nature portfolio

Peer Review File

An Expanded Registry of Candidate cis-Regulatory Elements

Corresponding Author: Dr Jill Moore

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Attachments originally included by the reviewers as part of their assessment can be found at the end of this file.

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)
Summary:
Moore et al. expanded the current ENCODE registry of cCREs to almost triple the number of regions identified in previous iterations for both human and mouse. This included increasing the number of cCREs with functional characterization of histone marks and CTCF binding. Additionally, incorporation of reporter assay data allowed for identification and characterization of silencer elements, providing a template for further studying these elements. Increasing understanding of silencer elements and the features that define them is of particular interest to the field and are generally understudied compared to other elements. The characterization of dual enhancer/silencer roles of cCREs in different cellular contexts is exciting and only possible with the comprehensive dataset available through the expanded registry and functional assays it comprises. The manuscript also details improvements to the interactive interface SCREEN and demonstrates how the registry can be used to combine datasets and draw functional conclusions. Several case studies are explored in more detail to illustrate the utility of the expanded registry, particularly with respect to silencer elements.

There are some areas where clarification or revision is recommended:

1. As a general comment, this is truly an accomplishment in terms of team science, data collection and analysis. Given the breadth of data, it is certainly challenging to synthesize it all neatly into one concise manuscript. However, it was very cumbersome to go back and forth between the main text/figures, extended figures, and supplementary notes and figures. The supplementary notes were themselves a single manuscript. Related to this, I found some of the more intriguing and novel findings detailed in the supplementary notes to be “main text worthy”. These results include the comparison of conservation scores across the expanded registries, the autoencoder data and the use cases for SCREEN. Most of this is overshadowed by the focus on silencers, which could itself be a stand-alone manuscript.

2. In results section “cCREs are classified into putative functional categories based on biochemical signatures and genomic context”
• The definitions of different cCRE classes based on genomic marks could benefit from adding hypothesized functions, particularly for the new categories CA-TF, CA, and TF. How many of these regions (or what proportion) are defined this way because the sample does not have data for H3K27ac, H3K4me3, CTCF, or some combination of the three, and how many are genuinely only bound by a TF or only CA, etc.? And how can we functionally interpret these regions specifically? Most of the functional data analysis do not focus on these regions.
• Last paragraph: The authors highlight that mouse biosamples are lacking histone modification data, and that 65% of CA cCREs are likely to be classified as enhancers with H3K27ac data – what percentage would be expected in human cCREs lacking H3K27ac data?

3. Supplementary Figure 1C demonstrates that TF anchors that do not overlap rDHS sites are more conserved than non-cCREs and the median conservation scores from the 100-way vertebrate alignment are similar across TFs and rDHSs. However, Supplementary Figure 3C shows that different groups representing the expanded registry show different levels of mammalian conservation (as opposed to 100-way vertebrate alignment). These results are interesting and may reflect

differences in conservation of elements contained within these groups. Is this because their content is different in terms of major regulatory feature? For example, do the TF cCREs represent more recently evolved elements?

4. Related to the previous comment, multi-mapping cCREs map to recently annotated gene duplications, but another obvious explanation is transposable element insertions. This should be characterized for those regions.

5. What measures are in place to control for differences in data generated by different labs, in different samples, and in data analysis? How generalizable are the findings derived from specific datasets?
• Supplementary note 2.2 – The differences between two STARR-seq data sets from different labs is discussed, were STARR scores better correlated between two different STARR-seq experiments from the same lab? Is there intrinsic variability between experiments or are differences mainly due to different methodologies between labs?

6. Supplementary Note 1.5: "The 894,934 cCREs added to the Registry due to an increase in biosample coverage were also more likely to have STARR-seq activity." This statement is a little mis-leading, since the STARR-seq activity is measured in only one cell type (K562) and the percent of overlap is low. It is also unclear what is driving that activity, are these promoter elements or enhancers? Activators or silencers. Does this explain why a higher percentage of the ENCODE v3 elements have STARR-seq activity given the differences in element content compared to the expanded registry?

