
Supplementary information

An expanded registry of candidate *cis*-regulatory elements

In the format provided by the
authors and unedited

Supplementary Information: An Expanded Registry of Candidate cis-Regulatory Elements

Table of Contents

Supplementary Notes	5
Supplementary Note 1 - Detailed updates to the Registry of cCREs	6
1.1 Defining element anchors	6
1.2 Multi-mapping sites	8
1.3 cCRE Data Collections	10
1.4 Classification of cCREs	11
1.5 Aggregate level cCRE annotations	13
1.6 Evolutionary conservation across cCRE classes	14
1.7 Expanded Registry of cCREs maintains specificity for functional elements	16
1.8 Using variational autoencoders to extract cCRE sequence features	19
1.9 Caveats and considerations when using the Registry of cCREs	22
Supplementary Note 2 - Functional characterization analysis	24
2.1 Integrating functional characterization data with the Registry of cCREs	24
Whole genome STARR-seq	24
MPRA	25
CRISPRi-FlowFISH	26
Additional CRISPR perturbations	26
Cell type-specific	27
2.2 The CAPRA Pipeline	27
2.3 Comparing whole genome STARR-seq datasets	31
2.4 STARR-seq activity of closed chromatin cCREs	33
2.5 Combinatorial effects of cCRE activity	35
2.6 TP53 activity in cancer cell lines	36
Supplementary Note 3 - Characterizing REST+ cCREs	37
3.1 Defining classes of REST+ cCREs	37
3.2 Examples of REST+ cCREs	41
Supplementary Note 4 - Identifying additional silencer cCREs and comparing them with existing collections of silencer elements	42
4.1 Reproducibility of STARR silencers	42
4.2 CRISPRi-FLOWFish validation of STARR silencers	43
4.3 Overlap of silencer cCREs with previously published collections	44
4.4 Chromatin accessibility enrichment	46
4.5 Overlap of K562 silencers with ChromHMM states	48

4.6 Conservation of K562 silencers and overlap with repeats	49
4.7 STARR silencer activity of non-cCRE regions	51
Supplementary Note 5 - MAFF and MAFK bound cCREs	51
5.1 Defining MAFF/MAFK+ cCREs	51
5.2 MAFF/MAFK+ cCREs do not have silencing activity	52
5.3 Gene ontology analysis	53
5.4 Enrichment of chromatin signatures at MAFF/MAFK+ enhancer cCREs	53
Supplementary Note 6 - Overlap of the Registry of cCREs with other classes of Encyclopedia annotations	54
6.1 Chromatin states (elements)	54
6.2 ENTEEx (sequence & elements)	55
6.3 Annotated GENCODE genes (genes)	55
6.4 Novel transcripts (genes)	56
6.5 Sites of nascent transcription (elements)	58
6.6 3D chromatin interactions (interactions)	58
6.7 Functional interactions (interactions)	59
6.8 Enhancer-gene interaction predictions (interactions)	59
Supplementary Note 7 - SCREEN	60
Supplementary Note 8 - Prioritizing variants at the RTBDN-MAST1 locus	61
Supplementary Note 9 - Additional Acknowledgements	65
9.1 Members of the ENCODE4 cCRE working group:	65
9.2 Members of the ENCODE4 TSS annotation working group:	65
9.3 Members of the ENCODE4 Biosamples working group:	65
Supplementary Methods	66
Estimating the number of mouse enhancer cCREs	66
Proximity of experimentally derived transcription start sites with rDHSs	66
Overlap of histone mark peaks with rDHSs	66
Overlapping transcription factor peaks with rDHSs	66
Identifying and characterizing transcription factors with low overlap with rDHSs	67
Calling Multi-mapping cCREs	67
Generating multi-level cCRE annotations	67
Characterizing the expansion of the Registry	68
Overlap with BICCN cCREs	69
Overlap with Loupe et al. union peaks	69
Overlap with other annotated CREs	70
Classification of STARR+ distal enhancers	70
Enrichment of transcription factor motifs in cCRE pairs	70
Gene expression near candidate silencers	71

Enrichment for transcription factor binding site at STARR silencers	71
Identifying allele-specific open chromatin	71
Calling allele-specific transcription factor binding in GM12878	72
Phylogenetic analysis of KLF1 missense mutation	73
Supplementary References	73
Supplementary Figures	77
Supplementary Figure 1 Expanding cCRE anchors to include transcription factor binding sites	78
Supplementary Figure 2 Including cCREs at multi-mapping loci identifies promoters for recently duplicated genes	80
Supplementary Figure 3 Size distribution of cCREs	82
Supplementary Figure 4 Evolutionary conservation of cCREs	84
Supplementary Figure 5 Characterizing the expanded Registry	86
Supplementary Figure 6 ENCODE4 cCREs have increased overlap with external regulatory element collections	88
Supplementary Figure 7 Distinct sequence features of promoter and enhancer cCREs	90
Supplementary Figure 8 Investigating GC content of promoter and enhancer cCREs	92
Supplementary Figure 9 Coverage of representative DNase Hypersensitive sites across tissues and organs	94
Supplementary Figure 10 cCREs are enriched for cell type-specific functional activity	96
Supplementary Figure 11 Overview of CAPRA pipeline	98
Supplementary Figure 12 CAPRA use cases	100
Supplementary Figure 13 Cell type-specific STARR activity is concordant with cell type-specific chromatin signatures	102
Supplementary Figure 14 Using CAPRA to dissect combinatorial activities of cCREs	104
Supplementary Figure 15 Characteristics of REST+ cCREs	106
Supplementary Figure 16 Characteristics of STARR silencer cCREs	108
Supplementary Figure 17 Silencer cCRE EH38E4193243 regulates PRDX2 via CTCF-Mediated chromatin interactions	110
Supplementary Figure 18 Comparisons with published silencers	112
Supplementary Figure 19 Defining MAFF/MAFK+ cCREs	114
Supplementary Figure 20 Characterizing MAFF/MAFK+ cCREs	116
Supplementary Figure 21 SCREEN	118
Supplementary Figure 22 Prioritizing genes and variants associated with red blood cell traits	120

Supplementary Notes

Supplementary Note 1 - Detailed updates to the Registry of cCREs

1.1 Defining element anchors

Similar to previous versions⁹, the ENCODE4 Registry of cCREs is primarily anchored on representative DNase hypersensitivity sites (rDHSs) derived from DNase-seq data^{64,65}. Briefly, to define rDHSs, we applied a uniform peak-calling and filtering strategy across all available high-quality DNase-seq profiles, followed by iterative clustering to select the strongest peak in each region and reduce redundancy. For ENCODE4, we aggregated 178,746,635 individual DHSs across 1,438 DNase profiles to annotate 2,863,480 rDHSs in human; for mouse we aggregated 54,542,655 individual DHSs across 466 DNase profiles to annotate 1,338,731 rDHSs.

In general, rDHSs capture the vast majority of biochemically active sites across the genome; they have high overlap with experimentally derived transcription start sites (TSSs) from CAGE and RAMPAGE (overlaps of 90 and 95%, respectively, **Supplementary Table 2a**), active histone mark peaks including H3K4me3 (average overlap of 96%, **Supplementary Table 2b**), H3K27ac (92%, **Supplementary Table 2c**), and H3K4me1 (91%, **Supplementary Table 2d**) and transcription factor ChIP-seq peaks (average overlap of 99%, **Supplementary Table 2d**). However, there was a subset of transcription factors with noticeably lower overlap with rDHSs (**Supplementary Figure 1a**). These included ZNF274 (on average 71% of peaks overlap with rDHSs), MAFK (73%), MAFF (73%), ATF2 (73%), ATF7 (84%), and CREB1 (85%). We computed the overlap between non-rDHS peaks for these transcription factors and defined five primary clusters based on hierarchical clustering (**Supplementary Figure 1b**). While there are a relatively few of these regions across the genome, they have higher conservation than non-cCRE regions, suggesting they may have a functional role ($p < 4.6 \times 10^{-16}$, pairwise, two-sided, Wilcoxon rank sum test with FDR correction, **Supplementary Figure 1c**). Therefore, to account for these putative regulatory regions, we supplemented our rDHSs with representative transcription factor clusters (rClusters)—clusters of peaks that represent binding sites for at least five transcription factors (**Supplementary Figure 1d**).

We selected a threshold of five ChIP-seq peaks to define transcription factor rClusters by performing several analyses aimed at balancing sensitivity and specificity. First, we generated

an elbow plot of the number of rClusters as a function of the number of supporting transcription factors, observing a clear inflection point at five factors, consistent with diminishing returns in signal-to-noise beyond this cutoff (**Supplementary Figure 1e**). Second we compared evolutionary conservation (PhastCons scores) for rClusters that did or did not overlap rDHSs, stratified by the number of supporting peaks (**Supplementary Figure 1d**). Interestingly, clusters supported by one to three transcription factors that did not overlap rDHSs showed slightly higher conservation than those that did (PhastCons score of 0.042 vs. 0.038, two-sided Wilcoxon test, $p < 2.2 \times 10^{-16}$) but this pattern reversed at five or more transcription factors, where rClusters overlapping rDHSs became more conserved—suggesting a more reliable regulatory signal beyond this threshold. Finally, to assess evolutionary constraint from a complementary perspective, we examined non-rDHS rClusters by their enrichment in great ape-specific regions based on the Zoonomia alignment triangle (**Supplementary Figure 1e**). This analysis revealed a steep drop in great ape-specific content beginning at four transcription factors, further motivating our use of a more conservative threshold of five. Collectively, these results support the heuristic cutoff of at least five transcription factors as a robust criterion for defining rClusters, balancing evolutionary conservation, cross-species constraint, and regulatory signal complexity.

For the human genome, we curated 821,672 transcription factor rClusters by combining 2,509 transcription factor ChIP-seq datasets; 86,748 (11%) did not overlap an rDHS and were included in our anchoring scheme. For the mouse genome, we curated 171,092 transcription factor rClusters by combining 167 transcription factor ChIP-seq datasets; 7,658 (4%) did not overlap an rDHS and were included in our anchoring scheme.

We scored all anchor regions (rDHSs and transcription factor rClusters) for DNase, ATAC, H3K4me3, H3K27ac and CTCF signals using our previously published z-score normalization method⁹. Briefly, we normalized epigenomic signals at rDHSs by computing Z-scores of the log10-transformed signal for each mark in each biosample. This approach accounts for differences in signal scaling across assays and biosamples, enabling direct comparability despite variation in raw read counts or fold-change values. Our final set of cCREs included all

rDHSs with a DNase maxZ score (maximum Z-score across all samples) greater than 1.64 and all transcription factor rClusters resulting in 2,348,854 cCREs in human and 926,843 in mouse.

1.2 Multi-mapping sites

Most ENCODE uniform processing pipelines, including those for DNase-seq and ChIP-seq, filter out reads that map to multiple genomic loci known as multi-mapping reads. This filtering is implemented to reduce false positive peak calls and mapping artifacts. However, focusing exclusively on uniquely mapping reads creates "holes" in the genome, particularly at recently duplicated genes.

To build a more comprehensive collection of cCREs, we used multi-mapping reads to call DHSs (and subsequently rDHSs) at a curated set of genomic loci. We identified these loci (1,036 for human and 1,470 for mouse) by analyzing high-confidence ATAC-seq peaks that did not overlap rDHSs (**Supplementary Table 3a,b**). Unlike DNase-seq, the ENCODE4 ATAC-seq pipeline allows reads to map to the genome up to four times, selecting the best matching location. When reads map equally well to multiple locations, they are randomly assigned. As a result, ATAC-seq can detect peaks in highly similar genomic regions, such as recently duplicated genes and repeats, missed by traditional DHS calling.

We called DHSs using multi-mapping reads specifically at these curated loci. This approach added an average of 2,292 DHSs per experiment (**Supplementary Table 3c,d**), resulting in a total of 24,160 additional rDHSs in the human genome and 40,537 in the mouse genome. These DHSs, which do not overlap existing cCRE anchors, are referred to as *multi-mapping* cCREs. This addition results in a total of 2,373,014 human and 967,380 mouse cCREs.

In humans, multi-mapping cCREs were enriched near recently duplicated genes. Acharya *et al.*⁶⁶ annotated 11,710 human genes as either arising from whole genome duplication events that took place early during vertebrate evolution (N=7,070, 60%) or more recent small scale duplication events (N=4,640, 40%). Using our standard collection of non-multi-mapping cCREs, we identified promoter cCREs for 96% of whole genome duplication genes but only 69% of small scale duplication genes (**Supplementary Figure 2a,b**, two-sided Fisher's exact

test, $p < 2.2 \times 10^{-16}$). By including *multi-mapping* cCREs, we now annotate promoters for 667 additional GENCODE V40 genes, of which 96 are small scale duplication genes and 12 are whole genome duplication genes (**Supplementary Figure 2a,b**).

As an example, *EIF3C* and *EIF3CL*, encode subunits of the eukaryotic translation initiation factor 3 (eIF3) complex and arose from a small scale duplicate event (**Supplementary Figure 2c,d**). Due to their high sequence similarities, the regions around their promoters have low mappability, as indicated by Umap mappability annotations from the Hoffman Lab⁶⁷. By including peaks called from multi-mapping reads, we can now annotate dozens of cCREs in the region, including the promoter cCREs for both genes, that were previously missed. This example highlights how including multi-mapping reads in our analysis provides a more complete annotation of cCREs, particularly at loci with low mappability, offering valuable insights into regulatory elements at recently duplicated genes that would otherwise remain obscured.

In addition to their enrichment near recently duplicated genes, multi-mapping cCREs were also more likely to overlap repetitive elements. We intersected multi-mapping cCREs, uniquely mapping cCREs, and non-cCRE genomic regions with RepeatMasker annotations⁵³ (**Supplementary Figure 2e, Supplementary Table 3e**). Compared to uniquely mapping cCREs, multi-mapping cCREs exhibited a significantly higher overlap with annotated repeats (68% vs 58%, two-sided Fisher's Exact Test, $p < 2.2 \times 10^{-16}$). This enrichment was primarily attributable to SINEs, particularly Alus, which overlapped 26% of multi-mapping cCREs versus 13% of uniquely mapping cCREs (two-sided Fisher's Exact Test, $p < 2.2 \times 10^{-16}$). These findings suggest that a substantial fraction of multi-mapping cCREs may have originated from transposable element insertions and highlight the utility of incorporating multi-mapping reads to recover regulatory elements embedded in repetitive genomic contexts.

Because most ENCODE data sets were processed using pipelines that exclude multi-mapping reads, the biochemical profiles underlying these multi-mapping cCREs are often incomplete. As a result, these elements cannot be reliably classified into standard regulatory categories such as promoter, enhancer, or CTCF-bound. We therefore recommend using the fully classified

collection of cCREs (N=2,348,854 for human and 926,843 for mouse) for most analyses and supplementing with multi-mapping cCREs when investigating loci with low mappability or recent duplication.

1.3 cCRE Data Collections

In both SCREEN's visualization, downloadable files, and UCSC genome browser track hubs, we annotate biosamples based on available data (**Supplementary Table 1a**).

Core collection

Thanks to the extensive coordination efforts by the ENCODE4 Biosample Working Group (**Supplementary Note 9.3**), 170 biosamples were profiled for DNase, H3K4me3, H3K27ac, and CTCF. We refer to these samples as the biosample-specific *Core Collection* of cCREs. These samples derive from a variety of organs and tissues and mainly comprise primary tissues and cells (**Extended Data Figure 1a,c**). We suggest that users prioritize these samples for their analysis as they contain all the relevant marks for the most complete annotation of cCREs.

Partial Data Collection

To supplement the Core Collection, 1,154 biosamples have DNase in addition to various combinations of the other marks (but not all three). Though we are unable to annotate the full spectrum of cCRE classes in these biosamples, having DNase enables us to annotate element boundaries with high resolution. Therefore, we refer to this group as the *Partial Data Collection*. In these biosamples, we classify elements using the available marks. For example, if a sample lacks H3K27ac and CTCF, its cCREs can only be assigned to the promoter, CA-H3K4me3, and CA groups, not the enhancer nor CA-CTCF groups. The *Partial Data Collection* contains unique tissues and cell states that are not represented in the *Core Collection*, such as embryonic brain tissue and stimulated immune cells that may be of high interest to some researchers. Therefore, if users are interested in cCRE annotations in such biosamples, we suggest leveraging the cell type-agnostic annotations or annotations from similar biosamples in the Core Collection, to anchor their analyses.

Ancillary Collection

For the 354 biosamples lacking DNase data, we do not have the resolution to identify specific elements and we refer to these annotations as the *Ancillary Collection*. In these biosamples, we simply label cCREs as having a high or low signal for every available assay. We highly suggest that users do not use annotations from the *Ancillary Collection* unless they are anchoring their analysis on cCREs from the *Core* or *Partial Data Collections*.

Custom Classification

While our classification schemes place each cCRE into specific, individual classes, the signal strengths for all recorded epigenetic features are retained for each cCRE in the Registry and can be used for customized searches by users. For example, users may want promoters that have high DNase, H3K4me3, and H3K27ac to distinguish from poised promoters that often lack H3K27ac signal. Additionally, by default, all chromatin accessibility annotations use DNase signal. If users prefer to use ATAC signals, this can also be accomplished using the SCREEN API.

1.4 Classification of cCREs

Many uses of cCREs are based on the regulatory role associated with their biochemical signatures. Analogous to GENCODE's catalog of genes, which are defined irrespective of their expression levels and transcript usage across cell types, we provide a general, cell type-agnostic classification of cCREs. This classification is based on each element's dominant biochemical signals across all available biosamples and its proximity to the nearest GENCODE transcription start site (**Figure 1b**).

- Promoter-like signatures (**promoter**) must fall within 200 bp of a TSS and have high chromatin accessibility and H3K4me3 signals.
- TSS-proximal enhancer-like signatures (**proximal enhancer**) have high chromatin accessibility and H3K27ac signals and are within 2 kb of an annotated TSS. If they are within 200 bp of a TSS, they must also have low H3K4me3 signal.
- TSS-distal enhancer-like signatures (**distal enhancer**) have high chromatin accessibility and H3K27ac signals and are farther than 2 kb from an annotated TSS.

- Chromatin accessibility + H3K4me3 (CA-H3K4me3) have high chromatin accessibility and H3K4me3 signals but low H3K27ac signals and do not fall within 200 bp of a TSS.
- Chromatin accessibility + CTCF (CA-CTCF) have high chromatin accessibility and CTCF signals but low H3K4me3 and H3K27ac signals.
- Chromatin accessibility + transcription factor (CA-TF) have high chromatin accessibility, low H3K4me3, H3K27ac, and CTCF signals, and are bound by a transcription factor.
- Chromatin accessibility (CA) have high chromatin accessibility and low H3K4me3, H3K27ac, and CTCF signals.
- Transcription factor (TF) have low chromatin accessibility, low H3K4me3, H3K27ac, and CTCF signals and are bound by a transcription factor.

In addition to these cell type-agnostic classifications, we also evaluated the biochemical activity of each cCRE in individual biosamples using the corresponding biosample-specific DNase, ATAC, H3K4me3, H3K27ac, and CTCF data. This allows us to annotate active cCREs in individual biosamples. Elements with low DNase Z-scores in individual biosamples are labeled with “Low Accessibility.”

Because of the uneven distribution of transcription factor data across biosamples and our desire to reduce false positive annotations, we do not specifically annotate the TF class of elements in individual cell types. We instead provide an aggregate list of these TF cCREs annotated with their supporting transcription factor ChIP-seq peaks. For cell types with transcription factor ChIP-seq data, we annotate CA-TF elements since the accessible chromatin corroborates the transcription factor binding.

These classes of cCREs vary in their genomic properties, including element size, which may reflect underlying biological features such as transcription factor occupancy, chromatin architecture, or nucleosome organization. Similar to the previous ENCODE3 release, ENCODE4 cCREs range in size from 150 to 350 bp, with a median width of 273 bp. We observe subtle but consistent differences across annotation classes: promoter cCREs tend to be wider (median 320 bp), while CA-CTCF elements are narrower (median 232 bp) (**Supplementary Figure 3**). These differences likely reflect variation in the strength and breadth of chromatin accessibility,

as broader DNase hypersensitive sites are generally associated with stronger regulatory activity.

1.5 Aggregate level cCRE annotations

To enable organ/tissue-specific interpretation of regulatory activity, we generated group-level cCRE annotations by aggregating biosamples based on organ/tissue of origin (**Supplementary Table 4a**). This approach expands upon our previously published strategy for defining brain-specific cCREs (b-cCREs; Pratt et al., 2024) by applying a consistent, consensus-based pipeline to dozens of new groups.

For each group, we scored cCREs using DNase-seq data from all biosamples within the group. We then retained cCREs that passed one or more of the following thresholds based on DNase Z-scores:

- Group-supported activity: $Z > 1.64$ (95th percentile) in ≥ 5 biosamples
- Moderately strong activity: $Z > 2.02$ (98th percentile) in ≥ 2 biosamples
- Highly strong activity: $Z > 2.32$ (99th percentile) in ≥ 1 biosample

These thresholds ensure inclusion of both shared and biosample-specific accessible regions. We refer to the resulting regions as group anchor cCREs.

Group anchor cCREs were then classified (promoter, proximal/distal enhancer, CA-H3K4me3, CA-CTCF, CA-TF, CA, Low Accessibility) using the same logic as the cell type-agnostic cCREs with maximum Z-scores for H3K4me3, H3K27ac, and CTCF signals and transcription factor ChIP-seq across all biosamples in the group.

To support more tailored analyses, we also provide the option to include or exclude cancer cell lines from each organ/tissue group. This flexibility enables researchers to generate annotations reflective of normal physiology or to capture tumor-associated regulatory programs, depending on their application.

This structured framework allows users to access cCRE annotations tailored to their biological context of interest while maintaining consistency with the broader ENCODE Registry. Group-level cCRE annotations are available for download through SCREEN.

1.6 Evolutionary conservation across cCRE classes

To assess the evolutionary constraint and cross-species preservation of cCREs, we employed three complementary approaches: (1) average PhastCons conservation scores, (2) positional conservation profiles using PhyloP, and (3) alignment breadth across mammals using Zoonomia-based conservation groups. Each method captures different dimensions of evolutionary conservation, enabling a more comprehensive assessment of regulatory element function and constraint.

1. Average conservation using PhastCons scores

We calculated the average PhastCons score across each cCRE using the 100-way vertebrate alignment. This score reflects the probability that each base is in a conserved element, integrating substitution rate and phylogenetic depth⁶⁸.

All cCRE classes showed significantly higher PhastCons scores than randomly selected accessible regions (**Supplementary Figure 4a**, pairwise, two-sided, Wilcoxon test with FDR correction, $p < 2.2 \times 10^{-16}$). Comparing between classes, promoter cCREs exhibited the highest levels of conservation (median = 0.151), consistent with their central role in transcription initiation. Proximal and distal enhancer cCREs showed moderate conservation (medians = 0.042 and 0.038, respectively), with proximal enhancers scoring slightly higher, reflecting their closer proximity to gene promoters. Among newly defined classes, CA-TF elements—characterized by strong chromatin accessibility and transcription factor binding—had the second highest conservation overall (median = 0.044), surpassing all other non-promoter classes. CA-CTCF and CA-H3K4me3 elements also showed higher conservation than background and were comparable to previously defined cCRE categories.

2. Positionally resolved conservation using PhyloP scores

To examine the positional structure of conservation around regulatory elements, we used PhyloP scores from the 100-way vertebrate alignment (**Supplementary Figure 4b**) and Zoonomia's 240-way mammal alignment (**Supplementary Figure 4c**). For each cCRE group, we generated aggregate plots showing the average PhyloP score centered on each cCRE. These plots provide insight into localized constraint patterns, such as the presence of highly conserved transcription factor binding sites at element centers. Because PhyloP profiles can be noisy for small sets, we restricted this analysis to sufficiently large groups ($n > 1,000$ elements per class or condition).

For both mammalian and vertebrate conservation, promoter cCREs showed the highest conservation, with maximum average scores of 0.83 and 0.45, respectively, consistent with their essential regulatory roles. Enhancer cCREs also exhibited strong conservation, particularly proximal enhancer cCREs (0.27 mammalian, 0.22 vertebrate), followed by distal enhancer (0.15 mammalian, 0.20 vertebrate). CA-TF and CA-H3K4me3 cCREs showed intermediate conservation levels, slightly higher than CA-CTCF and general CA elements, which had modest average conservation. Notably, TF cCREs had the lowest average conservation—comparable to background—in both alignments, suggesting recent evolutionary origin.

3. Mammalian alignment breadth using Zoonomia N1/N2 classification

To capture evolutionary dynamics at a broader phylogenetic scale, we used alignment data from the Zoonomia project⁵⁶, which includes reference-free alignments across 241 mammalian genomes (**Supplementary Figures 4d-k**). For each cCRE, we computed N1, the number of non-human mammals in which $\geq 90\%$ of the cCRE is aligned, and N2, the number with $< 10\%$ alignment²². These values define a triangle in N1–N2 space that captures conservation breadth and divergence. We binned cCREs into groups (**Supplementary Table 4b**):

- Group 1 (G1): Highly conserved ($N1 \geq 120$, $N2 \leq 25$)
- Group 2 (G2): Actively evolving ($20 \leq N1 \leq 50$, $N2 \leq 120$)
- Group 3 (G3): Primate-specific ($N1 \leq 50$, $N2 \geq 180$)

This classification enables cross-comparison of evolutionary constraints across cCRE classes and provides insight into the emergence of lineage-specific regulatory elements.

As expected, promoter cCREs were most enriched in the highly conserved category (G1; 53%) and least represented in the primate-specific category (G3; 7%, **Supplementary Figure 4d**). Proximal and distal enhancer cCREs also showed strong conservation, with 43–46% in G1 and 11–12% in G3 (**Supplementary Figure 4e,f**). Newly defined classes displayed broader evolutionary diversity. CA-TF and CA cCREs, for example, had a greater distribution across groups, with 28–30% in G1, 11–13% in G2, and 22% in G3—suggesting a mixture of ancient and lineage-specific regulatory roles. Notably, TF cCREs showed the strongest skew toward primate specificity, with 29% falling into G3. While non-cCRE background regions had an even higher overall G3 proportion (39%), these were predominantly concentrated at the bottom vertex of the triangle, corresponding to conservation solely amongst great apes ($N1 \leq 5$, $N2 \geq 235$). In the non-cCRE background, this represented 17% of all non-cCRE regions and 42% of their G3 group. In contrast, all cCRE classes were markedly depleted for the great-ape specific group, with fewer than 6% of any one class falling into this category (**Supplementary Table 4b**). For instance, only 3% of TF cCREs resided in this corner, despite their overall enrichment in G3. These results suggest cCREs in G3 are more likely to represent regulatory elements with conserved primate functions, rather than elements specific to humans or their closest relatives.

Together, these analyses provide a multidimensional view of conservation across the expanded Registry of cCREs and allow us to assess how classes assigned via biochemical signatures and genomic context relate to evolutionary constraint.

1.7 Expanded Registry of cCREs maintains specificity for functional elements

Our updated Registry includes 1,419,467 newly annotated cCREs. This increase can be attributed to both updates in our computational pipeline—which accounts for 524,533 new cCREs—and an increase in biosample coverage—which accounts for 894,934 new cCREs (**Supplementary Figure 5a**). Of the 926,535 cCREs annotated in ENCODE3, 906,307 (98%) are retained in ENCODE4 (**Supplementary Table 5a**). This includes 396,055 cCREs with identical boundaries (and therefore matching accessions) and 510,252 cCREs with slightly shifted boundaries, with a median shift of 29 base pairs. Of the 20,228 ENCODE3 cCREs not retained

in ENCODE4, 9,173 (45%) were included in the anchor list but did not meet the updated threshold of DNase max Z-score > 1.64 required for cCRE inclusion.

To assess the validity of the newly added cCREs, we examined their chromatin signatures, biosample activity using STARR-seq and evolutionary conservation. Although STARR-seq activity was only analyzed in one cell type (K562), the assay's episomal design enables detection of elements with latent enhancer potential even in the absence of chromatin accessibility in K562. Thus, this assay provides a useful albeit indirect functional benchmark for the newly added cCREs.

In previous versions of the Registry, we performed an additional filtering step as detailed in The ENCODE Project Consortium *et al.* Supplementary Note 1⁹. However, such stringent filtering resulted in underrepresentation of elements from cell types and tissues that were missing histone mark data, such as embryonic tissue samples profiled during the Roadmap Epigenomics project (**Supplementary Table 5b**). These “filtered” cCREs were enriched for STARR-seq activity (**Supplementary Figure 5b**, 4.3 fold enrichment, two-sided Fisher's Exact test, $p < 2.2 \times 10^{-16}$) and were evolutionarily conserved²² (average of 0.14 vs -0.03 at center, two-sided Wilcoxon test, $p < 2.2 \times 10^{-16}$, **Supplementary Figure 5c**) compared to non-cCRE regions. These “filtered cCREs” also had a nearly identical evolutionary profile to replicated cCREs in the Zoonomia N1/N2 framework (Group 1: 42% vs. 47%, Group 2: 12% for both, Group 3: 12% vs. 10%, **Supplementary Figure 5d,e**), further supporting their functional relevance. Because of these findings, we removed this additional filtering step in our updated pipeline

The 894,934 cCREs added to the Registry due to an increase in biosample coverage were also more likely to have STARR-seq activity (**Supplementary Figure 5b**, 5.4 fold enrichment, two-sided Fisher's Exact test, $p < 2.2 \times 10^{-16}$) and had higher evolutionary conservation (average of 0.04 at center, two-sided Wilcoxon test, $p < 2.2 \times 10^{-16}$, **Supplementary Figure 5c**) than non-cCRE regions. The Zoonomia conservation profile of these new cCREs was lower compared to the replicated or filtered cCREs (Group 1: 30%, Group 2: 14%, Group 3: 20%), reflecting their enrichment for distal enhancer, CA, and TF cCRE classes, which tend to exhibit

greater evolutionary divergence and cell type-specificity (**Supplementary Figure 5f, Supplementary Note 1.6**). In general, the differences in STARR-seq activity and evolutionary conservation between these three groups of cCREs can be attributed to different distributions of cCRE classes; for example, 83% of promoter cCREs were identified during ENCODE3 (**Supplementary Table 5a**), whereas the newly added cCREs from ENCODE4 are enriched for distal enhancer, CA and TF cCREs, which tend to show greater variability in activity and conservation. Collectively, these results demonstrate that despite the substantial expansion of the Registry, it continues to retain specificity for functionally and evolutionarily meaningful regulatory elements.

To further evaluate the comprehensiveness and specificity of the ENCODE4 Registry of cCREs, we compared our collection with cis-regulatory annotations from four studies^{69–72}. Li *et al.* performed single nucleus ATAC-seq on over one million cells from 42 brain regions. In total, they identified 107 distinct cell types and subsequently 544,729 CREs active in the brain⁶⁹. In their original report, they found only about half of their elements (51%) overlapped ENCODE3 cCREs. With the expanded Registry, 85% (N=462,009) of their elements overlap ENCODE4 cCREs (**Supplementary Table 5c**) and this increase in coverage was primarily due to newly annotated cCREs active in brain tissues (**Supplementary Figure 6a, Supplementary Table 5d**). On average, 62 thousand of the Li *et al.* brain CREs were active in an individual cell type and ENCODE4 cCREs overlap 96% of brain CREs active in one of the 107 cell types (**Supplementary Table 5c**). These results suggest that our collection covers the majority of regulatory regions active in a diverse collection of brain cell types.

Loupe *et al.* performed ChIP-seq for more than 100 transcription factors across nine brain regions, surveying both bulk tissue and nuclei sorted for neuronal markers⁷⁰. They previously reported that about 70% of their collection of 237,324 thousand transcription factor peaks (termed union peaks) overlapped ENCODE cCREs (**Supplementary Table 5e**). We compared their list of union peaks with our updated collection of cCREs and 223,548 of their elements (94%) overlap ENCODE4 cCREs (**Supplementary Table 5e**). Again, this increase in coverage was primarily due to our expansion in brain-specific cCREs (**Supplementary Figure 6b, Supplementary Table 5f**). On average, 85,000 union peaks were bound by transcription

factors in an individual brain region or cell type. ENCODE4 cCREs overlap 98% of these experiment-specific elements suggesting once again that our collection covers the majority of regulatory regions active in a diverse collection of brain cell types and regions (**Supplementary Table 5e**).

Liu *et al.* defined 1,032,273 accessible candidate CREs (acCREs) across 17 fetal organ systems using SHARE-seq⁷¹. ENCODE4 cCREs overlapped 894,581 (87%) of these elements, compared to only 517,764 (50%) for ENCODE3 (**Supplementary Table 5g**). The expanded overlap reflects improved coverage of embryonic cell and tissue biosamples, particularly brain, heart, lung, and limb tissues (**Supplementary Figure 6c, Supplementary Table 5h**), which were previously excluded in ENCODE3 due to the lack of histone mark support in those samples.

Finally, we compared ENCODE cCREs to transcription-associated CREs (tCREs) defined by Zhang *et al.* using single-cell 5' RNA-seq across 23 human tissues⁷². Of the 178,229 tCREs, 165,903 (93%) overlapped ENCODE4 cCREs, compared to 136,794 (77%) overlapping ENCODE3 (**Supplementary Table 5i**). This improvement was primarily driven by the inclusion of myeloid lineage immune samples—such as inflammatory and suppressor macrophages, dendritic cells, and CD14⁺ monocytes—which were among the top contributors to the increase in overlap (**Supplementary Figure 6d, Supplementary Table 5j**).

1.8 Using variational autoencoders to extract cCRE sequence features

We trained three cell type-specific Variational Autoencoders (VAEs) on the sequences of promoter, enhancer, and CA-H3K4me3 cCREs active in K562, HepG2, and HCT116 cells (**Extended Data Figure 2a**). The latent representations learned by the VAEs were then reduced to two dimensions using UMAP for visualization. In all three cell types, promoter and distal enhancer cCREs separated along the first dimension (**Extended Data Figure 2b, Supplementary Figure 7**). Proximal enhancers were evenly distributed across dimension one in K562 and centrally distributed for HepG2 and HCT116. CA-H3K4me3 cCREs clustered near promoters in K562 and HCT116 but were positioned more centrally in HepG2. Differences in

clustering for proximal enhancer and CA-H3K4me3 cCREs may reflect biological variations between cell types or could result from differences in data quality.

GC content, defined as the percentage of guanine (G) and cytosine (C) nucleotides in a cCRE, was strongly correlated with the first dimension across all three cell types with Pearson correlation coefficients of -0.93 for K562, -0.91 for HepG2, and -0.88 for HCT116, respectively ($p < 2.2 \times 10^{-16}$). Other sequence features had lower correlation coefficients such as CpG density (-0.78, -0.75, and -0.74, respectively) and normalized CG⁷³ (-0.61, -0.63 and -0.59, respectively). This indicates that GC content is a key sequence feature captured by our VAE models (**Extended Data Figure 2c**).

We then investigated how the regulatory potential of cCREs differs across subclasses of promoter and enhancer cCREs (**Extended Data Figure 2d**). We stratified promoter cCREs based on whether they were promoters for protein coding or lncRNA transcripts, as defined by GENCODE. Promoter cCREs of protein coding genes (N=17,080) had lower values for dimension one compared to promoter cCREs of lncRNA genes (N=4,175, median of 0.8 vs. 2.0, pairwise two-sided Wilcoxon test with FDR correction, $p < 2.2 \times 10^{-16}$). We also stratified distal enhancer cCREs into three distinct groups based on their overlap with transcription sites:

- 1) cCREs that overlapped the 5' ends of long RNA-seq reads, presumably the transcription start sites of novel, stable transcripts (N=1,895)
- 2) cCREs that overlap PRO-cap peaks but not the 5' end of long RNA-seq reads (N=3,916)
- 3) all other distal enhancers (N=29,677).

Enhancer cCREs that overlapped novel TSSs had the lowest dimension one values (5.5, $p < 1 \times 10^{-6}$) followed by enhancer cCREs with eRNAs (6.4, $p < 2.2 \times 10^{-16}$) and all other distal enhancer cCREs (8.3).

To further investigate whether GC content primarily reflects cellular regulatory function or sequence-intrinsic activity, we examined the correlation between GC content and the number of biosamples in which each cCRE was accessible (**Supplementary Figure 8a,b**). GC content was modestly correlated with ubiquity for promoters ($r = 0.28$) and more strongly correlated for

enhancers ($r = 0.43$), suggesting that GC-rich enhancers are more likely to be broadly active across cell types. However, the moderate strength of these associations indicates that GC content is not solely a proxy for accessibility and that other sequence features likely contribute to cell type-specific activity.

