**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)<br>ttggttGGgctagc<br>(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)<br>AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT |
| 465JvA | Barcoding Primer | **AAGCAATCTCTATACGGAGTTCAGT**AGGTTAACCAGATCCC<br>TTCCGGAATTCCAAGGTTG |
| 274JvA | Barcoding ultramer template | Agatcggaagagcgtcgtgtagggaaagagtgttagggataacagggtaatgcggcc<br>Gctggccgcaataaaatatctttatttcattacatctgtgtgttggttttttgtgtgaggatctgtg<br>actggagttcagacgtgtgctcttccgatct**ccagtgatgtgatgg**ttggccaaccttggaattccgg |
| 304JvA | GSP for reverse transcription | TACAGAGCTGACGTATCAGTACGGCCGCATTACCCTGTTATCCCTAACACTC |
| 285JvA | indexed illumina sequencing primer | CAAGCAGAAGACGGCATACGAGAT**ACAGCA**<br>*GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT*<br>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 |

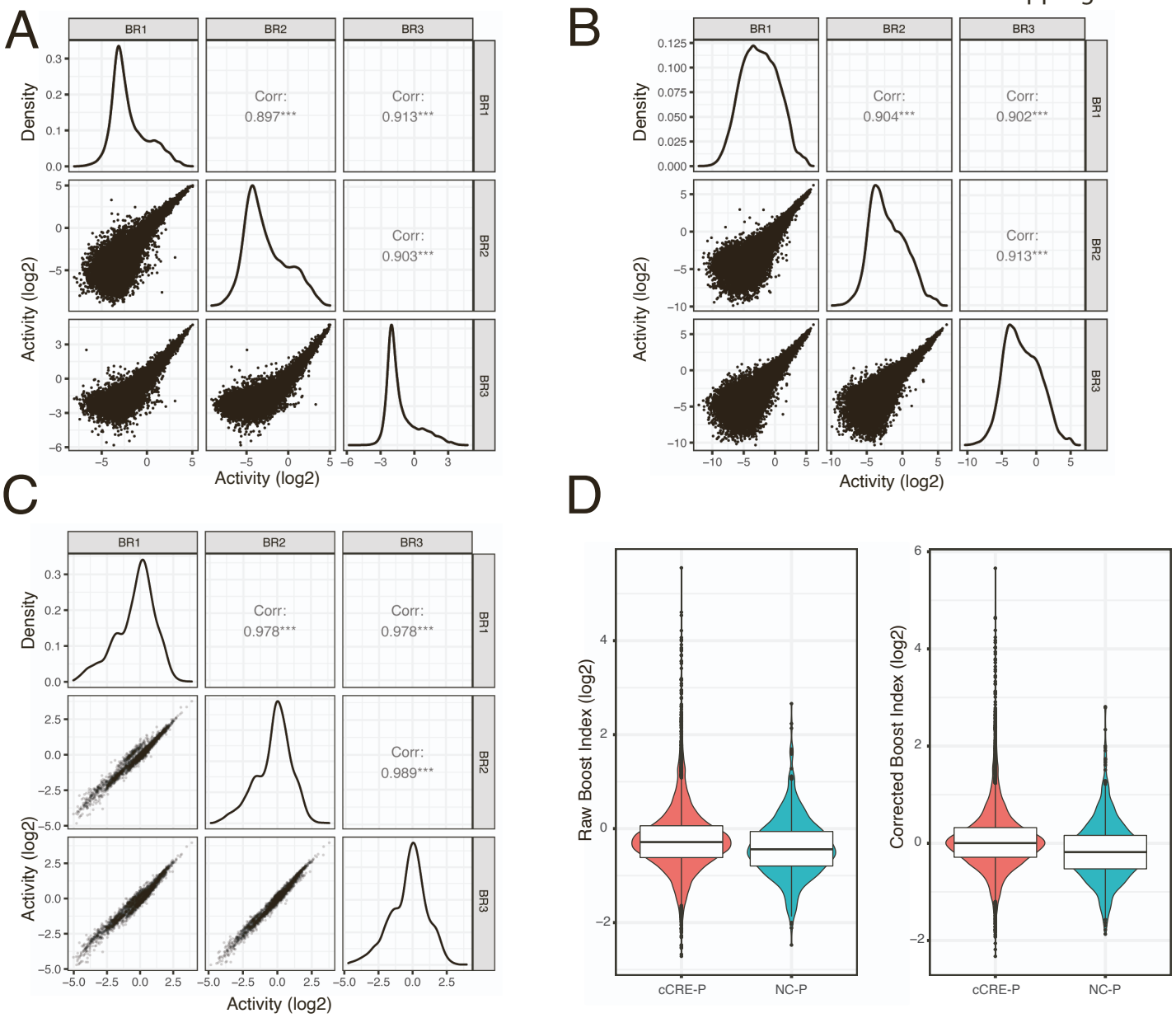


**Figure S1.** Reproducibility of data and boost index calculation. Related to Figure 2 and STAR methods. **(A-C)** Correlograms of the three biological replicates of each library pool. Lower left panels show pairwise scatterplots of the activities of all cCRE-P pairs per replicate. Middle panels show 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).

