Shigematsu, T., Gelinas, R., Stamatoyannopoulos, G., and Papayannopoulou, T. (1983). Human fetal to adult hemoglobin switching: changes in chromatin structure of the beta-globin gene locus. Proc. Natl. Acad. Sci. USA *80*, 7551–7555.

Haberle, V., Arnold, C.D., Pagani, M., Rath, M., Schernhuber, K., and Stark, A. (2019). Transcriptional cofactors display specificity for distinct types of core promoters. Nature *570*, 122–126.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.

Hounkpe, B.W., Chenou, F., De Lima, F., and De Paula, E.V. (2021). HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. Nucleic Acids Res. *49*, D947–D955.

Hsieh, T.S., Cattoglio, C., Slobodyanyuk, E., Hansen, A.S., Rando, O.J., Tjian, R., and Darzacq, X. (2020). Resolving the 3D landscape of transcription-linked mammalian chromatin folding. Mol. Cell *78*, 539. e8–553.e8.

Inoue, F., and Ahituv, N. (2015). Decoding enhancers using massively parallel reporter assays. Genomics *106*, 159–164.

Jing, H., Vakoc, C.R., Ying, L., Mandat, S., Wang, H., Zheng, X., and Blobel, G.A. (2008). Exchange of GATA factors mediates transitions in looped chromatin organization at a developmentally regulated gene locus. Mol. Cell *29*, 232–242.

Joshi, O., Wang, S.Y., Kuznetsova, T., Atlasi, Y., Peng, T., Fabre, P.J., Habibi, E., Shaik, J., Saeed, S., Handoko, L., et al. (2015). Dynamic reorganization of extremely long-range promoter-promoter interactions between two states of pluripotency. Cell Stem Cell *17*, 748–757.

Juven-Gershon, T., Hsu, J.Y., and Kadonaga, J.T. (2008). Caudal, a key developmental regulator, is a DPE-specific transcriptional factor. Genes Dev. *22*, 2823–2830.

Kermekchiev, M., Pettersson, M., Matthias, P., and Schaffner, W. (1991). Every enhancer works with every promoter for all the combinations tested: could new regulatory pathways evolve by enhancer shuffling? Gene Expr. *1*, 71–81.

Khan, A., and Zhang, X. (2016). dbSUPER: a database of super-enhancers in mouse and human genome. Nucleic Acids Res. *44*, D164–D171.

Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S.H. (2008). An extended transcriptional network for pluripotency of embryonic stem cells. Cell *132*, 1049–1061.

King, D.M., Hong, C.K.Y., Shepherdson, J.L., Granas, D.M., Maricque, B.B., and Cohen, B.A. (2020). Synthetic and genomic regulatory elements reveal aspects of *cis*-regulatory grammar in mouse embryonic stem cells. eLife *9*, e41279.

Kinoshita, K., Ura, H., Akagi, T., Usuda, M., Koide, H., and Yokota, T. (2007). GABPalpha regulates Oct-3/4 expression in mouse embryonic stem cells. Biochem. Biophys. Res. Commun. *353*, 686–691.

Kircher, M., Xiong, C., Martin, B., Schubach, M., Inoue, F., Bell, R.J.A., Costello, J.F., Shendure, J., and Ahituv, N. (2019). Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. Nat. Commun. *10*, 3583.

Klein, J.C., Agarwal, V., Inoue, F., Keith, A., Martin, B., Kircher, M., Ahituv, N., and Shendure, J. (2020). A systematic evaluation of the design and context dependencies of massively parallel reporter assays. Nat. Methods *17*, 1083–1091.

Köster, J., and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. Bioinformatics *28*, 2520–2522.

Kwon, D., Mucci, D., Langlais, K.K., Americo, J.L., Devido, S.K., Cheng, Y., and Kassis, J.A. (2009). Enhancer-promoter communication at the *Drosophila* engrailed locus. Development *136*, 3067–3075.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lettice, L.A., Heaney, S.J., Purdie, L.A., Li, L., De Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E., and De Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum. Mol. Genet. *12*, 1725–1735.

Li, X., and Noll, M. (1994). Compatibility between enhancers and promoters determines the transcriptional specificity of gooseberry and gooseberry neuro in the *Drosophila* embryo. EMBO J. *13*, 400–406.

Long, H.K., Prescott, S.L., and Wysocka, J. (2016). Ever-changing landscapes: transcriptional enhancers in development and evolution. Cell *167*, 1170–1187.

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell *161*, 1012–1025.

Merli, C., Bergstrom, D.E., Cygan, J.A., and Blackman, R.K. (1996). Promoter specificity mediates the independent regulation of neighboring genes. Genes Dev. *10*, 1260–1270.

Ohler, U., Liao, G.-C., Niemann, H., and Rubin, G.M. (2002). Computational analysis of core promoters in the *Drosophila* genome. Genome Biol. *3*. RESEARCH0087.

Pang, B., and Snyder, M.P. (2020). Systematic identification of silencers in human cells. Nat. Genet. *52*, 254–263.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Robson, M.I., Ringel, A.R., and Mundlos, S. (2019). Regulatory landscaping: how enhancer-promoter communication is sculpted in 3D. Mol. Cell *74*, 1110–1122.

Rossum, G.V., and Drake, F.L. (2009). Python 3 Reference Manual (CreateSpace).

Sahu, B., Hartonen, T., Pihlajamaa, P., Wei, B., Dave, K., Zhu, F., Kaasinen, E., Lidschreiber, K., Lidschreiber, M., Daub, C.O., et al. (2021). Sequence determinants of human gene regulatory elements. Nat. Genet. *54*, 283–294.

Schoenfelder, S., and Fraser, P. (2019). Long-range enhancer-promoter contacts in gene expression control. Nat. Rev. Genet. *20*, 437–455.

Segert, J.A., Gisselbrecht, S.S., and Bulyk, M.L. (2021). Transcriptional silencers: driving gene expression with the brakes on. Trends Genet. *37*, 514–527.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. Nat. Rev. Genet. *15*, 272–286.

Tak, Y.E., Horng, J.E., Perry, N.T., Schultz, H.T., Iyer, S., Yao, Q., Zou, L.S., Aryee, M.J., Pinello, L., and Joung, J.K. (2021). Augmenting and directing long-range CRISPR-mediated activation in human cells. Nat. Methods *18*, 1075–1081.

Tuan, D., Solomon, W., Li, Q., and London, I.M. (1985). The "beta-like-globin" gene domain in human erythroid cells. Proc. Natl. Acad. Sci. USA *82*, 6384–6388.

Vakoc, C.R., Letting, D.L., Gheldof, N., Sawado, T., Bender, M.A., Groudine, M., Weiss, M.J., Dekker, J., and Blobel, G.A. (2005). Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. Mol. Cell *17*, 453–462.

Van Arensbergen, J., Van Steensel, B., and Bussemaker, H.J. (2014). In search of the determinants of enhancer-promoter interaction specificity. Trends Cell Biol. *24*, 695–702.

van Arensbergen, J., FitzPatrick, V.D., de Haas, M., Pagie, L., Sluimer, J., Bussemaker, H.J., and van Steensel, B. (2017). Genome-wide mapping of autonomous promoter activity in human cells. Nat. Biotechnol. *35*, 145–153.

van Arensbergen, J., Pagie, L., FitzPatrick, V.D., de Haas, M., Baltissen, M.P., Comoglio, F., van der Weide, R.H., Teunissen, H., Võsa, U., Franke, L., et al. (2019). High-throughput identification of human SNPs affecting regulatory element activity. Nat. Genet. *51*, 1160–1169.

Wei, Z., Yang, Y., Zhang, P., Andrianakos, R., Hasegawa, K., Lyu, J., Chen, X., Bai, G., Liu, C., Pera, M., et al. (2009). Klf4 interacts directly with Oct4 and Sox2 to promote reprogramming. Stem Cells *27*, 2969–2978.

Wickham, H. (2016). ggplot2 Elegant Graphics for Data Analysis (Springer International Publishing).

You, F.M., Huo, N., Gu, Y.Q., Luo, M.-C., Ma, Y., Hane, D., Lazo, G.R., Dvorak, J., and Anderson, O.D. (2008). BatchPrimer3: a high throughput web application for PCR and sequencing primer design. BMC Bioinformatics *9*, 253.

Zabidi, M.A., Arnold, C.D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., and Stark, A. (2015). Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. Nature *518*, 556–559.

Zabidi, M.A., and Stark, A. (2016). Regulatory enhancer-core-promoter communication via transcription factors and cofactors. Trends Genet. *32*, 801–814.