7. In results section "Testing cCRE activity using functional assays"
• Paragraph 2: Given that only 8-12% of STARR seq libraries are used after the overlap restrictions, can the rest of the data be useful in any way? What percentage overlaps 3 cCREs or portions of multiple cCREs that can be used to ask questions about the function of specific portions of each cCRE?
• Last paragraph: The investigation of STARR seq fragments with two cCREs revealing enhancer cooperativity is interesting. The hypothesis that increased diversity of TF binding leads to this is supported by ChIP-seq experiments, but were these enhancers tested in the CRISPR perturbation experiments as well? Is this phenomenon also observed in endogenous chromatin contexts?

8. Supplementary Figure 5 and notes discuss that cCREs were more likely to be functionally active than non-cCREs, that around 80% of STARR-seq active regions overlap cCREs and that 50% of the cCREs tested had activity in STARR-seq; but in terms of expectation, wouldn't many of the cCREs be transcriptionally inert (CA-CTCF, CA-TF, etc)? This needs some clarification. Related to this, for Supplementary Figure 6, in the case of "inactive" cCREs, do they overlap CA-CTCF or other transcriptionally silent class of cCREs?

9. Extended Figure 5, were there any STARR-seq (CAPRA) defined silencers that did not overlap cCREs? To what extent do you expect silencers are missed as a results of 1) not being chromatin accessible or 2) not being covered by available TF chip data? This question relates to the idea that some repressors bind inaccessible DNA.

10. Supplementary note 4.2:
• Paragraph starting "Peng et al." – potential typo of percentage in last sentence, reporting the amount of overlap between Jayavelu silencers and silencer cCREs that were also Pang elements, 11/624 would be ~1.7% rather than 93%
• Reports of silencer and cCRE overlap in published work could benefit from a visual representation – A Euler plot with overlaps between proposed silencers, cCRE silencers etc. would make this section much easier to digest
• Would processing raw MPRA/STARR-seq data through the same computational pipeline and applying the same silencer defining criteria mitigate the differences observed between these reports?
• None of the STARR-silencer cCREs groups were enriched for repressive chromatin signatures in K562 cells, but this is contradictory to findings in Ngan et al., Cai et al., Pang et al., Hansen et al, etc. This point is somewhat addressed in Supplementary note 4.2 but the idea that the silencer cCREs identified in this study are their own distinct class of novel silencers is an overstatement if only based on the fact that they do not overlap silencers identified in other studies, especially considering the technical and contextual differences in the assays used to defined them. Even if not "enriched", was there any overlap with heterochromatin histone marks or bivalent chromHMM states, for example?
Minor comments:
• Supplementary figure 1 legend: after part d, K563 and Hepg2 typos – K562 and HepG2
• Supplementary 5i and main Figure 2d are the same data? Same figure?
• Figure 2c is missing a reference in the text. For this figure, it would be helpful to understand the relationship between the STARR-score of a region and the CAPRA quantification of solo and double fragments.
• In this statement, "This cell type-specificity was most dramatic for CRISPR perturbation experiments that disrupt elements in their native chromatin context," is this not expected given that CRISPR targets are chosen a priori?
• Extended data figure 4f: no embryo shown in this panel, would it be possible to include an example of the negative result?
• In the last results section: Asserting that KLF1 is a causal gene might be overstating the result. The evidence that it is involved is strong but could use additional validation of the connection between KLF1 expression levels and these red blood cell traits. Consider revising the language here.
• Extended data figure 6d: is the x-axis cut off? Should it state % of silencer cCREs with high chromatin accessibility?
• Supplementary note 5.3: promoter cCREs have been identified for 90% of protein coding genes, how much of an increase is this from the previous cCRE registry?

(Remarks on code availability)

Referee #2

(Remarks to the Author)
Moore et al. present an expanded catalog of candidate cis-regulatory elements derived from a combination of existing ENCODE datasets and new datasets released in ENCODE4. The manuscript describes the utility of this new resource, with a number of broad analyses and specific anecdotes. In my opinion, the primary value of the work lies in the resource it creates, which I believe to be well developed. The authors provide extensive supplementary materials, analytical code, and make the resource broadly usable. In this regard, I feel that the manuscript excels. I do, however, have concerns about some of the results presented and suggestions for how the authors could further improve the manuscript. I detail these below in sections for major and minor critique. I apologize if any of my suggestions are difficult to find in the manuscript but line numbers were not provided in the copy I received (this would be helpful in future versions).