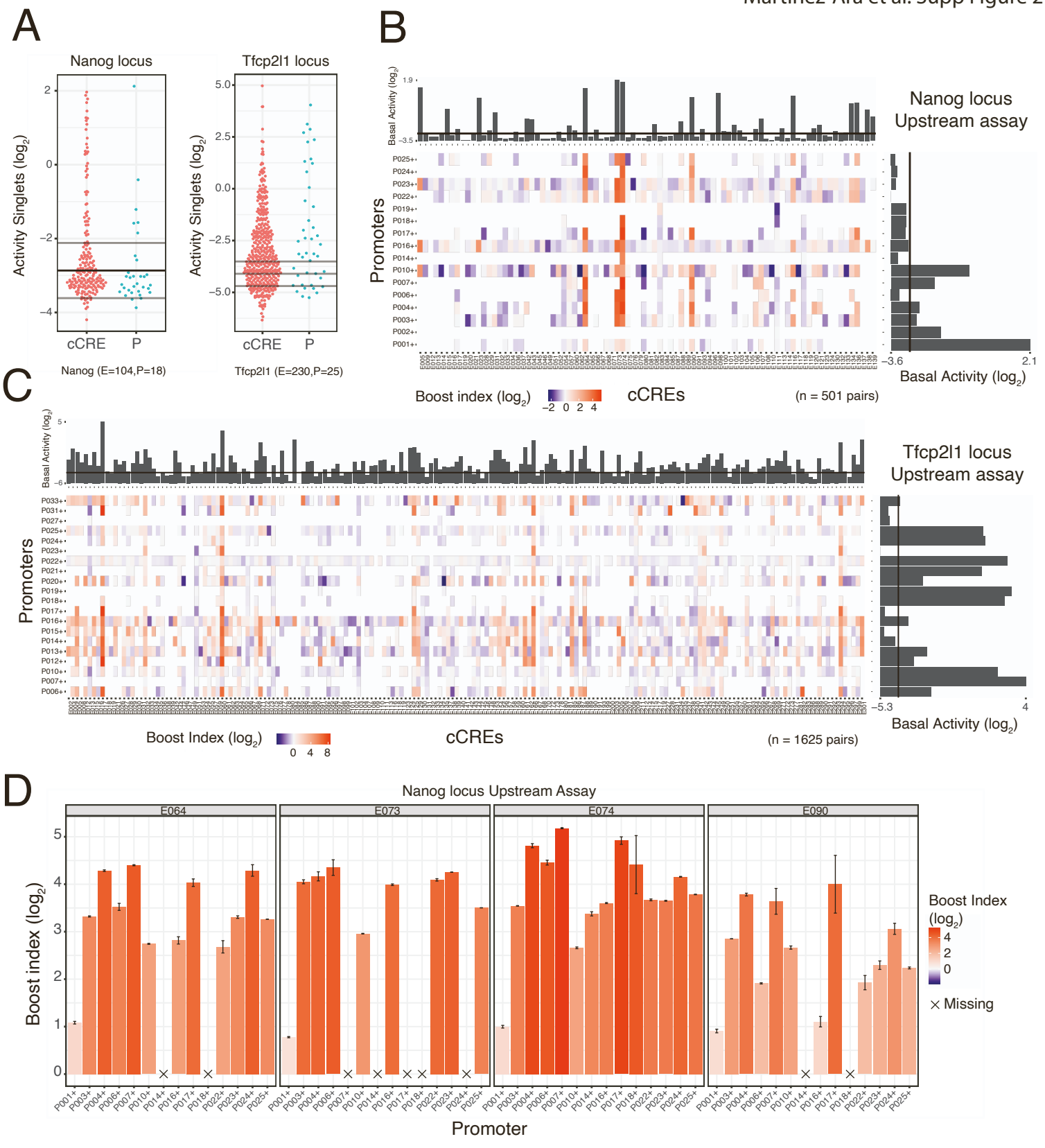


**Figure S2.** Element activities and boost indices obtained with Nanog and Tfcp2l1 Upstream libraries. Related to Figures 2 and 3. **A)** Transcriptional activities of cCREs and promoters. Each dot represents the mean activity of one singlet. Horizontal lines represent the average background activity 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.