Zhang, X., Yalcin, S., Lee, D.-F., Yeh, T.-Y.J., Lee, S.-M., Su, J., Mungamuri, S.K., Rimmelé, P., Kennedy, M., Sellers, R., et al. (2011). FOXO1 is an essential regulator of pluripotency in human embryonic stem cells. Nat. Cell Biol. *13*, 1092–1099.

Zorita, E., Cuscó, P., and Filion, G.J. (2015). Starcode: sequence clustering based on all-pairs search. Bioinformatics *31*, 1913–1919.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
e. cloni 10G supreme electrocompetent *E. coli*	Lucigen	Cat# 60081-1
5-alpha competent bacteria	NEB	Cat# C2987
Chemicals, peptides, and recombinant proteins		
My-Taq Red mix	Bioline	Cat# BIO-25044
Fast-link ligase	Lucigen	Cat# LK0750H
Klenow HC 3′→5′exo−	NEB	Cat# M0212L
CleanPCR magnetic beads	CleanNA	Cat# CPCR-0050
Gibson Assembly® Master Mix	NEB	Cat# E2611S
XmaJI (AvrII) restriction enzyme	Thermo Fischer	Cat# ER1561
XcmI restriction enzyme	NEB	Cat# R0533
NheI restriction enzyme	NEB	Cat# R0533
T4 DNA ligase	Roche	Cat# 10799009001
I-CeuI restriction enzyme	NEB	Cat# R0699S
I-SceI restriction enzyme	NEB	Cat# R0694S
TRIsure	Bioline	Cat# BIO-38032
DNase I recombinant, RNase-free	Roche	Cat# 04716728001
Maxima reverse transcriptase	Thermo Fischer	Cat# EP0743
Neurobasal medium	Gibco	Cat# 21103-049
DMEM-F12 medium	Gibco	Cat# 11320-033
N27	Gibco	Cat# 17504-044
B2	Gibco	Cat# 17502-048
LIF	Sigma-Aldrich	Cat# ESG1107
CHIR-99021	MedChemExpress	Cat# HY-10182
PD0325901	MedChemExpress	Cat# HY-10254
monothioglycerol	Sigma	Cat# M6145-25ML
Critical commercial assays		
ISOLATE II PCR and Gel Kit	Bioline	Cat# BIO-52059
End-It DNA End-Repair Kit	Epicentre	Cat# ER0720
Takara ligation kit version 2.1	Takara	Cat# 6022
PureLink HiPure Plasmid Filter Maxiprep Kit	Invitrogen	Cat# K210016
Mouse Embryonic Stem Cell Nucleofector Kit	Lonza	Cat# VPH-1001
GeneJET RNA Purification Kit	Thermo Fischer	Cat# K0732
MycoAlert Mycoplasma Detection Kit	Lonza	Cat# LT07-318
Deposited data		
Raw and Processed data	This study	GEO: GSE186265
Code	This Study	GitHub: https://github.com/vansteensellab/EPCombinations https://doi.org/10.5281/zenodo.6406791
Annotated mouse TSSs	GENCODE	Release vM16: https://www.gencodegenes.org/mouse/release_M16.html
mESC TAD coordinates	Bonev et al., 2017	Supplementary Table: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5651218/bin/mmc2.xlsx
Mouse reference genome mm10 genome build (GRCm38)	Genome Reference Consortium	https://www.ncbi.nlm.nih.gov/grc/mouse

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
DNAse Hypersensitivity mESCs 2i+lif	Joshi et al., 2015	GEO: GSE72164
DNAse Hypersensitivity mESCs serum	ENCODE	GEO: GSE37074
TF motif database	Diaferia et al., 2016	GitHub: https://github.com/vansteensellab/EPCombinations https://doi.org/10.5281/zenodo.6406791
RNA-seq mESCs 2i+lif	Joshi et al., 2015	GEO: GSE72164
ExP K562 count data	Bergman et al., 2022	GEO: GSE184426
MicroC mESCs	Hsieh et al., 2020	GEO: GSE130275
Experimental models: Cell lines		
E14TG2a mESCs	ATCC	Cat# CRL-1821
Oligonucleotides		
Table S3	This Study	N/A
Recombinant DNA		
SuRE vector (JvAp101)	van Arensbergen et al., 2017	N/A
Downstream Assay vector (JvAp102)	This Study	N/A
Software and algorithms		
Homer v4.10	Heinz et al., 2010	http://homer.ucsd.edu/homer/data/software/homer.v4.10.zip
BatchPrimer3 v1.0	You et al., 2008	https://wheat.pw.usda.gov/demos/BatchPrimer3/
R version 4.0.5 (2021-03-31)	R Core Team, 2021	https://www.r-project.org/
ggplot2	Wickham, 2016	https://ggplot2.tidyverse.org/
Bowtie2 version 2.3.4	Langmead and Salzberg, 2012	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Python version 3.6.2	Rossum and Drake, 2009	https://www.python.org/downloads/release/python-362/
Snakemake version 4.4.0	Köster and Rahmann, 2012	https://anaconda.org/bioconda/snakemake/files?version=4.4.0
Starcode version 1.1	Zorita et al., 2015	https://github.com/gui11aume/starcode
MEME suite version 5.0.2	Bailey et al., 2009	https://meme-suite.org/meme/doc/download.html
HRT Atlas v1.0	Hounkpe et al., 2021	www.housekeeping.unicamp.br
Other		
Hamilton Microlab® STAR	Hamilton	https://www.hamiltoncompany.com/automated-liquid-handling/platforms/microlab-star

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact Bas van Steensel (b.v.steensel@nki.nl).

Materials availability

Plasmids generated in this study are available without restriction upon request from the Lead Contact.

Data and code availability

- Raw sequencing data and processed data are available at Gene expression Omnibus (GEO): GSE186265. This paper also analyzes existing, publicly available data. The accession numbers for these datasets are listed in the key resources table. Lab journal records are available at https://osf.io/5a7h6/.
- Code of data processing pipelines and analysis scripts are available at https://github.com/vansteensellab/EPCombinations (DOI: 10.5281/zenodo.6406791).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell culture

All experiments were conducted in E14TG2a male mouse embryonic stem cells (mESC) (ATCC CRL-1821) cultured in 2i+LIF culturing media. 2i+LIF was made according to the 4DN nucleome protocol for culturing mESCs (https://data.4dnucleome.org/protocols/cb03c0c6-4ba6-4bbe-9210-c430ee4fdb2c/). The reagents used were Neurobasal medium (#21103-049, Gibco), DMEM-F12 medium (#11320-033, Gibco), BSA (#15260-037; Gibco), N27 (#17504-044; Gibco), B2 (#17502-048; Gibco), LIF(#ESG1107; Sigma-Aldrich), CHIR-99021 (#HY-10182; MedChemExpress) and PD0325901 (#HY-10254; MedChemExpress), monothioglycerol (#M6145-25ML; Sigma) and glutamine (#25030-081, Gibco). Monthly tests (#LT07-318; Lonza) confirmed that the cells were not contaminated by mycoplasma.

METHOD DETAILS

Selection of cCREs and promoters

For the design of the libraries we selected the cCREs and promoters from three TADs centered around each of the *Klf2*, *Nanog* and *Tfcp2l1* genes, using TAD coordinates from (Bonev et al., 2017). cCREs were selected based on DNase I hypersensitivity mapping data from mESCs in both 2i+LIF (Joshi et al., 2015) and serum (ENCODE Project Consortium, 2012) culturing conditions, which we reprocessed and aligned to the mm10 genome build. DNAse I hypersensitivity sites (DHSs) were called using Homer v4.10 (Heinz et al., 2010) with default parameters and peak style "factor". We defined cCREs as 450 bp windows centered on each peak. For promoters we used the Gencode mouse TSS annotation (Frankish et al., 2019). From each TSS we defined as promoters the -375 +75bp region. If the promoter regions overlapped with any cCRE then the promoter was redefined as the 450 bp region surrounding the center of the intersection of both elements. PCR primers were designed for each cCRE and promoter using the batch version of Primer3 (BatchPrimer3 v1.0) (You et al., 2008) allowing for primers to be designed on the 50 bps of each end. This yielded PCR products of ~400 bp for each element. Coordinates and sequences of all cCREs and Ps used as templates for primer design are included in Data S1.