Finally to further dissect the relationship between GC content and regulatory activity, we stratified 95 sequence-specific TFs based on two axes: the GC content of their canonical motifs and their binding bias toward promoter versus distal enhancer cCREs (**Supplementary Figure 8c, Supplementary Table 6**). We selected all transcription factors whose peaks overlapped either K562 promoters or enhancers by 25% (102 transcription factors across 149 experiments). This analysis revealed a moderate correlation between motif GC content and promoter preference (Pearson Correlation Coefficient = 0.42, $p = 9.1 \times 10^{-8}$), suggesting that sequence composition contributes to the genomic localization of transcription factor binding events. Breaking down by quadrants revealed distinct classes of transcription factors with consistent sequence and binding preferences. Quadrant I (high GC content, promoter preference) included 69 experiments covering canonical promoter-binding transcription factors such as SP1, MYC, NRF1, and E2F family members, consistent with their known affinity for GC-rich core promoter sequences^{74–77}. In contrast, Quadrant III (low GC content, enhancer preference) included 31 experiments for transcription factors such as GATA1/2, JUN/FOS, TAL1, and CEBPB, which are associated with lineage-specific distal enhancer activity and preferentially bind AT-rich motifs. Transcription factors in Quadrant IV (low GC content, promoter preference) include TATA-box-associated factors (e.g., TBP), core promoter-binding complexes (e.g., NF-Y) and stimulus-responsive regulators (e.g., ELK1, FOS). Their enrichment at AT-rich promoters suggests that certain regulatory programs rely on alternative promoter architectures, potentially for increased responsiveness or lineage-specific regulation. Quadrant II (high GC content, enhancer preference) included only a small number of TFs whose points were clustered near the origin, suggesting limited divergence in binding preference and potentially weaker or more context-dependent GC-driven enhancer associations. In general, these results suggest a general concordance between sequence preferences and cCRE class and support the model that GC content is a biologically relevant feature that shapes transcription factor binding specificity and regulatory element identity.

1.9 Caveats and considerations when using the Registry of cCREs

While the ENCODE Registry of cCREs provides a comprehensive resource for studying gene regulation across diverse human tissues and cell types, several important caveats should be considered when interpreting results derived from the Registry.

First, the Registry is inherently constrained by the availability and distribution of high-quality chromatin data. For example, biosamples from blood and brain tissues are extensively profiled due to coordinated efforts during ENCODE4 to study the regulatory landscape under immune cell stimulation and activation and the aging brain in the context of neurodegenerative disease. As a result, these tissues are well represented in the Registry, both in the number of biosamples and in the completeness of the four core assays (**Supplementary Table 1a**).

We observe a general correlation between the number of biosamples available for a given tissue or organ system and the number of rDHSs attributed to that group (**Supplementary Figure 9, Supplementary Table 6b**). However, this relationship is not strictly linear. For instance, despite having more blood-related biosamples (N=470) than brain biosamples (N=208), brain biosamples contribute a larger number of rDHSs (449,366 for brain vs. 381,194 for blood). Quality of DNase-seq data also impacts rDHS detection and attribution. For example, several high-depth DNase-seq datasets generated during ENCODE2 for genomic footprinting exhibit exceptional sensitivity in identifying open chromatin sites (**Supplementary Table 6b**). These datasets, often derived from *in vitro* cultured cell lines, detect a larger number of sites compared to primary tissue datasets with lower sequencing depth or quality. This suggests that both biosample number and signal quality influence the final annotations.

Assay completeness also impacts cCRE classification. While 170 biosamples contain all four core assays (DNase, H3K4me3, H3K27ac, and CTCF)—a substantial improvement over earlier versions of the Registry—the remaining 1,509 biosamples lack at least one assay. In particular, 951 biosamples have only DNase data, without accompanying histone marks or CTCF. Although DNase allows identification of accessible chromatin and supports general element detection, the absence of additional marks prevents confident classification, especially for

enhancers. As a result, many cCREs in these biosamples are labeled as CA (chromatin accessibility) due to partial data and may actually represent unannotated enhancers.

To estimate the potential impact, we performed both global and biosample-level analyses. We first estimated an upper bound by counting all CA cCREs that were exclusively derived from samples lacking H3K27ac. This suggested that up to 165,781 of human CA cCREs (67%) and 188,912 mouse CA cCREs (65%) may correspond to enhancers. Then for each DNase-only biosample, we calculated the number of CA cCREs that were classified as promoters, enhancers or CA-CTCF in other biosamples—i.e., "rescued" by additional chromatin marks elsewhere. On average, fewer than 1% of CA cCREs from DNase-only biosamples remained classified as CA in the cell type-agnostic classification, indicating widespread rescue by data from other samples (**Supplementary Table 6c**). However, we identified several exceptions.

Embryonic tissues had a significantly higher proportion of unrescued CA cCREs compared to adult tissues (median 1% vs. 0.1%, two-sided Wilcoxon test, $p < 2.2 \times 10^{-16}$), consistent with limited histone mark profiling in embryonic biosamples (e.g., only 24 embryonic samples have H3K27ac compared to 300 with DNase). This pattern was particularly noticeable in embryonic brain samples (median 2%, max of 6%) for which we have no H3K27ac data. Notably, several adult brain samples—including three from the head of the caudate nucleus—had higher percentages of unrescued CA cCREs (> 2%). This may reflect the distinct cell type composition of the caudate nucleus, which is enriched for medium spiny neurons, a population largely absent from other brain regions⁷⁸. Such variability highlights how samples derived from complex tissues can differ substantially in cell type representation, impacting annotation completeness. In contrast, blood-derived biosamples consistently showed near-complete rescue of CA cCREs, likely due to high coverage of these cell types across both primary tissues and established cell lines.

Additionally, like previous versions of the Registry of cCREs, our expanded ENCODE4 Registry of cCREs has far more human data than mouse data. In total, the current Registry integrates 3,202 experiments from 1,679 human biosamples compared to only 591 experiments across 367 mouse biosamples. There was no significant difference in the number of DHSs called from an individual human or mouse DNase-seq experiment with medians of 98,569 and 105,538, respectively (two-sided Wilcoxon test, $p = 0.54$), suggesting that similar numbers of cCREs

would be annotated across the human and mouse genomes if experiment and biosample coverage were the same.

To mitigate biases introduced by unequal biosample coverage or quality, we favor per-biosample analyses in downstream applications rather than relying exclusively on aggregated cCRE annotations. For example, in our REST+ cCRE analysis (**Supplementary Note 3**), we evaluated H3K27ac enrichment at REST-bound cCREs across individual biosamples. We then assessed whether tissue-specific enrichment patterns (e.g., in brain) exceed what would be expected by chance, given the number and distribution of biosamples. This approach helps control for imbalances in data representation and ensures that observed enrichments are biologically meaningful rather than artifacts of dataset composition.

In summary, while the Registry captures a broad landscape of regulatory activity, it may not fully represent tissues or contexts with limited or lower-quality data. We encourage users to consider these limitations and, when appropriate, to complement aggregate cCRE analyses with biosample-level assessments.

Supplementary Note 2 - Functional characterization analysis

2.1 Integrating functional characterization data with the Registry of cCREs

Whole genome STARR-seq

STARR-seq (Self-Transcribing Active Regulatory Region sequencing) is a high-throughput reporter assay that quantifies enhancer activity of DNA sequences by placing them in the 3' untranslated region (UTR) of a reporter gene downstream of a minimal promoter³. Active enhancers are part of the resulting transcript and the relative abundance of their transcripts is a measure of enhancer activity. For whole-genome STARR-seq, the input library of tested regions derives from fragmented genomic DNA. Candidate enhancers are traditionally identified through peak calling, looking for regions with enrichment for RNA fragments over input DNA fragments.

We integrated the Registry of cCREs with peak calls from seven whole genome STARR-seq experiments generated during ENCODE4 (**Supplementary Table 7a**). Between 69-91% of

STARR peaks overlap cCREs with variation in overlap attributed to peak calling method and STARR-seq protocol.

Because the median peak width of STARR peaks is greater than the width of cCREs (median widths of 467-533), STARR peaks can overlap multiple cCREs. Therefore to quantify the STARR activity of individual cCREs and annotate both positive and negative activity patterns, we developed a novel computational method (CAPRA, **Supplementary Note 2.2**) to annotate cCRE-specific STARR activity.

For STARR-seq data, we considered a cCREs *active* if it meet one of the following criteria:

- At least 50% of the cCRE overlaps a STARR-seq peak (N = 484,666)
- The cCREs has a significant STARR-score ($p < 0.05$) from CAPRA for either the solo or double quantifications (N = 186,339). Note that this included cCREs with STARR scores > 0 for all libraries and STARR scores < 0 (i.e. silencers) from one library (K562 from Reddy lab).

In total, we annotated 567,059 cCREs as having STARR-seq activity.

MPRA

Massively Parallel Reporter Assays (MPRAs) are high-throughput assays that measure the regulatory activity of thousands of DNA sequences simultaneously by linking each sequence to a reporter gene and quantifying its expression in cells⁷⁹. We integrated the Registry of cCREs with results from 62 ENCODE4 MPRA experiments spanning six cell types (**Supplementary Table 7b**). For each experiment, we compared the activity of tested MPRA regions that (1) fell entirely within a cCRE, (2) partially overlapped a cCRE or (3) did not overlap any cCRE. For MPRAs, we considered a tested region active if it had a $\text{Log}_2\text{FC} > 1$. Tested regions that fell entirely within a cCRE had the highest positivity rates (an average of 23% of tested regions had a $\text{Log}_2\text{FC} > 1$) compared to those that partially overlapped cCREs (14%) and those that did not overlap cCREs (7%, **Supplementary Table 7b**). These results demonstrate that cCREs are enriched for functional activity and that their boundaries effectively delineate regions of regulatory potential.

Additionally, for MPRA regions that fell entirely within cCREs, those that overlapped cCREs and had high chromatin accessibility were, on average, almost three times more likely to be active than those that overlapped cCREs with low levels of chromatin accessibility (**Supplementary Figure 10e-h, Supplementary Table 7b**). These results highlight the high specificity of the Registry of cCREs for capturing cell type-specific regulatory activity.

CRISPRi-FlowFISH

CRISPRi-FlowFISH is a high-throughput experimental method that combines CRISPR interference, RNA fluorescence in situ hybridization (FISH), and flow cytometry to systematically perturb distal regulatory elements in the genome, enabling quantitative assessment of their effects on target gene expression. The Sabeti lab used CRISPRi Flow-FISH screens with HCR-FlowFISH to identify CRE-gene pairs around 20 genes in K562 cells (**Supplementary Table 7c,d**)⁵. Guide RNAs were selected using two approaches (1) targeting features associated with CREs such as chromatin accessibility and histone mark peaks and (2) dense tiling to detect elements with functional effects in an unbiased manner. To evaluate the enrichment of K562 cCREs, all cCREs and non-cCRE regions in functionally active elements, we only compared within the set of cCREs that was tested (**Supplementary Figure 10i-l**).

The Engrietz lab used CRISPRi Flow-FISH screens with PrimeFlow readout⁶⁰ to identify CRE-gene pairs around 33 genes in K562 cells and, for some genes, up to four additional cell lines (**Supplementary Table 7c,d**). These experiments targeted DNase peaks and therefore nearly all tested elements were cCREs. One caveat to consider when interpreting results from either set of experiments is that a negative result in CRISPRi FlowFISH does not necessarily indicate that a cis-regulatory element (CRE) is inactive; it may regulate a gene that was not included in the measurement.

Additional CRISPR perturbations

Over 66 additional CRISPR-based experiments were performed by functional characterization centers during ENCODE4, 53 in human and 13 in mouse (**Supplementary Table 7c**). While there was a lot of variation in experimental design such as utilizing different CRISPR constructs—cutting, inhibition, activation—and guide RNA design approaches—full tiling vs

targeting peaks—we developed a systematic approach for characterizing cCRE activity in each of these assays. Briefly, we intersected cCREs with guide RNAs and quantified per cCRE levels in each experiment and then determined statistical significance using DESeq2.

Across CRISPR perturbations, we considered a cCREs *active* if it met one of the following criteria:

- For Sabeti lab data, tested cCRE overlapped a CASA defined element⁵
- For Engrietz lab data, tested cCRE overlapped an element-gene pair with FDR < 0.05 as defined in their lab's processing pipeline
- For other data, DESeq2 $p < 0.05$

In total, we annotated 25,362 cCREs as having activity via CRISPR perturbation.

Cell type-specific

Across all assays, cCREs were much more likely to be functionally active than sized matched non-cCRE genomic regions. Additionally cCREs with active chromatin signatures in a cell type more frequently tested positive in the same cell type than other surveyed cCREs (**Supplementary Figure 10**). This cell type-specificity was most dramatic for CRISPR perturbation experiments that disrupt elements in their native chromatin context (100-fold enrichment, two-sided Fisher's Exact Test, $p < 2.2 \times 10^{-16}$, **Supplementary Figure 10f**) compared to the episomal assays STARR-seq (35-fold enrichment, two-sided Fisher's Exact Test, $p < 2.2 \times 10^{-16}$, **Supplementary Figure 10d**) and MPRA (3-fold enrichment, two-sided Fisher's Exact Test, $p < 2.2 \times 10^{-16}$, **Supplementary Figure 10e**).

2.2 The CAPRA Pipeline

Identifying functional regulatory elements from STARR-seq data typically relies on peak-calling algorithms that assess RNA enrichment over input DNA. However, this approach often produces broad peaks that encompass multiple cCREs, limiting resolution and complicating interpretation in dense regulatory regions.

To overcome this limitation, we developed CAPRA (CRE-centric Analysis and Prediction of Reporter Assays), a method that uses the Registry of cCREs as predefined anchors to directly quantify STARR-seq activity without requiring peak calling (**Supplementary Figure 11**). CAPRA enables high-resolution, element-centric quantification of both enhancer and silencer activity by computing RNA-to-DNA fragment ratios for individual cCREs. It also supports quantification across cCRE pairs to assess potential combinatorial effects and a sweep mode for sub-cCRE functional profiling.

cCRE centric mode for characterizing element-level activity

The CAPRA pipeline includes three main steps:

1. **Data Acquisition and Preprocessing:** For each STARR-seq experiment, we downloaded aligned BAM files for input DNA and output RNA libraries from the ENCODE portal. These files were converted to BEDPE format, from which we extracted fragment coordinates and generated 4-column BED files for analysis.
2. **Fragment Quantification over cCREs:** We counted the number of full-length fragments overlapping either single cCREs or predefined adjacent cCRE pairs:
 - *Solo fragment analysis:* Fragments that overlapped exactly one cCRE (100%) were counted using BEDTools.
 - *Double fragment analysis:* Fragments that overlapped exactly two adjacent cCREs were identified and counted with respect to annotated cCRE pairs.These fragment counts were assembled into a matrix with cCREs or cCRE pairs as rows and fragment counts from DNA and RNA libraries as columns. Only elements with ≥ 1 DNA fragment were retained.
3. **Differential Enrichment Analysis with DESeq2:** We used DESeq2 to compare RNA and DNA fragment counts, modeling assay type (RNA vs. DNA) as the design variable. For each element, DESeq2 calculated a $\log_2$ fold change and p -value.

The p -value quantifies whether the RNA output for a given cCRE significantly deviates from what would be expected based on its DNA abundance. A significant p -value with a positive $\log_2$ fold change indicates enhancer-like activity (excess RNA), while a significant negative $\log_2$ fold change indicates silencer-like activity (depleted RNA). Elements with non-significant values

are likely inactive in the assay. These quantitative activity scores form the basis of our functional annotation of cCREs across biosamples, enabling systematic identification of candidate enhancers and silencers in their native sequence context.

To complement the peak-based overlap analysis and more precisely estimate the fraction of transcriptionally active cCREs, we used CAPRA to calculate STARR scores for each individual cCRE tested in the K562 STARR-seq library. We identified 11,181 cCREs with significant positive activity (STARR score > 0 and $p < 0.01$, STARR+). Of these, 4,249 STARR+ cCREs had high chromatin accessibility in K562, representing 3.2% of the 133,406 K562-active cCREs tested. The remaining 6,932 STARR+ cCREs had low chromatin accessibility in K562 and came from other biosamples, corresponding to only 0.36% of the 1,906,828 non-K562 cCREs tested.

These results underscore two key points: (1) functional activity in the STARR-seq assay is strongly enriched among cCREs that are active in the matched cell type, and (2) CAPRA provides a highly cell-type-specific and conservative measure of enhancer activity by requiring RNA enrichment relative to input DNA for each cCRE independently. This stands in contrast to peak-based overlap methods, which may overestimate activity when multiple cCREs are encompassed within a single STARR-seq peak.

Paired Sweep Mode for Sub-cCRE Functional Profiling

To extend CAPRA beyond cCRE-centric analysis and enable high-resolution mapping of regulatory activity between neighboring cCREs, we developed a complementary *Paired Sweep* analysis mode. This approach systematically evaluates partially overlapping STARR-seq fragments across the genomic interval between a pair of high-coverage cCREs to resolve fine-scale relationships between local sequence features and reporter activity.

For each cCRE pair, we performed two directional sweeps:

- Forward sweep: Starting with fragments that fully overlap the first cCRE, we progressively tiled 10 bp windows toward the second cCRE, quantifying RNA and DNA fragment coverage at each step.

- Reverse sweep: The same procedure was repeated in the opposite direction, starting from fragments that fully overlap the second cCRE and sweeping back toward the first.

To generate the activity profiles, we constructed a 10 bp resolution BED file spanning the interval between the two cCREs. For each window, we calculated fragment coverage across the input DNA and three replicate RNA libraries from the K562 STARR-seq dataset (ENCSR661FOW). Counts were normalized to library size (fragments per million), and summary profiles were generated for each sweep direction.

This analysis yields two directional profiles of RNA:DNA ratios across the cCRE pair, enabling visualization of activity gradients and identification of inflection points aligned with underlying sequence features. This mode is particularly informative in dense regulatory regions or cCRE clusters, where it can reveal local sequence-function relationships obscured in conventional element-centric approaches.

We demonstrate the utility of Paired Sweep mode using the cCRE pair EH38E2809119 and EH38E2809120 (**Supplementary Figure 12f**). In the forward sweep (from EH38E2809119 to EH38E2809120), the RNA:DNA ratio increased sharply near a non-canonical E-box motif (CAGCTG; green highlight), associated with binding by TFE and TFAP family members. A secondary increase in activity was observed near a cluster of STAT and zinc finger motifs (blue highlight). In the reverse sweep (from EH38E2809120 to EH38E2809119), the first increase in RNA:DNA ratio occurred near a GC-rich region with motifs for SP and KLF family transcription factors, followed by a pronounced increase downstream of a canonical E-box motif (CACGTG; orange highlight) corresponding to bHLH factors such as USF1, MYC, and MAX. These directional shifts in activity underscore the potential of this method to resolve regulatory logic at sub-cCRE resolution.

While this analysis is currently based on raw fragment counts rather than DESeq2-normalized values, it provides a powerful exploratory tool for dissecting motif-function relationships. Adequate coverage is essential to avoid artifacts due to low signal. Nonetheless, *Paired Sweep*

mode highlights the potential of fragment-centric strategies for high-resolution functional annotation and motivates future STARR-seq library designs tailored for such analyses.

2.3 Comparing whole genome STARR-seq datasets

The White and Reddy labs both generated whole-genome STARR-seq data as part of the ENCODE4 project. Although the labs used similar protocols, we observed notable differences in the CAPRA STARR scores of cCREs from the lab-specific datasets, which we attribute to subtle differences between methodologies. The correlation of STARR scores between K562 datasets generated by each lab were modest (**Supplementary Table 8a**, Pearson correlation coefficient=0.25) when considering all cCREs with at least ten reads in both input DNA libraries ($N = 507,547$, $p < 2.2 \times 10^{-16}$). We further compared the STARR scores across different classes of cCREs, particularly those with biochemical signatures in K562 cells (i.e., K562 cCREs). STARR scores were significantly correlated for K562 distal enhancers ($N = 4,085$, $R = 0.50$, $p < 2.2 \times 10^{-16}$), but not for K562 promoter cCREs ($N = 382$, $R = -0.01$, $p = 0.9$).

We hypothesize that these differences may be due to GC content, given our previous findings that GC content is the primary sequence feature distinguishing promoter and enhancer cCREs. To test this, we compared the correlation between STARR scores for K562 distal enhancer cCREs with either high or low GC content (**Supplementary Table 8a**). We observed that cCREs with high GC content had a lower correlation between lab datasets compared to those with lower GC content (correlation coefficients of 0.23 vs. 0.59, respectively). Similarly, high GC content promoters had a lower correlation between lab datasets compared to those with lower GC content (correlation coefficients of 0.14 vs. 0.63, respectively).

We also systematically evaluated inter-lab and inter-sample variability across six datasets from the White lab (performed on cell lines K562, HepG2, MCF-7, HCT116, A549 and SH-SY5Y) and one from the Reddy lab (K562). We performed principal component analysis (PCA) on 362,591 cCREs that were well covered (≥ 10 reads) across all DNA input libraries. Notably, the White lab K562 and HepG2 libraries were generated using the same input DNA pool, as were the MCF-7, HCT116, and A549 libraries, enabling us to assess variation both within and across experimental batches.

The first principal component (PC1) explained 32% of the variance and primarily separated MCF-7, A549, and HCT116 from SH-SY5Y and K562 (both White and Reddy) and HepG2. Analysis of the top 500 cCREs contributing to this component revealed three major groups (**Supplementary Figure 12d, Supplementary Table 8b,d**):

1. Group 1 (n = 126): cCREs with high activity in all samples except K562, enriched for p53 motifs—consistent with the absence of functional p53 in K562, as described in the main text
2. Group 2 (n = 361): cCREs enriched for JunD/B and FOS motifs, with elevated activity in HCT116, MCF-7, and A549. These TFs are components of the AP-1 complex, and their activity may reflect batch-specific differences in transfection efficiency or activation of cell stress response pathways in these cancer cell lines
3. Group 3 (n = 13): cCREs with low activity in HepG2 and White K562, enriched for REST motifs.

The second principal component (PC2) explained 28% of the variance and primarily captured inter-lab technical differences, particularly GC-content-related biases (**Supplementary Figure 12e, Supplementary Table 8c,d**). The top 500 cCREs contributing to PC2 grouped into two clusters:

1. Group 1 (n = 359): cCREs with high GC content and low activity in all White lab datasets but high activity in Reddy K562;
2. Group 2 (n = 141): cCREs with high activity in Reddy K562 and moderate activity in White MCF-7 and A549, enriched for NFY-family and GATA motifs.

These results highlight both technical and biological sources of variation in STARR-seq data. In particular, lab-specific effects such as GC-content bias and batch-specific responses likely linked to transfection conditions or cell stress can influence measured activity. Accordingly, we recommend within-batch comparisons (e.g., White lab K562 vs. HepG2, as shown in **Figure 2**) to identify biologically meaningful differences.

Due to the discrepancies between assays, we prioritized the data from each lab for specific types of analyses. Since the White lab performed STARR-seq across multiple cell types, we used their data to analyze cell type specificity of cCREs and identify sequence features predictive of activity (**Figure 2a,b, Extended Data Figure 3**). However, we recommend that for

primary comparisons, such as those in **Figure 2a,b**, analysis occurs between experiments with the same DNA library (e.g. K562 and HepG2). Conversely, because the Reddy lab data appeared less influenced by GC content, we used it to compare activity between elements, assess combinatorial effects (**Figure 2c-e**), and identify potential silencers (**Figure 3f**, **Extended Data Figure 5,6**).

Despite the differences in the assays, we consistently observed that the GFI1B motif was significantly associated with silencer activity in both datasets. In particular, cCREs with the GFI1B motif showed lower-than-expected activity in both the White lab (-0.23 vs. -0.11 , two-sided Wilcoxon test, $p < 2.2 \times 10^{-16}$) and Reddy lab (-0.16 vs. -0.04 , two-sided Wilcoxon test, $p < 2.2 \times 10^{-16}$) data. This convergence across datasets suggests that, despite methodological differences, common core regulatory features such as the GFI1B binding may underlie silencer activity.

These results underscore the robustness of certain regulatory elements across varying experimental conditions and highlight the importance of considering both shared and distinct characteristics in interpreting STARR-seq data. Future work should focus on resolving methodological differences and developing unified models to better capture the complex regulatory landscape reflected in these assays.

2.4 STARR-seq activity of closed chromatin cCREs

We investigated the regulatory logic underlying cCREs that exhibit strong STARR-seq activity in K562 despite lacking chromatin accessibility in the genome. Motif enrichment analysis revealed a distinct set of transcription factors that are enriched in these “closed-but-active” regions compared to those that are (1) both open and active and (2) had neutral STARR-seq activity (**Supplementary Table 9c**).

Notably, we observed strong enrichment ($\text{FDR} < 0.05$, $\log_2$ fold enrichment > 0.5 for both comparisons) for CUX1, ATF4, CEBPG, CEBPB, SRF, and FOXO3 motifs in the closed chromatin group:

- CUX1 functions as a pioneer transcription factor in hematopoietic cells by directly recruiting the BAF chromatin remodeling complex to lineage-specific enhancers, promoting nucleosome remodeling and increased chromatin accessibility⁸⁰.
- ATF4 is a central stress-responsive transcription factor that promotes cellular adaptation to conditions like amino acid deprivation, hypoxia, and ER stress by inducing survival pathways, and under persistent or excessive stress, it can trigger apoptosis through context-dependent gene regulation^{81,82}.
- CEBPG plays a critical role in the integrated stress response by heterodimerizing with ATF4 to regulate genes involved in oxidative stress and amino acid deprivation, positioning it as a key mediator of cellular stress adaptation and a potential contributor to oncogenesis^{82,83}.
- CEBPB is involved in immune response, cell differentiation, metabolism, and inflammation⁸⁴. It acts as a pioneer transcription factor by binding to DNA prior to stimulation, occupying a large number of regulatory regions in a relatively static fashion, and establishing an accessible chromatin landscape that facilitates the binding of other transcription factors during gene induction.
- SRF regulates immediate early gene expression and cofilin activity in response to acute stress, thereby playing a critical role in both short-term stress reactivity and long-term stress adaptation and resilience⁸⁵.
- FOXO3 plays a central role in cellular redox homeostasis by regulating antioxidant gene expression and modulating cell survival and apoptosis based on oxidative stress intensity⁸⁶.

Collectively, these findings suggest that specific transcription factors enriched in closed-but-active cCREs may have the capacity to initiate regulatory activity in chromatin contexts that are otherwise inaccessible, either through pioneer-like functions or in response to cellular stress. These results underscore the value of using MPRA-based assays to uncover latent enhancer potential that may be masked in native chromatin and provide mechanistic hypotheses for future experimental validation.

2.5 Combinatorial effects of cCRE activity

Using the CAPRA pipeline, we assessed the combinatorial effects of cCREs by comparing the activity of individual cCREs, derived from solo fragments (which span a single cCRE in its entirety), to that of double fragments (which span two cCREs in their entirety). Due to the strict overlap requirements, only a maximum of 26% cCREs were profiled with double fragments per experiment. We observed that the distribution of STARR scores for double fragments had a lower mean and smaller standard deviation compared to that of solo fragments (**Supplementary Figure 14a**, Wilcoxon test, $p < 2.2 \times 10^{-16}$), indicating potential fragment length biases in the STARR-seq data. Despite this, we chose not to normalize CAPRA STARR scores for fragment length, as our primary interest was in identifying instances of increased combinatorial activity between pairs. We opted for a more conservative approach to avoid overestimating such combinatorial effects.

When comparing STARR scores for cCREs using solo fragments versus double fragments, we found that activity levels for the majority of tested cCREs were correlated. This correlation strengthened as we applied more stringent thresholds for the minimum number of input DNA fragments, increasing from 0.16 for a minimum of one fragment to 0.42 for a minimum of ten fragments (**Supplementary Figure 14b-d**). However, there were notable exceptions where cCREs showed either higher or lower than expected activity. An example of higher-than-expected activity is highlighted in **Figure 2f-h** and discussed in the main text. Here, we focus on cases with lower-than-expected combined activities.

We identified 1,275 cCRE pairs with lower-than-expected combined activity, defined as instances where one cCRE exhibited high activity as a solo fragment (a positive STARR score with $p < 0.05$), but displayed lower activity when combined with another cCRE (double fragments), the STARR score dropped below zero. These cCRE pairs had lower chromatin accessibility and H3K27ac signals in K562, compared to other pairs of cCREs (two-sided Fisher's Exact Tests, $p < 1.7 \times 10^{-5}$, **Supplementary Table 9d**), suggesting that these regions are less likely to be active in their native chromatin context. We also analyzed sequence motifs overlapping each of these cCREs. The cCREs that exerted a dominant repressive effect were

enriched for motifs associated with transcription factors, some of which are known to have repressive activity, including GFI1B¹³, MEF2D⁸⁷, and ZNF121⁸⁸ (**Supplementary Table 9e**).

One notable example includes two distal enhancer cCREs, EH38E3837011 and EH38E3837012 (**Supplementary Figure 14e,f**). These elements are located approximately 8 kb downstream and within the body of the *TTPA* gene, which encodes the alpha tocopherol transfer protein that regulates vitamin E levels. *TTPA* is primarily expressed in the liver, and both EH38E3837011 and EH38E3837012 have active enhancer chromatin signatures exclusively in liver-derived biosamples. EH38E3837011 has a STARR score of 1.05 ($p = 0.005$), while EH38E3837012 has a STARR score of -0.31. When tested in combination, the combined region also has low activity, with a STARR score of -0.28, suggesting that EH38E3837012 exerts a dominant repressive effect on EH38E3837011. We hypothesize that this repression may be due to EH38E3837012 overlapping motifs for transcription factors with known repressive functions, including SMAD3⁸⁹, MAFK²⁴, and TCF7L2⁹⁰.

These findings highlight the complex interplay between cCREs, particularly in cases where combinatorial effects result in repressive activity that cannot be predicted from individual cCREs alone. Future work will prioritize a larger-scale investigation of such effects by designing STARR-seq experiments with varying fragment lengths to identify additional synergistic or antagonistic cCRE pairs. Additionally, we plan to explicitly test key hypotheses, such as the role of repressive motifs in modulating enhancer activity, through targeted experiments. This approach will help refine our understanding of the regulatory landscape and the mechanisms by which cCREs interact to influence gene expression.

2.6 TP53 activity in cancer cell lines

TP53 encodes the well characterized-tumor suppressor p53. K562 cells only have one copy of TP53 and this copy has a frame shift mutation leading to an N-terminal truncated protein that does not produce a protein product⁹¹. Although HepG2, MCF-7 and HCT116 are all cancer cell lines, they all have normal copies of p53⁹²⁻⁹⁴.

Supplementary Note 3 - Characterizing REST+ cCREs

3.1 Defining classes of REST+ cCREs

We defined a set of 5,850 REST+ cCREs by selecting non-promoter cCREs that both contained a REST motif and overlapped the summits of at least five REST ChIP-seq peaks. We excluded REST+ promoter cCREs from our analyses because these elements did not exhibit characteristics consistent with neuron-restrictive silencer elements (NRSEs). REST+ promoter cCREs demonstrated significantly higher than expected STARR scores, with a median value of 0.6 compared to 0.11 for REST- promoter cCREs and -0.02 for all cCREs (two-sided, Wilcoxon test, $p < 2.2 \times 10^{-16}$). Additionally, genes near REST+ promoter cCREs had higher expression levels (median TPM 16.8) than other annotated promoter cCREs in K562 cells (median TPM 11.7). These findings are inconsistent with a repressive role, and therefore, REST+ promoter cCREs were excluded from further analyses.

The REST+ cCREs spanned multiple cCRE classes, but were highly enriched for CA-TF 9.1-fold enrichment, chi-square test, $p < 2.2 \times 10^{-16}$, **Supplementary Figure 15 a,b, Supplementary Table 10d**), highlighting the importance of including this new category in the Registry. We characterized each class of REST+ cCREs (e.g. distal enhancer, CA-TF) using transgenic mouse enhancer assays to assess enhancer activity (**Supplementary Table 10e**, whole-genome STARR-seq to survey silencer activity, and additional histone ChIP-seq and gene expression datasets. We compared each REST+ cCRE class with other cCREs of the same class that lacked REST binding (REST- cCREs). Based on these analyses, we assigned each REST+ cCRE class to one of two categories:

- 1) REST+ enhancer/silencer cCREs, which function as neuronal enhancers but act as silencers in REST-expressing cells, or
- 2) REST+ silencer cCREs, which lack enhancer activity and act solely as silencers.

REST+ distal enhancer cCREs: Among the 5,850 REST+ cCREs, 2,516 (43%) were distal enhancer cCREs. These elements exhibited robust enhancer activity in vivo: 57% of those

tested were validated as enhancers in the transgenic mouse enhancer assay, a rate comparable to REST⁻ distal enhancer cCREs (60%, **Supplementary Figure 15c**).

Consistent with this, REST⁺ distal enhancer cCREs were significantly enriched for H3K27ac in brain tissues and neuron-derived biosamples (**Supplementary Table 10h**), supporting their role as neuronal enhancers. REST⁺ distal enhancer cCREs were also enriched for H3K27ac signals in endocrine tissues, such as the pancreas (**Supplementary Table 10h**), where REST is known to play a critical role in regulating pancreatic cell differentiation. For example, Rovira *et al.* demonstrated that REST inhibits pancreatic endocrine differentiation during embryonic development by repressing the formation of endocrine cells from progenitor cells⁹⁵. REST⁺ cCREs were also enriched for H3K27ac signal in HCT116 cells, a colon cancer cell line (**Supplementary Table 10h**). This included samples that were part of the ENCODE4 degran series, demonstrating the reproducibility of this result across multiple HCT116 experiments. Although the enrichment of H3K27ac signals in HCT116 cells is intriguing, the functional role of REST⁺ enhancer cCREs in this colon cancer cell line is not yet fully understood, and future studies could focus on elucidating the specific mechanisms through which REST⁺ enhancers contribute to gene regulation in HCT116.

Despite their enhancer activity in neurons, REST⁺ distal enhancer cCREs displayed silencer activity in REST-expressing K562 cells, with a median STARR-seq score of -0.10 (compared to -0.02 for all cCREs; two-sided Wilcoxon test, $p < 2.2 \times 10^{-16}$; **Figure 3e, Supplementary Figure 15d**). Furthermore, genes located near REST⁺ distal enhancer cCREs exhibited reduced expression in K562 (median TPM = 0.7) compared to genes near cCREs with no chromatin accessibility (median TPM = 2.3; two-sided Wilcoxon test, $p = 3.0 \times 10^{-6}$; **Supplementary Figure 15e**).

Together, these results support a dual-function model for REST⁺ distal enhancer cCREs: they act as enhancers in neuronal and select non-neuronal contexts (e.g., pancreas), but as silencers in REST-expressing cells such as K562. Based on these characteristics, we classified REST⁺ distal enhancer cCREs as REST⁺ enhancer/silencer cCREs.

REST⁺ proximal enhancer cCREs: REST⁺ cCREs comprised 1,080 proximal enhancers. While the majority of these elements do not exhibit features consistent with silencers, a subset

shows evidence of repressive activity. Of the six REST+ proximal enhancer cCREs tested in the transgenic mouse enhancer assay, five exhibited enhancer activity (**Supplementary Figure 15c**). Although the sample size was too small for robust statistical comparison, the validation rate was generally consistent with the REST- proximal enhancer cCRE control set (151 active and 105 inactive).

In the STARR-seq assay, these cCREs exhibited a slight but statistically significant shift toward silencer activity compared to class-matched controls (median = -0.03 vs 0.01; two-sided Wilcoxon test, $p = 0.02$, **Supplementary Figure 15d**). Nearby genes also showed modestly lower expression compared to other proximal enhancers (median TPM = 4.1 vs 13.8), but expression was not substantially lower than the background set of cCREs with low chromatin accessibility (TPM = 2.3, **Supplementary Figure 15e**).

Given their overall enhancer-like chromatin signatures and expression levels, we concluded that most REST+ proximal enhancer cCREs are not silencers. However, to capture elements with putative dual functionality, we classified REST+ proximal enhancers with statistically significant repressive STARR-seq activity ($N=17$, $p < 0.05$) as REST+ enhancer/silencer cCREs.

REST+ CA-TF cCREs: REST+ cCREs comprised 789 CA-TF cCREs. Of the 11 tested in the transgenic mouse enhancer assay, none showed enhancer activity (two-sided Fisher's Exact test, $p = 0.001$, **Supplementary Figure 15c**). By definition, CA-TF cCREs lack H3K27ac enrichment in all human biosamples, indicating they are not associated with active enhancer chromatin in any surveyed context.

REST+ CA-TF cCREs exhibited strong silencer activity in K562 cells, with a median STARR-seq score of -0.40, among the most repressive values observed across all cCRE classes (two-sided Wilcoxon test, $p < 2.2 \times 10^{-16}$, **Figure 3e**, **Supplementary Figure 15d**). Additionally, consistent with their repressive function, genes located near REST+ CA-TF cCREs showed minimal expression in K562 cells (median TPM = 0.1), markedly lower than the background expression of genes near low chromatin accessibility cCREs (median TPM = 2.3; two-sided Wilcoxon test, $p < 2.2 \times 10^{-16}$; **Supplementary Figure 15e**). Given their consistent lack of enhancer activity,

absence of active chromatin marks, strong silencer function, and association with low nearby gene expression, we classified REST+ CA-TF cCREs as REST+ silencer cCREs.

REST+ CA-CTCF cCREs: REST+ cCREs comprised 625 CA-CTCF cCREs. Of the four CA-CTCF cCREs tested in the transgenic mouse enhancer assay, two exhibited enhancer activity. Although suggestive, this sample size is too limited to assess enhancer potential with confidence. Although CA-CTCF cCREs lack H3K27ac signals across all human biosamples in the Registry, one of the active CA-CTCF cCREs, EH38E4261358, had an orthologous mouse cCRE, EM10E1516788, with strong H3K27ac enrichment in the developing mouse brain. This suggests that some CA-CTCF cCREs may function as enhancers in specific developmental contexts not represented in current human epigenomic datasets.