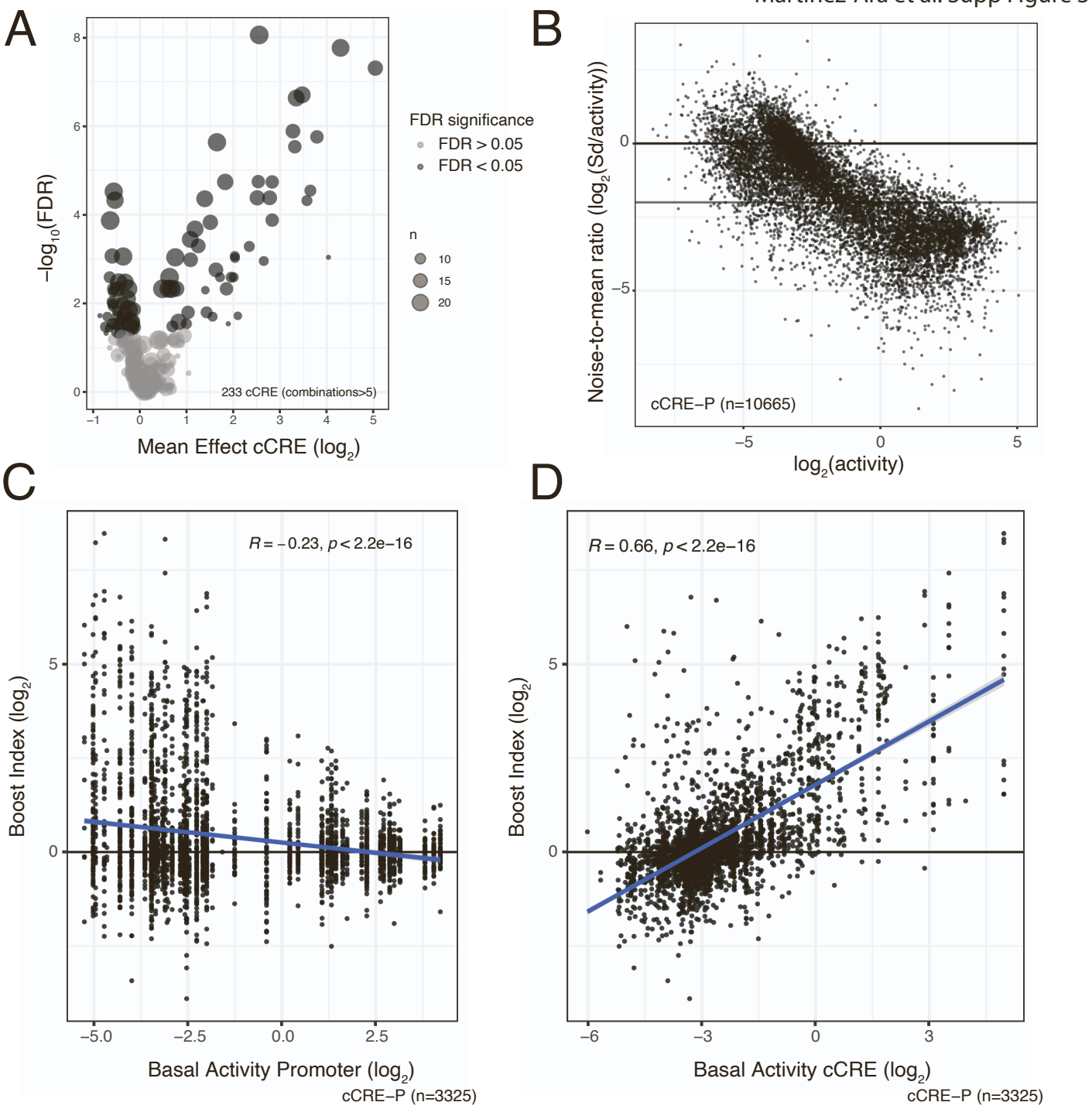


**Figure S3.** cCRE functional classification and activity influence on Boost indices. Related to Figures 2 to 4. **A)** Volcano plot of cCREs associated with activation or repression across promoters. A Wilcoxon test is performed per cCRE comparing the boost indices of all the cCRE-P combinations of 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.