Upstream assay library generation

For each locus, cCREs and promoters were amplified from mouse genomic DNA (extracted from E14TG2a mESCs, ATCC CRL-1821) by PCR using My-Taq Red mix (#BIO-25044; Bioline) in 384 well plates using automated liquid handling (Hamilton Microlab® STAR). PCRs were checked on gel and had a success rate between 60 and 90% depending on the locus. Equal volumes (10ul) of the resulting PCR products were mixed, and the resulting pool was purified by phenol-chloroform extraction followed by gel purification (BIO-52059; Bioline). The purified DNA fragments were then blunted and phosphorylated using End-It DNA End-Repair Kit (#ER0720; Epicentre). Part of the repaired pool was set apart for cloning of singlet libraries. The remainder was self-ligated using Fast-link ligase (LK0750H; Lucigen), after which duplets of ~800bp were excised from agarose gel and purified (BIO-52059; Bioline). Singlet and duplet pools were A-tailed using using Klenow HC 3′→5′ exo− (#M0212L; NEB).

The SuRE barcoded vector was prepared as previously described (van Arensbergen et al., 2017). First, we digested 10 μg of the SuRE vector with NheI (#R0131S; NEB) and XcmI (#R0533; NEB) and performed a gel-purification. Barcodes were generated by performing 10 PCR reactions of 100 μl each containing 5 μl of 10 μM primer 256JvA, 5 μl of 10 μM primer 264JvA and 1 μl of 0.1 μM template 254JvA (see Table S3 for oligonucleotide sequences). A total of 14 PCR cycles were performed using MyTaq Red Mix (#BIO-25043; Bioline), yielding ~30 μg barcodes. Barcodes were purified by phenol-chloroform extraction and isopropanol precipitation after which they were digested overnight with 80 units of AvrII (#ER1561; Thermo Fischer) and purified using magnetic bead purification (#CPCR-0050; CleanNA). The linear vector and the barcodes were then ligated in multiple 100 μl reactions containing 3 μg digested vector and 2.7 μg digested barcodes, 20 units NheI (#R0131S; NEB), 20 units XcmI, 10 μl of 10× CutSmart buffer, 10 μl of 10 mM ATP, 10 units T4 DNA ligase (#10799009001 Roche). A cycle-ligation of six cycles was performed (10 min at 22 °C and 10 min at 37 °C), followed by 20 min heat-inactivation at 80 °C. The ligations were purified by magnetic beads and digested with 40 units of XcmI (#R0533S; NEB) for 3 h, and size-selected by gel-purification.

Then singlet and duplet pools were separately ligated overnight into the SuRE barcoded vector using Takara ligation kit version 2.1 (#6022; Takara). Ligation products were purified using magnetic bead purification (#CPCR-0050; CleanNA). Next, 2 μl of the purified ligation products were electroporated into 20 μl of electrocompetent e. cloni 10G supreme (#60081-1; Lucigen). Each library was grown overnight in 500 ml of standard Luria Broth (LB) with 50 μl/ml of kanamycin and purified using a maxiprep kit (K210016, Invitrogen).

Downstream assay library generation

The Downstream assay vector was based on a pSMART backbone (Addgene plasmid # 49157; a gift from James Thomson). It was constructed using standard molecular biology techniques and contains a green fluorescent protein (GFP) open reading frame followed by a barcode, and a psiCheck polyadenylation signal (PAS) introduced during barcoding, followed by the cloning site for inserts and a triple polyadenylation site (SV40+bGH+psiCheckPAS).

The 10 highest expressing promoters of the *Klf2* Upstream library were selected to be cloned into the Downstream assay vector at the promoter position. These Promoters were amplified by PCR and individually inserted by Gibson assembly (#E2611S; NEB) into the Downstream assay vector. Then each of the 10 constructs were transformed into standard 5-alpha competent bacteria (#C2987; NEB) grown overnight in in 500 ml of standard Luria Broth (LB) with 50 μl/ml of kanamycin and purified.

Each of these promoter-containing vectors was then barcoded similarly as the SuRE vector (van Arensbergen et al., 2017). For this, we digested 10 μg of each vector with AvrII (#ER1561; Thermo Fischer) and XcmI (#R0533; NEB) and performed a gel-purification. Barcodes were generated by performing 10 PCR reactions of 100 μl each containing 5 μl of 10 μM primer 275JvA, 5 μl of 10 μM primer 465JvA and 1 μl of 0.1 μM template 274JvA (see Table S3 for oligonucleotide sequences). A total of 14 PCR cycles were performed using MyTaq Red Mix (#BIO-25043; Bioline), yielding ~30 μg barcodes. Barcodes were purified by phenol-chloroform extraction and isopropanol precipitation after which they were digested overnight with 80 units of NheI (#R0131S; NEB) and purified using magnetic bead purification (#CPCR-0050; CleanNA). Each vector variant and the barcodes were then ligated in one 100 μl reaction containing 3 μg digested vector and 2.7 μg digested barcodes, 20 units NheI (#R0131S; NEB), 20 units AvrII, 10 μl of 10× CutSmart buffer, 10 μl of 10 mM ATP, 10 units T4 DNA ligase (#10799009001 Roche). A cycle-ligation of six cycles was performed (10 min at 22 °C and 10 min at 37 °C), followed by 20 min heat-inactivation at 80 °C. The ligation reaction was purified by magnetic beads and digested with 40 units of XcmI (#R0533S; NEB) for 3 h, and size-selected by gel-purification, yielding ~1 μg barcoded vector for each variant.

The singlets were cloned into the enhancer position of the downstream barcoded vector variants by overnight ligation using Takara ligation kit version 2.1 (#6022; Takara). Ligation products were purified using magnetic bead purification (#CPCR-0050; CleanNA). Next, 2 μl of the purified ligation products were electroporated into 20 μl of electrocompetent e. cloni 10G supreme (#60081-1; Lucigen). Each library was grown overnight in 500 ml of standard Luria Broth (LB) with 50 μl/ml of kanamycin and purified using a maxiprep kit (K210016, Invitrogen).

Inverse PCR and sequencing to link inserted elements to barcodes

We identified barcode–insert combinations in the plasmid libraries by inverse-PCR followed by sequencing as described (van Arensbergen et al., 2017). In brief, the combination of barcode and element(s) was excised from the plasmid by digestion with I-CeuI(#R0699S, NEB); this fragment was circularised; remaining linear fragments were destroyed; and circular fragments were linearised again with I-SceI (#R0694S; NEB). These linear fragments were amplified by PCR with sequencing adaptors. The final product was sequenced on an Illumina MiSeq platform using 150-bp paired-end reads. This process was done separately for each of the libraries. In the singlet libraries the barcodes should be associated to only one insert and in the combinatorial libraries the barcodes should be associated with duplets.

Libraries transfection

E14TG2a mouse embryonic stem cells were transiently transfected using Amaxa nucleofector II, program A-30, and Mouse Embryonic Stem Cell Nucleofector™ Kit (#VPH-1001, Lonza). *Klf2* and *Nanog* loci Upstream assay libraries were mixed and transfected together, *Tfcp2l1* Upstream Assay libraries were transfected in separate experiments. All the Downstream assay sub-libraries were transfected as a mix. Three independent biological replicates were done for each library mix. For each biological replicate 16 million cells were transfected (4 million cells with 4 μg plasmid per cuvette)

RNA extraction and cDNA sequencing

RNA was extracted and processed for sequencing as described (van Arensbergen et al., 2017) with a few modifications. Cells were harvested 24 h after transfection, resuspended in TRIsure (#BIO-38032; Bioline) and frozen at -80 °C until further processing. From the Trisure suspension, the aqueous phase containing the RNA was extracted and loaded into RNA extraction columns (#K0732, Thermo Scientific). Total RNA was divided into 10 μl reactions containing 5 μg of RNA and was treated for 30 mins with 10 units of DNase I (#04716728001; Roche). Then DNase I was inactivated by addition of 1 μl of 25 mM EDTA and incubation at 70°C for 10 min.