In K562 cells, REST+ CA-CTCF cCREs exhibited strong silencer activity, with a median STARR-seq score of -0.35 (**Supplementary Figure 15d**). Nearby genes also showed significantly reduced expression (median TPM = 0.1 vs 2.3 for background; two-sided Wilcoxon test, $p < 2.2 \times 10^{-16}$; **Supplementary Figure 15e**), further supporting a repressive role.

Given the consistent evidence for strong silencer function across assays and limited evidence for enhancer activity, we classified the majority of REST+ CA-CTCF cCREs as REST+ silencer cCREs. The exception was the single element with dual lines of enhancer evidence—both activity in the transgenic mouse enhancer assay and H3K27ac enrichment in the developing mouse brain—which we instead classified as a REST+ enhancer/silencer cCRE.

REST+ CA-H3K4me3 cCREs: REST+ cCREs comprised 232 CA-H3K4me3 cCREs. None of the REST+ CA-H3K4me3 cCREs were tested in transgenic mouse enhancer assays, so their in vivo enhancer potential could not be directly assessed (**Supplementary Figure 15c**). However, consistent with their classification, these cCREs showed no evidence of enhancer-associated histone marks; they instead had strong silencer activity in the K562 STARR-seq assay (median score = -0.25 , Figure 3e, **Supplementary Figure 15d**). Genes located near REST+ CA-H3K4me3 cCREs had markedly reduced expression (median TPM = 0.1 vs 2.3 for background; two-sided Wilcoxon test, $p < 2.2 \times 10^{-16}$; **Supplementary Figure 15e**). Based on

their chromatin profile, silencer activity, and repressive effect on nearby gene expression, we classified REST+ CA-H3K4me3 cCREs as REST+ silencer cCREs.

REST+ TF cCREs: REST+ cCREs comprised 1,396 cCREs. None of the REST+ TF cCREs were tested in the transgenic mouse enhancer assay, so direct in vivo enhancer potential could not be evaluated (**Supplementary Figure 15c**). They demonstrated modest silencer activity in K562 cells, with a median STARR-seq score of -0.17 (Figure 3e, **Supplementary Figure 15d**). Genes located near REST+ TF cCREs also showed reduced expression (median TPM = 0.2 compared to 2.3 for background; two-sided Wilcoxon test, $p < 2.2 \times 10^{-16}$; **Supplementary Figure 15e**). Based on the absence of enhancer-associated chromatin features, silencer-like activity in STARR-seq, and their association with reduced gene expression, we classified REST+ TF cCREs as REST+ silencer cCREs.

3.2 Examples of REST+ cCREs

To illustrate the two types of REST+ cCREs, we highlight two representative examples (**Extended Data Figure 4**). using the enhancer/silencer cCRE EH38E2130108 and the silencer cCRE EH38E4127779.

The REST+ enhancer/silencer cCRE EH38E2130108 is located 2,199 bp downstream of a TSS for *MYT1* (Myelin Transcription Factor 1, [View on SCREEN](#)). *MYT1* encodes a neural-specific transcription factor that plays a critical role in the development and differentiation of neurons, particularly by regulating gene expression during neurogenesis^{96,97}. EH38E2130108 has active enhancer signatures in neurons, brain organoids, and the retinoblastoma cell line WERI-Rb-1 ([View on SCREEN](#)). In K562 cells, EH38E2130108 has high REST ChIP-seq and DNase signals (**Extended Data Figure 4a**, DNase Z-score = 2.84) and a low STARR score (-0.64, **Extended Data Figure 4b**). When tested in the transgenic mouse enhancer assay, EH38E2130108 had enhancer activity in mouse forebrain, midbrain, and hindbrain (**Extended Data Figure 4c**).

The REST+ silencer cCRE EH38E4127779 is located 57,160 upstream of *LINGO1*, which encodes a transmembrane protein expressed in the central nervous system that regulates

neuronal survival, axon regeneration, and myelination ([View on SCREEN](#))⁹⁸. EH38E4127779 is classified as a CA-TF cCRE and therefore does not have active enhancer signatures in any cell type ([View on SCREEN](#)). In K562 cells, EH38E4127779 has high REST ChIP-seq and DNase signals (**Extended Data Figure 4d**, DNase Z-score = 2.39) and a low STARR score (−0.94, **Extended Data Figure 4e**). When tested in the transgenic mouse enhancer assay, EH38E2130108 had no measurable activity (**Extended Data Figure 4f**).

Supplementary Note 4 - Identifying additional silencer cCREs and comparing them with existing collections of silencer elements

4.1 Reproducibility of STARR silencers

To assess the reproducibility of predicted silencers, we evaluated STARR scores for silencer cCREs defined using the Reddy lab's K562 STARR-seq dataset (ENCSR661FOW) across additional STARR-seq datasets generated by the White lab, including an independent K562 library (ENCSR858MPS) and six other cell types (HepG2, MCF-7, HCT116, A549 and SH-SY5Y).

Silencers identified in the Reddy dataset exhibited consistently negative STARR scores—indicative of silencer activity—across all six White lab datasets. This trend held across all four silencer classes: REST+ enhancer/silencers (median STARR scores −0.29 to −0.16) and REST+ silencers (−0.58 to −0.32) (both defined independently of STARR-seq) and robust (−0.3 to −0.44) and stringent STARR silencers (−0.75 to −0.51, **Supplementary Figure 16e**, Pairwise two-sided Wilcoxon tests with FDR correction, $p < 2.2 \times 10^{-16}$). Among these, stringent STARR silencers showed the strongest and most consistent repression. As a control, randomly selected cCREs displayed STARR scores near zero in all experiments.

Due to the pronounced GC content bias in the White lab datasets (**Supplementary Note 2.3**), we were unable to confidently define silencers *de novo* from those data: many canonical active promoters—which are typically GC-rich—exhibited negative STARR scores, confounding interpretation. As such, we relied on the Reddy lab K562 dataset for primary silencer identification and used the White lab datasets only for validation and cross-cell-type comparison.

We note that STARR-seq measures regulatory activity in an episomal context and may not fully recapitulate endogenous chromatin environments; however, the observed cross-dataset consistency demonstrates that our annotations are reproducible across cell type and experimental protocols.

4.2 CRISPRi-FLOWFish validation of STARR silencers

To assess STARR silencer cCRE activity in native chromatin contexts, we integrated results from K562 CRISPRi-FlowFISH experiments targeting 18 genes across four genomic loci⁶⁰. Of the 734 perturbed cCREs, two were STARR silencer cCREs (**Supplementary Table 11h**). The first, EH38E4193243, is a proximal enhancer for the *RTBDN* gene, which is preferentially expressed in the retina⁴³. EH38E4193243 has active enhancer signatures (high DNase and H3K27ac) exclusively in WERI-Rb-1, a retinoblastoma cell line, suggesting that this cCRE has highly cell type-specific enhancer activity. In other biosamples—including K562—it is bound by REST and *RTBDN* is not expressed. In addition to being a REST+ enhancer/silencer, EH38E4193243 is also a robust STARR silencer (STARR score = -1.02, $p = 0.05$).

CRISPRi perturbation of EH38E4193243 resulted in increased expression of *PRDX2* (effect size = -0.026, FDR = 0.002) (**Supplementary Note 4.2**). Though this cCRE lies 30 kb upstream of *PRDX2*, RNAPII and CTCF ChIA-PET interactions demonstrate that this cCRE is close to *PRDX2* in three dimensional space (**Supplementary Figure 17**). Interestingly, perturbing nearby cCRE EH38E3291318, resulted in a similar decrease in *PRDX2* expression (effect size = -0.030, FDR = 0.0004). In K562 cells, EH38E3291318 is a CA-CTCF cCRE that mediates 3D chromatin interactions. These CTCF-mediated interactions include a 14 kb loop with the *RTBDN* enhancer/silencer cCRE EH38E4193243 and a 130 kb loop with another CA-CTCF cCRE EH38E1939855 (**Supplementary Figure 17**). The long range interaction between EH38E3291318 and EH38E1939855 creates a domain encompassing seven actively expressed genes and 82 active cCREs in K562. Within this CTCF-mediated domain, the *RTBDN* enhancer/silencer cCRE EH38E4193243 is positioned near active promoters, which is evidenced by a RNAPII ChIA-PET link connecting it with the promoter of *JUNB*. CRISPRi-FlowFISH results demonstrate that inhibiting either the silencer activity of

EH38E4193243 or the 3D interactions maintained by CA-CTCF cCRE EH38E3291318 results in an increase in expression of *PRDX2*. These results suggest that silencers can impact the expression of multiple genes, including those tens-of-thousands of base pairs away. We can use our annotated STARR silencer and REST+ silencer cCREs to aid in future CRISPRi-FlowFISH study design to further investigate such long-range silencer interactions and validate additional silencer cCREs in their native chromatin contexts.

Despite silencer EH38E4193243 being proximal to the *JUNB* promoter in three dimensional space, there were no significant differences in *JUNB* expression when either silencer EH38E4193243 or CA-CTCF cCRE EH38E1939855 was perturbed. This could be due to several factors such as higher overall expression levels or false negative interactions. Understanding the rules governing such selectivity and specificity warrant future investigation and the Registry of cCREs serves as the ideal framework for such future studies.

The other STARR silencer cCRE tested by CRISPRi-FlowFISH was EH38E3019358. EH38E3019358 is a CA-TF cCRE in K562 that overlaps ChIP-seq peaks for erythroid transcription factors GATA1 and TAL1, as well as transcription factors with known repressive activities including TCF12⁹⁹, ZEB2¹⁰⁰, and IKZF1¹⁰¹ (**Supplementary Table 11h**). While we did not observe any significant changes in gene expression when this cCRE was targeted, this is likely because its target gene was not assayed in these particular CRISPRi-FlowFISH experiments. The two genes measured at this locus were *NFE2* (85 kb away) and *ITGA5* (209 kb away) and were not linked to EH38E3019358 via any 3D chromatin interactions. Instead, possible target genes for EH38E3019358 may include *SMUG1*, *CBX5*, or the *HOXC* locus which are all connected to EH38E3019358 via Hi-C loops or ChIA-PET interactions (**Supplementary Table 11h**).

4.3 Overlap of silencer cCREs with previously published collections

Our collection of silencer cCREs has fairly low overlap with previously published collections of silencers, suggesting that we identified thousands of novel silencers in the human genome. Below are detailed comparisons of our silencer cCREs with these previously published collections.

Jayavelu *et al.*¹⁶ tested 7,539 “uncharacterized CREs”—that is open chromatin regions that were at least 2.5kb from a TSS and did not overlap either a H3K4me3 or CTCF ChIP-seq peak—in an MPRA/STARR-seq assay. We refer to these elements as *Jayavelu elements*. They identified silencers by selecting all elements with a mean activity level below control regions, resulting in a total of 2,997 candidates, which we refer to as *Jayavelu silencers*. In total, 83% of Jayavelu elements and 81% of Jayavelu silencers overlapped cCREs. Of the 2,586 cCREs that overlapped Jayavelu silencers, only 232 (9%) were annotated as silencer cCREs (**Supplementary Figure 18a,b**). However, these 2,586 cCREs did have lower STARR scores than cCREs that overlapped non-silencer Jayavelu elements (median STARR score of -0.06 vs 0.14, $p < 2.2 \times 10^{-16}$) and randomly selected cCREs (-0.06 vs. -0.02, $p < 2.2 \times 10^{-16}$), suggesting that we may identify concordant STARR silencer cCREs with less stringent thresholds. It is also worth noting that 250 of our silencer cCREs overlapped Jayavelu elements and 232 (93%) were Jayavelu silencers. Stratifying by specific categories of silencer cCREs, 100% of tested stringent STARR silencer cCREs, 95% of tested robust STARR silencer cCREs, 96% of REST+ silencer cCREs, and 90% of REST+ enhancer/silencer cCREs were called as Jayavelu silencers (**Supplementary Figure 18a**).

Pang *et al.*¹⁷ tested 177,046 open chromatin regions identified through FAIRE-seq, which we refer to as *Pang elements*, in a repressive ability of silencer elements (ReSE) functional screen from which they identified 2,662 candidate silencers, which we refer to as *Pang silencers*. In total, 56% of Pang elements and 57% of Pang silencers overlapped cCREs. Of the 1,793 cCREs that overlapped Pang silencers, only 11 (0.6%) were annotated as silencer cCREs. These 1,793 cCREs actually had slightly higher STARR scores than cCREs that overlapped non-silencer Jayavelu elements (median STARR score of 0.06 vs. 0.05, $p = 0.07$) and significantly higher STARR scores than randomly selected cCREs (0.06 vs. -0.02, $p < 2.2 \times 10^{-16}$). Of the 624 silencer cCREs that overlapped Pang elements, only 11 (1.8%) were Jayavelu silencers.

Huang *et al.*¹⁵ identified silencers by selecting DNase hypersensitivity sites that overlapped H3K27me3 and were negatively correlated with gene expression. They identified 8,084

candidate silencers, which we refer to as *Huang silencers*. In total, 93% of Huang silencers overlapped cCREs. Of the 22,465 cCREs that overlapped Huang silencers, only 135 (0.6%) were annotated as silencer cCREs. Additionally the 22,465 cCREs that overlapped Huang silencers had slightly higher STARR scores than randomly selected cCREs (0.02 vs. -0.02, $p < 2.2 \times 10^{-16}$).

Cai *et al.*¹⁸ identified silencers by annotating H3K27me3-rich regions (MRRs). They identified 1,677 MRRs, which we refer to as *Cai silencers*. In total, nearly 100% of Cai silencers overlapped cCREs. Of the 222,109 cCREs that overlapped Pang silencers, 1,138 (0.5%) were annotated as silencer cCREs. Additionally the 1,138 cCREs that overlapped Cai silencers had significantly higher STARR scores than randomly selected cCREs (0.004 vs. -0.02, $p < 2.2 \times 10^{-16}$).

Overall, our analysis indicates that silencer cCREs identified through our approach represent a distinct class of regulatory elements compared to previous studies. While Jayavelu *et al.* use of a similar MPRA/STARR-seq-based method resulted in a higher degree of overlap, the divergence seen with other datasets, such as those from Pang *et al.*, Huang *et al.*, and Cai *et al.*, is explored in greater detail in **Supplementary Notes 4.5-4.6**.

4.4 Chromatin accessibility enrichment

Only 14% of silencer cCREs had high levels of chromatin accessibility in K562 cells (z-score > 1.64, **Supplementary Figure 18c**). Stratifying by subclasses, 32% of REST+ enhancer/silencer cCREs had high chromatin accessibility, followed by REST+ silencer cCREs (17%), robust STARR silencer cCREs (8%) and stringent STARR silencer cCREs (6%). Silencer cCREs did have slightly higher levels of chromatin accessibility in K562 compared to randomly selected cCREs (median z-score 0.26 vs. -0.08, two-sided Wilcoxon test, $p < 2.2 \times 10^{-16}$). These results suggest silencer cCREs maintain a baseline level of accessibility that may be sufficient for the binding of repressive transcription factors, even if not as broadly accessible as promoters or enhancers.

Since silencer cCREs had low chromatin accessibility in the cell type in which they act as silencers—whether defined by STARR-activity or REST binding—we were interested in determining in which cell types these cCREs had high DNase signals. Compared to non-silencer cCREs, silencer cCREs were enriched for chromatin accessibility in early-stage cell types such as pluripotent cell lines (ESCs and iPSCs) and some *in vitro* differentiated cell lines (**Extended Data Figure 6a,b, Supplementary Table 12a**). Tissues from embryonic samples were more likely than adult tissues to have open chromatin at silencer cCREs, presumably because embryonic tissues comprise fewer mature cell types. This pattern was particularly significant for REST+ silencer cCREs (**Extended Data Figure 3c, Supplementary Table 12b**). Embryonic tissues had significantly higher chromatin accessibility at REST+ silencer cCREs compared to matched adult tissue, except for brain tissues, where REST+ silencer cCREs had low chromatin accessibility in both embryonic and adult tissues.

There were also cancer cell lines, such as Ishikawa (endometrial adenocarcinoma) and T47D (breast cancer), that showed enriched chromatin accessibility for silencer cCREs (**Extended Data Figure 6a,b, Supplementary Table 12a**). This may be due to the tendency of certain cancers to revert to a more pluripotent-like state, characterized by the reactivation of developmental pathways¹⁰² and perhaps a broader chromatin landscape. Future work will aim to determine whether these changes in chromatin accessibility at silencers are specific to particular cancer types and to assess their functional impact on gene regulation and cellular behavior.

It is important to note that the presence of DNase signal in embryonic and early-stage biosamples does not necessarily indicate that silencer cCREs are actively functioning as silencers in those contexts. Instead, we interpret this enrichment as evidence that these elements are accessible at specific points during cellular differentiation, providing an opportunity for transcriptional repressors such as REST to bind and initiate chromatin remodeling or repression programs. The low accessibility observed in mature cell types like K562 is consistent with silencer function and the establishment of a repressed chromatin state.

4.5 Overlap of K562 silencers with ChromHMM states

To further contextualize our silencer cCREs, we analyzed their chromatin state annotations in K562 using a 15-state ChromHMM model (**Supplementary Figure 18d, Supplementary Table 12c**). We compared these annotations to those for silencer sets from previously published studies, including Jayavelu *et al.*¹⁶, Pang *et al.*¹⁷, Huang *et al.*¹⁵ and Cai *et al.*¹⁸

REST+ enhancer/silencer cCREs and REST+ silencer cCREs showed a high proportion of overlap with the *Quiescent* chromatin state (47% and 72%, respectively), similar to STARR-stringent (64%) and STARR-robust (66%) silencers. These values were comparable to cCREs with low chromatin accessibility (presumably inactive, 65%), suggesting that many silencers reside in regions lacking strong histone modification signals in K562 cells. REST+ and STARR silencers also exhibited modest enrichment for the *Repressed Polycomb (ReprPC)* state (19–26%), although not to a greater extent than low accessibility cCREs (20%), and well below the levels observed in silencer datasets defined by H3K27me3 enrichment (e.g., Cai: 57.5%; Huan: 53.8%). This supports the conclusion that while some silencers may overlap canonical repressive marks such as H3K27me3, this overlap is not uniquely predictive of silencer function and does not necessarily distinguish them from other inactive elements.

We also examined overlap with the *Bivalent* state, characterized by co-occurrence of H3K27me3 and H3K4me1/3. REST+ enhancer/silencer cCREs showed a modest Bivalent signal (5.8%), slightly higher than low accessibility cCREs (2.3%) and considerably more than STARR stringent (0.7%) or REST+ silencer cCREs (1.5%). In contrast, the Huan *et al.* set had the highest Bivalent overlap (15.2%), consistent with their reliance on H3K27me3-rich regions in embryonic tissues. Overall, these results suggest that bivalency is not a widespread feature of our silencer cCREs.

In summary, while some REST+ and STARR silencers overlap repressive or bivalent chromatin states, these features are neither consistent nor uniquely enriched relative to low accessibility cCREs. This contrasts with silencer sets specifically defined by H3K27me3 enrichment, which show high overlap with *ReprPC* and *Bivalent* states. Our results indicate that functionally

defined silencers—particularly those identified via STARR-seq—may represent a distinct class of regulatory elements that are not readily identifiable by chromatin state annotations alone.

4.6 Conservation of K562 silencers and overlap with repeats

To better understand the sequence origins and evolutionary dynamics of silencer elements, we compared both transposable element (TE) content and evolutionary conservation profiles across our annotated silencers and those from the previously published studies^{15–18}. We also compared these features to representative non-silencer cCRE classes (promoters, enhancers, low accessibility) in K562 cells.

Repeat Element Overlap

We evaluated the fraction of each silencer set overlapping major classes of transposable elements (SINEs, LINEs, and LTRs), using RepeatMasker annotations. Notably, distinct silencer annotations showed divergent transposable element enrichments (**Extended Data Figure 6d, Supplementary Table 12d**).

REST+ silencer cCREs (27%) and STARR silencer cCREs (29% and 30% for stringent and robust sets, respectively) were enriched for LINEs. This pattern was notably stronger than in REST+ enhancer/silencer cCREs (16%) and all non-silencer control groups, suggesting that LINE-associated sequences may play a functional role in silencing or contribute to sequence turnover. Among published annotations, Jayavelu *et al.* (22.5%) and Cai *et al.* (18.5%) also showed elevated LINE overlap, while Pang *et al.* (11.9%) and Huan *et al.* (12.7%) had lower LINE content.

Pang silencers exhibited the highest SINE overlap (27%), followed by Cai and Jayavelu (22–23%). Our REST+ silencers (17%) and STARR-seq silencers (19–23%) fell in an intermediate range, while K562 promoters (5%) and low accessibility cCREs (22%) provided lower and higher baselines, respectively.

Jayavelu silencers showed the strongest LTR enrichment (26%), consistent with prior reports that some LTRs harbor transcriptional repressor motifs. REST+ silencers (11%) and STARR

silencers (16%) had modest LTR content, while most other annotations, including Huan and Pang silencers and K562 control cCREs, fell in the 8–14% range.

These patterns suggest that different silencer definitions may reflect distinct evolutionary mechanisms, with REST+ and STARR silencer cCREs particularly enriched in LINEs—especially L1 elements—while Pang silencers emphasize SINE-derived regions and Jayavelu silencers capture LTR-associated sequences.

Zoonomia-Based Evolutionary Conservation

We further evaluated evolutionary conservation of silencer cCREs using N1/N2 metrics derived from Zoonomia’s 241-species mammalian alignment as described in **Supplementary Note 1.6** and complementing our silencer cCRE analyses in **Supplementary Notes 3 and 4.6 (Supplementary Figure 18e-h, Supplementary Table 12d)**. Among published silencers, Huan silencers showed relatively high conservation (Group 1: 50%, Group 2: 13%, Group 3: 6%), where as Jayavelu silencers were more primate-specific (Group 1: 32%, Group 2: 12%, Group 3: 22%). Notably, Pang silencers had the second highest proportion of elements in Group 2 (19%), consistent with their enrichment in SINE elements and signatures of intermediate evolutionary constraint.

Together, these results reveal that silencer annotations differ markedly in both repeat element content and evolutionary conservation, likely reflecting a mixture of biological function, annotation strategy, and the underlying sequence features used in each method. The observed enrichment of specific TE families (e.g., L1 in REST+ and STARR-seq silencers, LTRs in Jayavelu, and SINEs in Pang) points to the possibility that different silencer subclasses may arise from distinct evolutionary mechanisms.

Overlap with common genetic variation

To assess evolutionary constraint within human populations, we intersected our silencer sets with common genetic variants from dbSNP (build 155) and calculated variant densities. REST+ enhancer/silencer cCREs exhibited slightly higher variant density than canonical enhancer classes and promoters (0.00372 vs. 0.00331–0.00340, **Supplementary Table 12f**), but

remained below the levels observed for REST+ silencers (0.00398) and both robust and stringent STARR silencers (0.00381–0.00389). These densities were also lower than those seen in published silencer sets such as Cai *et al.* (0.00426), Pang *et al.* (0.00395), and Jayavelu *et al.* (0.00339). Collectively, these results suggest that putative silencer elements tend to show reduced selective constraint, consistent with a greater tolerance for genetic variation.

4.7 STARR silencer activity of non-cCRE regions

To evaluate the extent to which silencer activity may be missed due to absence from the current Registry of cCREs, we conducted a specialized CAPRA analysis on a background set of genomic regions that do not overlap any annotated cCRE. We randomly selected a size-matched set of genomic regions that overlapped similar regions of mappability (UMAP K24 from Hoffman lab⁶⁷). Using CAPRA, we quantified STARR-seq activity for these non-cCRE regions and identified 161 stringent silencers ($p < 0.01$) and 1,743 robust silencers ($p < 0.05$, **Supplementary Table 12g**). While there were about half as many of these regions identified (0.01% vs. 0.02% and 0.1% vs. 0.2%, respectively) and these regions showed weaker repression on average than silencer cCREs (median STARR score of -1.32 for cCREs vs -1.15 for non-cCREs, two-sided Wilcoxon test, $p < 2.2 \times 10^{-16}$) they nonetheless exhibited statistically significant depletion of RNA relative to input DNA. Motif analysis of the stringent non-cCRE silencers revealed enrichment for GFI1B binding sites—a transcriptional repressor with no corresponding ChIP-seq data included in the Registry—suggesting that some non-cCRE silencers may reflect bona fide repressive elements missed due to limited chromatin accessibility or incomplete transcription factor profiling (**Supplementary Table 12h**). These results highlight the value of STARR-seq for identifying repressive potential in regions not captured by chromatin-based annotations and motivate the development of complementary strategies to prioritize non-cCRE silencers for follow-up.

Supplementary Note 5 - MAFF and MAFK bound cCREs

5.1 Defining MAFF/MAFK+ cCREs

MAFF and MAFK are small Maf transcription factors that typically act as heterodimeric partners with other bZIP proteins to regulate gene expression in response to cellular stress^{103,104}. Lacking intrinsic transactivation domains, they function primarily as modulators, contributing to

transcriptional repression or context-specific activation. Both factors have been implicated in oxidative stress responses, inflammation, and hematopoietic differentiation.

We defined MAFF/MAFK+ cCREs as cCREs that overlap the ChIP-seq peak summits of both transcription factors. In K562, we identified 17,387 MAFF/MAFK+ cCREs, of which 10,061 (58%) exhibited low chromatin accessibility (**Supplementary Figure 19a,b, Supplementary Data 5, Supplementary Table 13a**). In HepG2, we annotated 33,236 MAFF/MAFK+ cCREs, with 27,559 (82%) classified as low chromatin accessibility (**Supplementary Figure 19c,d, Supplementary Data 6, Supplementary Table 13a**). Although overall overlap between cell lines was moderate (11,182 shared cCREs; 64% of the K562 set), concordance was much higher within the low-accessibility subset: 85% of low chromatin accessibility MAFF/MAFK+ cCREs in K562 were also present in HepG2 (**Supplementary Figure 19e,f**). The majority of low-accessibility MAFF/MAFK+ cCREs fell into two cell type-agnostic categories: distal enhancer (44% in K562, 40% in HepG2) and TF (41% and 43%, respectively, (**Supplementary Figure 19b,d, Supplementary Table 13a**). We focused subsequent analyses on the low-accessibility MAFF/MAFK+ cCREs shared between these two cell types (N = 8,455).

5.2 MAFF/MAFK+ cCREs do not have silencing activity

Given prior evidence that MAFF and MAFK can act as transcriptional repressors, we hypothesized that they may bind silencers. However, unlike REST+ cCREs, MAFF/MAFK+ cCREs did not show significantly reduced STARR-seq activity (−0.2 vs −0.2 for all cCREs **Supplementary Figure 19g**). Nearby gene expression was only modestly lower than that of other low-accessibility cCREs in K562 (median TPM 1.6 vs. 2.4; two-sided Wilcoxon test, $p = 0.0007$; **Supplementary Figure 19h**), and significantly higher than expression near REST+ or STARR silencers (0.1 and 0.5, respectively, two-sided Wilcoxon test, $p < 2.2 \times 10^{-16}$). These results suggest that while MAFF/MAFK+ cCREs may contribute to transcriptional repression, they likely operate through a distinct mechanism from our silencer cCREs.

5.3 Gene ontology analysis

To investigate the biological relevance of MAFF/MAFK+ cCREs, we performed gene ontology enrichment analysis for genes located near these elements. The results revealed strong enrichment for developmental and morphogenetic processes, including limb and appendage morphogenesis, cardiac and renal development, camera-type eye formation, and neural differentiation and connectivity (**Supplementary Table 13b**). We also observed overrepresentation of pathways related to neurogenesis, synaptic organization, axon guidance, epithelial-mesenchymal transition, cell polarity, and Wnt and TGF- β signaling—hallmarks of dynamic developmental regulation. These findings support the hypothesis that MAFF/MAFK+ cCREs mark a subset of latent or stimulus-responsive enhancers poised to regulate lineage-specific gene expression programs in response to developmental or environmental cues.

We further analyzed the subset of MAFF/MAFK+ TF cCREs that lack active enhancer chromatin signatures. Genes near these elements were enriched for sensory perception pathways, particularly those related to chemical detection and olfaction (**Supplementary Table 13c**). These functions are typically restricted to specialized cell types, such as olfactory sensory neurons, that are not yet represented in the Registry of cCREs. These findings support the idea that MAFF/MAFK+ TF cCREs represent context-dependent enhancers active in unprofiled cellular or developmental contexts, and that the integration of multi-modal data enables the Registry to capture latent regulatory potential even in the absence of direct functional characterization.

5.4 Enrichment of chromatin signatures at MAFF/MAFK+ enhancer cCREs

We next identified biological contexts in which the MAFF/MAFK+ enhancer cCREs had active signatures, that is high DNase and H3K27ac signals (**Supplementary Table 13d,e**). The highest enrichment for DNase and H3K27ac signals at MAFF/MAFK+ enhancer cCREs was in the SK-N-MC cell line with fold enrichments > 3 for both signals (median fold enrichments across all samples were 0.5 for DNase and 0.7 for H3K27ac). MAFF/MAFK+ enhancer cCREs were also enriched for active signatures in primary myeloid-derived cell types such as monocytes, dendritic cells, and neutrophils, but showed minimal signal in most lymphoid lineages,

including T and B cells (median DNase fold enrichment 1.3 vs 0.3, $p < 2.2 \times 10^{-16}$). Similarly, fibroblasts showed significantly higher enrichment compared to other biosamples (median DNase fold enrichment 1.7 vs 0.5, $p = 1.8 \times 10^{-15}$). Finally, we observed strong H3K27ac enrichment at MAFF/MAFK+ enhancer cCREs in parathyroid adenoma tissue, a benign neoplastic sample. Enrichment was notably higher than in any other primary tissue (median fold enrichment 1.5 vs 0.7).

Supplementary Note 6 - Overlap of the Registry of cCREs with other classes of Encyclopedia annotations

The ENCODE Encyclopedia comprises four levels of annotations: sequence, elements, genes, and interactions. The Registry of cCREs is an integral component of the element annotations and, when combined with other annotations, provides a broader view of gene regulation by capturing the interplay between regulatory elements and their functional contexts.

6.1 Chromatin states (elements)

We previously demonstrated that cCREs and chromatin states, such as those generated by ChromHMM, are highly concordant⁹. These two approaches for identifying cis-regulatory regions are complementary, with cCREs providing high spatial resolution and chromatin states providing finer-grained histone mark combinations. Annotating cCREs with chromatin states can reveal subtle differences in regulatory element activity, while maintaining high resolution elements. For example, Borsari *et al.*, integrated temporally derived chromatin states with cCREs to study the transdifferentiation of human pre-B cells into macrophages and discovered that a precise ordering of histone modifications was required for gene activation³⁴.

Therefore, to complement the Registry of cCREs, we generated ChromHMM states for 330 epigenomes consisting of H3K4me3, H3K27ac, H3K4me1, H3K36me3, H3K27me3, and H3K9me3 histone marks (**Supplementary Table 14**). Because we observed differences in signal thresholding between models trained on ChIP-seq or Mint ChIP-seq data, we trained two assay-specific models (**Supplementary Methods**). We trained a ChIP-seq model on 20 epigenomes, which were generated across all project stages, and used the model to generate chromatin states for 210 epigenomes. For Mint ChIP-seq data, we trained the model on 20

epigenomes derived from a combination of primary cells, *in vitro* differentiated cells and cell lines, which we used to generate chromatin states for 120 epigenomes. These chromatin states can be viewed and are available for download on SCREEN.

6.2 ENTEEx (sequence & elements)

The EN-TEEx resource mapped functional genomic datasets to diploid genomes from four individuals, revealing over a million allele-specific loci, demonstrating the importance of local nucleotide sequences, and enabling improved predictive modeling of eQTLs in challenging tissues¹⁰⁵. As part of the EN-TEEx resource we annotated cCREs with allele-specific biochemical activity and tissue-specific activity. Here we updated these annotations for V4 cCREs.

Of the 514,427 cCREs overlapping variants that were heterozygous in at least one of the four ENTEEx donors, 26% (131,215) had allelic imbalance for either one of the six surveyed histone marks (H3K4me3, H3K4me1, H3K27ac, H3K36me3, H3K27me3, and H3K9me3) or one of the three DNA binding proteins (CTCF, EP300, POLR2A). For 28 tissues, we also annotated active cCREs via overlap with the histone modifications H3K4me1, H3K4me3 and H3K27ac, annotating 31% (N = 717,095) of V4 cCREs. These annotations are available to view and download via SCREEN.

6.3 Annotated GENCODE genes (genes)

We identified promoter cCREs for 90% of protein coding GENCODE genes (**Supplementary Figure 2a**). Our coverage of lncRNA genes, which tend to be much more cell type specific (Gini coefficient of 0.56 compared to 0.37 for protein coding genes, $p < 2.2 \times 10^{-16}$), was less at only 57%. Within specific cell types—specifically deeply profiled cell lines—an average of 82% of expressed genes (TPM > 1, as measured by short read RNA-seq) have an annotated promoter cCRE in that cell type (**Supplementary Table 15a**). This percentage increased to 90% and 92% when we analyzed genes with higher expression levels of 5 and 10 TPM, respectively. These results demonstrate that the Registry of cCREs covers the majority of promoters for genes expressed in the included cell types.

We then compared the expression of four classes of genes: (1) genes with an active promoter cCRE (high DNase and high H3K4me3) with high H3K27ac signal, (2) genes with an active promoter cCRE that had low H3K27ac signal, (3) genes with an inactive promoter cCRE (low DNase and low H3K4me3), and (4) genes without a promoter cCRE (**Supplementary Table 15b**). Across all 13 tested cell types, genes with active promoter cCREs with high H3K27ac signal had the highest expression with an average median TPM of 10.3. Genes with promoter cCREs that had low H3K27ac signal but high chromatin accessibility and H3K4me3 signals in the tested cell type had lower levels of expression with average median TPMs of 2.0, suggesting these promoter cCREs are likely poised with basal levels of transcription. Genes in the last two groups were generally not expressed with average median TPMs < 0.1.

To assess improvements in coverage, we compared the proportion of protein-coding genes with at least one annotated promoter cCRE in ENCODE3 (GENCODE v24) and ENCODE4 (GENCODE v40). In ENCODE3, 83% of protein-coding genes were associated with a promoter cCRE, compared to 90% in ENCODE4. Gene Ontology analysis of the newly covered genes showed enrichment for biological processes such as sensory perception (e.g., bitter taste and phototransduction), immune responses (e.g., humoral and antimicrobial), and hormone and steroid metabolism (**Supplementary Table 15c**). Notably, although taste receptor cells remain underrepresented in the Registry, many of these genes were captured through weaker promoter signals in related biosamples—particularly newly added brain and in vitro-derived neuronal samples—suggesting that partial regulatory activity in related cell types can help extend promoter annotations to loci that are otherwise difficult to profile.

6.4 Novel transcripts (genes)

Our previous work demonstrated that while the GENCODE annotations are fairly comprehensive, transcription assays that capture 5' ends of transcripts such as RAMPAGE and long read RNA-seq identify novel TSSs, particularly those for highly tissue specific transcripts¹⁰⁶.

RAMPAGE

We previously curated a set of 50 thousand transcription sites by integrating RAMPAGE data across 115 cell and tissue types¹⁰⁶. We classified these transcription sites, which we referred to as RAMPAGE representative peaks (rPeaks), based on their overlap with current GENCODE annotations and the genomic locations of the 3' ends of paired end reads. TSSs of annotated GENCODE genes, whether verified or novel, had high overlap with cCRE (96 and 92%, respectively, **Supplementary Table 15d**). TSSs of novel transcripts had slightly lower overlap (86%) and rPeaks that we deemed not be true TSSs and rather the result of local transcription had the lowest overlap with cCREs at 71%.

Long read RNA-seq

The Mortazavi lab generated 98 long read RNA-seq experiments from human biosamples and 65 from mouse biosamples¹⁰⁷. Most samples (N = 137) were sequenced on the Pacific Biosciences (PacBio) Sequel II platform with a smaller number sequenced on the original PacBio Sequel platform (N = 23) and limited number sequenced on the Oxford Nanopore MinION platform (N = 3). These datasets were complemented by eight long read datasets generated by the LRGASP project which assayed three human biosamples (six experiments) and one mouse biosample (two experiments) using different library construction methods and sequencing platforms.

For each long read experiment, we generated TSS clusters by merging 5' ends of all reads within 25 bp of one another. We then selected all clusters with at least 5 reads per million (RPM). This resulted in an average of ~14,000 TSS clusters per human experiment and ~13,000 TSS clusters per mouse experiment (**Supplementary Table 15e, f**). On average, 87% of human TSS clusters were within 200 bp of a human cCRE center and 81% of mouse TSS clusters were within 200 bp of a mouse cCRE center. We observed differences in human cCRE overlap depending on the sequencing platform with an overlap of 87% of TSS clusters from PacBio Sequel II data, 90% of TSS clusters from PacBio Sequel data and 80% of TSS clusters from Oxford Nanopore MinION data. These results support previous findings that the PacBio platform more readily identifies true 5' ends of transcripts¹⁰⁸. We also hypothesize that TSS clusters that are not within 200 bp of a cCRE center are likely false positives as they are less likely to also overlap PRO-cap peaks (**Supplementary Table 15e**). On average 68% of TSS

clusters that overlap a cCRE also overlap a PRO-cap peak in that same cell type compared to only 2% of TSS clusters that do not overlap cCREs (two-sided Fisher's Exact Test, $p < 2.2 \times 10^{-16}$).