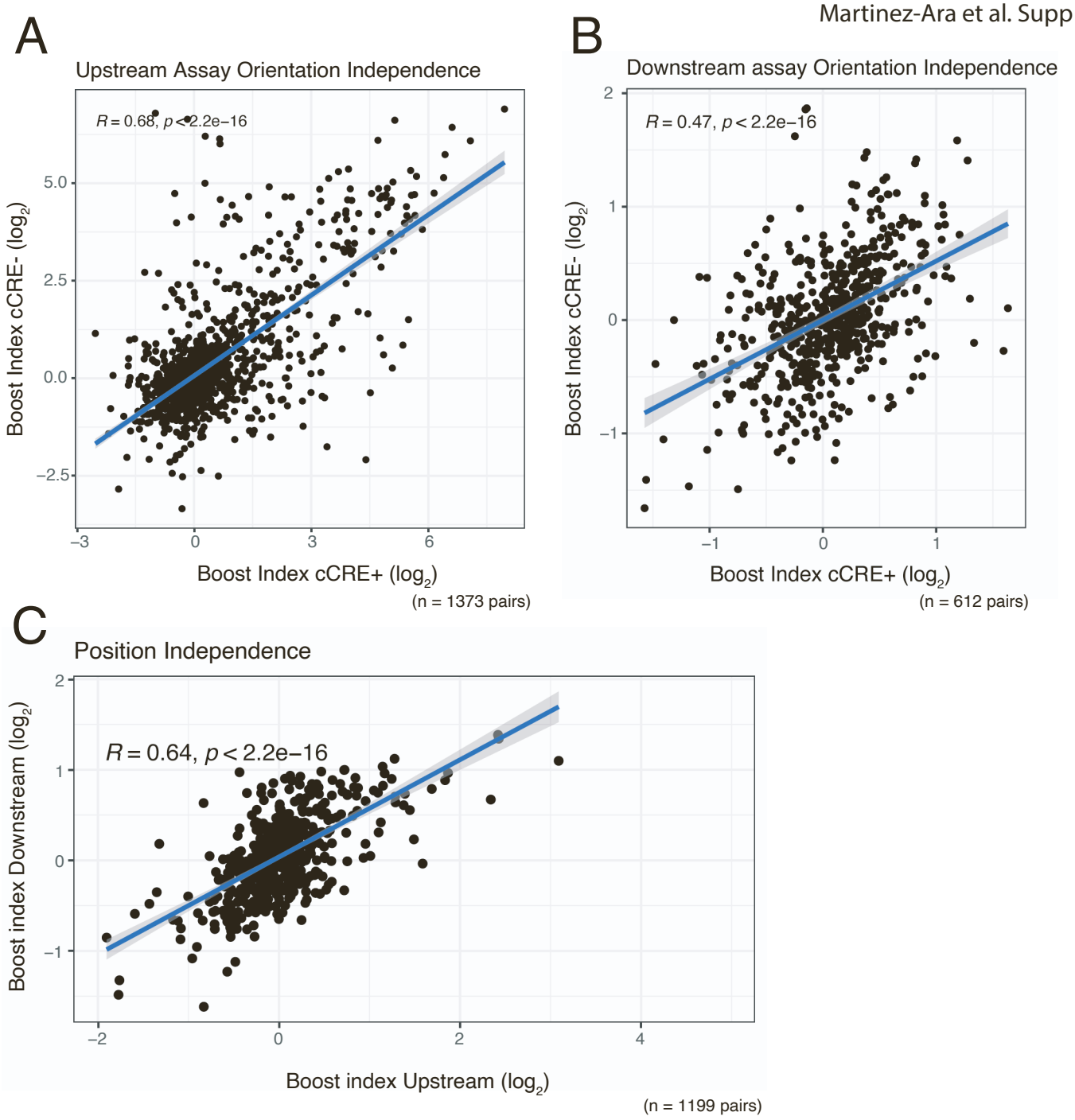


**Figure S4.** Orientation and position independence of cCREs. Related to STAR methods. **(A-B)** Correlation between boost indices of both cCRE orientations of the same cCRE-P combination, in the **(A)** Upstream assay and **(B)** Downstream assay. Data are from the Klf2 locus libraries. Note that "+" 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.