Major critiques:

1. I'm curious about the regions of low chromatin accessibility that show ChIP-seq signal for TFs. I wonder if the authors could explore this class of regions more in-depth to come up with a rationale for how this happens and whether this represents a distinct biological entity or a technical artifact.

2. The finding that GC content is highly different between protein coding promoters and distal enhancers is interesting. I wonder if the authors could provide concrete evidence that demonstrates that this does not coincide with higher density of CpG regions in the vicinity of promoters. Meaning, is this observation a reflection of the propensity of these regions to be methylated or is it a reflecting of some other GC-related mechanism such as changes to TF binding.

3. Does the CAPRA method for STARR-seq analysis somehow account for the length of the cCRE sequence? If not, wouldn't this confound the analysis (longer sequences have more chance to have more TF binding sites)?

4. The enrichment results presented in Figure 2e show that cell type-specific TFs are enriched in the distal enhancer cCREs with differentially high STARR seq scores across K562 and HepG2. What was the background set of regions for this enrichment test? It would seem like the most appropriate set would be the cCRE regions for the individual experiments. Otherwise, the enrichment signal might just represent the enrichment of TF motifs in the cCRE regions used for testing in the assay.

5. I feel that the authors have an opportunity to attempt to understand why certain sequences that should not have MPRA/STARR-seq activity do have activity. Could the authors use the tested regions that show "unexpected" activity and try to understand why they have that activity (motifs etc)? These assays have become widely used and such guidance could be very helpful in the design and analysis of future libraries.

6. I was somewhat confused by some of the writing in the section on REST. Firstly, the authors don't really define the REST-negative subset. Secondly, REST is meant to be active in non-neuronal cells but the authors state that "the REST+ subset of distal enhancer cCREs where more likely to be active in brain tissue…which was further supported by enrichment for H3K27ac signals in neuron-related biosamples". These results seem to make sense to me but only after connecting a lot of dots. If I understand it correctly, the authors are implying that in non-neuronal cells, the REST elements are repressed (no activity in the mouse embryo) but in neuronal cells, the same elements are active because that is what REST is supposed to do (suppress activity of regulatory elements which are normally active in neurons). I feel like the authors could be more explicit in this section to make this easier to follow. I think some of this is compounded by the last sentence of the paragraph where the authors present this as a new finding even though it seems to me that the underlying biology of REST necessitates this observation.

7. Related to the REST section, I don't really understand the difference in what is presented for the REST+ Distal Enhancers and the REST+ CA-TF regions. Why would REST bind to an element that isnt active in neurons? Is the conclusion (that the CA-TF regions act as strong silencers in REST-bound cell types) different from the conclusion on REST+ Distal Enhancers? In general, I'm finding it very hard to glean the underlying biology and meaning from this section, especially with regard to what is novel.

8. Could the authors put their silencer findings in the context of the entire gene regulatory landscape? How common are silencers compared to enhancers? Do they show increased or decreased levels of genetic variation compared to enhancers? Do they tend to be more "universal" / have the same effect across multiple cell types compared to distal enhancers which are highly cell type specific? Are they more frequently observed in tissues during development compared to adulthood (as implied by the text).

9. There are some visual complexities to the display of the RTBDN locus in both Ext Data Fig 5 and main Fig 4. Minimally, the directly to the right of RTBDN is labeled "MAST2" in one figure and "MAST1" in the other. MAST2 is on an entirely different chromosome. This raised some major alarm bells for me because I'm unsure how such an error occurs. There are also different naming conventions employed, in one plot antisense genes are labeled with an AS-* name and in the other they appear to use an ENSG name.