6.5 Sites of nascent transcription (elements)

The Lis and Yu labs conducted 24 PRO-cap experiments to measure transcriptional initiation using cap-selection-based nascent RNA sequencing¹⁰⁹ (**Supplementary Table 15g**). To evaluate the role of cap selection, they also performed two experiments in K562 cells without the cap selection step. In experiments with cap selection, an average of 98% of bidirectional PRO-cap peaks overlapped cCREs, compared to only 74% in experiments without cap selection. When restricting the analysis to cell type-specific cCREs, 90% of peaks from cap-selected experiments overlapped active cCREs in the corresponding cell type, compared to just 46% from experiments without cap selection. These findings highlight the critical importance of cap selection in the PRO-cap protocol and confirm that cCREs capture the vast majority of transcriptionally active sites in a given cell type.

6.6 3D chromatin interactions (interactions)

ChIA-PET

The JAX lab generated 67 replicated ChIA-PET experiments—30 targeting RNAPII and 39 targeting CTCF—across a diverse set of cell and tissue types (**Supplementary Table 16a,b**). Prior to intersecting with cCREs, we selected all interactions that were reproducible between biological replicates with anchors greater than 8 kb apart, merging interactions with overlapping anchors. This resulted in an average of 23 thousand interactions for RNAPII experiments and 71 thousand interactions for CTCF experiments (**Supplementary Table 16a,b**).

Nearly all ChIA-PET interactions (>99%) had at least one anchor overlapping a cCRE, with most interactions (96%) having both anchors overlapping a cCRE (**Supplementary Table 16a,b**). However, when considering cell type-specific cCRE annotations, the percentages of both anchors overlapping a cCRE dropped to an average of 68% for RNAPII interactions and 50% for CTCF interactions (**Supplementary Table 16c**). This trend was particularly pronounced

in H1 embryonic stem cells: over 80% of RNAPII and CTCF interactions had at least one anchor overlapping an active cCRE in H1 cells, but fewer than 35% of interactions had both anchors overlapping. These results suggest that some interactions involve regions with regulatory potential that lack active chromatin signatures in the surveyed cell type.

Hi-C

The Aiden lab generated 136 intact Hi-C experiments, identifying an average of ~33,000 loops per experiment (**Supplementary Table 16d**). Consistent with ChIA-PET results, over 99% of Hi-C loops had at least one anchor overlapping a cCRE, and 97% had both anchors overlapping a cCRE. However, in specific cell types, only 43% of loops had both anchors overlapping cCREs active in those cell types (**Supplementary Table 16c**). These findings further support the idea that loops often connect active regulatory elements to regions that may be poised for activity.

6.7 Functional interactions (interactions)

As previously mentioned (**Supplementary Note 2.1**), CRISPRi-FlowFISH is an experimental technique that perturbs CREs to assess their impact on gene expression, enabling systematic mapping and prediction of CRE-gene interactions. We integrated these functional interactions with the Registry of cCREs resulting in 233 cCRE-gene pairs corresponding to 197 cCREs and 15 genes, which can be found on SCREEN.

6.8 Enhancer-gene interaction predictions (interactions)

The ENCODE Distal Regulation working group curated a resource of over 13 million enhancer-gene regulatory interactions across 352 cell types and tissues, integrating models, chromatin state measurements, 3D contacts, and genetic perturbation data¹¹⁰. We integrated the Registry of cCREs with this resource, making the results available on SCREEN. For biosamples in the *Core Collection*, we intersected cell type-specific cCREs with predicted interactions from cell-type-matched rE2G models and evaluated the percentage of cCRE with predicted genes. Though we performed the analysis for all classes of cCREs (**Supplementary Table 16e**), distal enhancer cCREs were of greatest interest. On average, 56% of distal enhancer cCREs were annotated with at least one target gene, with each enhancer linked to

1.9 genes on average. In 75% of cases, the predicted target corresponded to the nearest protein coding gene.

Supplementary Note 7 - SCREEN

We previously developed a web portal, SCREEN (Search Candidate cis-Regulatory Elements by ENCODE) to allow other researchers to query and visualize data from the Registry and the ENCODE Encyclopedia⁹. As part of this work, we redesigned and refactored the entire site to improve usability and performance while also adding new features and datasets (**Supplementary Figure 21**).

SCREEN allows users to search the Registry using a variety of methods, including (1) genomic region, (2) individual cCRE accessions, (3) SNP rsID, or (4) gene name. SCREEN also supports users who are interested in exploring multiple regions through BED file upload and intersection. After searching, users are presented with an interactive table summarizing the cCREs that meet their search criteria. This table can be further filtered based on specific biosample activity, biochemical signal scores, cCRE conservation, or linked genes.

Clicking on individual elements in the results table allows users to see detailed information about each cCRE. Along with basic information such as coordinates and classification (e.g., promoter, proximal enhancer, CA-H3K4me3, etc), each cCRE details page also includes information about nearby genes and SNPs, gene expression, and overlap with transcription factor binding sites. New features include (1) ChromHMM states, which are presented via an interactive genome browser and table, (2) functional characterization data, including quantifications from our new CAPRA quantification method, and (3) expanded cCRE-gene links, which include data from 3D chromatin interactions, CRISPRi-FlowFISH experiments, and computational predictions. Alongside the search functionality, SCREEN also hosts files for users to download and use in their own analysis including the entire set of V4 cCREs, subsets of cCREs based on genomic region, classification, biosample, and biochemical signals, and large matrices containing normalized signals for each of the core biochemical marks across all biosamples and cCREs.

We developed the frontend of SCREEN using Next.js, an open-source React framework that provides robust support for server-side rendering. We have also heavily utilized components from Material UI, an open-source React component library, which implements Google's Material Design. The data for this project are stored in a Google Cloud storage bucket and Google Cloud PostgreSQL database. It is provided to our frontend through a GraphQL application program interface (API). All data on SCREEN can also be downloaded or programmatically accessed through our GraphQL API for downstream analysis.

Supplementary Note 8 - Prioritizing variants at the RTBDN-MAST1 locus

Our initial analysis whittled down the initial list of 63 RBC variants (3 lead and 60 in high linkage disequilibrium) to a list of ten variants that overlapped K562 cCRES (**Figure 4, Supplementary Table 17c**). To further prioritize the RBC variants for future functional follow-up, we implemented a rank-based integrative scoring method. We applied a “rank averaging” approach that collated evidence across multiple data sources, allowing flexible weighting of each category based on its biological relevance to gene regulation and erythroid lineage function (**Supplementary Table 17f**).

We scored each variant across six categories:

1. Impact on coding sequence with the following prioritization: nonsense > missense > synonymous > noncoding [Weight = 1]
2. Allele-specific chromatin accessibility in erythroid progenitors (e.g., erythroblasts, megakaryocyte-erythroid progenitors, **Supplementary Table 17g**) [Weight = 2]
3. Allele-specific chromatin accessibility in other hematopoietic lineages (**Supplementary Table 17g**) [Weight = 0.5]
4. Linkage to KLF1 based on experimental evidence prioritizing: CRISPR perturbation > chromatin contacts & linear proximity > no association (**Supplementary Table 17c**). [Weight = 2]
5. Overlap with transcription factor ChIP-seq peaks in K562 (**Supplementary Table 17h**) [Weight = 1]

6. Evidence of allele-specific binding in GM12878, a lymphoid cell line used as a surrogate due to the lack of heterozygosity and allele-resolved data in K562 (**Supplementary Table 17i**) [Weight = 0.5]

This strategy enabled us to integrate diverse sources of evidence and rank variants based on their potential functional impact within the context of *KLF1* regulation. Our top three candidates are further detailed below.

rs2072597

Our top ranking variant, rs2072597, is a missense variant in the second exon of *KLF1*, resulting in a substitution of Serine (TCC) with Proline (CCC) at the 102nd amino acid. Despite the substitution of two biochemically distinct amino acids, we hypothesize that this mutation does not dramatically impact the function or stability of KLF1 because the amino acid change occurs within a disordered region of the protein, rather than in a functional domain, such as one of the annotated zinc finger domains. Additionally, analysis of the Zoonomia 240-way alignment across mammals shows that this position is not conserved; proline is the most common residue at this site (137 species, including chimpanzee and gorilla), followed by serine (65 species), and threonine (25 species; **Supplementary Figure 22e**). This suggests that the substitution of Serine with Proline is tolerated and may not disrupt protein function across species through additional functional testing is required to confirm.

In addition to changing protein sequence, rs2072597 may influence *KLF1* expression through cis-regulatory mechanisms. The variant lies within a proximal enhancer cCRE located just downstream of the *KLF1* transcription start site, and exhibits allele-specific chromatin accessibility in hematopoietic stem cells (HSCs, $p = 0.03$), multipotent progenitors (MPPs, $p = 0.01$), megakaryocyte-erythroid progenitors (MEPs, $p = 0.03$), common myeloid progenitors (CMPs, $p = 0.01$), and granulocyte-monocyte progenitors (GMPs, $p = 0.01$)—consistent with the erythroid lineage specificity of *KLF1* (**Supplementary Table 17g**). Seven transcription factors had allele-specific binding at this locus in GM12878 and 46 transcription factors were bound in K562. While these findings may support a model in which rs2072597 modulates *KLF1* transcriptional output, it is also possible that the observed allele-specific patterns reflect

downstream consequences of haplotype-specific promoter activity influenced by other nearby causal variants. Further experimental dissection will be required to resolve the individual and combinatorial effects of variants within this locus.

rs2280742

Rs2280742 overlaps distal enhancer cCRE EH38E3291349 which lies within 5 kb of the *KLF1* TSS. EH38E3291349 is classified as an enhancer in K562 and myeloid-derived biosamples such as CD14⁺ monocytes and the promyeloblast cell line HL-60. It is also classified as an enhancer in a small subset of tissue biosamples, including heart, lung, and muscle (**Supplementary Figure 22c, Supplementary Table 17j**). We observed allele specific chromatin accessibility in erythroblasts ($p = 0.02$, **Supplementary Table 17g**), further confirming that this enhancer is also active in the erythroid lineage and that we observe an allelic skew at this variant. Rs2280742 overlaps 102 transcription factor ChIP-seq peaks in K562 lying directly adjacent to the binding sites of E-box family transcription factors including NFE2, MAFK, MAFG, JUN, JUND, and FOSL1 (**Supplementary Figure 22d, Supplementary Table 17h**). Previous work also identified allelic imbalance of transcription factor ChIP-seq reads at this locus¹¹¹. Interestingly, EH38E3291349 is also a MAFF/MAFK+ cCRE (**Supplementary Note 6**) with evidence of MAFF and MAFK binding in both K562 and HepG2 but only enhancer signatures in K562; in HepG2 it has low chromatin accessibility. This may indicate that this enhancer is only active in specific cell states or conditions, which may have interesting implications for functional activity and should be studied in future testing.

rs2290688

Rs2290688 is a synonymous variant located within the seventh exon of the *MAST1* gene. It also overlaps the CA-CTCF cCRE EH38E3291318, a CTCF-mediated chromatin anchor that is active across nearly all surveyed cell types. This cCRE overlaps CTCF ChIA-PET loops in 100% of biosamples, Hi-C loops in 82%, and shows strong CTCF ChIP-seq signal in 99% of biosamples (**Supplementary Table 17k-m**). Although rs2290688 does not disrupt the core CTCF motif, it overlaps ChIP-seq peaks for 88 transcription factors in K562 (**Supplementary Table 17h**) and exhibits allele-specific binding for seven transcription factors in GM12878 (**Supplementary Table 17i**), suggesting potential allele-specific regulatory effects. We also

observed allele-specific chromatin accessibility in CD8⁺ T cells ($p = 0.02$, **Supplementary Table 17g**), although this may be less relevant given the limited overlap with erythroid-related cell types.

The Engreitz lab tested CA-CTCF cCRE EH38E3291318 in seven CRISPRi-FlowFISH experiments, measuring the readout of *JUNB*, *KLF1*, *LYL1*, *PRDX2*, *RAD23A*, *RNASEH2A*, and *WDR83OS*, two of which, *KLF1* and *PRDX2* had significant results. As previously mentioned in **Supplementary Note 4.1**, targeting EH38E3291318 resulted in a significant increase in expression of *PRDX2* (effect size = -0.03 , $p = 4.1 \times 10^{-4}$) where we hypothesized that inhibiting EH38E3291318 disrupted 3D chromatin interactions between the silencer cCRE EH38E4193243 and *PRDX2*. Targeting EH38E3291318 also resulted in the decreased expression of *KLF1* (effect size = 0.04 , $p = 6.7 \times 10^{-6}$). This duality in effect sizes aligns with the notion that disrupting EH38E3291318 interferes with the 3D chromatin architecture, rewiring enhancer-promoter interactions, thereby leading to contrasting regulatory outcomes for genes within the affected genomic neighborhood.

Other variants and caveats

While rs2072597, rs2280742, and rs2290688 emerged as the top candidates in our integrated ranking framework, other variants within this haplotype may also contribute to the regulation of *KLF1* and related genes. Several lower-ranked variants showed evidence of allele-specific chromatin accessibility or transcription factor binding, along with links to *KLF1*, suggesting potential functional roles that warrant further investigation. The regulatory architecture at this locus is highly complex, featuring dense clusters of cCREs, multiple 3D chromatin interactions, and widespread linkage with expression changes across numerous genes and tissues. Indeed, RBC-associated variants within this region are linked to expression quantitative trait loci (eQTLs) for at least 11 genes in 17 distinct tissues (**Supplementary Table 17n**), underscoring the challenge of pinpointing causal mechanisms. Moreover, this haplotype has a relatively high allele frequency in global populations—ranging from 0.68 in Europeans to 0.39 in East Asians (**Supplementary Table 17a**)—which may reflect a modest but widespread regulatory effect. These caveats should be carefully considered when designing future validation experiments, as

combinatorial perturbations or assays capable of detecting subtle expression changes may be necessary to fully resolve the regulatory contributions of individual variants.

Supplementary Note 9 - Additional Acknowledgements

9.1 Members of the ENCODE4 cCRE working group:

Greg Andrews, Alex Barrera, Mike Beer, Brad Bernstein, Beatrice Borsari, Stephanie Calluori, You Chen, Alan Du, Shaimae Elhajjajy, Jesse Engreitz, Chuck Epstein, Kaili Fan, Weixiang Fang, Adam Frankish, Adam Frankish, Yu Fu, Idan Gabdank, Mingshi Gao, Mark Gerstein, Daniel Gilchrist, Will Greenleaf, Gamze Gursoy, Ben Hitz, Hongkai Ji, Graham Johnson, Brendan Kearney, Anshul Kundaje, Lazaros Lataniotis, Christina Leslie, John Lis, Mats Ljungman, Brian Magnuson, Georgi Marinov, Ariel McShane, Eyal Metzl Raz, Wouter Meuleman, Jill Moore, Stephanie Morris, Ali Mortazavi, Kristy Mualim, Abdullah Ozer, Mike Pazin, Henry Pratt, Soumya Raychaudhuri, Milad Razavi-Mohseni, Thomas Reimonn, Xingjie Ren, Narges Rezaie, Marina Ruiz-Romero, Richard Sandstrom, Yin Shen, Mike Snyder, John Stam, Yi Wang, Eve Wattenberg, Annika Weimer, Zhiping Weng, Kevin White, Shannon White, Jin Woo Oh, Wang Xi, Jinrui Xu, Li Yao, Ingrid Youngworth, Haiyuan Yu, Michelle Yu, Ken Zaret, Junke Zhang, Lina Zheng

9.2 Members of the ENCODE4 TSS annotation working group:

Greg Andrews, Gabby Balderrama, Silvia Carbonell, Kelly Cochran, Daniel Kim, Anshul Kundaje, Julien Lagarde, Ishita Mangla, Jill Moore, Ali Mortazavi, Vivek Ramalingam, Milad Razavi-Mohseni, Liz Rebboah, Fairlie Reese, Jacob Schreiber, Laksshman Sundaram, Joshua Theisen, Jacob Tome, Diane Trout, Li Yao, Ingrid Youngworth

9.3 Members of the ENCODE4 Biosamples working group:

Ali Mortazavi, Annika Weimer, Ariel McShane, Aviva Presser Aiden, Barbara Wold, Beatrice Borsari, Ben Hitz, Bill Noble, Brad Bernstein, Briana Nuñez, Brian St. Hilaire, Brian Williams, Carrie Davis, Chad Nusbaum, Chuck Epstein, Cricket Sloan, Dan Gilchrist, Eileen Cahill, Elise Feingold, Erez Lieberman Aiden, Eric Mendenhall, Eva Gega, Fabio Navarro, Gabriela Balderrama Gutierrez, Haiyuan Yu, Huiya Gu, Idan Gabdank, Ishwarya Venkata, Ivan Bochkov, Jason Hilton, Jennifer Jou, Jenn Wineski, Jess Halow, Jill Moore, Jing Zhang, Joel Rozowsky,

John Stamatoyannopoulos, Karan Bedi, Kathy Rader, Konor von Kraut, Laura Reinholdt, Liz Gaskell, Mark Gerstein, Mark Mackiewicz, Mats Ljungman, Meenakshi, Mike Pazin, Mike Snyder, Minji Kim, Neva Durand, Nina Farrell, Patrick Gallagher, Ping Wang, Raj Kaul, Rick Myers, Rob Spitale, Sam Moore, Saurabh Agarwal, Shin Lin, Stephanie Morris, Tim Reddy, Ulrike Litzenburger, Weiwei Zhong, Will Greenleaf, Yiing Lin, Yijun Ruan, Yunhai Luo, Zack Myers, Zhiping Weng

Supplementary Methods

Estimating the number of mouse enhancer cCREs

To estimate the upper bound of the number of possible “missing” human and mouse enhancer cCREs, we subtracted the number of TSS distal CA cCREs annotated in biosamples with both DNase and H3K27ac data from the total number of annotated distal CA cCREs. This left only CA cCREs in biosamples lacking H3K27ac data which have the potential to be classified as enhancer cCREs if H3K27ac were profiled in those biosamples.

Proximity of experimentally derived transcription start sites with rDHSs

We downloaded FANTOM CAGE¹¹² peaks from the [Fantom Database](#) and calculated their distance to the nearest rDHS using BEDTools closest⁵⁵. We then calculated the percentage of CAGE peaks within 250 bp of an rDHS (**Supplementary Table 2a**). We performed the same analysis for RAMPAGE rPeaks¹⁰⁶ that were associated with GENCODE genes.

Overlap of histone mark peaks with rDHSs

We downloaded peaks from the ENCODE portal for all experiments targeting H3K4me3, H3K27ac, and H3K4me1 (**Supplementary Table 2b-d**). We intersected these peaks with rDHSs using BEDTools intersect⁵⁵.

Overlapping transcription factor peaks with rDHSs

We downloaded peaks for all transcription factor experiments from the ENCODE portal that met the following criteria (**Supplementary Table 2e**):

- Transcription factor was sequence specific, i.e., has an annotated motif in HOCOMOCOv11⁵⁷
- FRiP score (fraction of reads within peaks) > 0.3
- Peak files were labeled as “Preferred default peaks”

We intersected each of these 535 peak sets with rDHSs using BEDtools⁵⁵ and calculated the percentage of total overlapping peaks. We selected all peak sets with less than 90 percent overlap for additional analysis (N = 50).

Identifying and characterizing transcription factors with low overlap with rDHSs

For the 50 experiments, we selected all peaks that did not overlap rDHSs. Then, for each pair of experiments, we calculated the overlap coefficient between the non-rDHS peaks. We then performed hierarchical clustering using the default parameters of hclust to identify common sets of transcription factors.

For groups with at least three thousand sites, which included the MAFK/MAFF, ATF/CREB1 and CEBPB clusters, we calculated the average PhastCons 100-way vertebrate conservation scores across each region. We compared scores to all rDHSs, TF anchors, and non-cCRE regions. We calculated statistical significance using a Wilcox test.

Calling Multi-mapping cCREs

We called DHSs with the Hotspot2 peak caller¹¹³ using unfiltered BAM alignment files from the ENCODE portal (**Supplementary Table 3c**). Instead of calling DHSs genome wide, we called DHSs on a subset of the genome centered on high confidence ATAC-seq peaks that did not overlap rDHSs (**Supplementary Table 3a,b**). Using our rDHS pipeline, clustered the multi-mapping DHSs and identified representative elements which we called multi-mapping cCREs.

Generating multi-level cCRE annotations

To generate tissue-, life stage-, and disease-specific cCRE annotations, we developed a pipeline that builds upon our previously described consensus approach for brain-specific cCREs¹¹⁴. For each tissue group, we began with a curated list of biosamples with available

DNase-seq data. We identified a consensus set of anchor cCREs by selecting cCREs that exhibited DNase Z-scores > 1.64 (one-sided $p < 0.05$) in at least 5 biosamples, or passed more stringent thresholds ($Z > 2.02$ in ≥ 2 biosamples or $Z > 2.32$ in ≥ 1 biosample). These Z-scores correspond to approximately the 98th and 99th percentiles, respectively, and were chosen to retain both broadly accessible and strong biosample-specific elements.

We then computed the maximum Z-score across biosamples for each mark per anchor cCRE. Elements were then classified using the standard Registry cCRE criteria based on these maximum values and overlap with transcription factor ChIP-seq peak summits

Characterizing the expansion of the Registry

We stratified ENCODE4 cCREs into three categories:

1. Replicated from ENCODE3. These are cCREs that are present in both versions of the Registry. These cCREs can either have the same exact position, in which case their accession remains the same, or their boundaries can slightly shift.
2. Called in ENCODE3 but filtered out due to incomplete data. We previously filtered out cCREs deriving from biosamples lacking certain combinations of data (see Supplementary Note 1 from ⁹). However, with our expansion to include new classes of cCREs, we now include these elements in our Registry.
3. New in ENCODE4 are cCREs never previously called.

We characterized these three groups by intersecting them with whole genome STARR-seq peaks from K562 (ENCFF908UFR) and calculating the average per-base conservation across each cCRE using 240-way Mammalian phyloP scores from the Zoonomia consortium.

To identify which biosamples contributed most to the increased Registry, we calculated chromatin accessibility enrichment for groups 2 & 3 (called in ENCODE3 and new in ENCODE4). In each biosample, we determined the fraction of new V4 cCREs with high signal and compared it to the fraction of all cCREs with high signal.

Overlap with BICCN cCREs

We downloaded BICCN cCRE annotations from Li *et al.*⁶⁹ via catlas.org, which contained a consensus set of elements and elements called for individual cell types (N = 107). Using BEDTools, we intersected each peak set with three versions of cCREs: ENCODE3 cCREs (V2, 2020), a preliminary version of ENCODE4 cCREs (V3, 2022), and our updated ENCODE4 cCREs (V4, 2024). We then calculated the percentage of peaks that overlapped cCREs for each version.

To identify which cCREs contributed to the increased overlap, we calculated chromatin accessibility enrichment for newly added V4 cCREs that overlapped BICCN cCREs. For each biosample, we determined the fraction of overlapping V4 cCREs with high signal and compared it to the fraction of all cCREs with high signal. Biosamples were ranked based on enrichment and stratified by their tissue or organ of origin.

Overlap with Loupe *et al.* union peaks

We downloaded [Supplementary Table 4](#) from Loupe *et al.*,⁷⁰ which contains the union of peaks called for individual brain regions and cell types. For each brain region and cell type combination (N = 12), we generated BED files of the peak locations. Using BEDTools, we intersected each peak set with three versions of cCREs: ENCODE3 cCREs (V2, 2020), a preliminary version of ENCODE4 cCREs (V3, 2022), and our updated ENCODE4 cCREs (V4, 2024). We then calculated the percentage of peaks that overlapped with cCREs for each version.

To identify which cCREs contributed to the increased overlap, we focused on the union peaks from Dorsolateral Prefrontal Cortex (DLPFC) NeuN+ cells. Specifically, we calculated chromatin accessibility enrichment for newly added V4 cCREs that overlapped these union peaks. For each biosample, we determined the fraction of overlapping V4 cCREs with high signal and compared it to the fraction of all cCREs with high signal. Biosamples were ranked based on enrichment and stratified by their tissue or organ of origin.

Overlap with other annotated CREs

Similarly we downloaded acCREs from Liu *et al.*⁷¹ and tCREs from Zhang *et al.*⁷² and using BEDTools, intersected each peak set with three versions of cCREs: ENCODE3 cCREs (V2, 2020), a preliminary version of ENCODE4 cCREs (V3, 2022), and our updated ENCODE4 cCREs (V4, 2024). We then calculated the percentage of peaks that overlapped with cCREs for each version.

We then performed the same enrichment analysis as described above for the V4 cCREs.

Classification of STARR+ distal enhancers

We classified STARR+ distal enhancers specific to K562 and HepG2 cells based on their biochemical signatures as defined in the cCRE workflow. Each STARR+ region was assigned to one of the following cCRE classes: distal enhancer (dELS), CA-H3K4me3, CA-CTCF, CA-TF, CA, or low chromatin accessibility. For each cell type, we calculated the fraction of STARR+ cCREs in each class.

To assess differences in cCRE classifications between K562 and HepG2, we used a two-sided Fisher's exact test. This analysis was performed for all K562 and HepG2 STARR+ distal enhancers and for subsets of these enhancers that overlapped motifs for transcription factors GATA1, HNF4A, p53, and GFI1B.

Enrichment of transcription factor motifs in cCRE pairs

To look for motif enrichment in negative combinatorial pairs of cCREs. We selected all cCRE pairs meeting the following requirements (N = 1,275):

- One cCRE of the pair had independent STARR activity with a positive solo STARR score and $p < 0.05$
- The other cCRE of the pair had a negative solo STARR score
- The double STARR score of the pair was negative

We compared the overlap of motifs between cCREs with the positive and negative solo STARR scores, considering all transcription factors expressed in K562 (TPM > 5, N = 181) and using a two-sided Fisher's Exact test with FDR correction.

Gene expression near candidate silencers

For each group of cCREs, we identified the closest protein coding gene, as defined by linear distance to the transcription start site, using BEDTools. We then calculated expression of these genes in transcripts per million using K562 total RNA-seq data from the ENCODE portal (ENCFF421TJX). We performed two-sided Wilcoxon tests to evaluate significance.

Enrichment for transcription factor binding site at STARR silencers

To evaluate transcription factor enrichment at STARR silencers, we calculated the enrichment of transcription factor motifs and ChIP-seq peak summits across four groups of cCREs:

- Stringent STARR silencers, negative log2 fold change and $p < 0.01$
- Robust STARR silencers, negative log2 fold change and $p < 0.05$
- Active cCREs, positive log2 fold change and $p < 0.01$
- Neutral cCREs, $-0.1 < \log_2 \text{fold change} < 0.1$

For motif analysis, we quantified the number of cCREs in each group overlapping transcription factor motifs identified by FIMO (see **Methods**). We assessed statistical significance using a two-sided Fisher's exact test with FDR correction. For ChIP-seq data, we intersected peak summits with cCREs using BEDTools, calculating the number of cCREs in each group overlapping each TF summit, and determined significance using a two-sided Fisher's exact test with FDR correction.

For transcription factors significantly enriched in STARR silencers ($\text{FDR} < 0.05$ for all comparisons), we identified overlapping cCREs and computed pairwise Overlap Coefficients (see **Supplementary Methods**). These coefficients were used for hierarchical clustering (via hclust) to identify transcription factors with common binding sites.

Identifying allele-specific open chromatin

To identify allele-specific open chromatin regions, we analyzed ATAC-seq data from two publicly available datasets: Calderon *et al.* (GSE118189)¹¹⁵ and Corces *et al.* (GSE74912)¹¹⁶. All

samples were processed using the ENCODE ATAC-seq pipeline to obtain filtered and aligned BAM files.

For each variant of interest, we extracted reads overlapping the variant site using samtools. For each donor, we pooled BAM files across technical replicates. BAM files from the same cell type and study were merged using samtools merge and indexed with samtools index.

We used a custom Python script to count base calls at the variant position in the merged BAM files. We aggregated base-level read counts by donor and filtered for heterozygous donors, defined as those with >10% of total reads for each of the two most abundant alleles at the position.

For each cell type, we then pooled allele-specific read counts across heterozygous donors. To assess allelic imbalance, we performed a two-sided Fisher's exact test comparing observed allele counts to the null expectation of a 50:50 ratio. We restricted the analysis to cases with at least 25 total reads across the two major alleles.

Calling allele-specific transcription factor binding in GM12878

We used Alleleseq2¹¹⁷ (<https://github.com/trgaleev/AlleleSeq2>) to perform allele-specific transcription factor binding in GM12878. First, we built the personal genome using makePersonalGenome.mk, with the unphased NA12878 genotype VCF and BAM files as input. For each individual ChIP-seq experiment, we use the fastq files of all available replicates to run PIPELINE.mk to identify ASB events with the following parameters: ALIGNMENT_MODE=ASB RM_DUPLICATE_READS=on FDR_CUTOFF=0.10. For experiments with more than 1 replicates, PIPELINE_aggregated_counts.mk was used to pool counts across replicates. To identify AS chromatin accessibility, we used the following parameter: ALIGNMENT_MODE=ASCA RM_DUPLICATE_READS=on FDR_CUTOFF=0.10.

Briefly, Alleleseq2 quantified the number of reads overlapping hetSNVs and tested the significance of allelic imbalance using only hetSNVs that had ≥ 8 reads with beta-binomial distribution. A simulation-based FDR was generated for each experiment as previously

described (167). For each experiment, we examined the reference allelic ratio distribution for all hetSNVs with ≥ 8 reads. Two experiments targeting IRF5 and E4F1 respectively were manually excluded from further analysis because their allelic ratio (number of reads aligned to reference allele / number of reads aligned to both alleles) distributions had poor fit for beta-binomial distribution.

Phylogenetic analysis of *KLF1* missense mutation

We downloaded the 241-way mammalian multiple sequence alignment and phylogenetic tree from the Zoonomia consortium^{56,118,119}. We next used hal2maf from the Comparative Genomics Toolkit¹²⁰ to extract the alignment at the *KLF1* missense mutation. We then parsed the resulting MAF file and converted each codon to its corresponding amino acid. Finally, we plotted the phylogenetic tree colored by amino acid in R using the ggtree library¹²¹.

Supplementary References

64. Vierstra, J., Meuleman, W. & Stamatoyannopoulos, J. A comprehensive catalogue of accessible DNA elements encoded by the human genome.
65. Vierstra, J. & Stamatoyannopoulos, J. The accessible DNA landscape of the mouse genome.
66. Acharya, D. & Ghosh, T. C. Global analysis of human duplicated genes reveals the relative importance of whole-genome duplicates originated in the early vertebrate evolution. *BMC Genomics* **17**, 71 (2016).
67. Karimzadeh, M., Ernst, C., Kundaje, A. & Hoffman, M. M. Umap and Bimap: quantifying genome and methylome mappability. *Nucleic Acids Res.* **46**, e120 (2018).
68. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
69. Li, Y. E. *et al.* A comparative atlas of single-cell chromatin accessibility in the human brain. *Science* **382**, eadf7044 (2023).
70. Loupe, J. M. *et al.* Multiomic profiling of transcription factor binding and function in human brain. *Nat. Neurosci.* **27**, 1387–1399 (2024).
71. Liu, B. B. *et al.* Dissecting regulatory syntax in human development with scalable

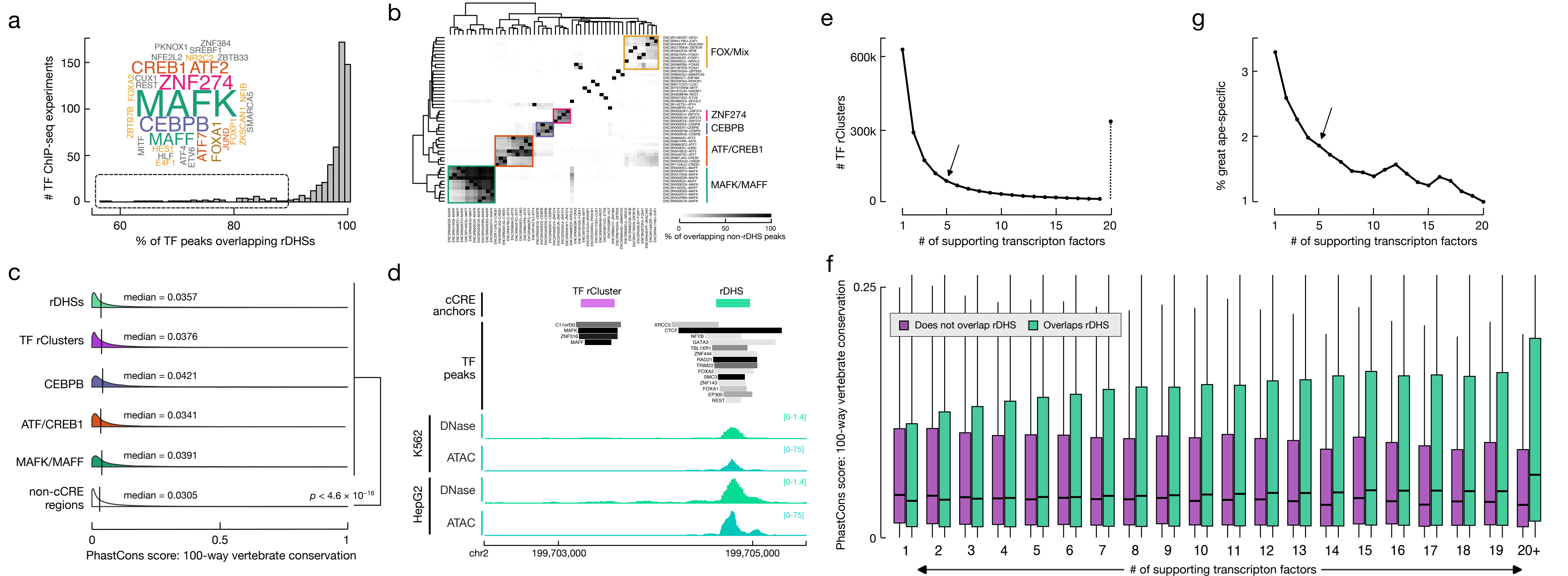
- multiomics and deep learning. *Genomics* (2025).
72. Moody, J. *et al.* A single-cell atlas of transcribed cis -regulatory elements in the human genome. *Genomics* (2023).
 73. Landolin, J. M. *et al.* Sequence features that drive human promoter function and tissue specificity. *Genome Res.* **20**, 890–898 (2010).
 74. Hasegawa, Y. & Struhl, K. Different SP1 binding dynamics at individual genomic loci in human cells. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2113579118 (2021).
 75. Sabò, A. & Amati, B. Genome recognition by MYC. *Cold Spring Harb. Perspect. Med.* **4**, a014191–a014191 (2014).
 76. Liu, K. *et al.* Molecular mechanism of specific DNA sequence recognition by NRF1. *Nucleic Acids Res.* **52**, 953–966 (2024).
 77. Attwooll, C., Lazzerini Denchi, E. & Helin, K. The E2F family: specific functions and overlapping interests. *EMBO J.* **23**, 4709–4716 (2004).
 78. Driscoll, M. E., Bollu, P. C. & Tadi, P. Neuroanatomy, nucleus caudate. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2025).
 79. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**, 159–164 (2015).
 80. Liu, W. *et al.* CUX1 regulates human hematopoietic stem cell chromatin accessibility via the BAF complex. *Cell Rep.* **43**, 114227 (2024).
 81. Wortel, I. M. N., van der Meer, L. T., Kilberg, M. S. & van Leeuwen, F. N. Surviving stress: Modulation of ATF4-mediated stress responses in normal and malignant cells. *Trends Endocrinol. Metab.* **28**, 794–806 (2017).
 82. Huggins, C. J. *et al.* C/EBP γ is a critical regulator of cellular stress response networks through heterodimerization with ATF4. *Mol. Cell. Biol.* **36**, 693–713 (2015).
 83. Renfro, Z., White, B. E. & Stephens, K. E. CCAAT enhancer binding protein gamma (C/EBP- γ): An understudied transcription factor. *Adv. Biol. Regul.* **84**, 100861 (2022).
 84. Ren, Q. *et al.* C/EBP β : The structure, regulation, and its roles in inflammation-related diseases. *Biomed. Pharmacother.* **169**, 115938 (2023).
 85. Zimprich, A. *et al.* Serum response factor (SRF) ablation interferes with acute stress-associated immediate and long-term coping mechanisms. *Mol. Neurobiol.* **54**, 8242–8262 (2017).