For the Upstream Assay the cDNA was produced using Maxima reverse transcriptase (#EP0743; ThermoFisher Scientific) in 20 ul reactions and amplified by PCR as described (van Arensbergen et al., 2017). Per biological replicate 8 to 10 reactions were carried out in parallel in order to cover enough barcode complexity of the library. For the Downstream Assay the RNA was extracted and processed the same way until cDNA production. Here, cDNA was produced using a specific primer (304JvA sequence in Table S3 for oligonucleotide sequences) using Maxima reverse transcriptase (#EP0743; ThermoFisher Scientific) in 20ul reactions using the same conditions previously described (van Arensbergen et al., 2017). Primer 304JvA introduces an adaptor sequence 5' to the primer sequence which is targeted in the first PCR (see below) to ensure strand specific amplification of barcodes. Then cDNA was amplified in 2 steps (nested PCRs) in order to make the reaction strand-specific. The first PCR reaction was run for 10 cycles (1 min 96 °C, 10 times (15 s 96 °C, 15 s 60 °C, 15 s 72 °C)) using (index variants of) primers 285JvA (containing the S2, index and p7 adaptor) and 305JvA (targeting the adapter introduced by 304JvA). Each 20 μl RT reaction was amplified in a 100-μl PCR reaction with MyTaq Red mix. The second PCR reaction was performed using 10ul of the product of the previous reaction in 100 μl reactions (1 min 96 °C, 8×(15 s 96 °C, 15 s 60 °C, 15 s 72 °C)) using the same index variant primer and primer 437JvA (containing the S1, and p5 adaptor). For both Upstream and Downstream assays, the resulting PCR products were sequenced on an Illumina 2500 HiSeq platform with 65bp single end reads.

Plasmid DNA (pDNA) barcode sequencing

For normalisation purposes, barcodes in the plasmid pools were counted as follows. For both assays the process was the same. For each library 1 μg of plasmid was digested with I-SceI in order to linearise the plasmid. Then, barcodes were amplified by PCR from 50 ng of material using the same primers and reaction conditions as in the amplification of cDNA in the Upstream assay, but only 9 cycles of amplification were used (1 min 96 °C, 9 times (15 s 96 °C, 15 s 60 °C, 15 s 72 °C)). For each library, two technical replicates were carried out by using different index primers for each replicate. Samples were sequenced on an Illumina 2500 HiSeq platform with 65bp single end reads.

QUANTIFICATION AND STATISTICAL ANALYSIS

The data generated by the calculation of boost indices calculation was further processed and analyzed in R (R Core Team, 2021) the figures were generated using ggplot2 (Wickham, 2016). The main packages and software are listed in the key resource table and an extensive bibliography is available in all the scripts (https://github.com/vansteensellab/EPCombinations). The processes described in Linking barcodes to element singlets or duplets and Pre-Processing of cDNA and pDNA reads were performed using a custom Snakemake (Köster and Rahmann, 2012) pipeline. Statistical details for individual experiments have been provided in the main text, figure legends, and Method Details.

Linking barcodes to element singlets or duplets

For each library the iPCR data was locally aligned using bowtie (version 2.3.4) (Langmead and Salzberg, 2012) with very sensitive parameters (–very-sensitive-local) on a custom bowtie genome. This custom genome was generated using bowtie. It consists of virtual chromosomes corresponding to each cCRE or a P from each locus. Bam alignment files were processed using a custom python (Rossum and Drake, 2009) script that identifies from read 1 the barcode and cCRE or P element, and from read 2 the cCRE or P element. In case of singlet libraries both reads should identify the same element, whereas in combinatorial libraries read 1 is derived from the barcode-proximal element and read 2 from the barcode distal element. In the combinatorial libraries we cannot distinguish between a combination of one element with itself in the same orientation or a single element, therefore these were removed from combinatorial libraries. In the Downstream Assay both reads identify the only element cloned in the downstream position. If no element was found, the barcode was assigned as empty vector. The resulting barcode-to-element(s) lists were clustered using Starcode (version 1.1) (Zorita et al., 2015) to remove errors from barcode sequencing. Finally, barcodes present in multiple libraries or matched with multiple element combinations were removed from the data.

Pre-Processing of cDNA and pDNA reads

For each replicate of each library pool transfection barcodes were extracted from the single end reads by using a custom python script that identifies the constant region after the barcode. Near-identical barcodes were pooled using Starcode (version 1.1) (Zorita et al., 2015) to remove errors from barcode sequencing, and barcode counts were summarized. The process was the same for cDNA and pDNA counts and for Upstream and Downstream data.

Post processing of cDNA and pDNA counts

For each transfection, barcodes identified in the cDNA were matched to the barcodes in the iPCR data, and all barcodes were counted in cDNA and pDNA replicates. Barcode counts were normalised to the total number of barcode reads from each sample. Activity per barcode was then calculated as a cDNA:pDNA ratio of normalised counts. Next, activities from multiple barcodes belonging to the same element singlet or combination were averaged, requiring a minimum of 5 barcodes per singlet or combination and at least 8 pDNA counts per barcode. The mean activity of each singlet or combination across replicates was calculated as the geometric mean of the three replicates. Activities for all pairs including cCRE-cCRE, P-P and P-cCRE pairs are included in Data S2.

Calculation of boost indices

We initially calculated raw boost indices simply as a $\log_2$ ratio of the activity of each cCRE–P pair over the activity of the corresponding P alone. However, 20 negative controls that we included in the *Klf2* libraries, consisting of randomly generated DNA sequences of similar size and G/C content as the cCREs (Data S1), generally showed a negative boost index by this measure (median value -0.45 when inserted upstream) (Figure S1D). We therefore calculated corrected boost indices as the $\log_2$ ratio of cCRE-P activity over the median cCRE-P activity per promoter (Figure S1D). Importantly, in the *Klf2* library data this largely removed the negative bias that we observed with the negative controls; we thus assume that this correction is adequate and therefore also applied it to the boost indices obtained with the other libraries. For the analyses in Figures 2, 3, 4, 5, 6, 7, S2–S5, and S7 except Figures S4A and S4B the boost indices of cCREs were averaged over both orientations of the cCREs as boost indices correlated between orientations (Figure S4). Boost indices of all cCRE-P pairs are included in Data S3 and S4.

Identification of activating and repressive cCREs

To classify cCREs according to their general effect on all promoters we performed a Wilcoxon test on the boost indices of each cCRE combined with every promoter against the rest of the population. P-values were corrected for multiple hypothesis testing using the

Benjamini-Hochberg method and an FDR cutoff of 5% was chosen. cCREs significantly activating or repressing at the 5% FDR cutoff were classified as activators or repressors respectively. The rest of cCREs were classified as ambiguous.

Analysis of selectivity

We performed a Welch's ANOVA (or Welch F-test) on the calculated Boost indices to assess the selectivity of each cCRE with more than 5 cCRE-P combinations. For this purpose, each replicate of each orientation of the cCRE-P was used as a datapoint and each cCRE-P combination was used as a group. P-values were corrected for multiple hypothesis testing using the Benjamini-Hochberg method and an FDR cutoff of 5% was chosen. The Welch F-test was chosen over the classic ANOVA due to heteroscedasticity of the data.

Analysis of Houskeeping and non-housekeeping promoter selectivity

We classified the promoters from the Klf2 locus as Housekeeping and non-housekeeping according to the publicly available HRT Atlas v1.0 database (Hounkpe et al., 2021). We used this classification to overlay it with a clustered and reduced version of the *Klf2* locus boost index matrix. This matrix was clustered using hierarchical clustering. The complete *Klf2* locus boost index matrix was used to asses the similarities between promoters by correlating each promoter using pearson correlation. We then separated correlations in within Housekeeping promoters, within Non-Housekeeping promoters and between Housekeeping and Non-housekeeping.

Analysis of selectivity on K562 ExP data

The cDNA and pDNA count data from (Bergman et al., 2022) was processed the same way as our data and boost indices were calculated using the same procedure. Because of the higher number of promoters available in the data we randomly down-sampled the dataset 10 times to have 10 subsets of 230 Enhancers and 20 promoters. This was done in order to limit the degrees of freedom to be able to apply the Welch F-test.

On each of the processed down-sampled subset we performed a Welch's ANOVA (or Welch F-test) on the calculated Boost indices to assess the selectivity of each E with more than 5 E-P combinations. For this purpose, each replicate of E-P combination was used as a datapoint and each E-P combination was used as a group. P-values were corrected for multiple hypothesis testing using the Benjamini-Hochberg method and an FDR cutoff of 0.1% was chosen. This cutoff was more stringent than on the analysis performed on our dataset as the number of tests performed was one order of magnitude higher. Again, the Welch F-test was chosen over the classic ANOVA due to heteroscedasticity of the data.