10. The KLF1 GWAS example leaves a little to be desired, especially given that the authors are amongst the world's experts in this type of biological analysis. In my opinion, this type of variant prioritization analysis in a paper like this should seek to set the standard in the field for plausible effects. The authors put forward rs2290688, stating that it overlaps a CA-CTCF cCRE and implying that it disrupts CTCF binding and thus re-wires 3D interactions in this locus. But does the variant reside

in a CTCF motif? If not, what is the proposed mechanism of action? The variant is present in ~30% of individuals (across many ancestries) and thus it seems likely that there are many publicly available datasets of chromatin accessibility (including in ENCODE) where the authors could look in data from blood cells to see if there is any evidence for allele-specific effects (i.e. allelic imbalance). Moreover, this variant and the others proposed as potentially impactful (rs2280742 and rs2072597) aren't shown on the track present in Fig 4a, nor are any 3D chromatin interaction tracks, which we know exist given what is shown in Ext Data Fig 5. Overall, I find this figure to be quite lacking and not to meet the standard in the field for compelling evidence of a GWAS mechanism. While I appreciate the citation of CRISPRi-FlowFISH data, it is unsurprising that CRISPRi of a CTCF site has an effect on gene expression and I feel that much more is needed here. My sentiments are compounded by the discussion of KLF1 as a "novel causal gene" which seems very overstated.

Minor Critiques:

11. Abstract – "STARR-seq, MPRA, CRISPR perturbation, and transgenic mouse assays now cover over
12. 90% of human cCREs". The word "cover" makes it sound like these assays are definitive when in fact most of them have been in immortalized cell lines whose activity may not reflect the primary cell types. Perhaps the authors could say "… assays have profiled over 90% of human cCREs"

13. Suggestion - Define "ENCODE" at first use

14. What portion of the cCREs annotated in ENCODE3 are captured in ENCODE4? Given that the methodology for identifying these elements is purportedly novel, I'm curious about the consistency with previous iterations.

15. Ext Data Fig 1 – It is unclear what the numbers in the upper right of the figure refer to and how/why they are different. ("12,800+ experiments across ENCODE & Roadmap Epigenomics" and "5,000+ experiments DNase, ATAC, H3K4me3, H3K27ac, CTCF"). I assume this represents the existing data (ENCODE3) and the data that is new to ENCODE4 but this could be made clearer.

16. First paragraph of results – you use the abbreviation "CRE" instead of "cCREs". Is this intentional?

17. The extensive Supplementary Notes are commendable. However, at least in my copy, the clickability of the table of contents (specifically the figures/tables/etc) did not work as expected. Given its size, please ensure this works to assist readers.

18. Sup Fig 5 – Include key for pink/black/grey colors in figure (rather than only explaining this in the legend)

19. Figure 2c is not cited in the text

20. One number that the authors don't report but would seem to me to be highly citable would be the percentage of the genome covered by the regulatory elements discovered in ENCODE4.

(Remarks on code availability)

Referee #3

(Remarks to the Author)
The manuscript by Moore, et al. details a significantly expanded resource of candidate cis-regulatory elements (cCREs) in human and mouse derived from ENCODE datasets, the vast majority of which have supporting data from one or more of four functional assays. Highlighted in this manuscript is their special focus on annotating silencers, which are particularly difficult to validate. Lastly, they demonstrate the utility of this resource researching GWAS variants. This is an extremely valuable and important resource that provides critical annotations to the genomics community. My concerns detailed below are not so much with the identification of these cCREs, but mainly with annotation decisions and some analyses, the lack of some annotations that would significantly enhance the use of the cCREs by the genomics community, and some overly strong claims. In general, the manuscript is well-written, but there are some inconsistencies that show it requires more scrutiny to represent a well-integrated whole, and there is a lack of description in some places that would make this difficult to interpret for scientists without specific expertise. Overall, though, this represents a valuable dataset that will be of great use to a broad spectrum of researchers.

I. Annotation scheme: I appreciate the added categories in this new version that provide more nuance in the function of different cCREs. I understand that this is challenging given the uneven distribution of both biosamples and data available for each biosample, but given that most regulatory elements, especially distal elements, have some form of tissue specificity, it seems important to try and capture that as much as possible, even given the limitations of the data.