86. Bernardo, V. S., Torres, F. F. & da Silva, D. G. H. FoxO3 and oxidative stress: a multifaceted role in cellular adaptation. *J. Mol. Med.* **101**, 83–99 (2023).
87. Kim, Y.-J. *et al.* The transcription factor Mef2d regulates B:T synapse-dependent GC-TFH differentiation and IL-21-mediated humoral immunity. *Sci. Immunol.* **8**, eadf2248 (2023).
88. Luo, A. *et al.* ZNF121 interacts with ZBRK1 and BRCA1 to regulate their target genes in mammary epithelial cells. *FEBS Open Bio* **8**, 1943–1952 (2018).
89. Vincent, T. *et al.* A SNAIL1-SMAD3/4 transcriptional repressor complex promotes TGF-beta mediated epithelial-mesenchymal transition. *Nat. Cell Biol.* **11**, 943–950 (2009).
90. Wang, H. & Matise, M. P. Tcf7l2/Tcf4 transcriptional repressor function requires HDAC activity in the developing vertebrate CNS. *PLoS One* **11**, e0163267 (2016).
91. Law, J. C., Ritke, M. K., Yalowich, J. C., Leder, G. H. & Ferrell, R. E. Mutational inactivation of the p53 gene in the human erythroid leukemic K562 cell line. *Leuk. Res.* **17**, 1045–1050 (1993).
92. Bressac, B. *et al.* Abnormal structure and expression of p53 gene in human hepatocellular carcinoma. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 1973–1977 (1990).
93. Liu, Y. & Bodmer, W. F. Analysis of P53 mutations and their expression in 56 colorectal cancer cell lines. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 976–981 (2006).
94. Lim, L. Y., Vidnovic, N., Ellisen, L. W. & Leong, C.-O. Mutant p53 mediates survival of breast cancer cells. *Br. J. Cancer* **101**, 1606–1612 (2009).
95. Rovira, M. *et al.* REST is a major negative regulator of endocrine differentiation during pancreas organogenesis. *Genes Dev.* **35**, 1229–1242 (2021).
96. Kim, J. G. & Hudson, L. D. Novel member of the zinc finger superfamily: A C2-HC finger that recognizes a glia-specific gene. *Mol. Cell. Biol.* **12**, 5632–5639 (1992).
97. MYT1 myelin transcription factor 1 [Homo sapiens (human)] - Gene - NCBI. <https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=DetailsSearch&Term=4661>.
98. LINGO1 leucine rich repeat and Ig domain containing 1 [Homo sapiens (human)] - Gene - NCBI. <https://www.ncbi.nlm.nih.gov/gene/84894>.
99. Lee, C.-C. *et al.* TCF12 protein functions as transcriptional repressor of E-cadherin, and its overexpression is correlated with metastasis of colorectal cancer. *J. Biol. Chem.* **287**, 2798–2809 (2012).
100. Postigo, A. A., Depp, J. L., Taylor, J. J. & Kroll, K. L. Regulation of Smad signaling through

- a differential recruitment of coactivators and corepressors by ZEB proteins. *EMBO J.* **22**, 2453–2462 (2003).
101. Ichiyama, K. *et al.* Transcription factor Ikzf1 associates with Foxp3 to repress gene expression in Treg cells and limit autoimmunity and anti-tumor immunity. *Immunity* **57**, 2043–2060.e10 (2024).
 102. Nwabo Kamdje, A. H. *et al.* Developmental pathways associated with cancer metastasis: Notch, Wnt, and Hedgehog. *Cancer Biol. Med.* **14**, 109–120 (2017).
 103. Katsuoka, F. *et al.* Genetic evidence that small maf proteins are essential for the activation of antioxidant response element-dependent genes. *Mol. Cell. Biol.* **25**, 8044–8051 (2005).
 104. Kannan, M. B., Solovieva, V. & Blank, V. The small MAF transcription factors MAFF, MAFK and MAFK: current knowledge and perspectives. *Biochim. Biophys. Acta* **1823**, 1841–1846 (2012).
 105. Rozowsky, J. *et al.* The EN-TEEx resource of multi-tissue personal epigenomes & variant-impact models. *Cell* **186**, 1493–1511.e40 (2023).
 106. Moore, J. E. *et al.* Integration of high-resolution promoter profiling assays reveals novel, cell type-specific transcription start sites across 115 human cell and tissue types. *Genome Res.* **32**, 389–402 (2022).
 107. Reese, F. *et al.* The ENCODE4 long-read RNA-seq collection reveals distinct classes of transcript structure diversity. *Genomics* (2023).
 108. Calvo-Roitberg, E., Daniels, R. F. & Pai, A. A. Challenges in identifying mRNA transcript starts and ends from long-read sequencing data. *Genome Res.* **34**, 1719–1734 (2024).
 109. *Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing.*
 110. Gschwind, A. R. *et al.* An encyclopedia of enhancer-gene regulatory interactions in the human genome. *bioRxiv* (2023) doi:10.1101/2023.11.09.563812.
 111. Downes, D. J. *et al.* An integrated platform to systematically identify causal variants and genes for polygenic human traits. *Genetics* (2019).
 112. Abugessaisa, I. *et al.* FANTOM5 CAGE profiles of human and mouse reprocessed for GRCh38 and GRCm38 genome assemblies. *Sci Data* **4**, 170107 (2017).
 113. *hotspot2: Implementation of hotspot2 by Eric Rynes.* (Github).
 114. Pratt, H. E. *et al.* Using a comprehensive atlas and predictive models to reveal the complexity and evolution of brain-active regulatory elements. *Sci Adv* **10**, eadj4452 (2024).

115. Calderon, D. *et al.* Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nat. Genet.* **51**, 1494–1505 (2019).
116. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
117. Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.* **7**, 522 (2011).
118. Zoonomia Consortium. A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020).
119. Sullivan, P. F. *et al.* Leveraging base-pair mammalian constraint to understand genetic variation and human disease. *Science* **380**, eabn2937 (2023).
120. *ComparativeGenomicsToolkit*. (Github).
121. Yu, G. Using ggtree to Visualize Data on Tree-Like Structures. *Curr. Protoc. Bioinformatics* **69**, e96 (2020).

Supplementary Figures



Supplementary Figure 1 | Expanding cCRE anchors to include transcription factor binding sites

a, Histogram depicting the percentage of transcription factor (TF) peaks overlapping rDHSs. TFs with less than 90% overlap constituted a distinct set of TFs (depicted in the word cloud) such as MAFK, CEBPB, and ZNF274. Size of TF name is proportional to the number of TF experiments with less than 90% overlap. Full results can be found in **Supplementary Table 2e**.

b, Heatmap depicting the overlap coefficient between TF peaks, which do not overlap rDHSs, between experiments with low rDHS overlap as defined in the dashed box in **a**. Five primary TF experiment clusters were defined using hierarchical clustering.

c, Density plots depicting the average 100-way vertebrate alignment PhastCons conservation score for rDHSs (green) and proposed TF rCluster anchors (purple). Three sub-categories of TF rClusters, as defined in **b**, are shown including CEBPB sites (blue), ATF/CREB1 sites (orange), and MAFK/MAFF sites (teal). Non-cCRE regions are shown in white. All groups were significantly more conserved than non-cCRE regions (pairwise, two-sided Wilcoxon rank sum test with FDR correction, $p < 4.6 \times 10^{-16}$), supporting their inclusion in the Registry of cCREs.

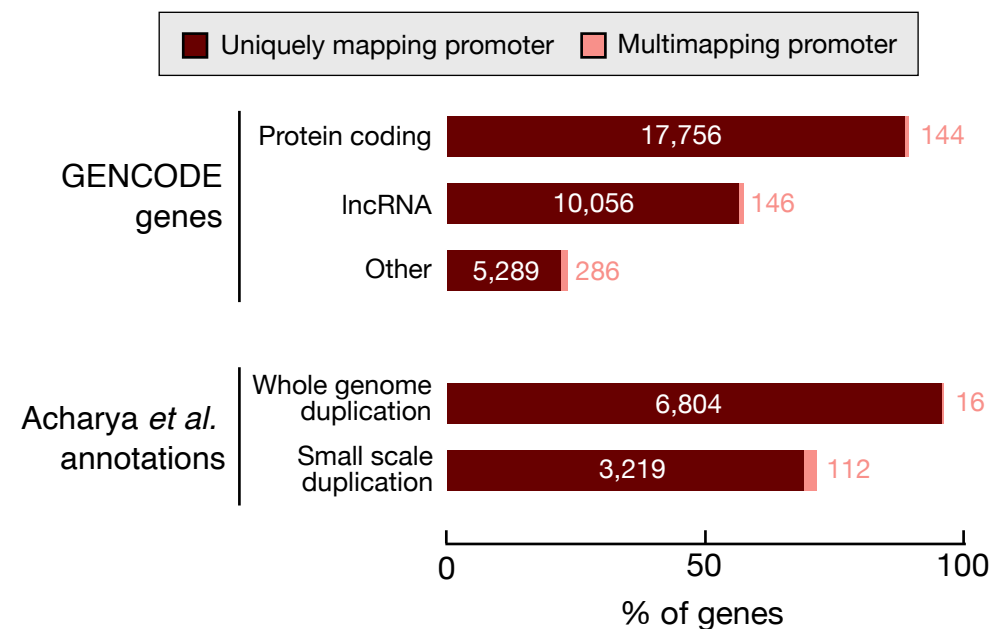
d, Example of rDHS (green) and TF rCluster (purple). The rDHS has high chromatin accessibility in K562 and HepG2 as measured by DNase-seq and ATAC-seq data. The TF rCluster has low chromatin accessibility. Both sites overlap TF ChIP-seq peaks with the TF rCluster overlapping peaks for MAFK, MAFF, ZNF316, and C11orf30.

e, Line plot depicting the number TF rClusters we identify stratified by the number of transcription factors supporting that cluster. Arrow denotes our threshold ≥ 5 .

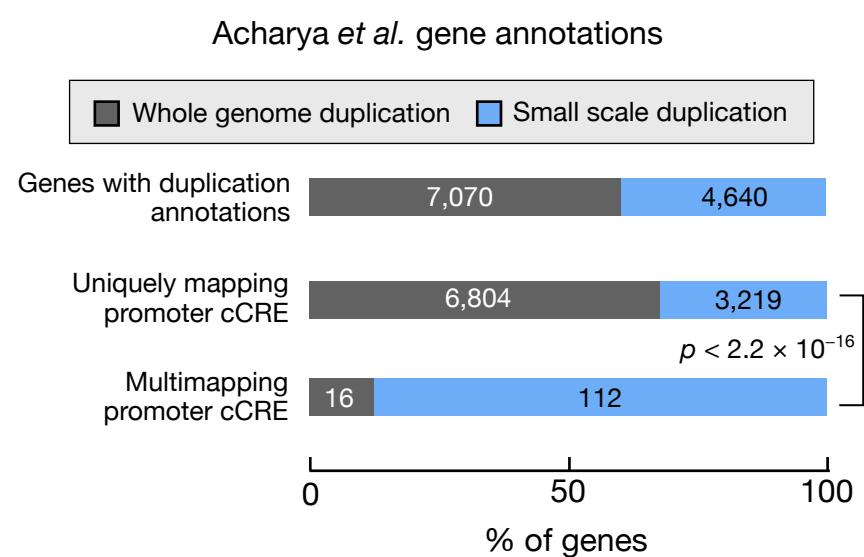
f, Boxplots depicting average PhastCons 100-way vertebrate conservation for TF rClusters stratified by (1) number of supporting transcription factors (x-axis) and (2) overlap with rDHSs (green for overlapping, purple for non-overlapping).

g, Line plot depicting the percent of TF rClusters in the great-ape specific group as defined by Zoonomia conservation stratified by the number of transcription factors supporting that cluster. Arrow denotes our threshold ≥ 5 .

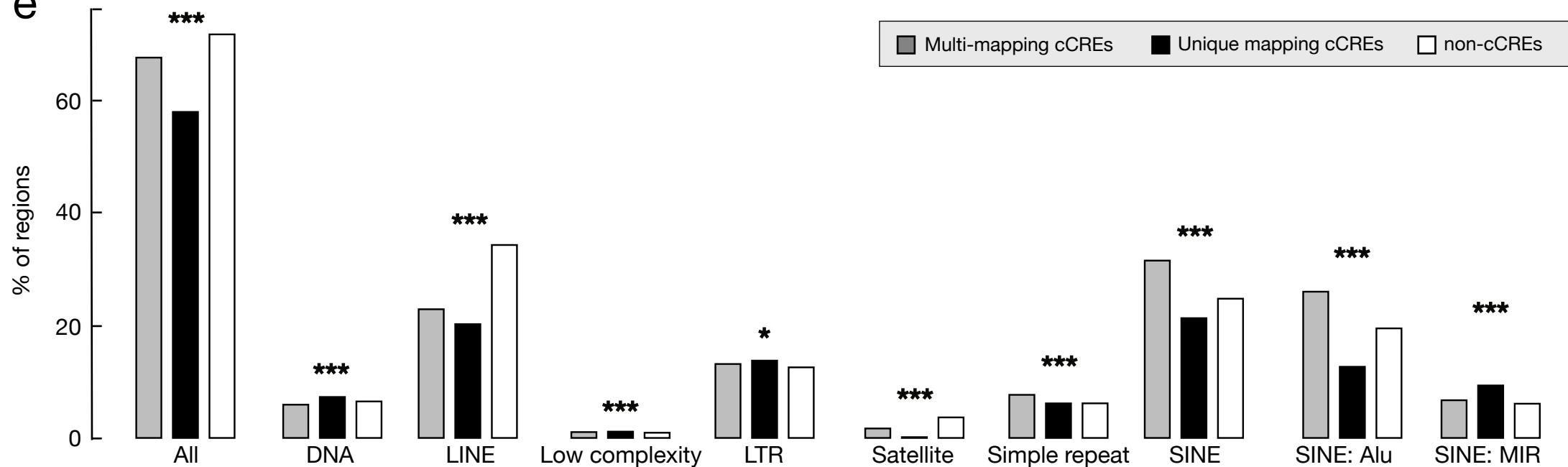
a



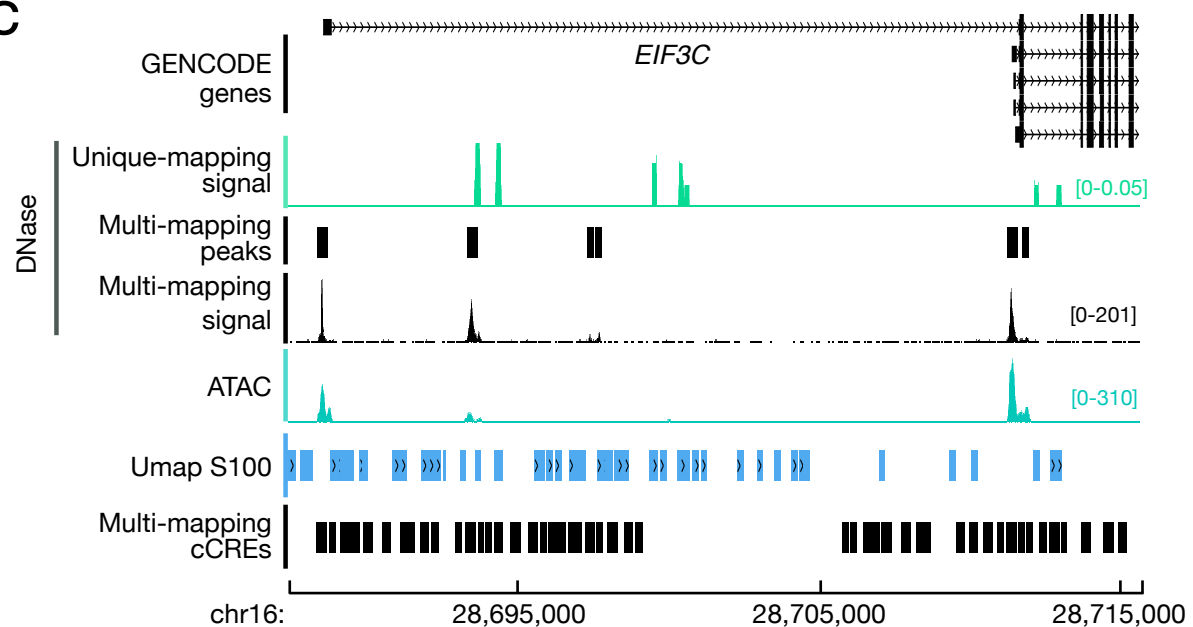
b



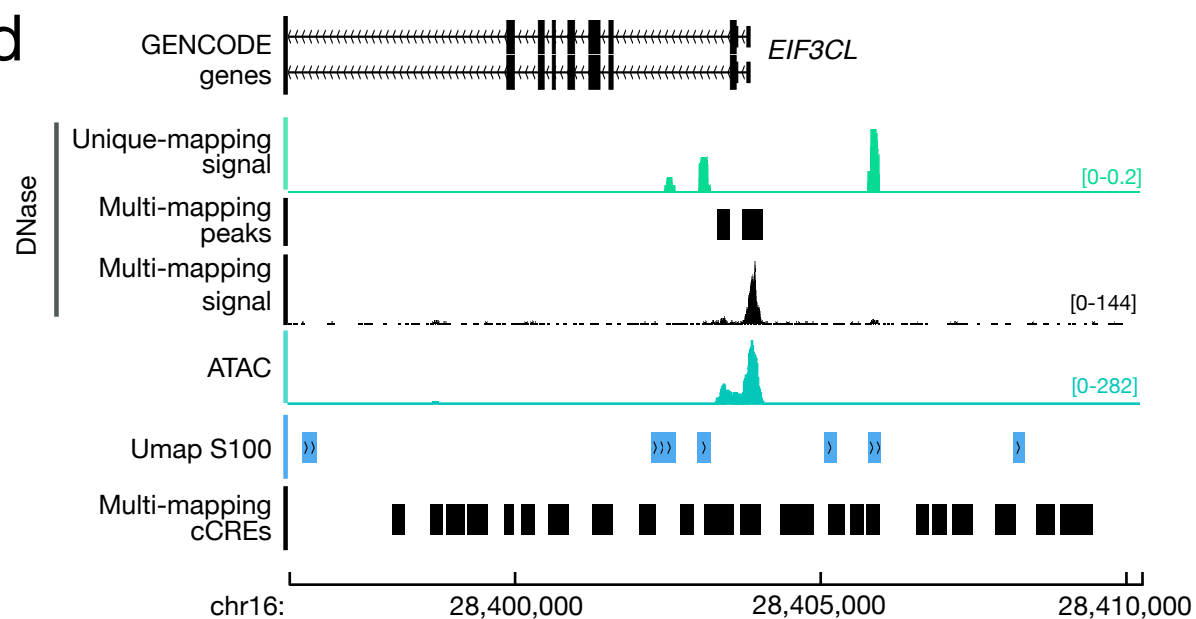
e



C

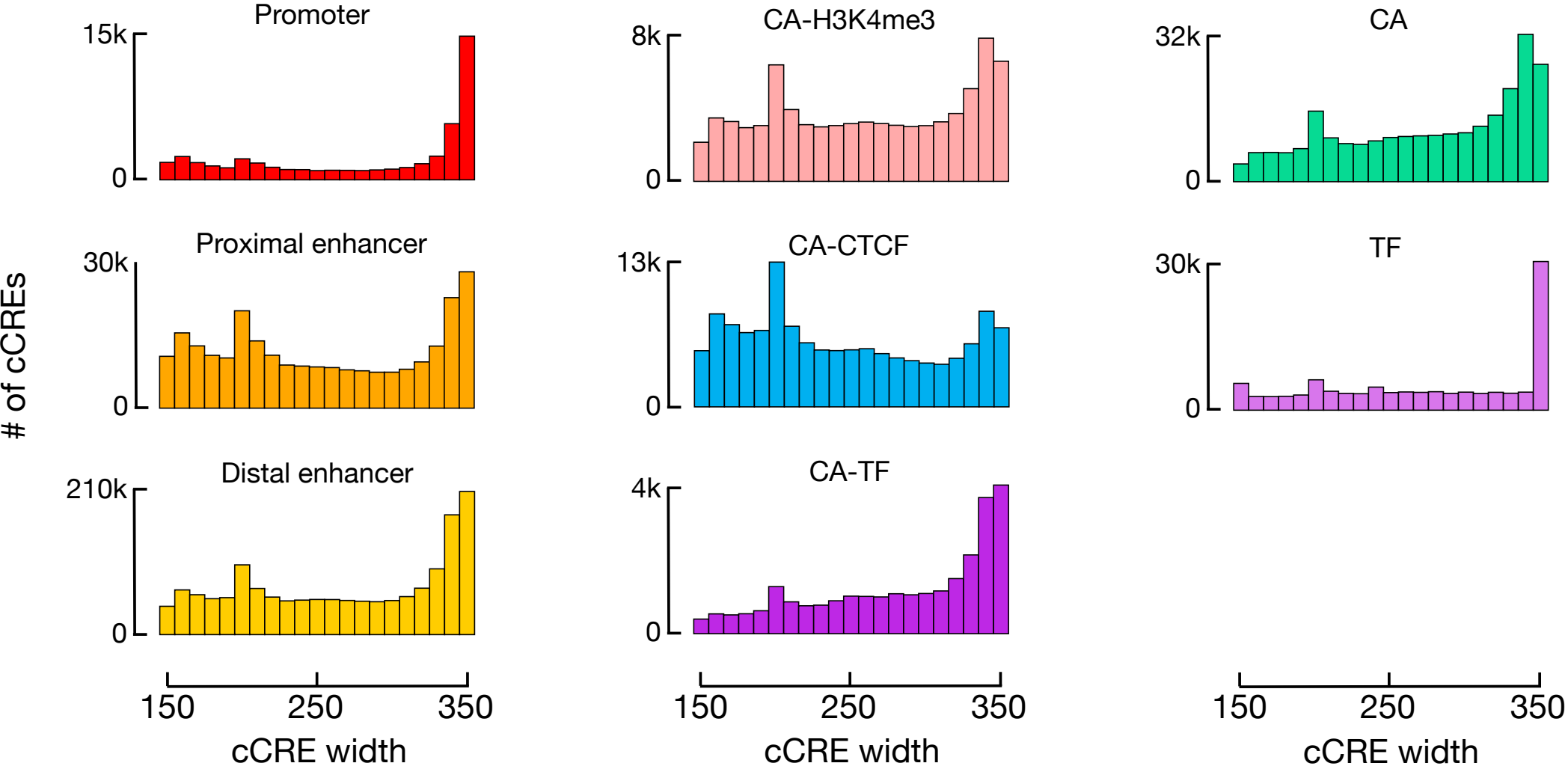


d



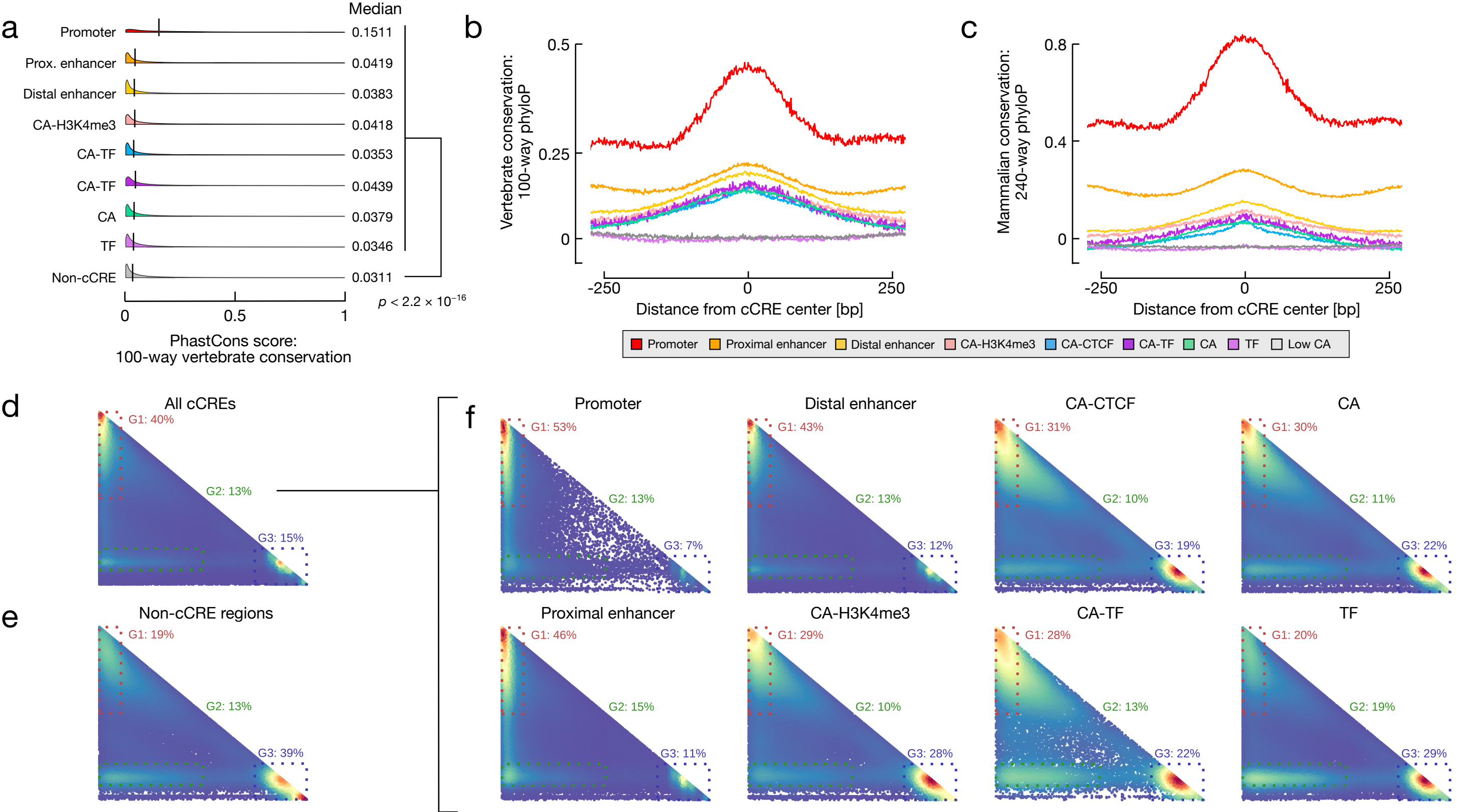
Supplementary Figure 2 | Including cCREs at multi-mapping loci identifies promoters for recently duplicated genes

a, Barplots depicting the percentage of genes with annotated promoter cCREs. Genes are stratified based on GENCODE annotations (top) or classification by Archarya *et al.*⁶⁶. Promoter cCREs are stratified by whether they arose from uniquely mapping cCREs (dark red) or multi-mapping cCREs. **b**, Barplots showing the proportion of genes originating from whole genome duplication (gray) or more recent small scale duplication (blue) as annotated by Archarya *et al.* All annotated genes are shown (top) along with genes with uniquely mapping promoters (middle) and genes with multi-mapping promoters (bottom). **c**, Genome browser view of the *EIF3C* locus with the default unique-mapping DNase (green) and ATAC (teal) tracks. Multi-mapping DNase-seq signal is shown in black along with DHSs called from this signal. Umap S100 mappability tracks are shown in blue and multi-mapping cCREs are shown in black. **d**, Genome browser shot of the *EIF3CL* locus with the same tracks as defined in **c**. **e**, Barplots depicting the percentage of multi-mapping cCREs (gray), uniquely mapping cCREs (black), or non-cCRE regions overlap repetitive elements stratified by class. Stars indicate statistically significant differences between multi-mapping cCREs compared with unique mapping cCREs and non-cCRE regions based on two-sided Fisher's exact tests with FDR correction, *** for $p < 0.001$ and * for $p < 0.05$. Full table of p -values can be found in **Supplementary Table 3e**.



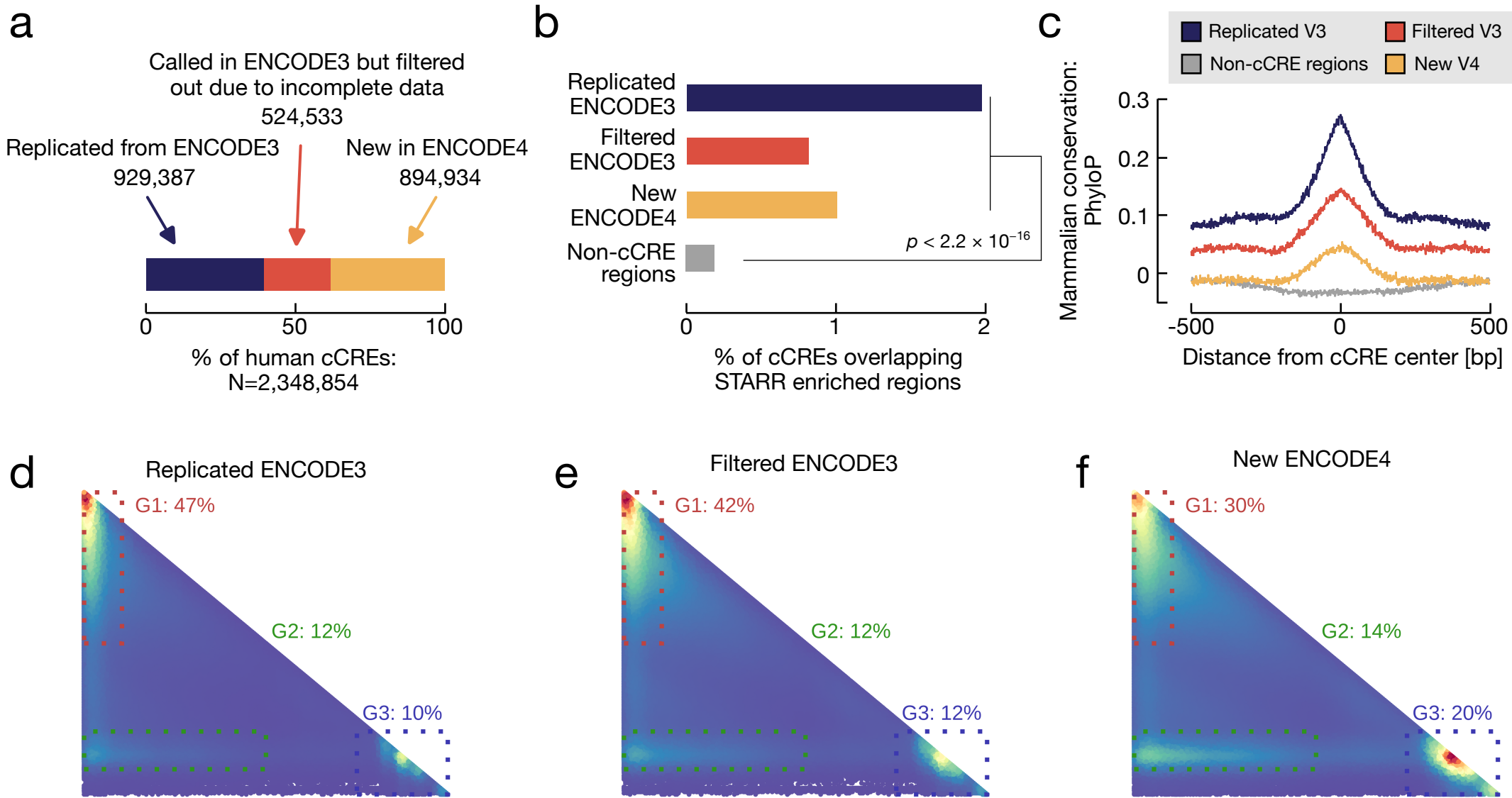
Supplementary Figure 3 | Size distribution of cCREs

Histograms depicting the distributions of cCRE width stratified by cCRE class. All cCREs range in width from 150 to 350 bp.



Supplementary Figure 4 | Evolutionary conservation of cCREs

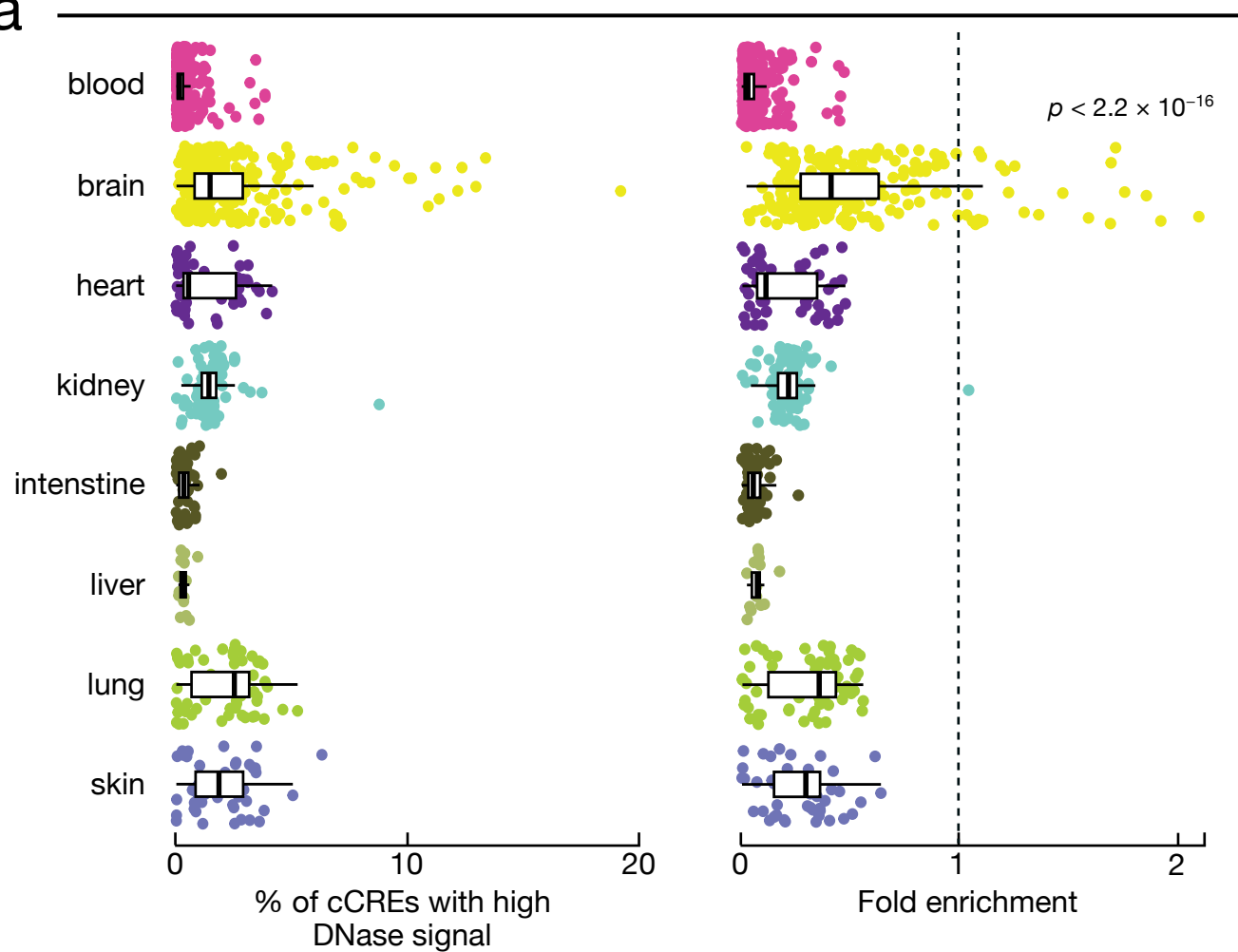
a, Density plots showing the distribution of PhastCons scores from the 100-way vertebrate alignment, stratified by cCRE class. Vertical bars indicate median values, which are also listed at right. P-value corresponds to pairwise, two-sided Wilcoxon test with FDR correction. **b**, Line plots of average phyloP conservation scores (100-way vertebrate alignment) $\pm$ 250 bp around the center of each cCRE, stratified by cCRE class. **c**, As in **b**, but using phyloP scores from the 240-way mammalian alignment. **d**, Heatscatter plot showing the distribution of human ENCODE4 cCREs based on their alignment across Zoonomia mammalian genomes. N1 denotes the number of species in which $\geq 90\%$ of a cCRE aligns; N2 denotes the number of species in which $\leq 10\%$ of a cCRE aligns. Rectangles indicate predefined groups: Group 1 (G1), highly conserved ($N1 \geq 120$, $N2 \leq 25$); Group 2 (G2), actively evolving ($20 \leq N1 \leq 50$, $N2 \leq 120$); and Group 3 (G3), primate-specific ($N1 \leq 50$, $N2 \geq 180$). The percentage of cCREs in each group is indicated. **e**, Heatscatter plot as in **d**, but for size-matched non-cCRE regions. **f**, Heatscatter plots as in **d**, stratified by cCRE class.