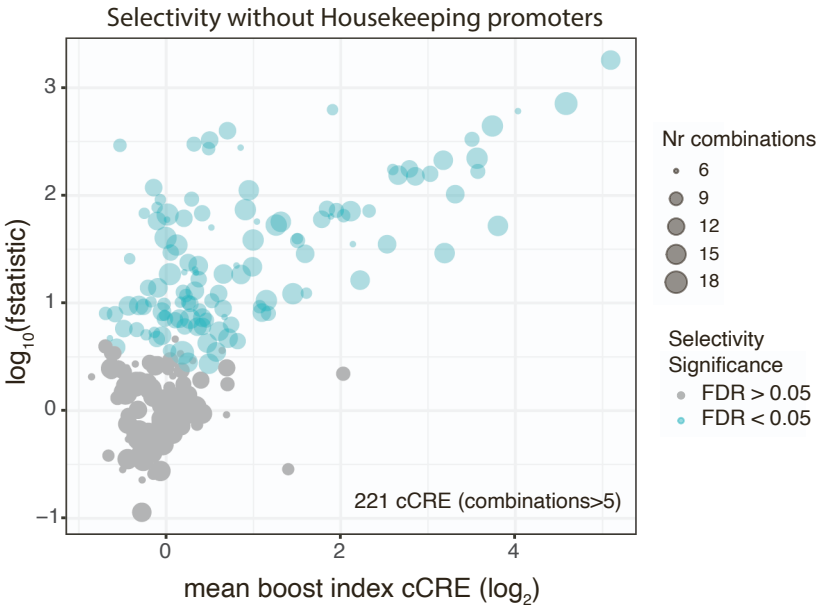


**Figure S5.** Selectivity is widespread among non-housekeeping promoters. Related to Figure 5. Results of selectivity analysis as performed in Figure 4C, but excluding housekeeping promoters [48]. Data are averages over 3 independent biological replicates.