1) It is disappointing that annotations are only available either at the aggregated level or at the individual biosample level. I believe most researchers will approach these annotations with a particular tissue of interest, and neither of these

annotations will provide accurate (aggregate) or easily summarized (biosample) information about their particular tissue. While it is understandable that some tissues do not have many biosamples and/or comprehensive data across experiments, being able to evaluate tissue-specific information vs the aggregate would be extremely valuable. An excellent example is the dual-function elements that are highlighted later in the paper.
2) Based on the description of how the aggregate annotations were assigned and the fact that the data is highly biased towards originating in a few tissues, especially blood, then the “aggregate” annotations seem like they will primarily reflect these most abundant tissue types. Was there any effort made to try and normalize for this imbalance in tissue? Maybe determine a tissue-specific signal, where possible, and then aggregate over these?
3) Somewhat related to the above, obviously these annotations are limited by the data available, both for the signals used to make these annotations (i.e. H3K4me3, H3K27ac) and uneven number of samples across tissues. There is some discussion about these limitations in reference to the mouse cCREs, but a discussion of how these limitations may affect the human annotations, and their accuracy, seems warranted
4) Given the emphasis on discovery of silencers in the paper, I find it curious that none of the annotations identify silencer elements. And there doesn’t seem to be a way in the Screen tool to known which elements have evidence of acting as a silencer. I understand that these span the annotation categories and that traditional repressive histone marks are not available for most samples, but it would seem providing that information would be valuable.
5) At the very least, I think it would be informative to know how ubiquitous a cCRE is across the tissues, even if only considering the open chromatin data. As the cCREs expand in number and span larger portions of the genome, it becomes more important to be able to see which are relevant to a particular tissue of interest. More on this below.
6) The GC content analysis is not surprising. I wonder whether this mainly reflects that promoters are more ubiquitously accessible, and high GC content reflects the DNA sequence environment that helps maintain their accessibility. If you were to, say, focus only on promoters and label these based on how ubiquitous they were across tissues, would you also see a similar relationship with GC content? More generally, is GC content more associated with how ubiquitous an element is rather than its annotated function?
7) The idea of a “novel TSS” is suddenly introduced in Ext Data Fig 2d without previously being mentioned or defined. “eRNAs” are likewise mentioned without definition of these data or annotations.
8) It would be great to know the distribution of sizes of the cCREs, perhaps broken down by annotation category, and the total genome coverage of cCREs.
9) In the ENCODE4 cCRE Pipeline, there are several heuristic parameters that understandably cannot be exhaustively evaluated but for which no real justification is provided nor a sense for how much or little they affect the annotations (i.e. DHSs are first filtered by width (>= 150 bp), signal (> than 10%tile), and FDR(<= 0.001); DHSs are first filtered by width (>= 150 bp), signal (> than 10%tile), and FDR(<= 0.001); We then select all rPeaks that overlap the summits of at least 5 peaks and do not overlap an rDHS). It would be great to show at least some bit of sensitivity analysis to indicate these parameter decisions are fairly robust.

II. Testing with functional assays: This was a monumental effort that is much appreciated. Interestingly, this section also emphasizes the cell-type specific nature of cCREs, especially distal elements. The paired cCRE analysis was interesting.

1) It states that 93% of cCREs were tested, but what percentage showed activity in at least one assay?
2) A clearer explanation of CAPRA is needed. Unless you are familiar with MPRAs, simply knowing how it calculates a p-value doesn’t help you understand the significance of the p-value. The Methods description does not shed light on this either.
3) Given the amount of TF ChIP-seq data generated for the K562 and HepG2 cells, why was this not used to investigate cell-specific STARR+ cCREs instead of just motifs?
4) This functional data is powerful validation, but it doesn’t seem to be integrated into the SCREEN application? It would be great for users to know which elements have been validated and using what assay.

III. Silencer analysis: These are difficult to functionally validate, and so a focus on this is greatly appreciated.

1) An intuition for how the REST+ cCREs were tested using “enhancer” transgenic assays is needed. Presumably, the hypothesis is that since REST is only active in neural cells, then you would not expect to see transgenic activity in non-neural cells, but you would in neurons.
2) Similarly for STARR-seq, since this is typically used to detect enhancers, how do you test for silencer activity? How do you calculate a p-value for these? This does not appear to be described in the main paper nor the supplementary notes.
3) Again it would be great to have results from these silencer assays incorporated into SCREEN.