TF motif Survey

We used a custom TF motif database provided by the lab of Gioacchino Natoli containing 2,448 TF motifs which was built on top of a previously published version (Diaferia et al., 2016) (Dataset composition and sources available at https://github.com/vansteensellab/EPCombinations). TF motifs were filtered for expression of TFs in mESCs cultured in 2i+LIF according to published RNA-seq (higher expression than 1RPM) (Joshi et al., 2015). We scored presence or absence of a TF motif in each cCRE using FIMO (MEME suite, version 5.0.2) (Bailey et al., 2009). We then searched for motifs associated with (1) general enhancer activity, (2) self-compatibility and (3) duplets of self-compatible motifs. In (1), for each TF motif we compared the general cCRE-P population to combinations where the TF motif was present at the cCRE. In (2), for each TF motif we compared the cCRE-P combinations where the TF motif was present at the cCRE to the combinations where it was present at both the cCRE and the promoter. In (3), we took all the significant TF motifs at a 1% FDR and an effect size higher than 0.1 (n=66). Then we tested all pairwise non-repeated TF motif duplets. Per TF motif duplet we compared the cCRE-promoters where both TF motif were present at the cCRE to the combinations where both were present at both the cCRE and the promoter. In all comparisons a Wilcoxon test was applied to the boost indices of each group and the effect size was calculated a difference of median boost indices. In each analysis p-values were corrected for multiple hypothesis testing using the Benjamini-Hochberg method. We required a minimum of 50 cCRE-promoter combinations per group.

Micro-C data correlation

Processed publicly available Micro-C data was obtained from (Hsieh et al., 2020). Contact scores between cCRE-P pairs were averaged across bins overlapping a +-500 bp window from the location of each element using 400 bp bins. Pearson correlation was calculated on the relationship between Boost indices and contact frequencies of cCRE-P pairs.

Molecular Cell, Volume *82*

Supplemental information

Systematic analysis of intrinsic enhancer-promoter compatibility in the mouse genome

Miguel Martinez-Ara, Federico Comoglio, Joris van Arensbergen, and Bas van Steensel

SUPPLEMENTARY TABLES

Supplementary table 1. Numbers of tested Promoters (Ps), cCREs and cCRE–P pairs in each combinatorial MPRA library. Related to Figure 1.

Library	Ps present	cCREs present	cCRE–P pairs tested	cCRE–P pairs (orientation-independent)
Klf2 Upstream	23	82	3758	1400
Nanog Upstream	18	88	1321	595
Tfcp2l1 Upstream	25	198	5599	2490
Klf2 Downstream	10	84	1364	752

Supplementary table 2. Other combinations of cCREs and Ps in each MPRA library.
Related to Figure 1.

Library	cCRE-cCRE	cCRE-cCRE (orientation-independent)	P-P	P-P (orientation-independent)	P-cCRE	P-cCRE (orientation-independent)
Klf2 Upstream	10626	4284	1335	441	4067	1439
Nanog Upstream	10536	4769	155	82	1511	713
Tfcp2l1 Upstream	44515	21149	626	274	5239	2386
Klf2 Downstream	0	0	420	225	0	0

Supplementary table 3. Oligonucleotide sequences. Related to STAR methods.

Name	Type	Sequence (5' -> 3')
275JvA	Barcoding Primer	(N:25252525) ttggttGGgctagc (N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N) AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
465JvA	Barcoding Primer	**AAGCAATCTCTATACGGAGTTCAGT**AGGTTAACCAGATCCC TTCCGGAATTCCAAGGTTG
274JvA	Barcoding ultramer template	Agatcggaagagcgtcgtgtagggaaagagtgttagggataacagggtaatgcggcc Gctggccgcaataaaatatctttatttcattacatctgtgtgttggttttttgtgtgaggatctgtg actggagttcagacgtgtgctcttccgatct**ccagtgatgtgatgg**ttggccaaccttggaattccgg
304JvA	GSP for reverse transcription	TACAGAGCTGACGTATCAGTACGGCCGCATTACCCTGTTATCCCTAACACTC
285JvA	indexed illumina sequencing primer	CAAGCAGAAGACGGCATACGAGAT**ACAGCA** *GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT* CGTGGAGGAGCTGCACAGCAACAC
305JvA	adapter primer first PCR	TACAGAGCTGACGTATCAGTACG

437JvA	illumina sequencing adapter	AATGATACGGCGACCACCGAGATCT*ACACTCTTT CCCTACACGACGCTCTTCCGATCT*
256JvA	Barcoding Primer	tgtgatggttggccaaccttggaattccggaagggatctggttaaccttggaacc
264JvA	Barcoding Primer	(N:25252525) Ttggttcctagg (N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N) AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
254JvA	Barcoding ultramer template	Aagggatctggttaaccttggaaccttggccaacgtacgactggagatcggaagagcacacg tctgaactccagtcactagggataacagggtaatacactctttccctacacgacgctcttccgatct

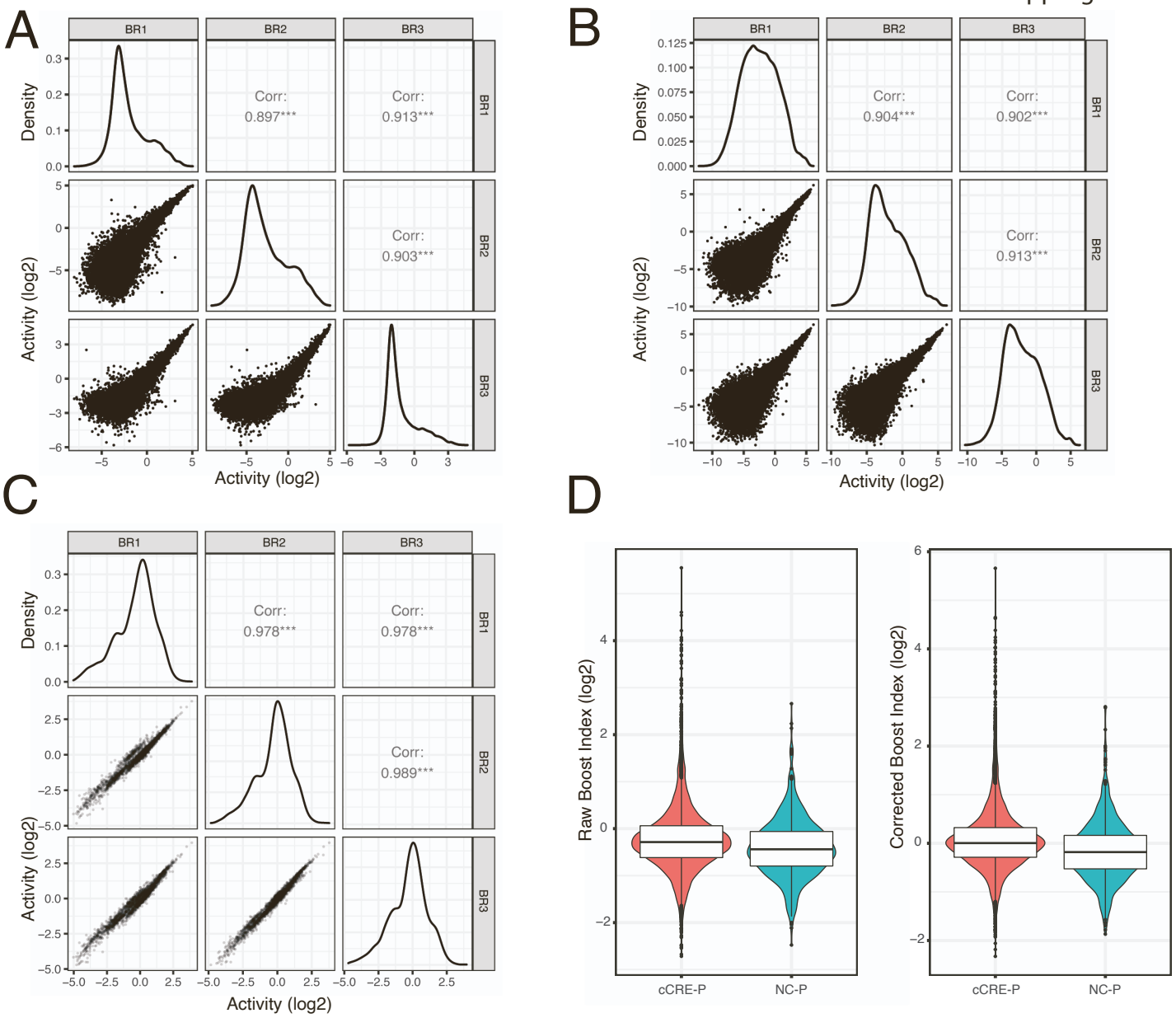


Figure S1. Reproducibility of data and boost index calculation. Related to Figure 2 and STAR methods. **(A-C)** Correlograms of the three biological replicates of each library pool. Lower left panels show pairwise scatterplots of the activities of all cCRE-P pairs per replicate. Middle panels show the density of data distribution in each replicate and upper right panels show the Pearson correlation coefficients. **A)** Klf2 and Nanog Upstream libraries. **B)** Tfcp2l1 Upstream library. **C)** Klf2 Downstream libraries. **D)** Upstream assay boost index distributions for cCRE-P and negative controls – promoter (NC-P) combinations. Left panel: raw boost indices; right panel: boost indices after correction for negative bias (see Methods).