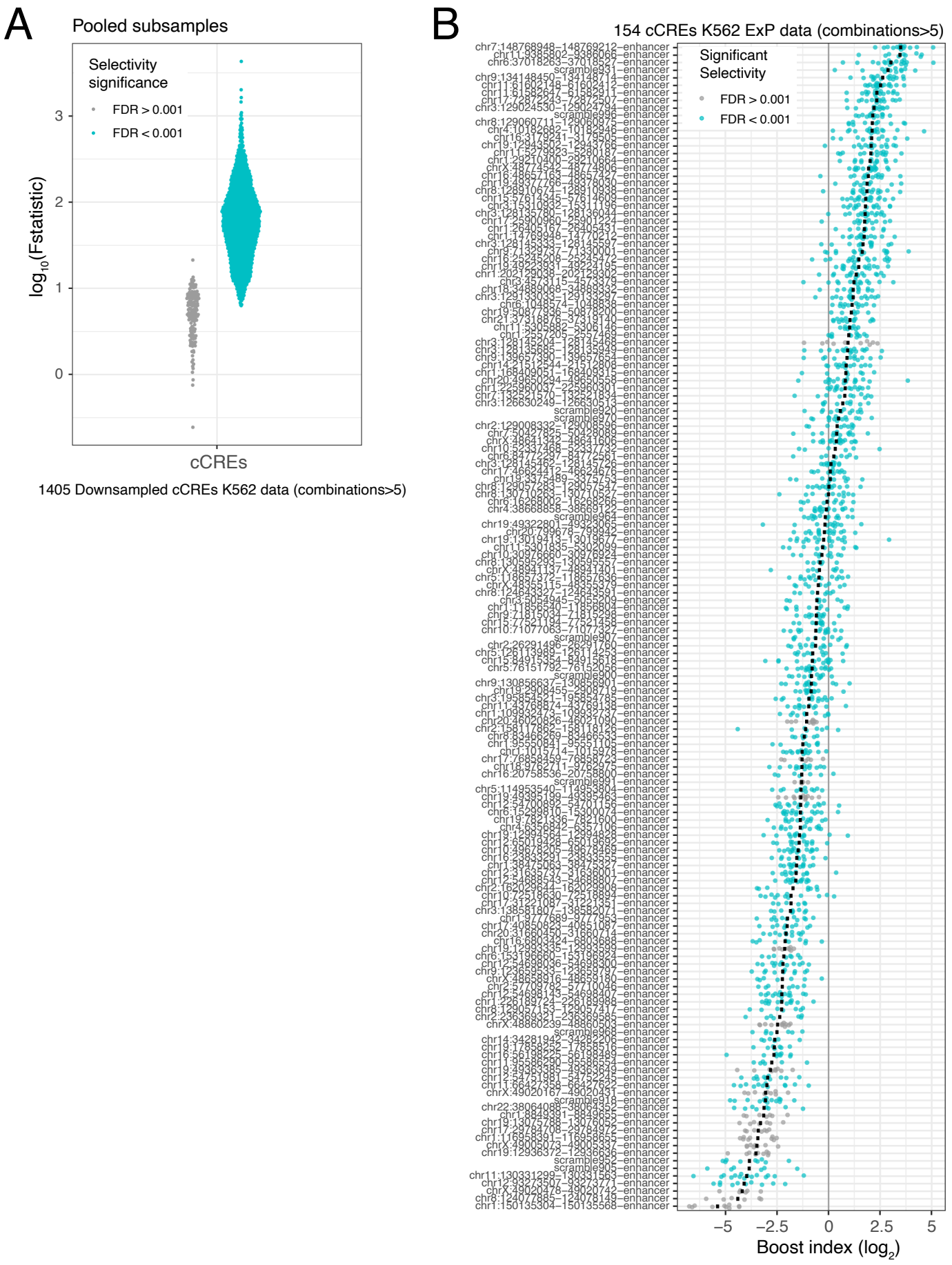


**Figure S6.** Promoter selectivity of cCREs analysis on ExP K562 data (Bergman et al., 2021). Related to Figure 4. A) Summary of Welch F-test selectivity analysis results for all the 10 subsamples of cCREs from the ExP K562 data. From the whole dataset cCRE-P combinations were downsampled 10 times to 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).