IV: Disease associated genes

1) Since ENCODE datasets are so highly enriched in assays performed in blood cell types, it is unsurprising that the example hear is related to a blood cell trait. Given as well that only aggregate annotations are given without tissue information, it brings into question whether the current cCRE annotation is nearly as useful for non-blood cell traits as is demonstrated for a blood cell trait. While the example provided shows the power of what can be done when all of these types of data are available for a relevant tissue, unless it can be similarly demonstrated in a non-blood cell trait, or even attempted, then this limitation must be acknowledged.

V: Other comments

1) An obviously important characteristic of a regulatory element is its target gene(s). While I understand within ENCODE this is the focus of a different effort within the Interactions group, I note that some of the leaders of this effort are co-authors on this manuscript. It is disappointing that these target annotations are not better integrated within the SCREEN resource. Right now, SCREEN does have a “Linked Genes” drop down for each cCRE and provides evidence for links based on individual assays, but predictions from the ENCODE Enhancer-Gene predictions group (rE2G, ABC, GraphReg, EPIraction) are not integrated. As these efforts are part of the same consortium, and the Abstract of this paper even touts that “Integrating the Registry with other ENCODE annotations facilitates genetic variation interpretation and trait-associated gene identification”, there should be better integration across groups to more easily provide all relevant information in a single resource.
2) The analysis comparing the overlap in these new ENCODE cCREs with independently defined brain-specific regulatory elements does support the these cCREs are more comprehensive than the previous version, but this again is performed in a well-represented tissue in the datasets used. The statement “These results demonstrate that the ENCODE4 Registry is comprehensive while still maintaining biological specificity.” has some truth, but the comprehensiveness is necessarily uneven across tissues. This should either be acknowledged, or the comprehensiveness should be demonstrated in a larger range of tissues.

(Remarks on code availability)
We have reviewed the code, though not thoroughly. The authors should be commended for comprehensively providing all of their scripts, though overall these are poorly annotated, except for previously published software like CAPRA. The readme for the overall repository (https://github.com/weng-lab/ENCODE-cCREs/blob/master/README.md) is not informative.

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)
Thank you to the authors for their thoughtful and thorough responses to my comments and concerns. The extensive new analyses and clarifications in this revision were substantive and helpful in strengthening the findings and addressing gaps in the previous version. In particular the analysis of silencer annotations across studies was informative. I also appreciate the effort the authors made to create a companion website that would make navigation of the paper easier.

(Remarks on code availability)
I have reviewed the github for clarity in documentation and organization. The analyses for generated figures are clearly labeled and a comprehensive list of links to tools is provided.

Referee #2

(Remarks to the Author)
The authors have done an excellent job in responding to my initial critiques. I appreciate the in-depth responses to the questions relating to the sections on REST/silencers and on the restructuring of the GWAS example. I do not have any additional substantial concerns and believe this will be a great resource for the community.

(Remarks on code availability)

Referee #3

(Remarks to the Author)
The revised manuscript from Moore, et al, is significantly improved from the original submission. As stated previously, the most significant contribution of this work is a significantly expanded resource of candidate cis-regulatory elements (cCREs) in human and mouse derived from ENCODE datasets. A notable secondary contribution is the careful attention to and annotation of silencer elements, which are historically underrepresented in regulatory element databases. This revision adds additional supporting evidence for their resource and improved access to these annotations through their SCREEN online tool. All my previous concerns have been addressed, and I believe concerns of the other reviewers have likewise been addressed. I commend the authors on their thoroughness in this respect. I have no further concerns.

(Remarks on code availability)
I performed a cursory review of the code provided on their gitlab page. As the data required to replicate most of these analyses is too large to reasonably download, I did not run any of their code. The code, those, is pretty well documented, with scripts associated with the generation of figures present. The README for the overall pipeline and comments within the code provide sufficient detail for use by others if they would want to apply these to other datasets.