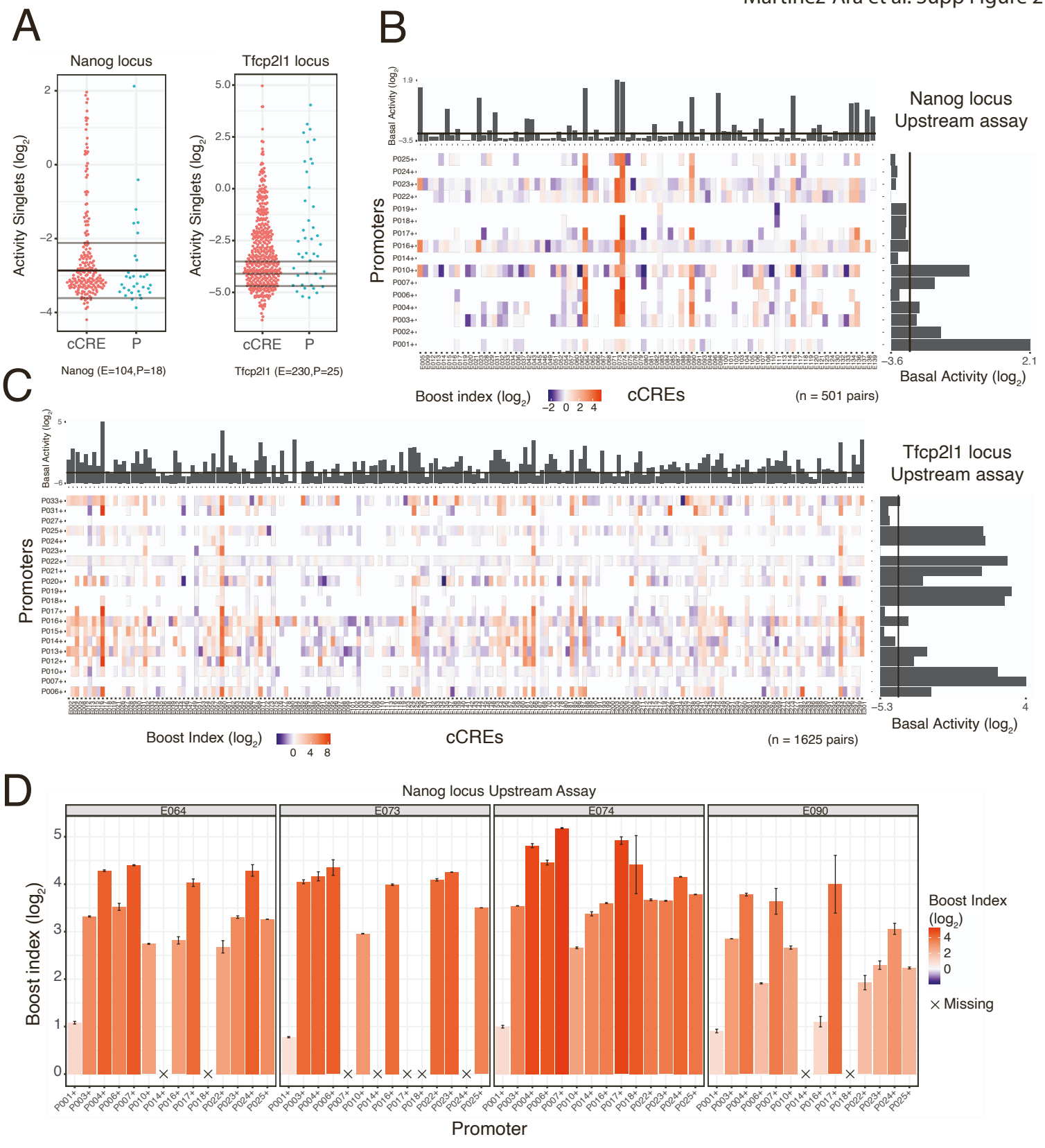


Figure S2. Element activities and boost indices obtained with Nanog and Tfcp2l1 Upstream libraries. Related to Figures 2 and 3. **A)** Transcriptional activities of cCREs and promoters. Each dot represents the mean activity of one singlet. Horizontal lines represent the average background activity of empty vectors (black line) plus or minus two standard deviations (grey lines). Elements with activities more than two standard deviations above the average background signal are defined as active. **B-C)** Boost index matrices for cCRE–P pairs from Nanog and Tfcp2l1 loci (both Upstream assays). White tiles indicate missing data. Barplots on the right and top of each panel show basal activities of each tested P or cCRE, respectively, with the black line indicating the background activity of the empty vector. **D)** Examples of cCRE-P combinations for cCREs E064, E073, E074 and E090 of the Nanog locus. Barplots represent the mean boost index of each combination, vertical lines represent the standard deviation of each boost index. All data are averages over 3 independent biological replicates.