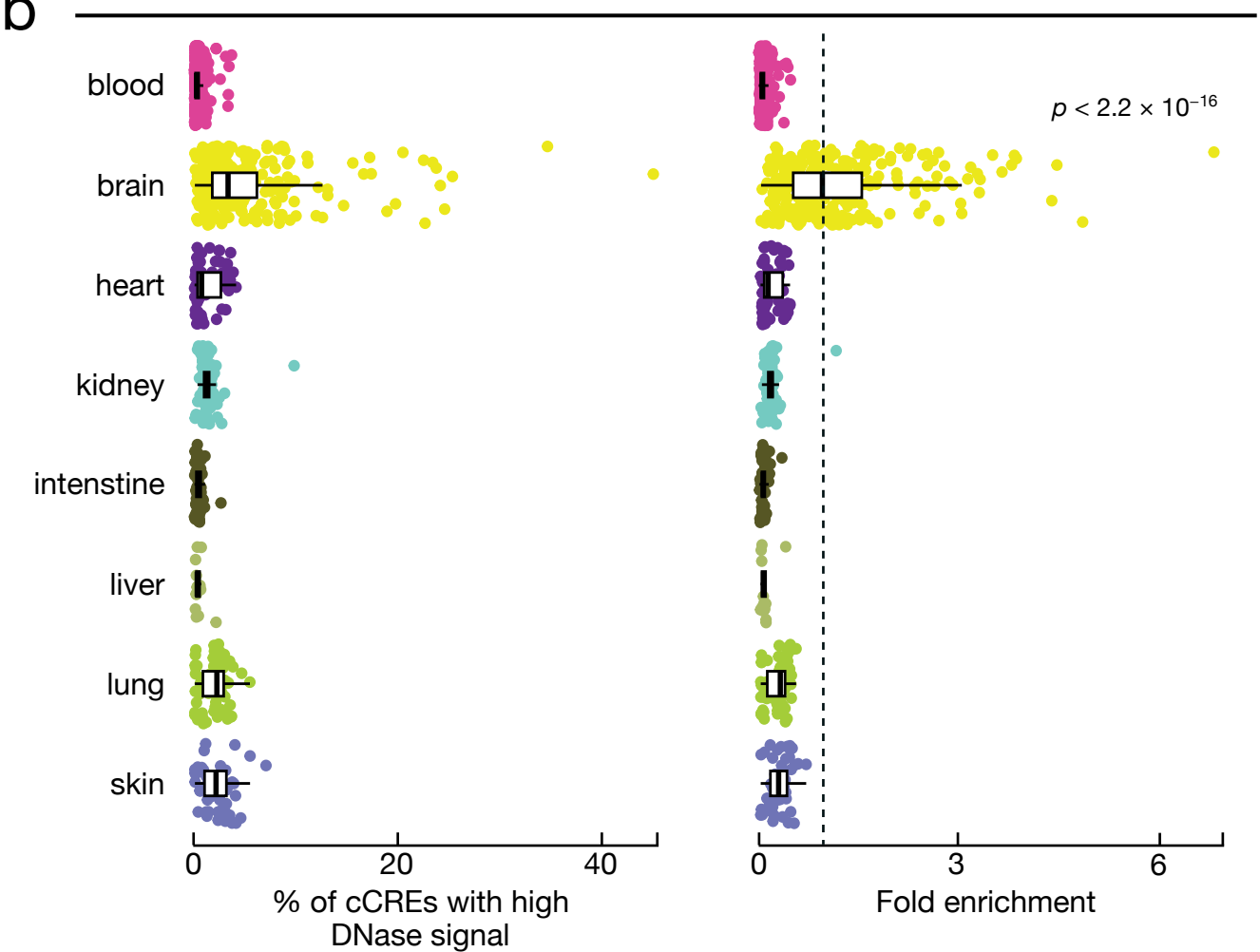
Supplementary Figure 5 | Characterizing the expanded Registry

a, Barplot showing the percentage of human cCREs in the ENCODE4 Registry that were: previously in the ENCODE3 registry (blue), initially called in ENCODE3 but filtered out (orange), new to ENCODE4 (yellow). **b**, Barplots showing the percentage of cCREs (as defined in **a**) and size-matched non-cCRE regions (gray) that overlap K562 STARR-seq peaks. **c**, Line plots depicting average phyloP conservation scores calculated from 240-way Mammalian alignment for +/- 500 bp around each cCRE center stratifying by cCRE group as defined in **a**. **d**, Heatscatter plot showing the distribution of ENCODE4 cCREs replicated from ENCODE3 based on their alignment across Zoonomia mammalian genomes. N1 denotes the number of species in which $\geq 90\%$ of a cCRE aligns; N2 denotes the number of species in which $\leq 10\%$ of a cCRE aligns. Rectangles indicate predefined groups: Group 1 (G1), highly conserved ($N1 \geq 120$, $N2 \leq 25$); Group 2 (G2), actively evolving ($20 \leq N1 \leq 50$, $N2 \leq 120$); and Group 3 (G3), primate-specific ($N1 \leq 50$, $N2 \geq 180$). **e**, Heatscatter plot as in **d**, but for ENCODE4 cCREs that were filtered out in ENCODE3. **f**, Heatscatter plot as in **d**, but for cCREs new in ENCODE4.

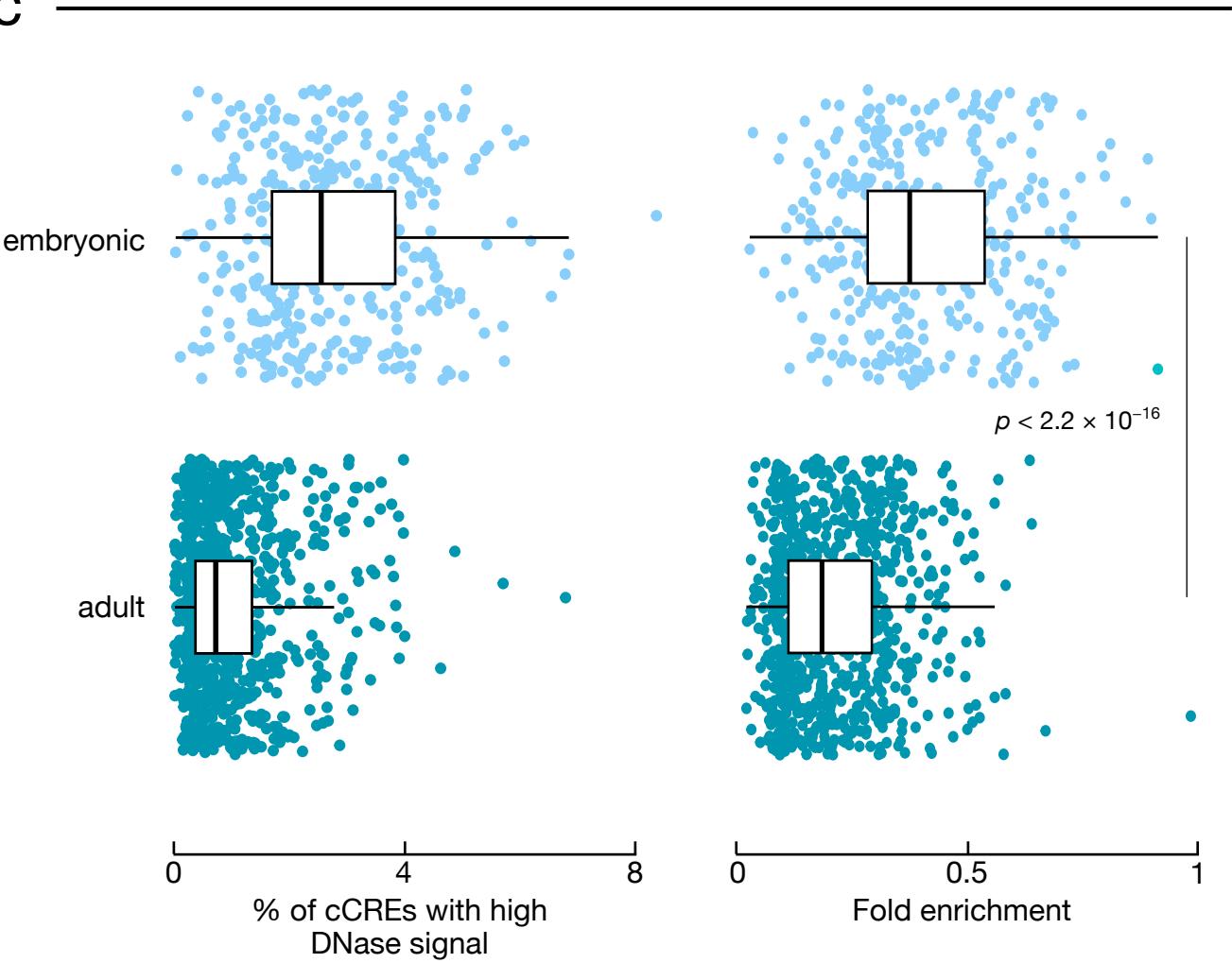
Li *et al.* BICCN cCREs



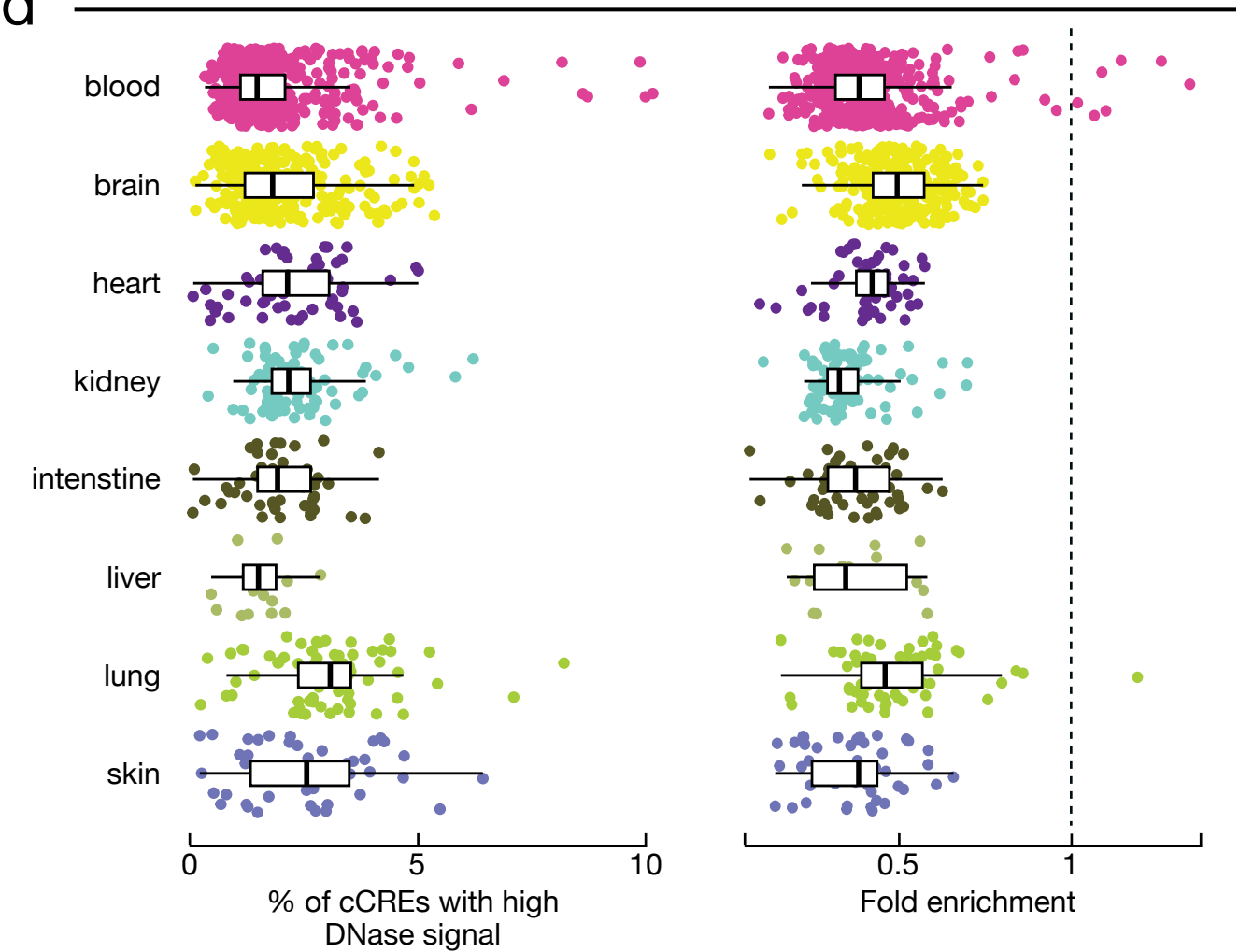
Loupe *et al.* union peaks



Zhang *et al.* acCREs

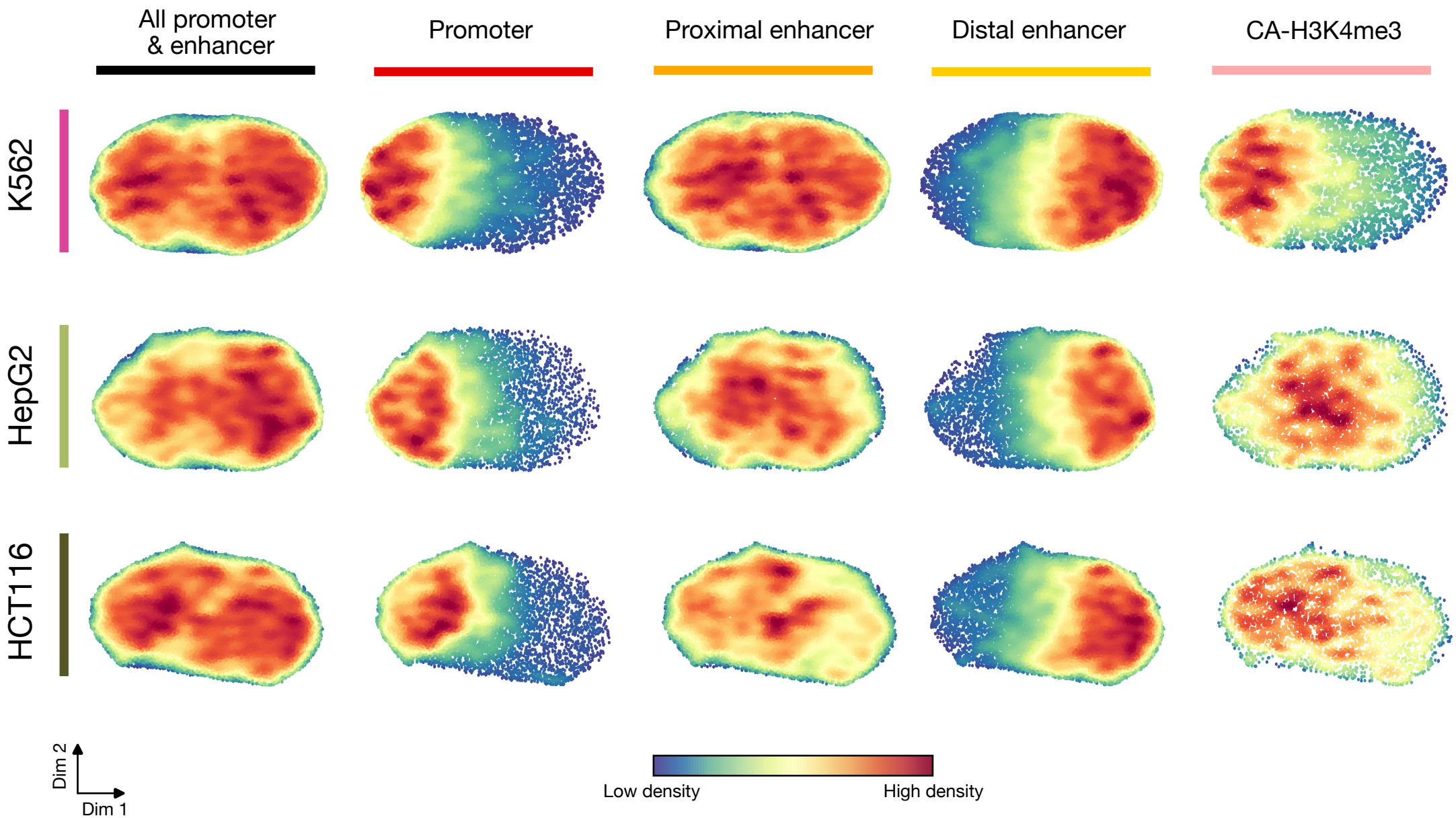


Liu *et al.* tCREs



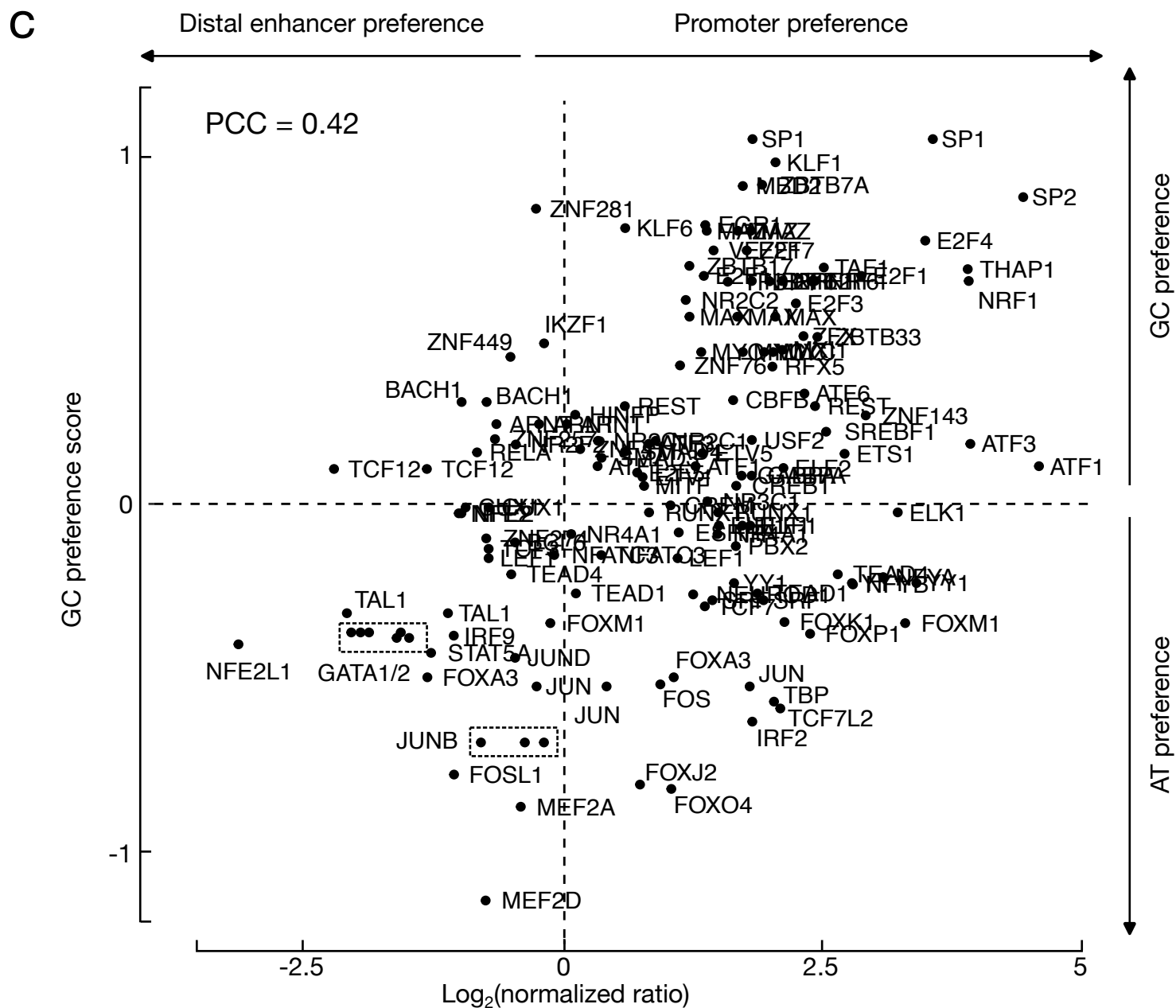
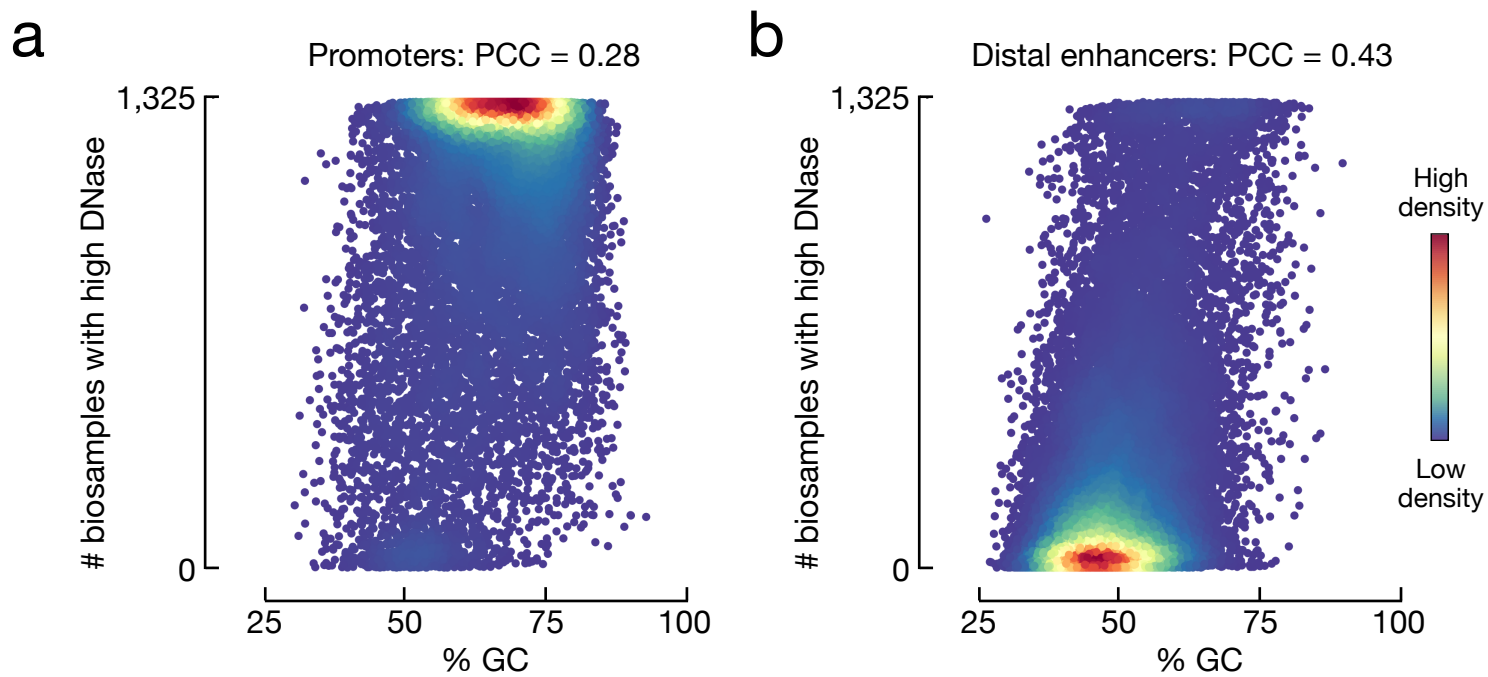
Supplementary Figure 6 | ENCODE4 cCREs have increased overlap with external regulatory element collections

a, Boxplots showing the percentage (left) and enrichment—compared to all cCREs—(right) of ENCODE4-specific cCREs that overlap BICCN cCREs with high DNase signals in different biosamples, stratified by tissue/organ of origin. Each point represents an individual biosample. **b**, Boxplots showing the percentage (left) and enrichment (right) of ENCODE4-specific cCREs that overlap union peaks from Loupe *et al.* with high DNase signals in different biosamples, stratified by tissue/organ of origin. **c**, Boxplots showing the percentage (left) and enrichment (right) of ENCODE4-specific cCREs that overlap tCREs from Li *et al.* with high DNase signals in different biosamples, stratified by tissue/organ of origin. **d**, Boxplots showing the percentage (left) and enrichment (right) of ENCODE4-specific cCREs that overlap acCREs from Zhang *et al.* with high DNase signals in different biosamples, stratified by biosample lifestage (embryonic vs. adult).



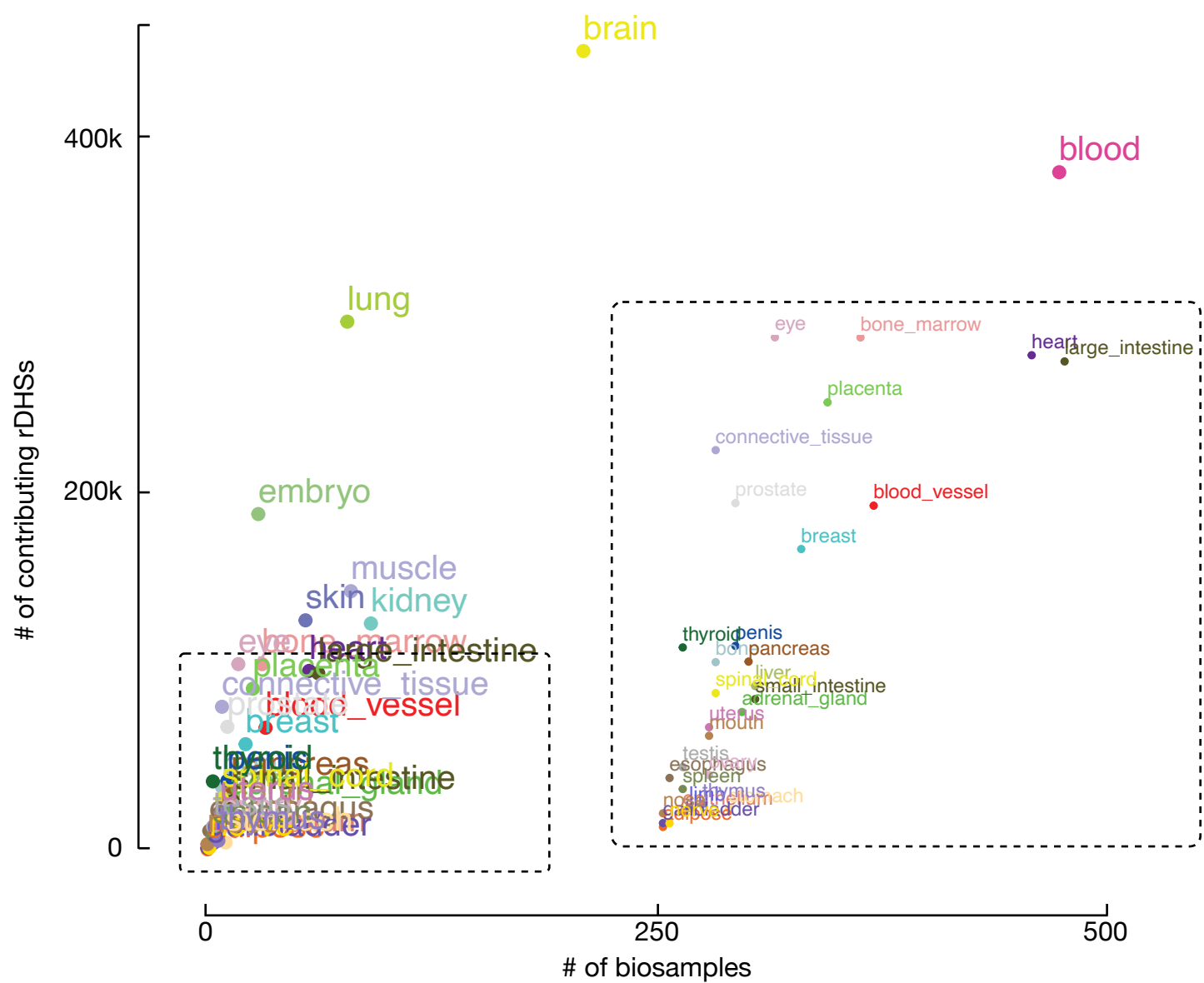
Supplementary Figure 7 | Distinct sequence features of promoter and enhancer cCREs

Density scatter plots showing the latent dimensions of cCRE sequence projected onto two dimensions by UMAP from variational autoencoders trained on three cell lines, K562 (top pink), HepG2 (middle green), and HCT116 (bottom olive). All promoter and enhancer cCREs are shown in column one followed by promoter (column 2) proximal enhancer (column 3), distal enhancer (column 4) and CA-H3K4me3 cCREs (column 5), in subsequent columns.



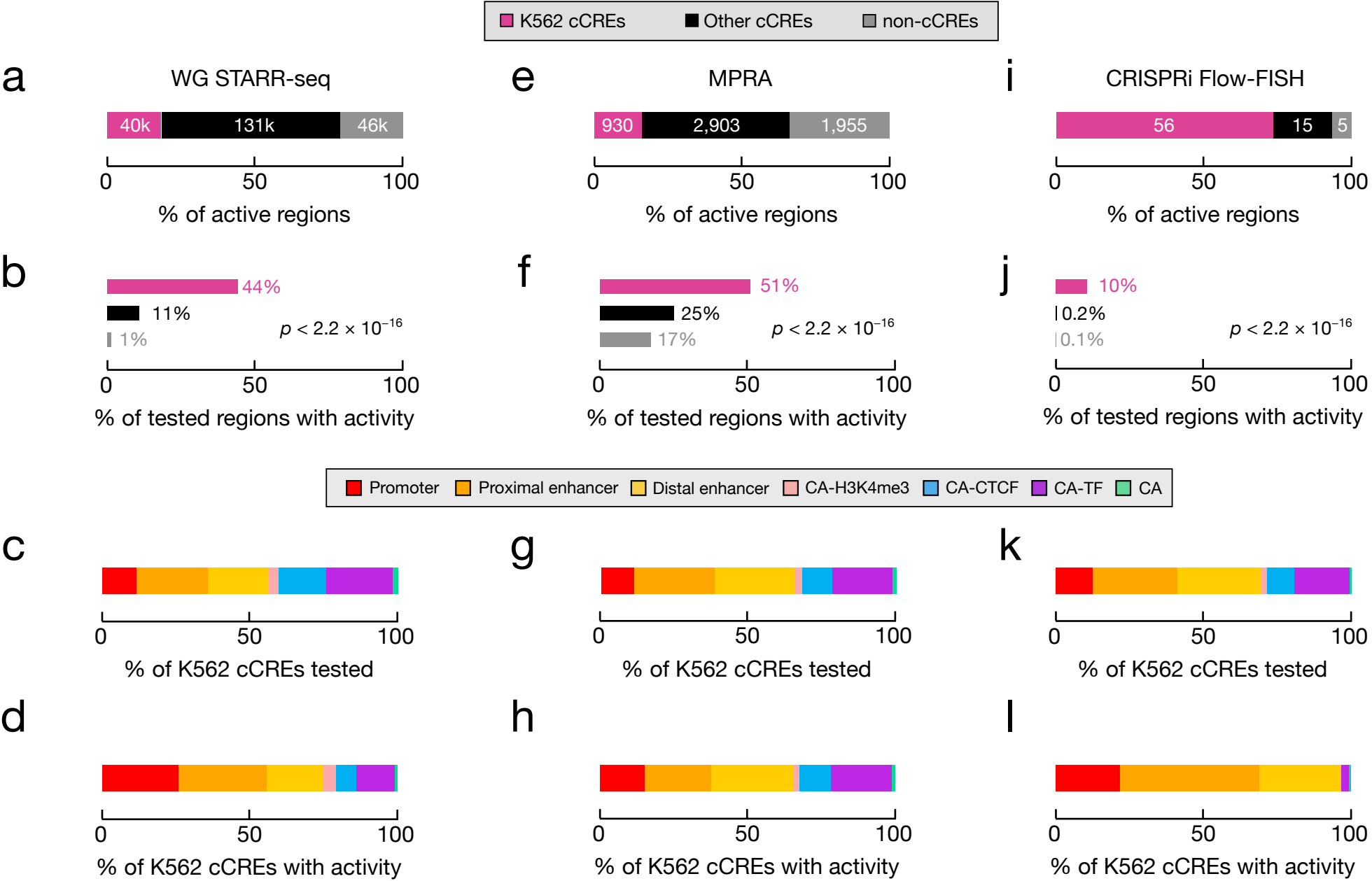
Supplementary Figure 8 | Investigating GC content of promoter and enhancer cCREs

a, Heatscatter depicting the number of biosamples with high DNase signal vs. GC content for promoter cCREs. **b**, Heatscatter plot as in **a**, but for distal enhancer cCREs. **c**, Scatterplot showing the relationship between GC preference score and promoter-binding preference for sequence-specific transcription factors, with each point representing a single ChIP-seq experiment in K562. The x-axis represents promoter-binding preference, defined as the $\log_2$ ratio of (% of peaks overlapping promoters / % overlapping enhancers) normalized by the promoter-to-enhancer ratio in K562. The y-axis shows the GC preference score, calculated by identifying the maximum log-odds score for guanine (G) or cytosine (C) at each motif position and averaging across all positions.



Supplementary Figure 9 | Coverage of representative DNase Hypersensitive sites across tissues and organs

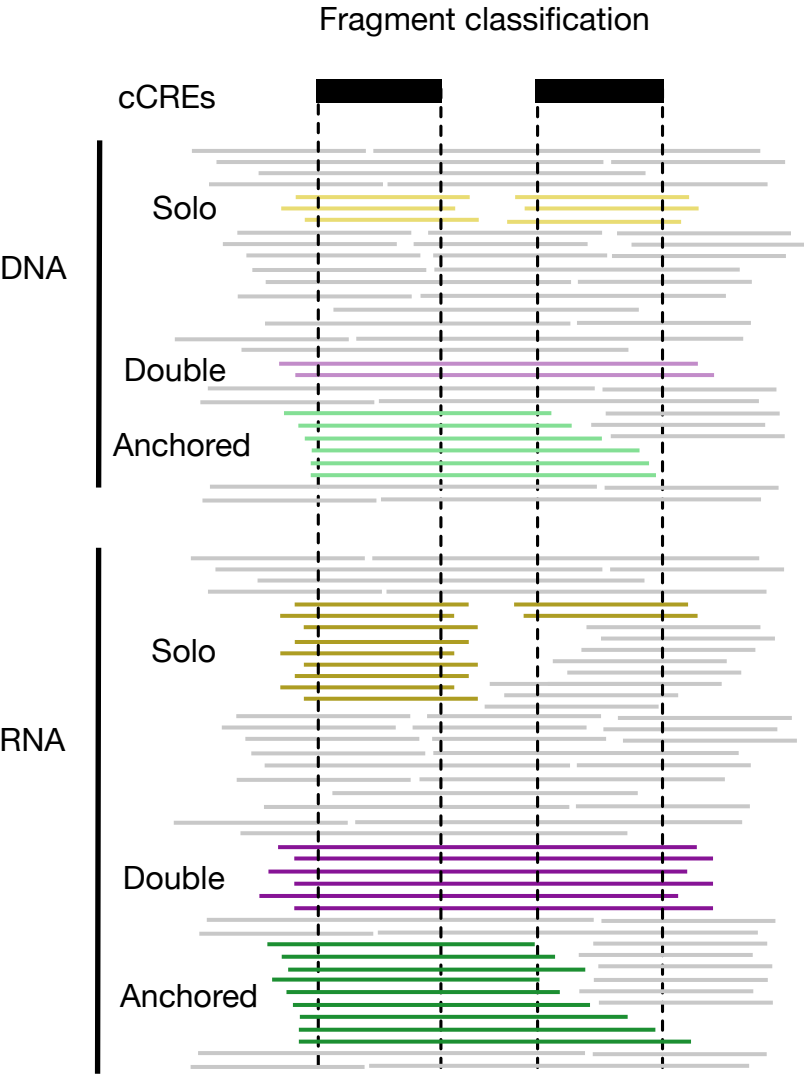
Scatterplot showing the relationship between the number of representative DNase hypersensitive sites (rDHSs) attributed to each biosample and the number of biosamples per organ or tissue group in the Registry of cCREs. The dashed box highlights the region shown in the inset.



Supplementary Figure 10 | cCREs are enriched for cell type-specific functional activity

a, Barplot showing the percentage of active whole genome STARR-seq region (called peaks) that overlap K562 cCREs (pink), other cell type cCREs (black), and non-cCRE regions (gray). Numbers designate the number of regions in each category. **b**, Barplot depicting the percentage of cCRE or non-cCRE regions (as defined in **a**) that were tested in the assay that overlapped active whole genome STARR-seq regions. **c**, Barplot showing the percentage of K562 cCREs belonging to each class that were tested by STARR-seq. **d**, Barplot showing the percentage of STARR-seq active K562 cCREs belonging to each class.. **e**, Barplot showing the percentage of active MPRA regions ($\log_2$ fold change greater than one) that overlap K562 cCREs (pink), other cell type cCREs (black), and non-cCRE regions (gray). Numbers designate the number of regions in each category. **f**, Barplot depicting the percentage of cCRE or non-cCRE regions (as defined in **e**) that were tested in the assay that overlapped active MPRA regions. **g**, Barplot showing the percentage of K562 cCREs belonging to each class that were tested by MPRA. **h**, Barplot showing the percentage of MPRA active K562 cCREs belonging to each class. **i**, Barplot showing the percentage of active CRISPRi Flow-FISH regions ($\text{FDR} < 0.05$) that overlap K562 cCREs (pink), other cell type cCREs (black), and non-cCRE regions (gray). Numbers designate the number of regions in each category. **j**, Barplot depicting the percentage of cCRE or non-cCRE regions (as defined in **i**) that were tested in the assay that overlapped active CRISPRi Flow-FISH regions. **k**, Barplot showing the percentage of K562 cCREs belonging to each class that were tested by CRISPRi Flow-FISHc. **l**, Barplot showing the percentage of CRISPRi active K562 cCREs belonging to each class.