Referees' comments:

Referee #1 (Remarks to the Author):

Summary:
Moore et al. expanded the current ENCODE registry of cCREs to almost triple the number of regions identified in previous iterations for both human and mouse. This included increasing the number of cCREs with functional characterization of histone marks and CTCF binding. Additionally, incorporation of reporter assay data allowed for identification and characterization of silencer elements, providing a template for further studying these elements. Increasing understanding of silencer elements and the features that define them is of particular interest to the field and are generally understudied compared to other elements. The characterization of dual enhancer/silencer roles of cCREs in different cellular contexts is exciting and only possible with the comprehensive dataset available through the expanded registry and functional assays it comprises. The manuscript also details improvements to the interactive interface SCREEN and demonstrates how the registry can be used to combine datasets and draw functional conclusions. Several case studies are explored in more detail to illustrate the utility of the expanded registry, particularly with respect to silencer elements.

There are some areas where clarification or revision is recommended:

1. As a general comment, this is truly an accomplishment in terms of team science, data collection and analysis. Given the breadth of data, it is certainly challenging to synthesize it all neatly into one concise manuscript. However, it was very cumbersome to go back and forth between the main text/figures, extended figures, and supplementary notes and figures. The supplementary notes were themselves a single manuscript. Related to this, I found some of the more intriguing and novel findings detailed in the supplementary notes to be "main text worthy". These results include the comparison of conservation scores across the expanded registries, the autoencoder data and the use cases for SCREEN. Most of this is overshadowed by the focus on silencers, which could itself be a stand-alone manuscript.

Thank you for your thoughtful comments and for recognizing the scope and collaborative effort behind this work. We completely agree that synthesizing the breadth and depth of the data into a single manuscript posed a significant challenge, and we appreciate the reviewer's feedback on the difficulty of navigating between the main text, extended figures, and supplementary materials.

In response, we have made several edits to improve the clarity and flow of the manuscript, including adding additional cross-references between the main text and supplementary content to better guide readers to relevant analyses. To help readers more easily access and explore specific results, we also have developed a companion website (temporary URL: https://sites.google.com/view/encode4-ccres) that will collate all information related to a given

topic (e.g., conservation, silencers, or SCREEN use cases) in one place. This resource will supplement the manuscript and improve accessibility for both computational and experimental users.

2. In results section "cCREs are classified into putative functional categories based on biochemical signatures and genomic context"

- **The definitions of different cCRE classes based on genomic marks could benefit from adding hypothesized functions, particularly for the new categories CA-TF, CA, and TF. How many of these regions (or what proportion) are defined this way because the sample does not have data for H3K27ac, H3K4me3, CTCF, or some combination of the three, and how many are genuinely only bound by a TF or only CA, etc.? And how can we functionally interpret these regions specifically? Most of the functional data analysis do not focus on these regions.**

We appreciate your interest in the interpretation of new cCRE categories, particularly CA, CA-TF, and TF-only elements. Indeed, cell type-agnostic CA cCREs often reflect limitations in histone mark availability, particularly for H3K27ac, but CA-TF and TF cCREs should be confidently classified.

CA cCREs and biosample coverage
CA cCREs often reflect limitations in histone mark availability, particularly for H3K27ac. To distinguish true CA-only elements from putative enhancers with missing data, we performed a systematic analysis of cCREs called in biosamples lacking H3K27ac, H3K4me3, and CTCF. We then assessed how many of these biosample-specific CA cCREs were reclassified in the cell-type agnostic classification. We have updated **Supplementary Note 1.9** with this analysis and caveat.

Supplementary Note 1.9:

> *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 then 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%, Wilcoxon $p < 2.2 \times 10^{-16}$), 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 regions[123]. 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.

CA-TF and TF cCREs

Most biosamples (82%) with transcription factor data also had DNase and H2K27ac data; therefore, in most cases we are not calling CA-TF and TF cCREs in biosamples lacking promoter and enhancer associated chromatin marks. Therefore, these cCREs represent classes with high-confidence TF binding and chromatin accessibility (CA-TF) or TF binding alone (TF).

Furthermore, a substantial proportion of REST+ silencer and STARR silencer cCREs fall into these two categories, suggesting they play a key role in transcriptional repression. Additionally, many CA-TF and TF cCREs exhibit signatures of regulatory function—including enrichment for transcription factor motifs and elevated sequence conservation relative to background (**Supplementary Note 1.6)**. TF cCREs are also more likely to fall into primate-specific conservation groups, consistent with recent evolutionary emergence.