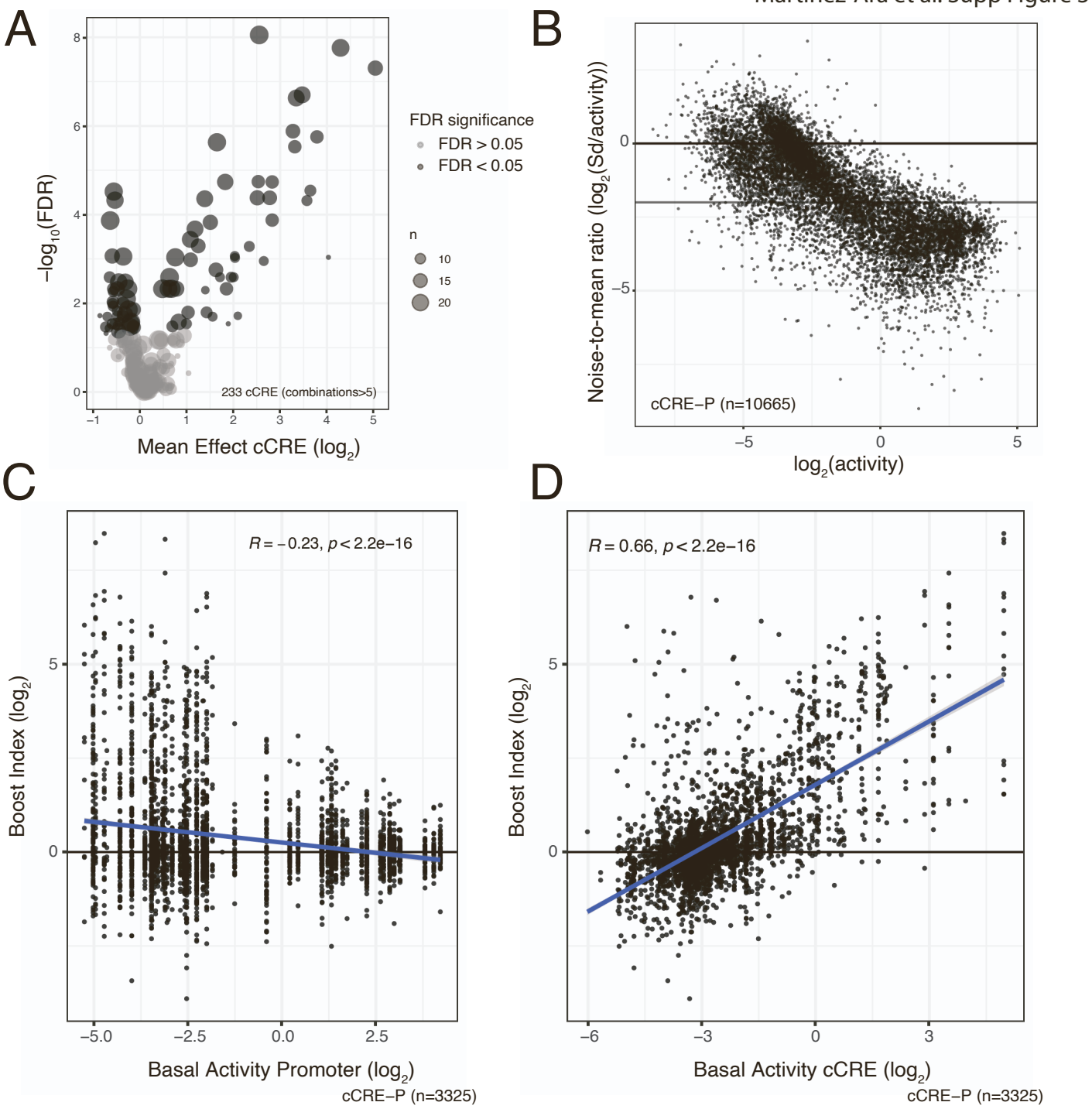


Figure S3. cCRE functional classification and activity influence on Boost indices. Related to Figures 2 to 4. **A)** Volcano plot of cCREs associated with activation or repression across promoters. A Wilcoxon test is performed per cCRE comparing the boost indices of all the cCRE-P combinations of that cCRE against the rest of cCRE-P combinations. A minimum of 6 combinations is required per cCRE. P-values are corrected for multiple hypothesis testing using the Benjamini-Hochberg method (FDR). **B)** Relationship between noise-to-mean ratio (Standard Deviation/mean Activity) and mean activity of cCRE-Ps. Horizontal lines represent noise-to-mean ratios of 1 and of 4 in log2 scale. **C)** Relationship between boost indices and basal (singlet) P activity. Each column of dots shows the data of cCRE–P pairs for one P. Data are from Upstream assays of all three loci combined. **D)** Relationship between boost indices and basal (singlet) cCRE activity. All data are averages over 3 independent biological replicates. R is pearson correlation and p its corresponding p-value.

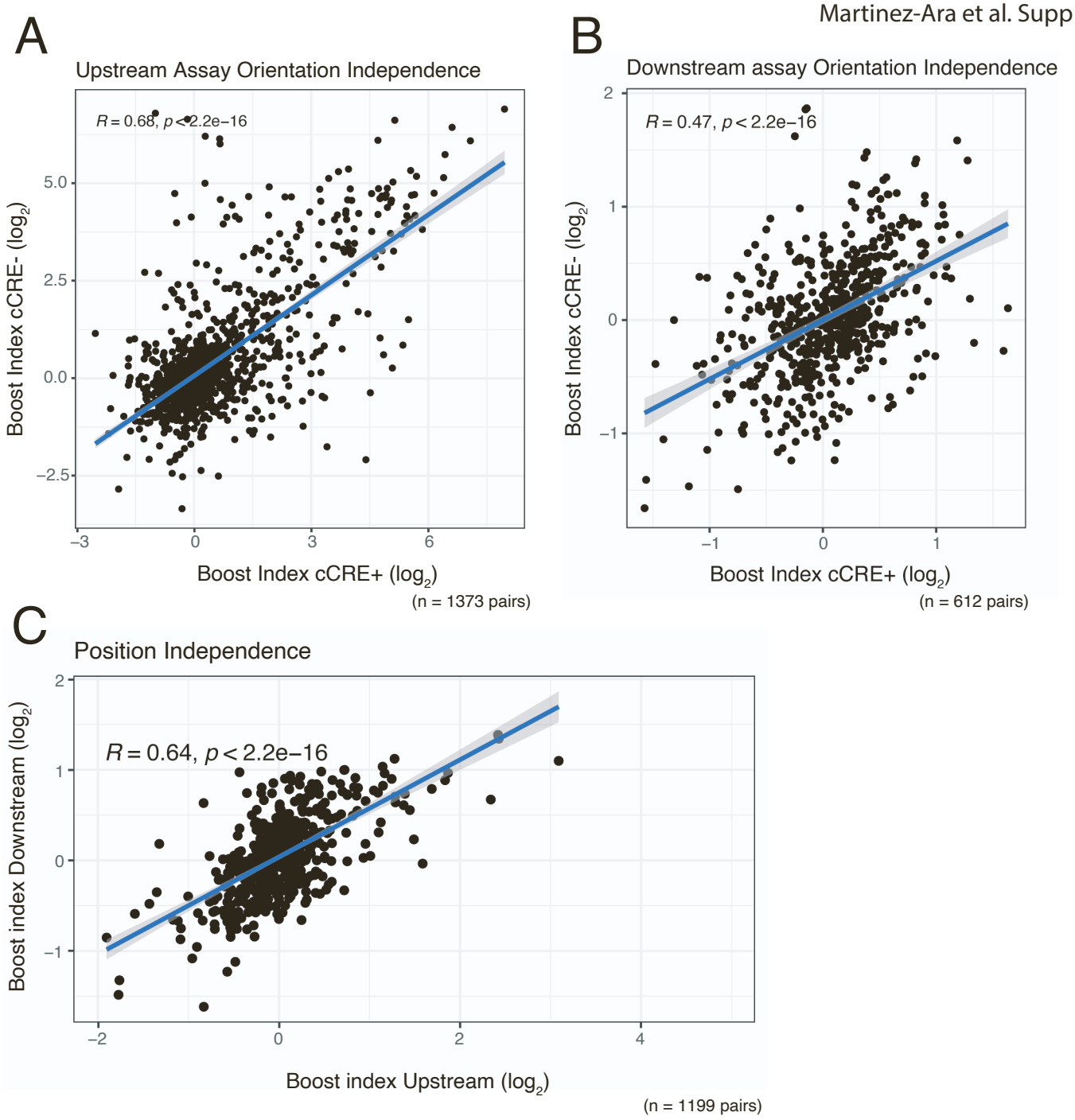


Figure S4. Orientation and position independence of cCREs. Related to STAR methods. **(A-B)** Correlation between boost indices of both cCRE orientations of the same cCRE-P combination, in the **(A)** Upstream assay and **(B)** Downstream assay. Data are from the Klf2 locus libraries. Note that "+" and "-" orientations are arbitrary labels, because cCREs do not have an intrinsic orientation. **(C)** Correlation between boost indices of cCRE-P combinations shared between the Upstream and Downstream assays of the Klf2 locus. In all panels R is the Pearson correlation coefficient. All data are averages over 3 independent biological replicates. In C Boost indices are averaged over cCRE orientations. R is pearson correlation and p its corresponding p-value.

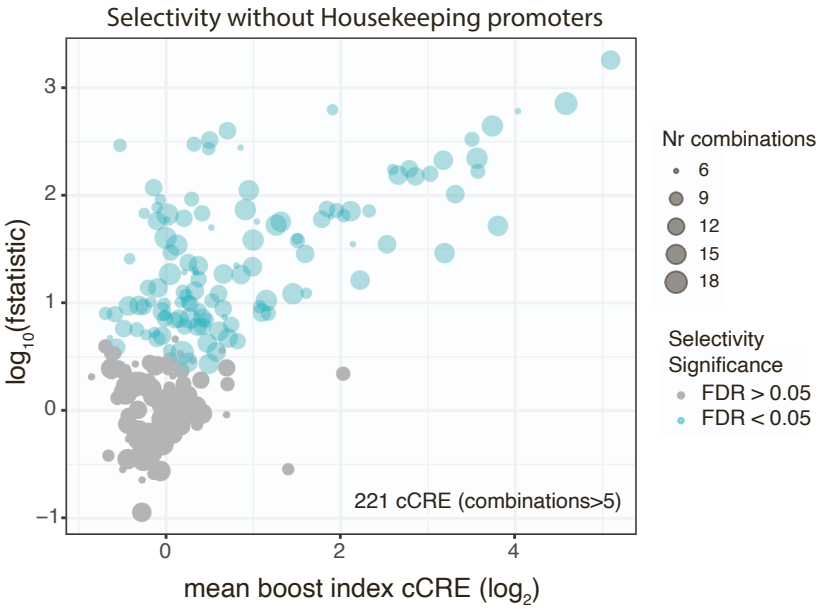


Figure S5. Selectivity is widespread among non-housekeeping promoters. Related to Figure 5. Results of selectivity analysis as performed in Figure 4C, but excluding housekeeping promoters [48]. Data are averages over 3 independent biological replicates.