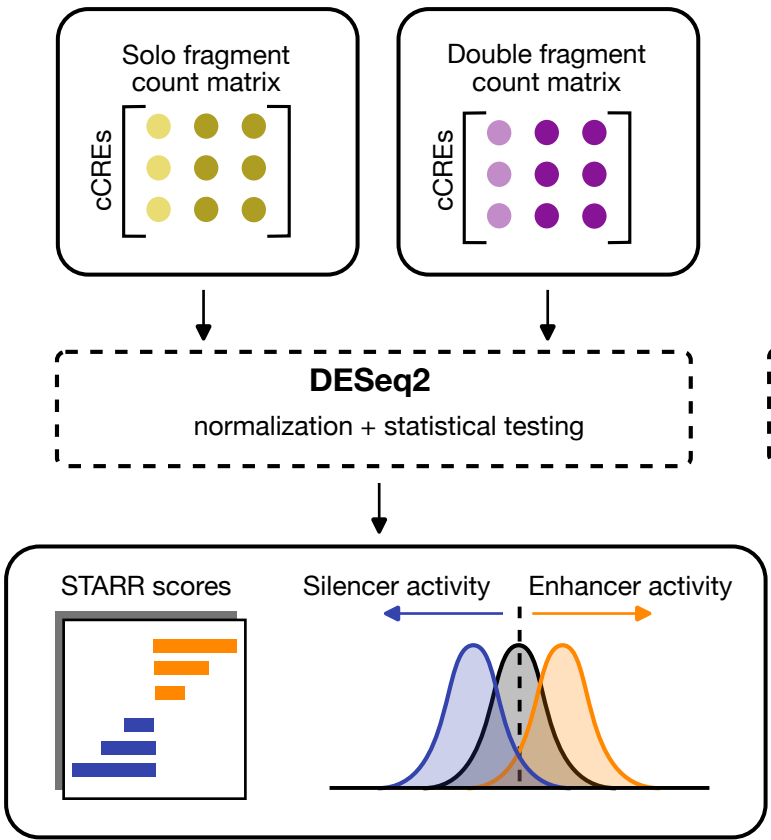
a



CRE-centric Analysis and Prediction of Reporter Assays

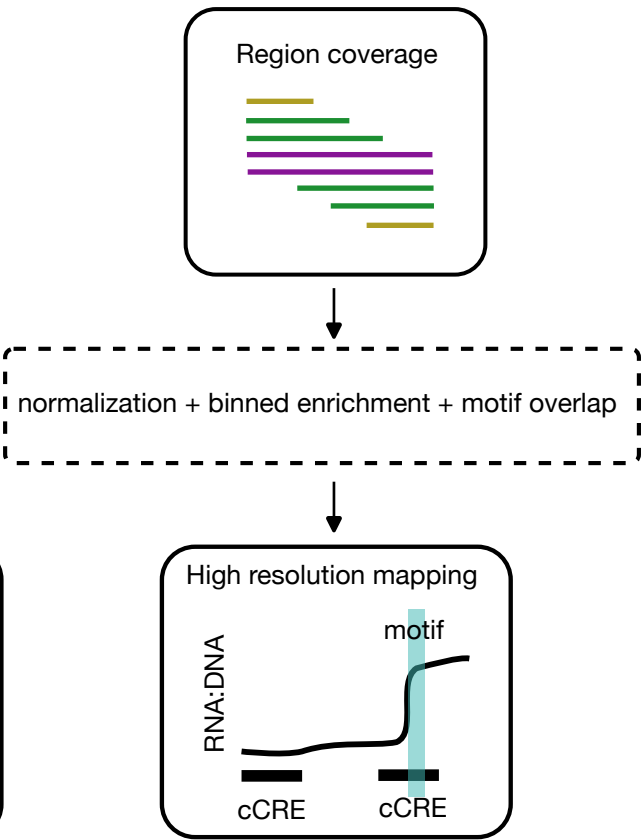
b

Mode: cCRE centric



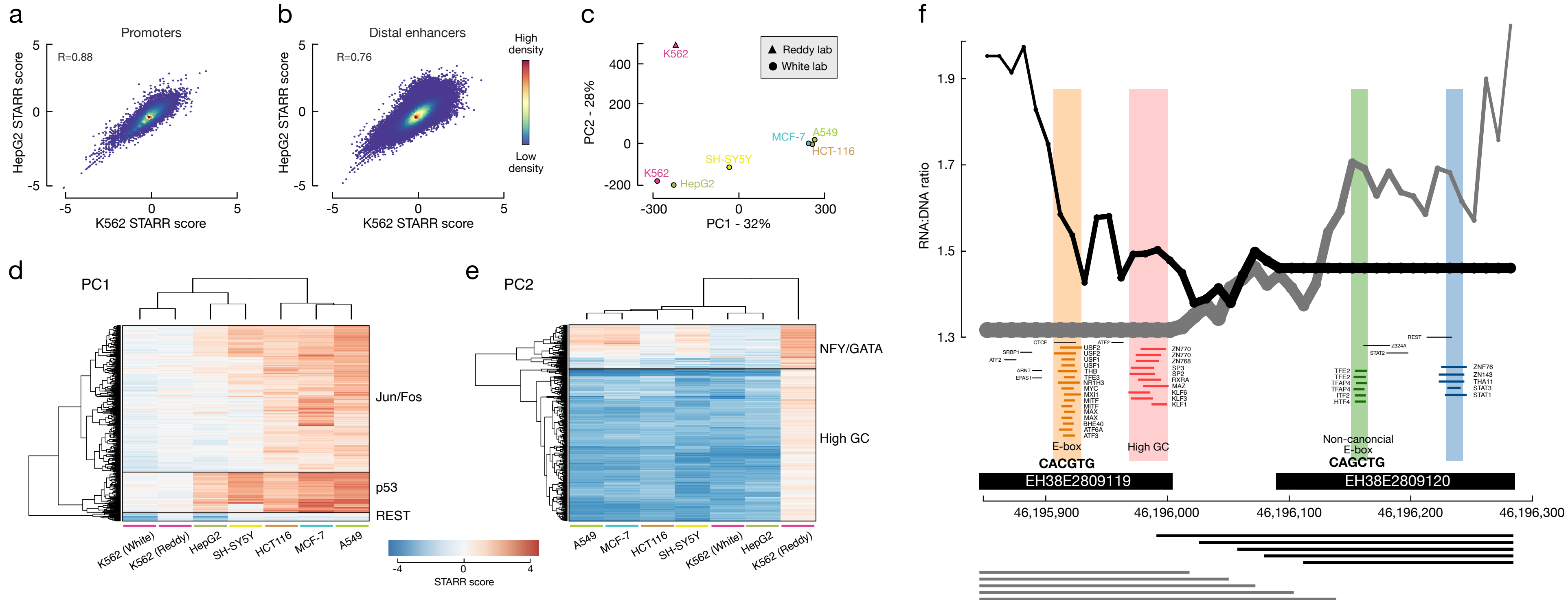
c

Mode: paired sweep



Supplementary Figure 11 | Overview of CAPRA pipeline

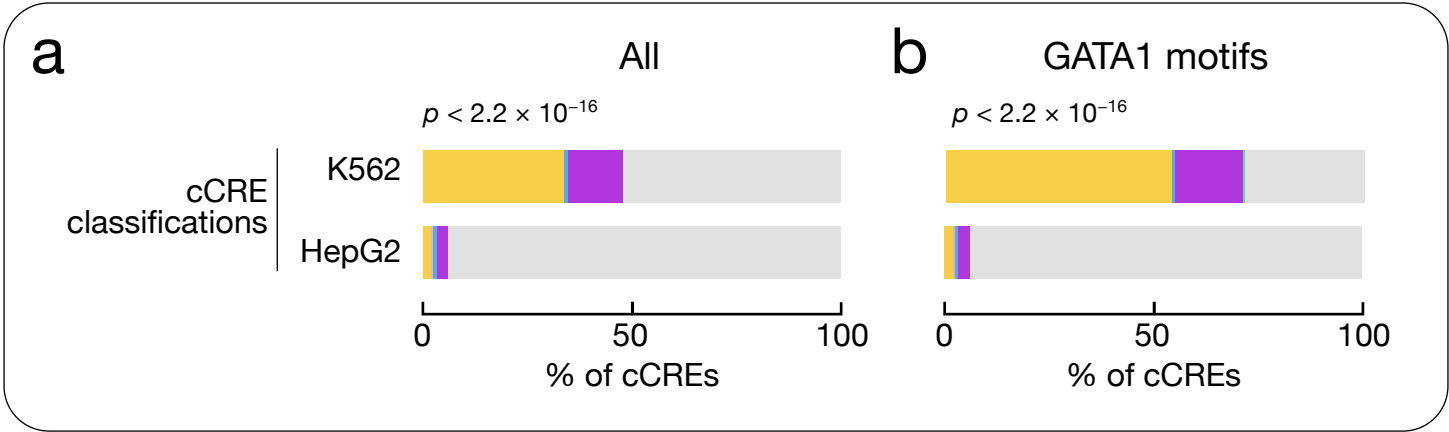
a, Schematic illustrating the types of STARR-seq fragments used in the CAPRA pipeline. *Solo* fragments (yellow) fully overlap a single cCRE, *double* fragments (purple) fully overlap a pair of adjacent cCREs, and *anchored* fragments (green) fully overlap one cCRE and partially overlap an adjacent cCRE. **b**, Schematic of *cCRE-centric* mode of CAPRA. Solo and double fragments are counted across input DNA and RNA libraries. Counts are input into DESeq2, which models RNA vs. DNA abundance using a negative binomial framework and estimates $\log_2$ fold changes and significance values. These values quantify enhancer-like (positive, orange) or silencer-like (negative, blue) activity at individual cCREs (*solo* fragments) or cCRE pairs (*double* fragments). **c**, Schematic of *paired sweep* mode of CAPRA. Solo, double, and anchored fragments are normalized by library size and aggregated into 10 bp bins across the interval between two cCREs. This is done in both directions, sweeping from each cCRE toward the other. Resulting RNA:DNA ratio profiles are compared to underlying sequence features to identify motifs associated with regulatory activity transitions.



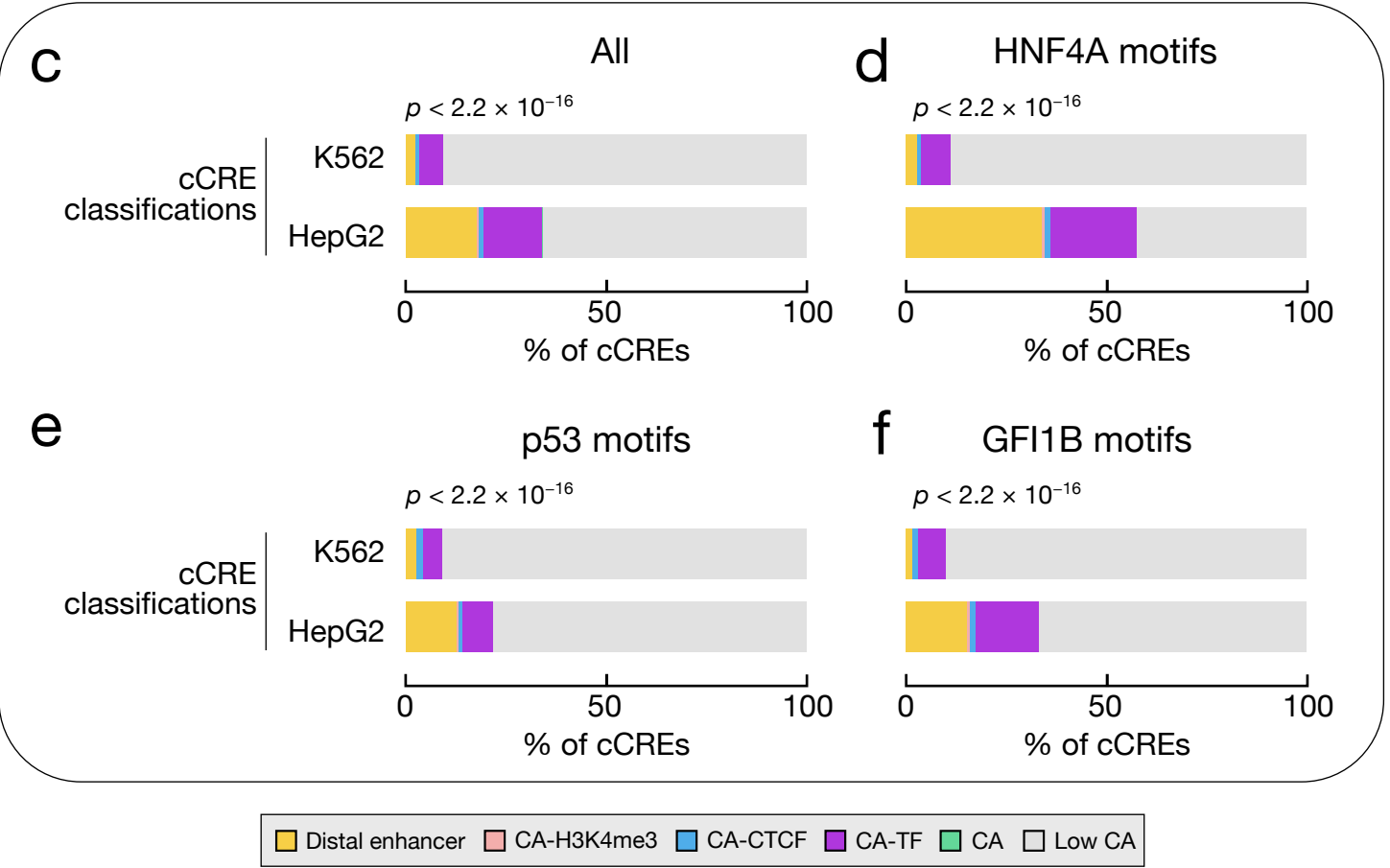
Supplementary Figure 12 | CAPRA use cases

a, Density scatterplot showing the correlation of CAPRA-calculated STARR scores for promoter cCREs in K562 (x-axis) and HepG2 (y-axis). **b**, Density scatterplot showing the correlation of STARR scores for distal enhancer cCREs in K562 (x-axis) and HepG2 (y-axis). These data are also presented in Figure 2a but are colored here by point density rather than cell type specificity. **c**, Principal component analysis (PCA) of STARR scores across 362,591 cCREs from seven STARR-seq libraries. Each point represents a library, with color indicating cell type and shape indicating lab origin (triangles for Reddy lab, circles for White lab). PC1 and PC2 explain 32% and 28% of the variance, respectively. **d**, Heatmap of STARR scores for the top 500 cCREs contributing to PC1 from panel **c**. Three major clusters are shown, identified by hierarchical clustering; enriched sequence motifs for each cluster are listed at right. **e**, Heatmap of STARR scores for the top 500 cCREs contributing to PC2 from panel **c**. Two major clusters are shown, with corresponding enriched motifs listed at right. **f**, Line plots showing RNA:DNA ratios for the cCRE pair EH38E2809119 and EH38E2809120 using CAPRA's Paired Sweep mode. Gray and black lines represent fragments anchored at EH38E2809119 and EH38E2809120, respectively; line width is proportional to the number of input DNA fragments. FIMO-identified motifs across the region are shown below the plots, with four highlighted motif groups: canonical E-box (orange), high GC content (red), non-canonical E-box (green), and ZNF/STAT motifs (blue).

K562 STARR+ distal enhancer cCREs

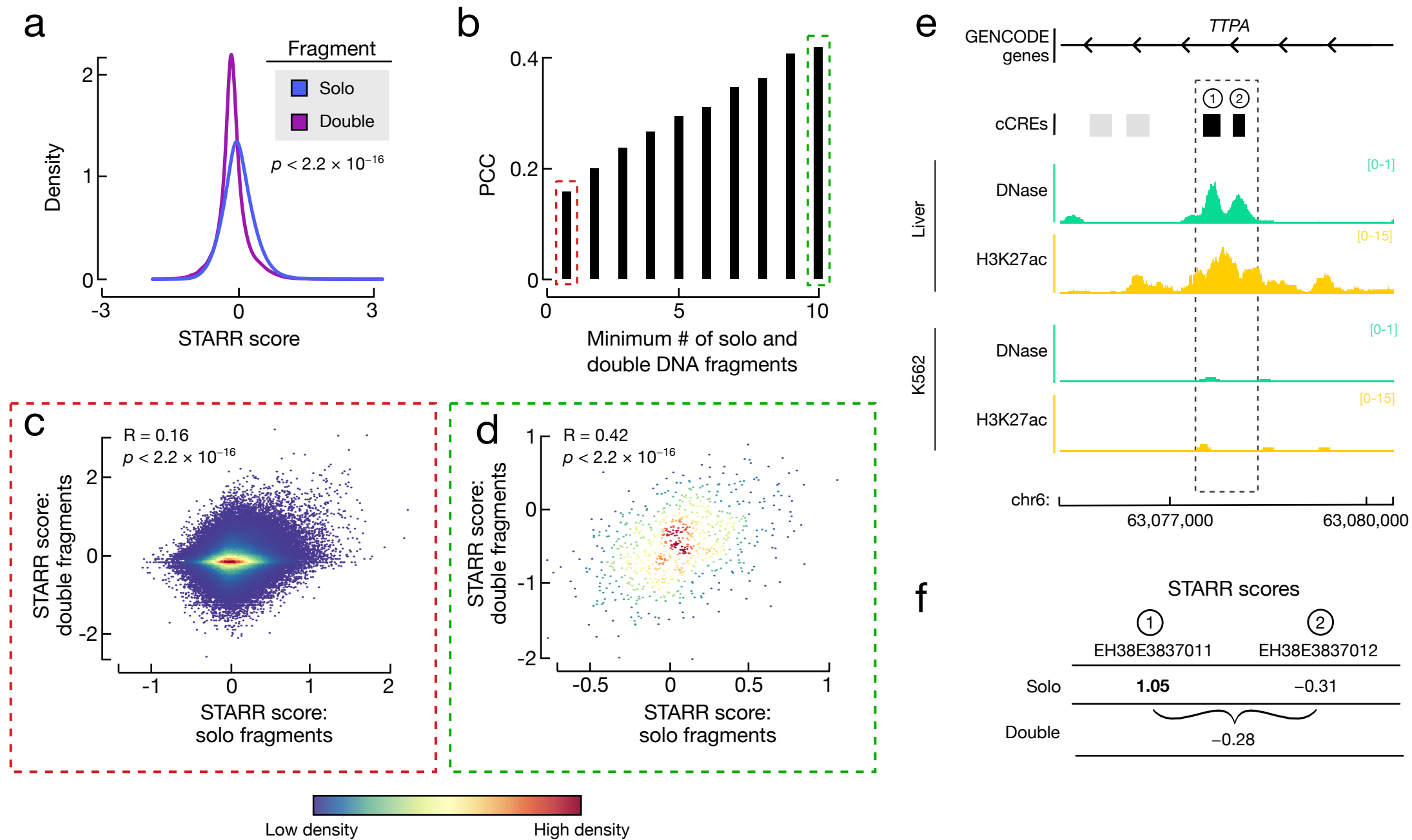


HepG2 STARR+ distal enhancer cCREs



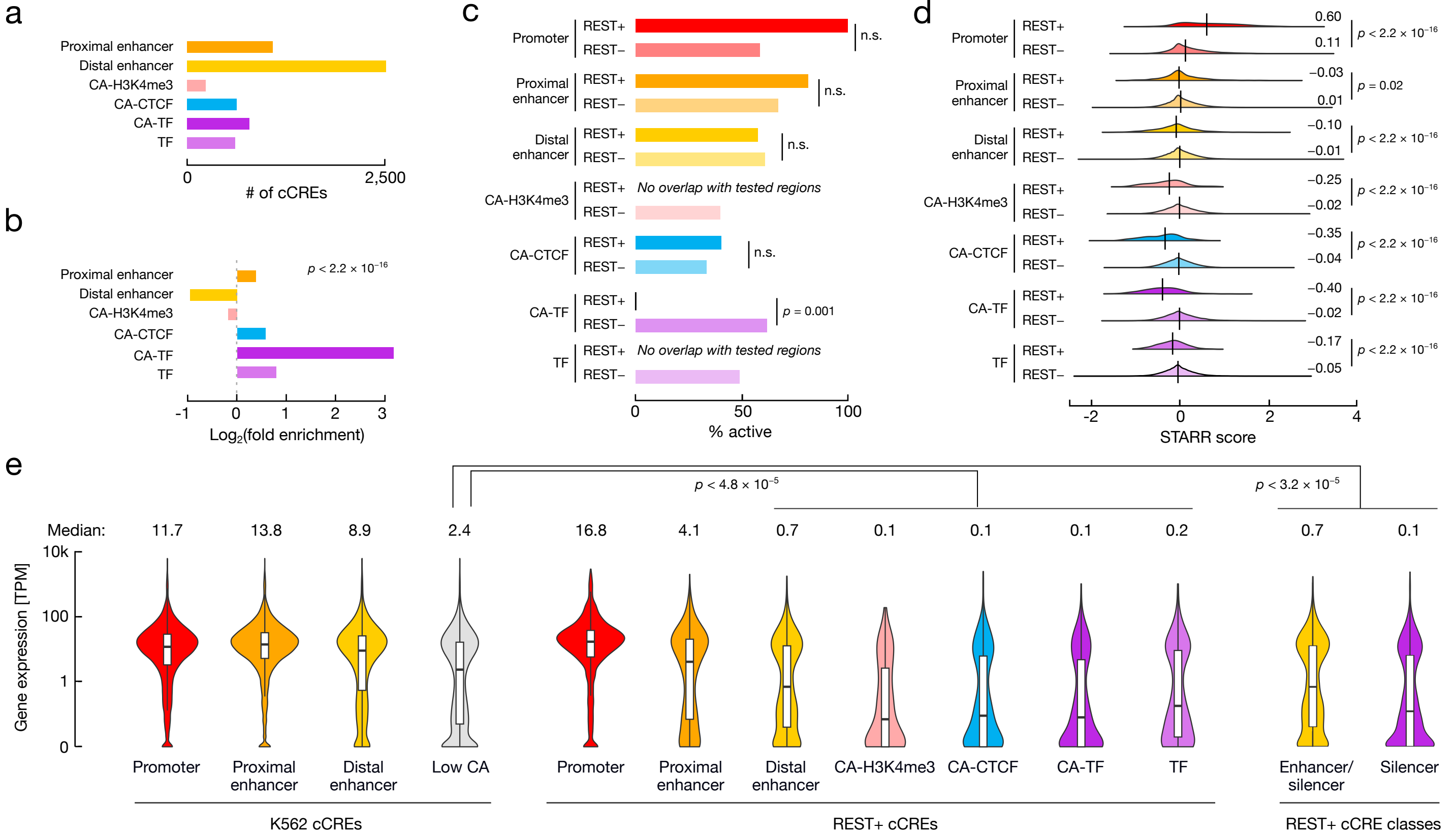
Supplementary Figure 13 | Cell type-specific STARR activity is concordant with cell type-specific chromatin signatures

a, Barplots showing the percentage of distal enhancer cCREs with K562-specific STARR activity (K562 STARR+) belonging to each cCRE class called in K562 cells (using K562 chromatin signatures) and in HepG2 cells (using HepG2 chromatin signatures). **b**, Barplots showing the percentage of K562 STARR+ distal enhancer cCREs that overlap GATA1 motifs belonging to each cCRE class called in K562 cells and in HepG2 cells. **c**, Barplots showing the percentage of distal enhancer cCREs with HepG2-specific STARR activity (HepG2 STARR+) belonging to each cCRE class called in K562 cells (using K562 chromatin signatures) and in HepG2 cells (using HepG2 chromatin signatures). **d**, Barplots showing the percentage of HepG2 STARR+ distal enhancer cCREs that overlap HNF4A motifs belonging to each cCRE class called in K562 cells and in HepG2 cells. **e**, Barplots showing the percentage of HepG2 STARR+ distal enhancer cCREs that overlap p53 motifs belonging to each cCRE class called in K562 cells and in HepG2 cells. **f**, Barplots showing the percentage of HepG2 STARR+ distal enhancer cCREs that overlap GFI1B motifs belonging to each cCRE class called in K562 cells and in HepG2 cells. In all plots, the *** denotes p less than 0.001 as calculated by a chi-square test.



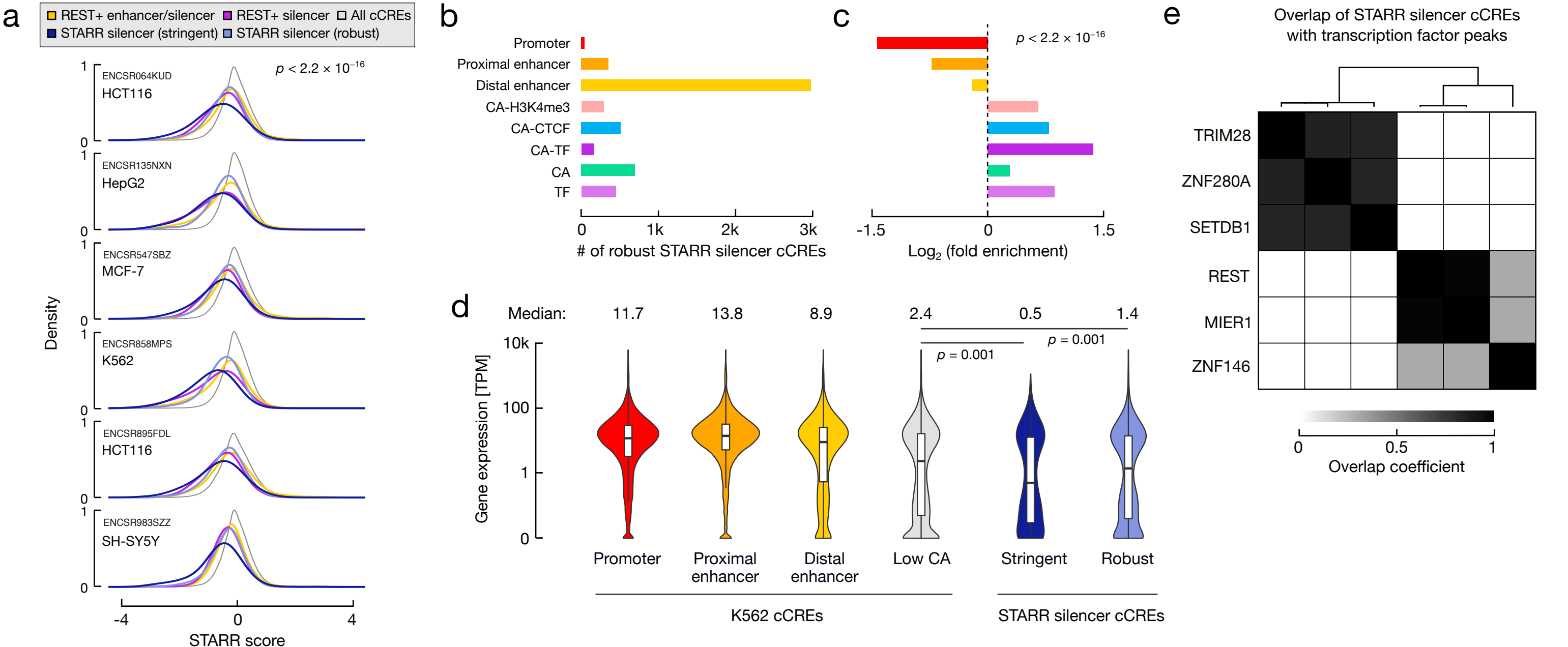
Supplementary Figure 14 | Using CAPRA to dissect combinatorial activities of cCREs

a, Density plots showing the distribution of STARR scores calculated by CAPRA from solo (blue) and double (purple) fragments. **b**, Barplots showing the pearson correlation coefficient (PCC) for cCRE pairs from the average STARR scores calculated from testing individual cCREs (solo DNA fragments) and the STARR scores calculated from testing multiple elements (double DNA fragments) as the threshold for the minimum number of input DNA fragments increases. Representative scatterplots for a cutoff of one (red box) and ten (green box) are shown in **c** and **d**. **c**, Density scatterplot of STARR scores for pairs of cCREs with a minimum of one DNA input fragment calculated by averaging quantifications from solo fragments or double fragments. **d**, Density scatterplot of STARR scores for pairs of cCREs with a minimum of ten DNA input fragments calculated by averaging quantifications from solo fragments or double fragments. **e**, Example of two cCREs (1: EH38E3837011 and 2: EH38E3837012) within the TTPA gene with lower than expected combinatorial effects. These cCRE have high DNase and H3K27ac signals in liver but low signals in K562 cells. **f**, STARR scores for EH38E3837011 and EH38E3837012 deriving from solo fragments (top) and double fragments (bottom). Scores with $p < 0.05$ are in bold.



Supplementary Figure 15 | Characteristics of REST+ cCREs

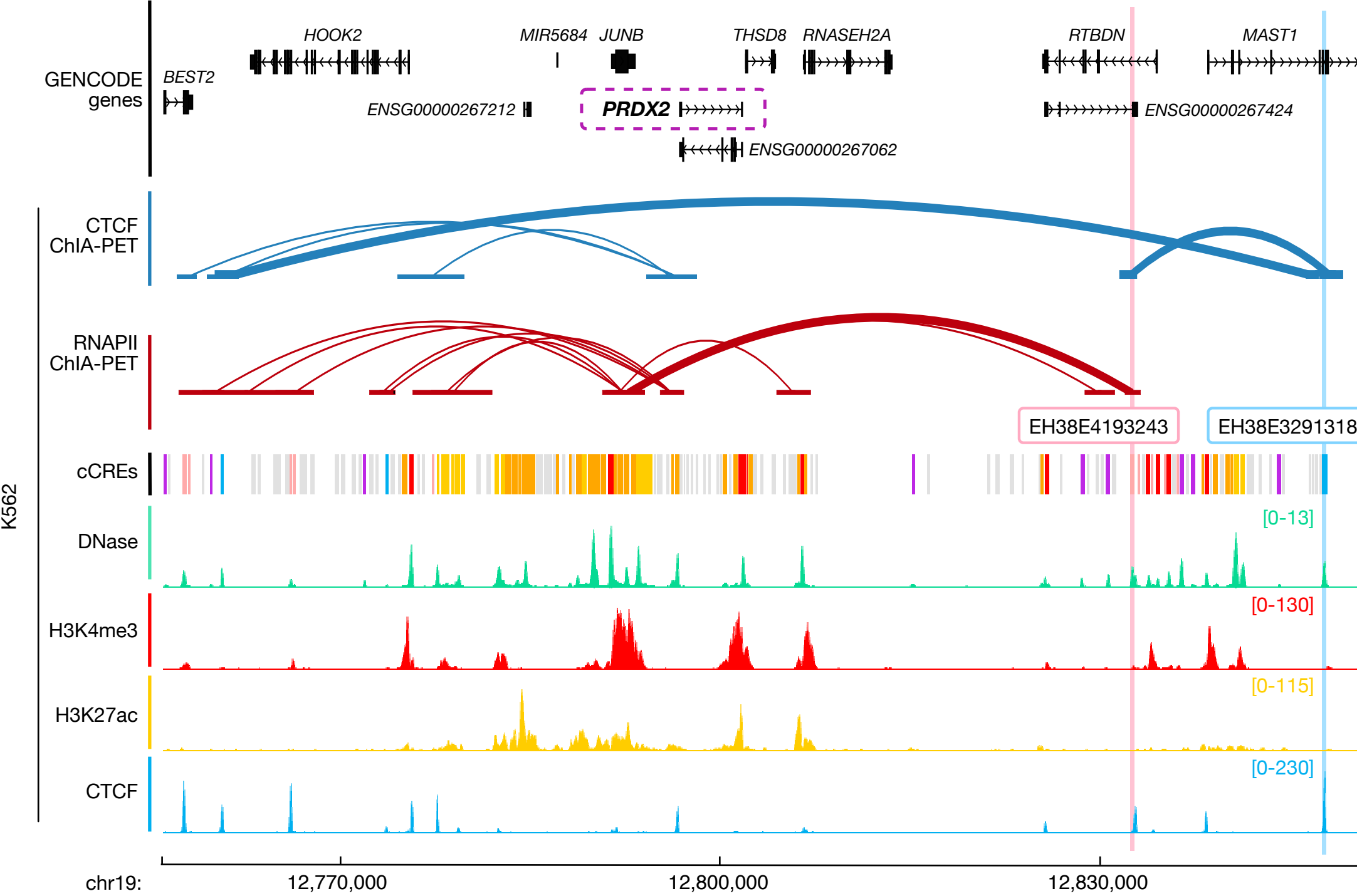
a, Barplots depicting the number of REST+ cCREs stratified by cCRE class. **b**, Barplots depicting the enrichment for cCRE classes of REST+ cCREs compared to the entire Registry. **c**, Barplots depicting the percentage of VISTA regions with enhancer activity in a transgenic mouse enhancer assay. Regions are stratified by cCRE class and overlap with REST ChIP-seq peaks (REST+). p-values are shown for two-sided Fisher's exact tests. **d**, Density plots showing the distribution of K562 STARR scores for each class of cCRE stratified by REST binding (REST+). Vertical lines denote medians which are also displayed to the right of each distribution. p-values correspond to Wilcoxon tests. **e**, Nested violin-boxplots showing gene expression in K562 cells for the nearest gene of different classes and categories of cCREs. On the left are cCREs annotated in K562 based on K562-specific chromatin signatures including promoter, proximal enhancer, distal enhancer and low chromatin accessibility cCREs. In the middle are REST+ cCREs stratified by cell type agnostic classification. On the right are the final REST+ cCRE categories: enhancer/silencer and silencer. Medians are shown above each distribution.