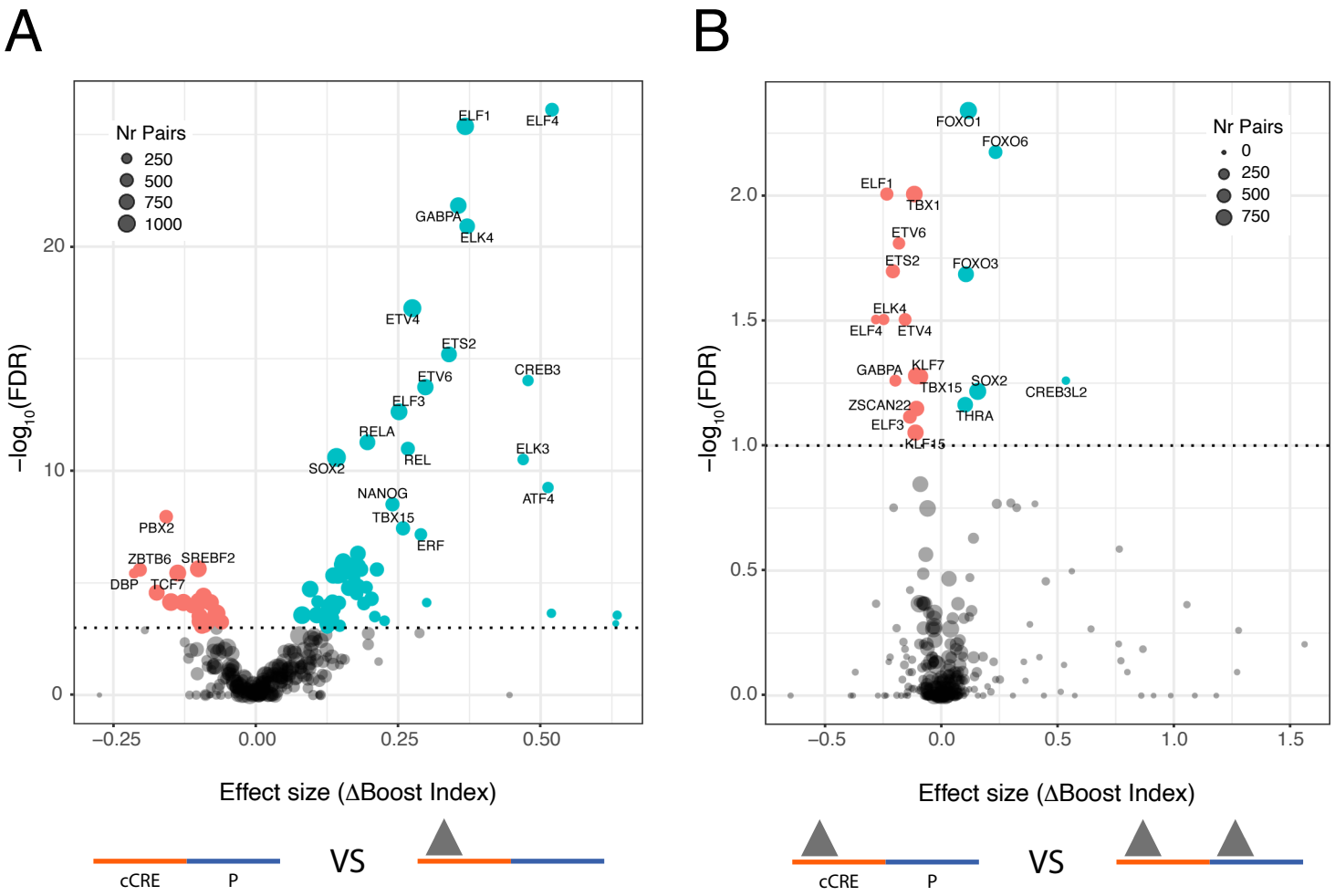


**Figure S7.** Identification of single TF motifs that correlate with boost indices. Related to Figure 6. **(A)** TF motifs in cCREs associated (at 1% FDR cutoff) with activation (turquoise) or repression (red). **(B)** Motifs of putative self-compatible TFs, i.e. motifs that predict increased or reduced 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].