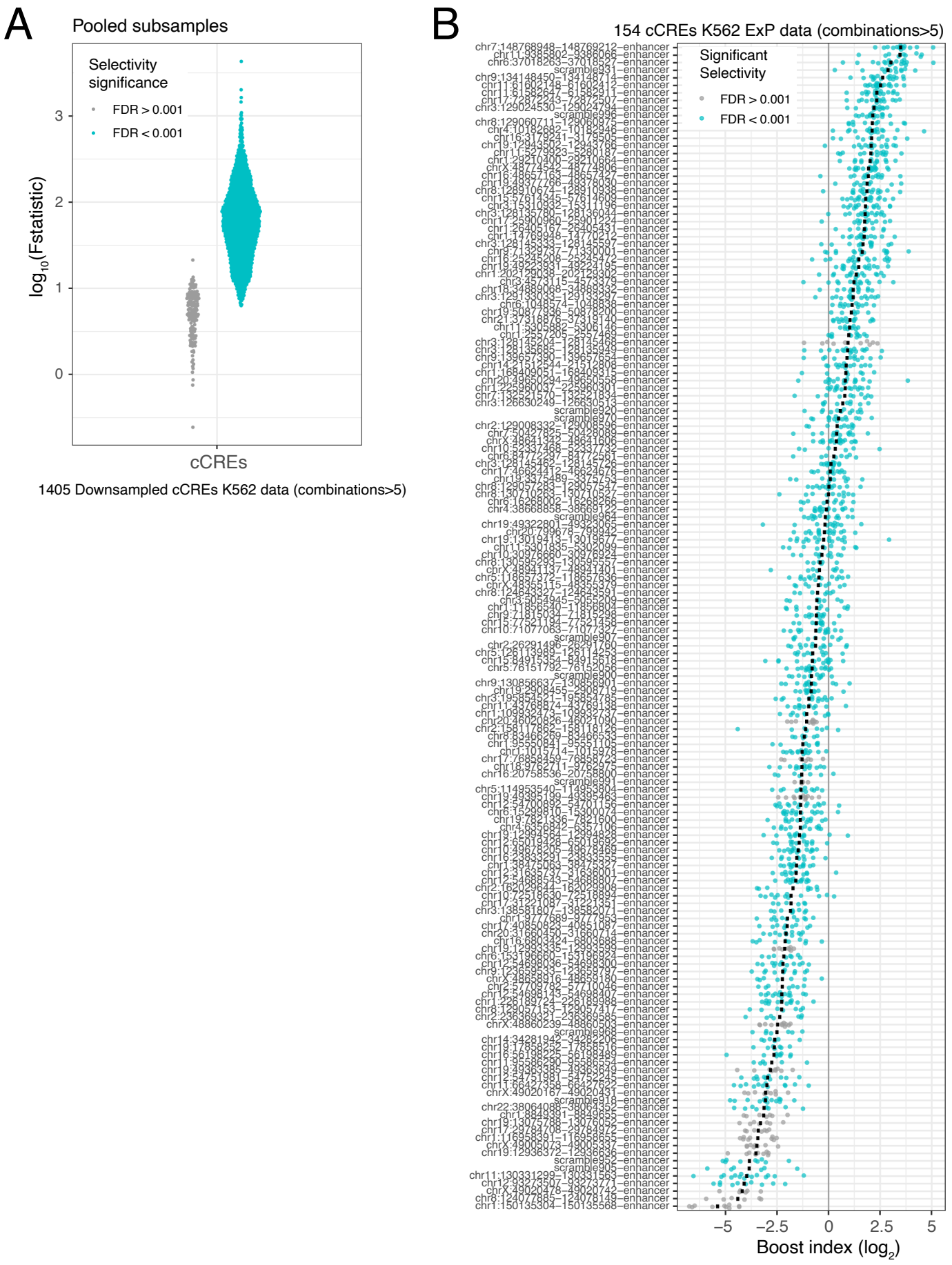


Figure S6. Promoter selectivity of cCREs analysis on ExP K562 data (Bergman et al., 2021). Related to Figure 4. A) Summary of Welch F-test selectivity analysis results for all the 10 subsamples of cCREs from the ExP K562 data. From the whole dataset cCRE-P combinations were downsampled 10 times to have 10 subsamples of similar complexity as the dataset generated in this paper. Each dot represents one cCRE with more than 5 cCRE–P combinations in the subsample. Significantly selective cCREs (0.1% FDR cutoff) are highlighted in turquoise. B) Boost index distributions for each cCRE from one of the 10 subsamples. Each dot represents one cCRE–P combination; black bar represents the mean. Turquoise colouring marks cCREs that have a larger variance of their boost indices than may be expected based on experimental noise, according to the Welch F-test after multiple hypothesis correction (0.1% FDR cutoff).

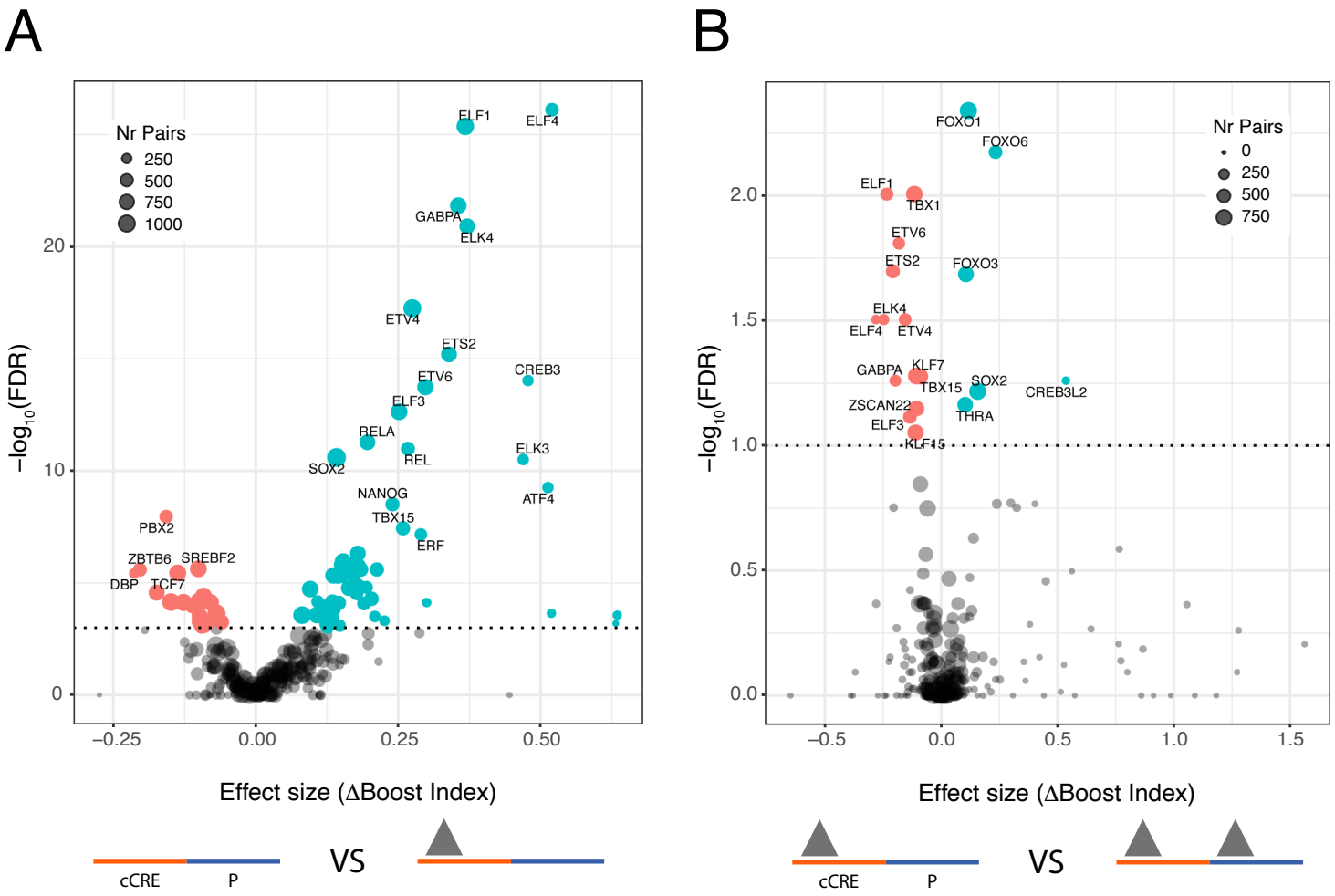


Figure S7. Identification of single TF motifs that correlate with boost indices. Related to Figure 6. **(A)** TF motifs in cCREs associated (at 1% FDR cutoff) with activation (turquoise) or repression (red). **(B)** Motifs of putative self-compatible TFs, i.e. motifs that predict increased or reduced boosting indices when present both at the cCRE and P, compared to being present only at the cCRE. TF motifs associated with higher or lower boost indices at a 1% FDR cutoff are highlighted. We note that TF motifs with multiple hits from the same family, such as for ELK, FOXO and ELF factors, may in fact be due to the activity of one TF motif of that family [69].