Supplementary Figure 16 | Characteristics of STARR silencer cCREs

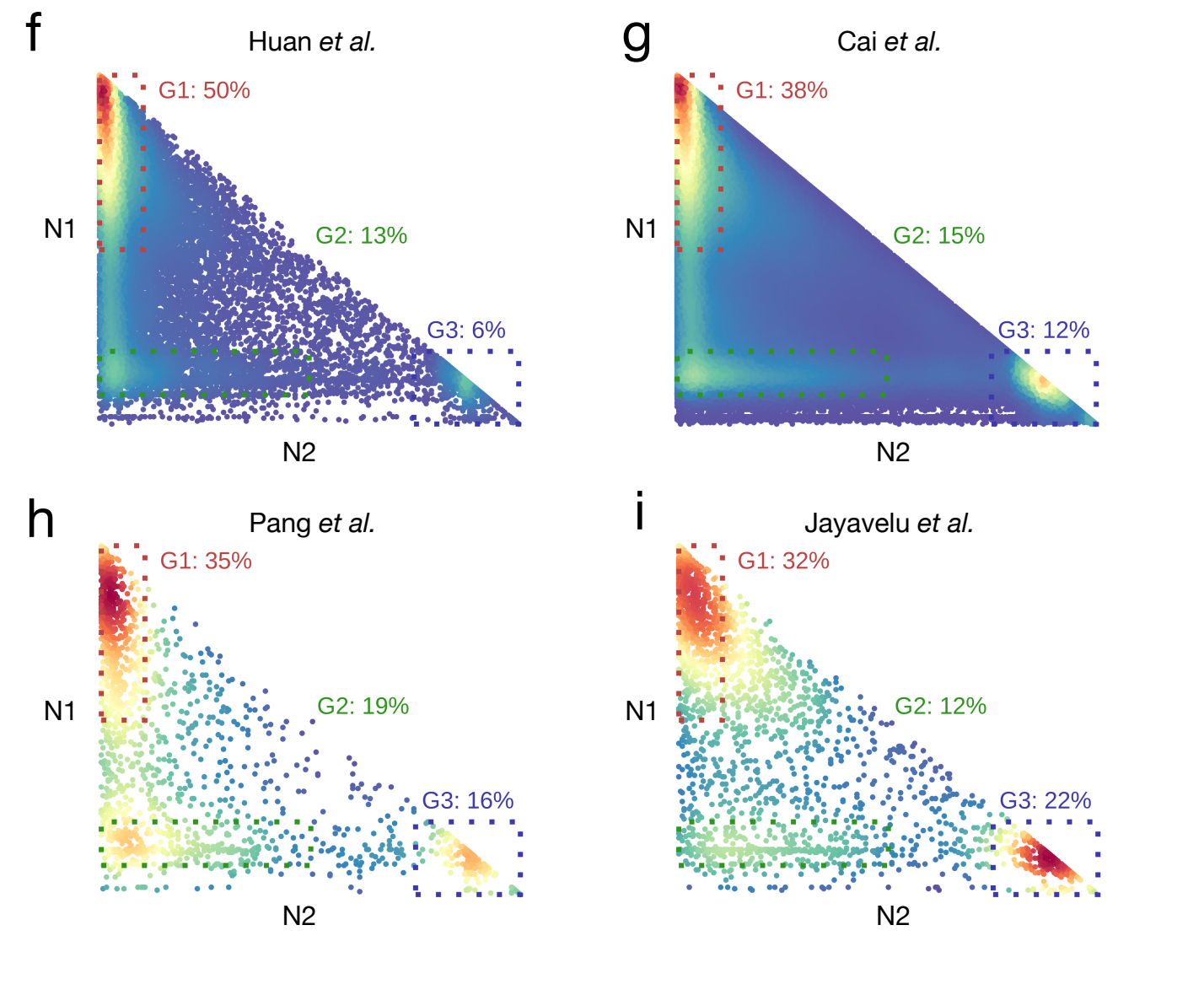
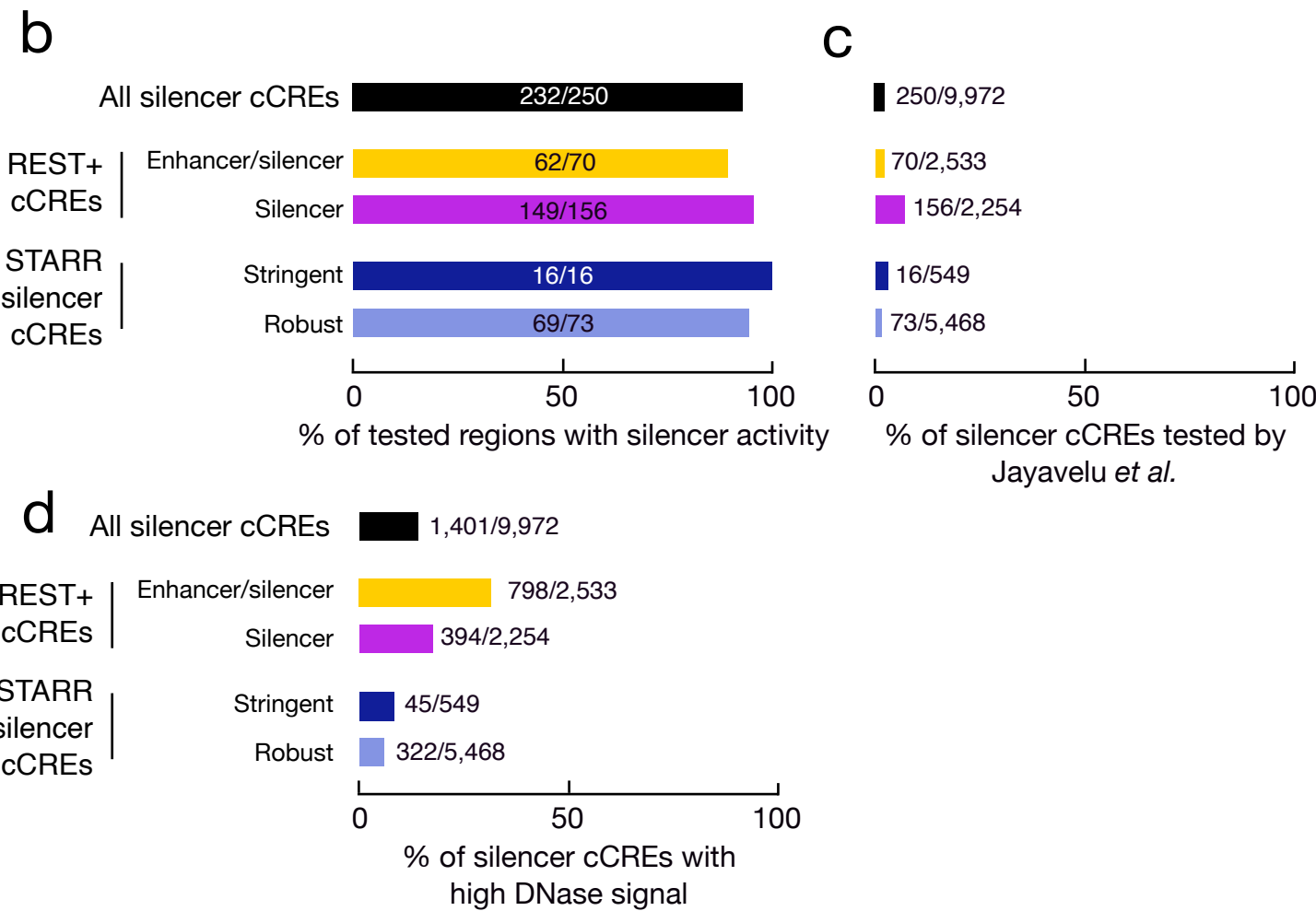
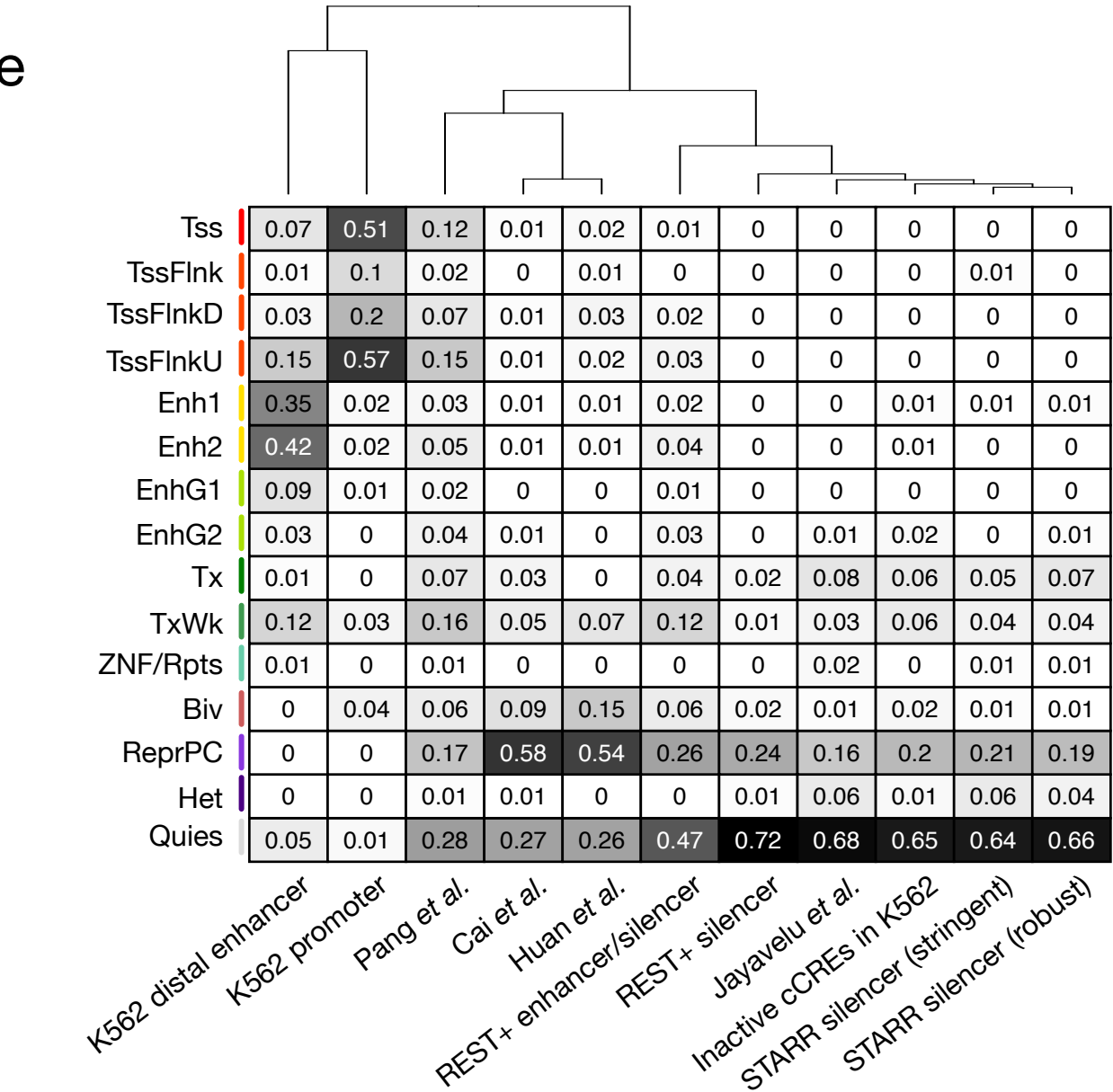
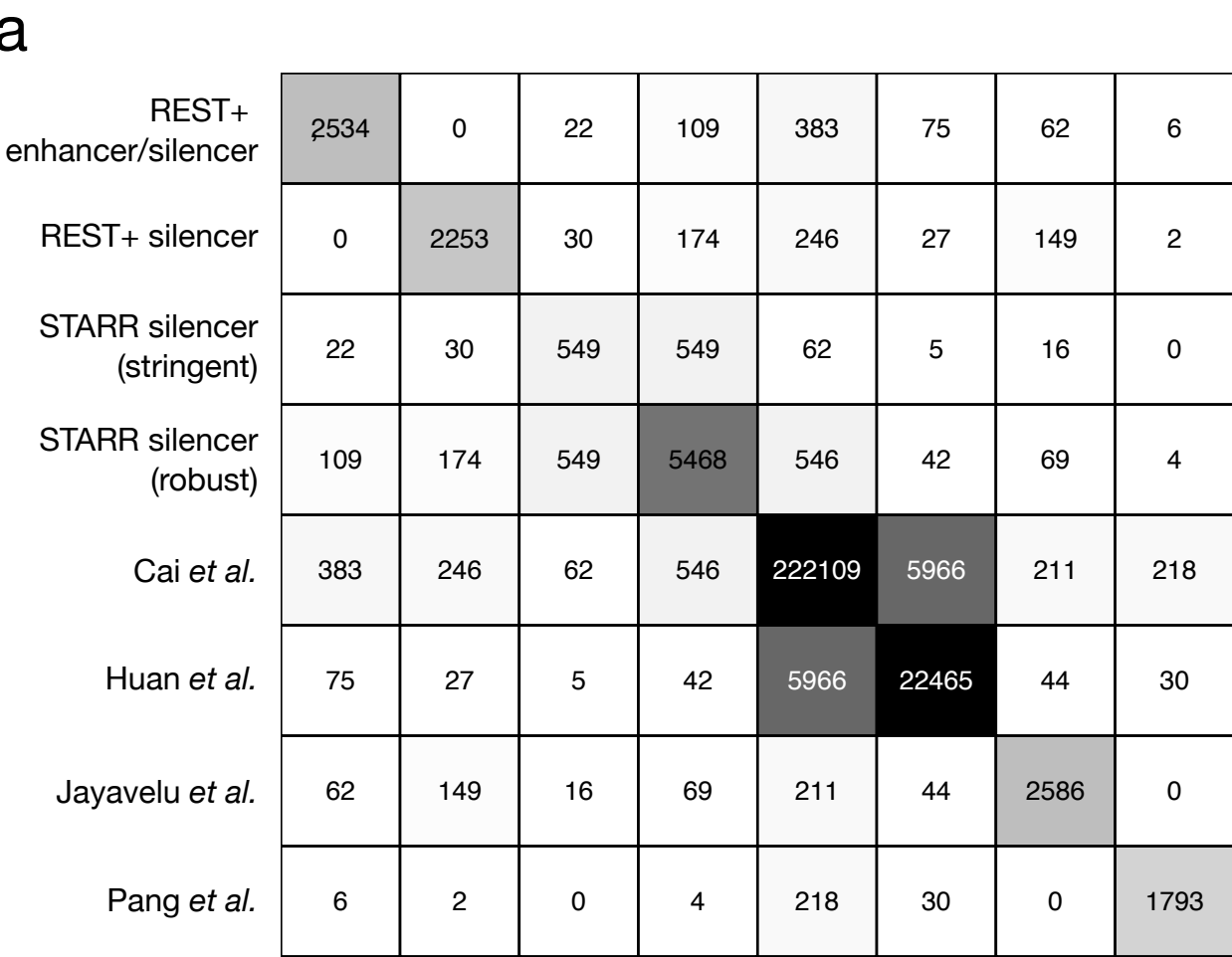
a, Density plots depicting STARR scores for silencer cCREs in six STARR-seq libraries stratified by category: REST+ enhancer/silencer (yellow) REST+ silencer (purple), stringent STARR silencer (dark blue), robust STARR silencer (light blue). All silencer distributions are significantly shifted compared to the background of all cCREs (gray). P-values, all $< 2.2 \times 10^{-16}$, are calculated from a pairwise, two-sided Wilcoxon rank sum test with FDR correction. **b**, Barplots depicting the number of robust STARR silencer cCREs stratified by cCREs class. **c**, Barplots depicting the enrichment for cCRE classes of robust STARR silencer cCRE compared to the entire Registry. **d**, Heatmap depicting the overlap of STARR silencer cCREs with binding sites for TRIM28, ZNF280A, SETDB1, REST, MIER1 and ZNF146 (DNA binding proteins enriched in STARR silencers, **Supplementary Table 11d**). Hierarchical clustering reveals two distinct clusters of STARR silencer cCREs based on ChIP-seq peak overlap. **e**, Nested violin-boxplots showing gene expression in K562 cells for the nearest gene of different classes and categories of cCREs. On the left are cCREs annotated in K562 based on K562-specific chromatin signatures including promoter, proximal enhancer, distal enhancer and low chromatin accessibility (CA) cCREs. On the right are STARR silencer cCREs stratified by threshold categories. Medians are shown above each distribution. P-values are calculated from a pairwise, two-sided Wilcoxon rank sum test with FDR correction.

Moore *et al.* Supplementary Figure 17



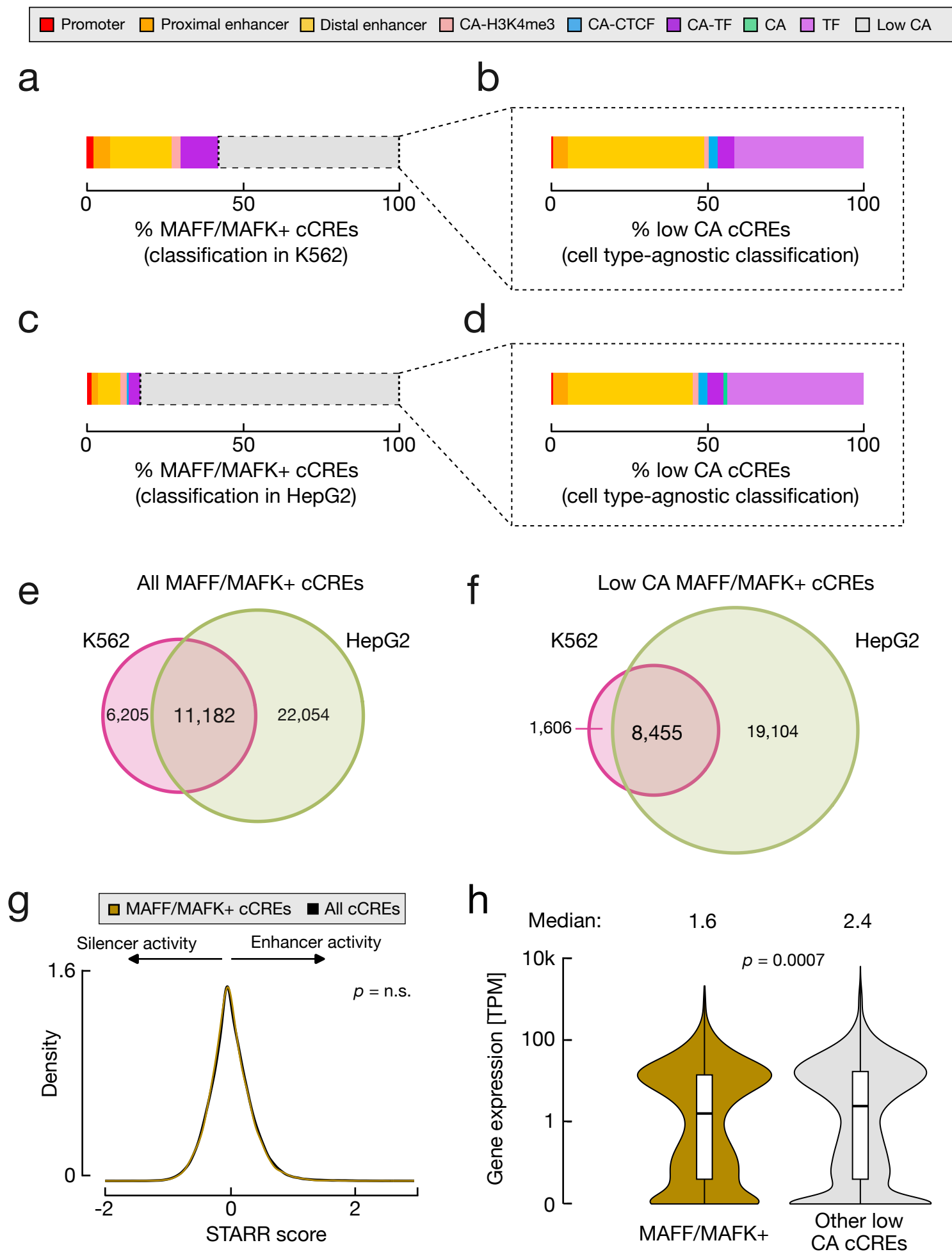
Supplementary Figure 17 | Silencer cCRE EH38E4193243 regulates PRDX2 via CTCF-Mediated chromatin interactions

Genome browser view showing physical 3D genome interactions between silencer cCRE EH38E4193243 (pink highlight) and CTCF-anchor EH38E3291318 (blue highlight) with a domain of expressed genes. Targeting either silencer EH38E4193243 or CTCF-anchor EH38E3291318 results in increased expression of *PRDX2* (dashed box). CTCF ChIA-PET links (dark blue) directly connect silencer EH38E4193243 with EH38E3291318. CTCF-anchor EH38E3291318 is also connected with upstream cCREs, forming a domain which brings silencer EH38E4193243 near the active promoters, as shown by RNAPII ChIA-PET links (dark red). Biochemical signals (DNase: green, H3K4me3: red, H3K27ac: yellow, CTCF: blue) and cCRE classifications for K562 cells are displayed.



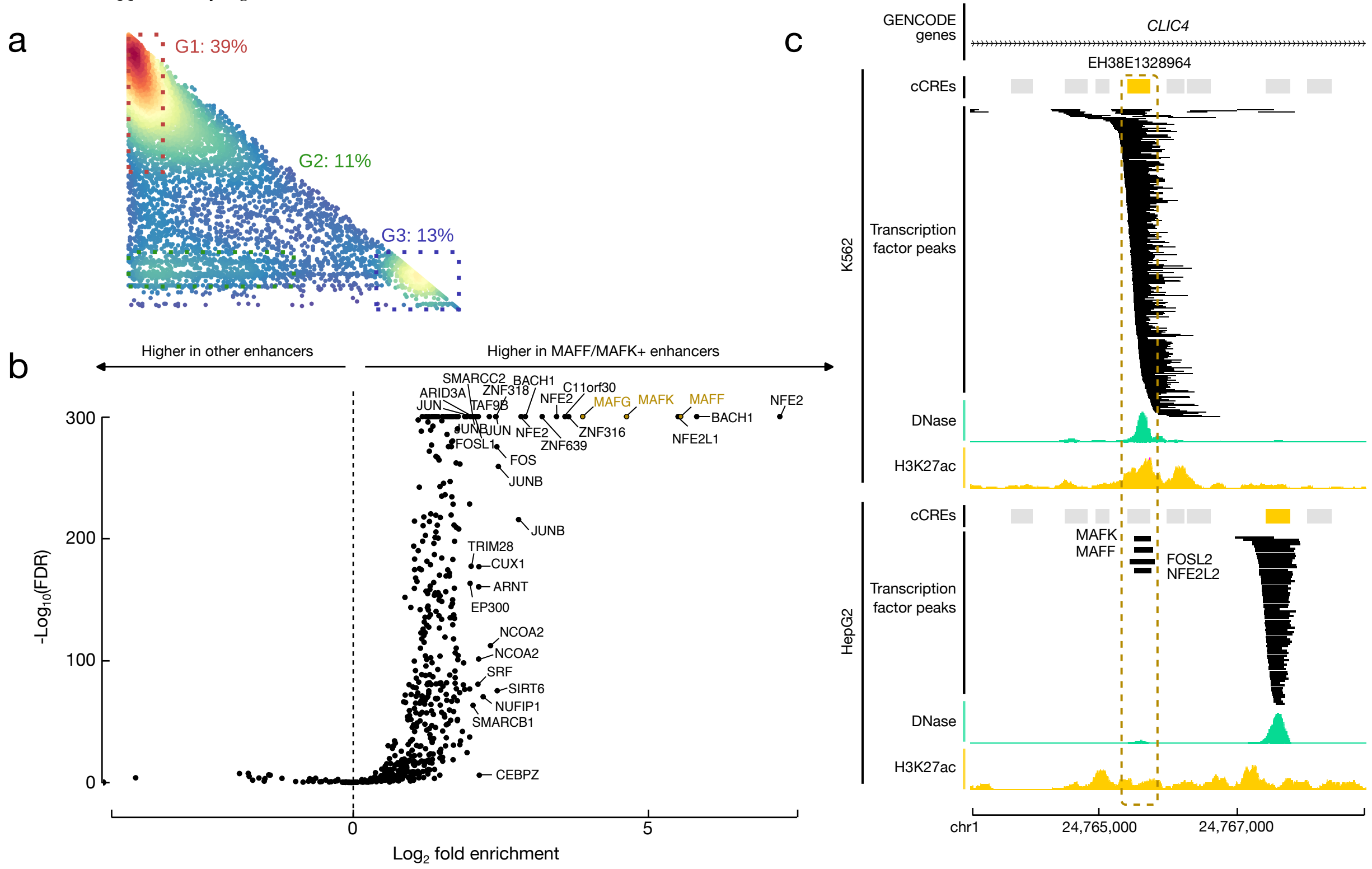
Supplementary Figure 18 | Comparisons with published silencers

a, Heatmap depicting shared cCREs across ENCODE silencer categories (REST+ enhancer/silencer, REST+ silencer, STARR silencer (stringent), and STARR silencer (robust) cCREs) and cCREs that overlap previously published silencer annotations from Cai *et al.*, Huan *et al.*, Jayavelu *et al.*, Pang *et al.* **b**, Barplots depicting the percentage of silencer cCREs tested by Jayavelu *et al.* that were deemed to have silencer activity in that study. Results for all silencer cCREs are in black, REST+ enhancer/silencer cCREs in yellow, REST+ silencer cCREs in purple, stringent STARR silencer cCREs in dark blue, and robust STARR silencer cCREs in light blue. **c**, Percentage of ENCODE silencer cCREs tested by Jayavelu *et al.* stratified as described in **b**. **d**, Percentage of ENCODE silencer cCREs with high DNase signal in K562 cells, stratified as described in **b**. **e**, Heatmap showing the fraction of each silencer cCRE group (as defined in **a**), K562 promoter, K562 distal enhancer, and cCREs with low chromatin accessibility in K562 that overlap K562 ChromHMM states. State details are described in **Supplementary Table 14c**. **f**, Heatscatter plot showing the distribution of cCREs that overlap Huan *et al.* silencers based on their alignment across Zoonomia mammalian genomes. N1 denotes the number of species in which $\geq 90\%$ of a cCRE aligns; N2 denotes the number of species in which $\leq 10\%$ of a cCRE aligns. Rectangles indicate predefined groups: Group 1 (G1), highly conserved ($N1 \geq 120$, $N2 \leq 25$); Group 2 (G2), actively evolving ($20 \leq N1 \leq 50$, $N2 \leq 120$); and Group 3 (G3), primate-specific ($N1 \leq 50$, $N2 \geq 180$). **g**, Heatscatter plot as in **f**, but for cCREs that overlap Cai *et al.* **h**, Heatscatter plot as in **f**, but for cCREs that overlap Pang *et al.* **i**, Heatscatter plot as in **f**, but for cCREs that overlap Jayavelu *et al.*



Supplementary Figure 19 | Defining MAFF/MAFK+ cCREs

a, Bar plot depicting the percentage of MAFF/MAFK+ cCREs in K562 belonging to each cCRE class. **b**, Bar plot depicting the percentage of K562 low chromatin accessibility (CA) cCREs (gray bar in **a**) belonging to each cell type-agnostic class. **c**, Bar plot as defined in **a** for MAFF/MAFK+ cCREs in HepG2. **d**, Bar plot as defined in **b** in HepG2. **e**, Venn diagram depicting the overlap of MAFF/MAFK+ cCREs in K562 (pink) and HepG2 (green). **f**, Venn diagram depicting the overlap of low CA MAFF/MAFK+ cCREs in K562 (pink) and HepG2 (green). **g**, Density plot of STARR scores for MAFF/MAFK+ cCREs (gold) compared to a background of all cCREs (black). *P*-value corresponds to a two sided Wilcoxon rank sum test. **h**, Nested violin-boxplots showing gene expression in K562 cells for the nearest gene of MAFF/MAFK+ cCREs with low CA in K562 versus genes near other low CA cCREs in K562. *P*-value corresponds to a two sided Wilcoxon rank sum test.



Supplementary Figure 20 | Characterizing MAFF/MAFK+ cCREs

a, , Heatscatter plot showing the distribution of MAFF/MAFK+ cCREs based on their alignment across Zoonomia mammalian genomes. N1 denotes the number of species in which $\geq 90\%$ of a cCRE aligns; N2 denotes the number of species in which $\leq 10\%$ of a cCRE aligns. Rectangles indicate predefined groups: Group 1 (G1), highly conserved ($N1 \geq 120$, $N2 \leq 25$); Group 2 (G2), actively evolving ($20 \leq N1 \leq 50$, $N2 \leq 120$); and Group 3 (G3), primate-specific ($N1 \leq 50$, $N2 \geq 180$). **b**, Scatter plot depicting the enrichment of transcription factor peak summits at MAFF/MAFK+ enhancers in K562 compared to other K562 enhancers. Top enriched transcription factors are labeled. **c**, Genome browser shot of MAFF/MAFK+ cCRE EH38E1328964 that is a distal enhancer in K562 (top) but has low chromatin accessibility in HepG2.

cCRE details

Visualizations

Applets

Biosample activity

cCRE Details

Biosample Activity

Linked Genes

Nearby Genomic Features

Orthologous cCREs in Other Species

Associated Gene Expression

Associated Transcript Expression

Functional Data

TF Motifs and Sequence Features

ChromHMM States

Configure UCSC Genome Browser

ENTEx

Table View

Genome Browser View

EH38E3291318

EH38E3291318

chr19:12,847,611-12,847,959

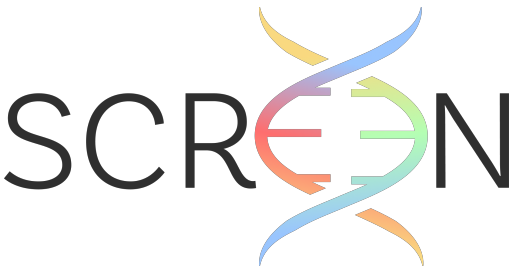
Cell type agnostic classification

Cell Type	DNase max-Z ↓	ATAC max-Z	H3K4me3 max-Z	H3K27ac max-Z	CTCF max-Z	Classification
cell type agnostic	5.39	4.46	3.57	2.45	8.19	Distal Enhancer

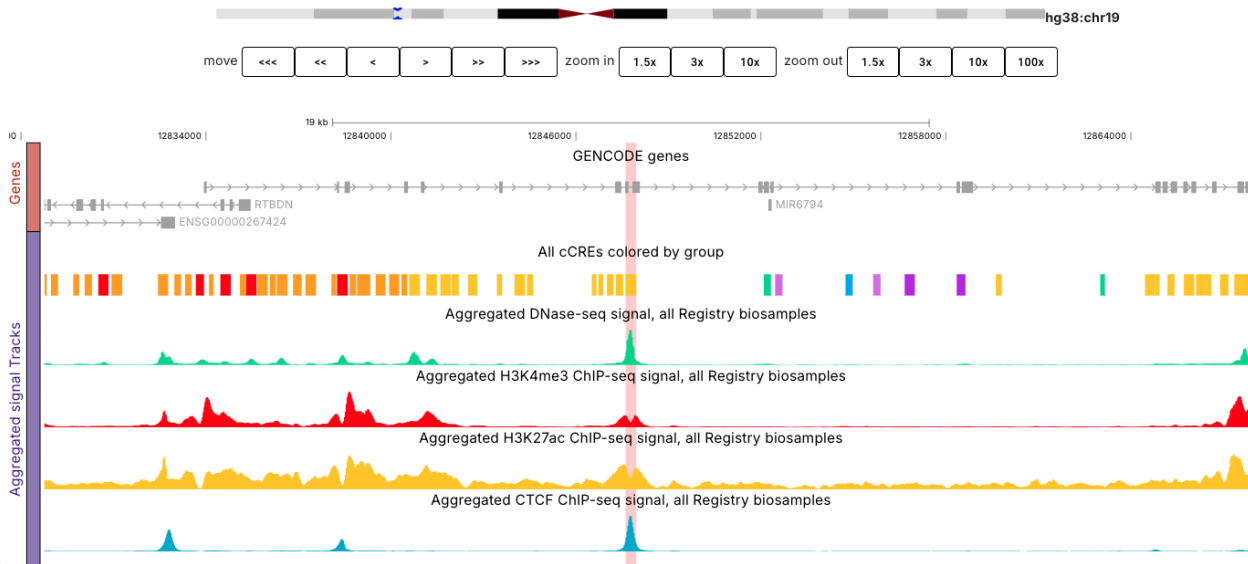
Core Collection

Cell Type	DNase Z-score ↓	ATAC Z-score	H3K4me3 Z-score	H3K27ac Z-score	CTCF Z-score	TF	Classification
CD14-positive monocyte, female	4.26	--	1.95	0.59	4.89	Yes	Chromatin Accessible with H3K4me3
hepatocyte (in vitro differentiated cells)	4.02	--	0.40	0.01	5.12	Yes	Chromatin Accessible with CTCF
endothelial cell (in vitro differentiated cells)	4.01	--	0.62	-0.63	4.79	Yes	Chromatin Accessible with CTCF
neural progenitor cell (in vitro differentiated cells)	3.79	--	1.36	-1.22	4.98	Yes	Chromatin Accessible with CTCF
K562	3.73	3.07	1.54	0.71	4.96	Yes	Chromatin Accessible with CTCF

Rows per page: 5 1-5 of 171



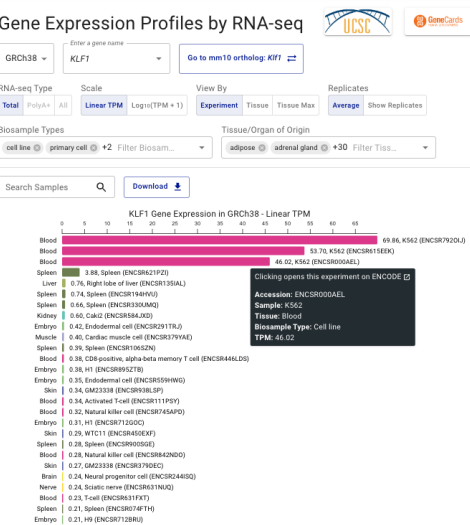
UMass Genome Browser



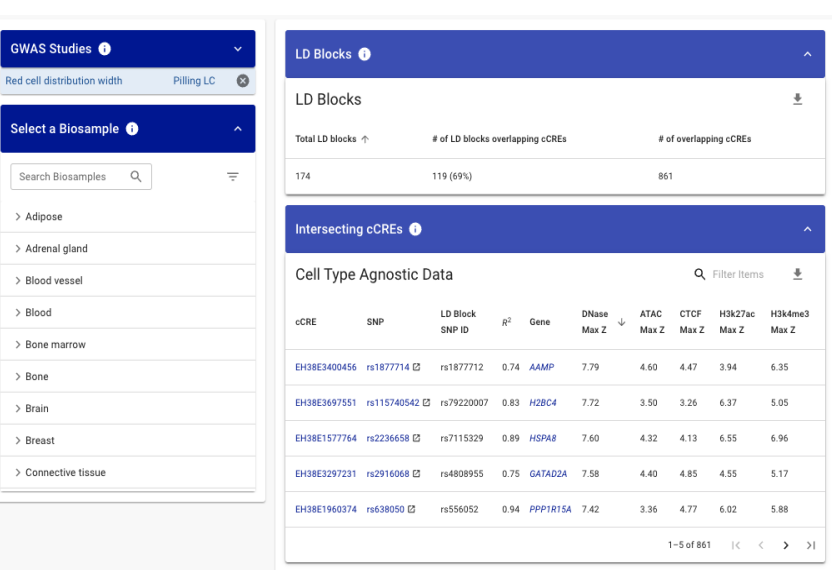
Interactive UMAP



Gene expression

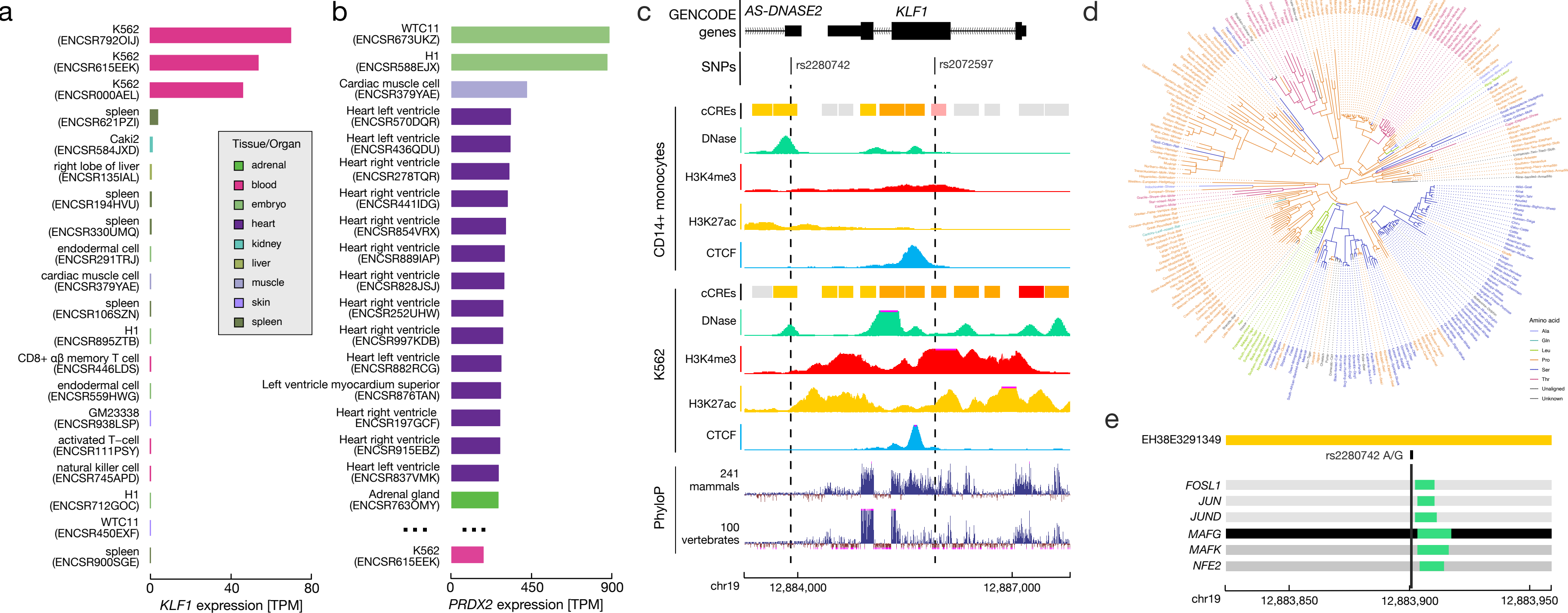


GWAS



Supplementary Figure 21 | SCREEN

SCREEN (Search candidate cis-Regulatory Elements by ENCODE) is an online resource for exploring and visualizing cCREs. Each cCRE has an element specific details page (top) that contains information about biosample activity and overlapping Encyclopedia annotations such as gene links, nearby gene expression, transcription factor binding sites, and conservation. SCREEN presents various visualizations (middle) including an embedded genome browser to enable seamless exploration along the genome (left) and interactive UMAPs to further explore relationships between biosamples and select subsets of samples. Finally, SCREEN hosts two applets gene expression (left) and GWAS (right).



Supplementary Figure 22 | Prioritizing genes and variants associated with red blood cell traits

a, Barplots depicting *KLF1* expression across the top 20 biosamples with highest expression in transcripts per million (TPM). Bars are colored by tissue/organ of origin as defined on SCREEN. **b**, Barplots depicting *PRDX2* expression in TPM across the top 18 biosamples with highest expression plus K562. Bars are colored by tissue/organ of origin as defined on SCREEN. **c**, Genome browser view of two variants rs228072 and rs2072597 in high linkage disequilibrium with variants associated with Red Blood Cell traits. Data for two cell types, CD14+ monocytes and K562 cells are shown including cell type specific cCREs and signals for DNase (green), H3K4me3 (red), H3K27ac (yellow) and CTCF (blue). On the bottom are PhyloP conservation scores from mammalian and vertebrate alignments. **d**, Circular dendrogram displaying phylogeny for 240 placental mammals from the Zoonomia project showing amino acid usage at position 102 of *KLF1* based on each species reference genome. **e**, Genome browser view of rs2280742 and overlapping cCRE EH38E3291349 (yellow) and transcription factor peaks with corresponding position of motif (green).

Supplementary Tables

- Supplementary Table 1 | Summary of datasets in Version 4 of the Registry of cCREs
- Supplementary Table 2 | Overlap of CRE-associated biochemical annotations and rDHSs
- Supplementary Table 3 | Multi-mapping cCREs
- Supplementary Table 4 | Aggregate level cCREs and Zoonomia conservation
- Supplementary Table 5 | Enrichment of Version 4-specific cCREs
- Supplementary Table 6 | Characterizing sequence features and biosample contributions of cCREs
- Supplementary Table 7 | Functional characterization of cCREs
- Supplementary Table 8 | Comparisons of whole genome STARR-seq libraries
- Supplementary Table 9 | Characterizing cell type and element-specific STARR-seq activity
- Supplementary Table 10 | Characteristics of REST+ cCREs
- Supplementary Table 11 | Characteristics of STARR-silencer cCREs
- Supplementary Table 12 | Comparison of silencer cCREs
- Supplementary Table 13 | Characteristics of MAFF/MAFK+ cCREs
- Supplementary Table 14 | ENCODE Encyclopedia Element Annotations
- Supplementary Table 15 | ENCODE Encyclopedia Transcription Annotations
- Supplementary Table 16 | ENCODE Encyclopedia Interactions
- Supplementary Table 17 | Identifying genes associated with red blood cell traits

Supplementary Data

- Supplementary Data 1 | ENCODE4 GRCh38 cCREs
- Supplementary Data 2 | ENCODE4 mm10 cCREs
- Supplementary Data 3 | CAPRA Quantifications (Solo)
- Supplementary Data 4 | CAPRA Quantifications (Double)
- Supplementary Data 5 | K562 MAFF/MAFK+ cCREs
- Supplementary Data 6 | HepG2 MAFF/MAFK+ cCREs