- **Last paragraph: The authors highlight that mouse biosamples are lacking histone modification data, and that 65% of CA cCREs are likely to be classified as enhancers with H3K27ac data – what percentage would be expected in human cCREs lacking H3K27ac data?**

We thank the reviewer for this important question. We have now performed a similar analysis in human biosamples lacking H3K27ac data, in addition to the biosample-specific analysis described above. Based on this updated evaluation, we estimate that approximately 67% (165,781/245,985) of CA cCREs in human could be reclassified as enhancers if H3K27ac data were available—consistent with our previous estimate in mouse. This finding, along with the corresponding caveats and implications for enhancer annotation, is now included in **Supplementary Note 1.9**.

3. Supplementary Figure 1C demonstrates that TF anchors that do not overlap rDHS sites are more conserved than non-cCREs and the median conservation scores from the 100-way vertebrate alignment are similar across TFs and rDHSs. However, Supplementary Figure 3C shows that different groups representing the expanded registry show different levels of mammalian conservation (as opposed to 100-way vertebrate alignment). These results are interesting and may reflect differences in conservation of elements contained within these groups. Is this because their content is different in terms of major regulatory features? For example, do the TF cCREs represent more recently evolved elements?

Thank you for this insightful comment. We have now systematically harmonized our conservation analysis across the manuscript using a three-pronged approach described in **Supplementary Note 1.6**. These methods include (1) average PhastCons scores from the 100-way vertebrate alignment, (2) positionally resolved conservation using PhyloP (both 100-way vertebrate and 240-way mammal), and (3) evolutionary breadth using Zoonomia-based alignment metrics (N1/N2; N1 denotes number of mammalian species that align ≥ 90% of a human cCRE, and N2 denotes number of mammalian species that align ≤10% of a human cCRE).

Indeed, TF cCREs represent more recently evolved elements. This is evident in the Zoonomia N1/N2 analysis, where TF cCREs are the most enriched (34%) for primate-specific (Group 3) conservation and the least represented among highly conserved elements. In contrast, promoter and proximal enhancer cCREs show much deeper conservation across mammals and vertebrates. These patterns suggest that the observed differences in conservation reflect the underlying regulatory content of each group, with TF cCREs likely representing a more dynamic and primate-specific component of the regulatory genome.

Supplementary Note 1.6:

> *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 ≤ 5, N2 ≥ 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.

4. Related to the previous comment, multi-mapping cCREs map to recently annotated gene duplications, but another obvious explanation is transposable element insertions. This should be characterized for those regions.

You are right! We intersected multi-mapping cCREs, unique mapping cCREs, and non-cCRE genomic regions with RepeatMasker annotations. Multi-mapping cCREs are significantly enriched for SINE elements relative to the other groups. Further analysis revealed that this enrichment is primarily driven by Alu elements, suggesting that many multi-mapping cCREs may have originated from transposable element insertions. We have added this analysis as new panels in **Supplementary Figure 2** and summarized the enrichment statistics in **Supplementary Table 3e**. The main text and Supplementary Note 1.2 have been updated accordingly as shown below:

Main Text

*We also used ATAC-seq data to annotate chromatin accessibility in biosamples lacking DNase-seq data. Finally, we made technical improvements to our pipeline to define 24,160 human and 40,537 mouse cCREs located in regions previously difficult to map, including recently duplicated genomic loci and repetitive elements such as Alus (**Supplementary Note 1.2, Supplementary Figure 2, Supplementary Table 3, Supplementary Data 3, 4**).*

Supplementary Note 1.2

*In addition to their enrichment near recently duplicated genes, multi-mapping cCREs were also more likely to overlap repetitive elements. To assess this, we intersected multi-mapping cCREs, uniquely mapping cCREs, and non-cCRE genomic regions with RepeatMasker annotations (**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%, Fisher's Exact Test, $p < 2.2 \times 10^{-16}$). This enrichment was primarily attributable to SINE elements, particularly Alu elements, which overlapped 26% of multi-mapping cCREs versus 13% of uniquely mapping cCREs (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.*

5. What measures are in place to control for differences in data generated by different labs, in different samples, and in data analysis? How generalizable are the findings derived from specific datasets?

We agree that technical variability is an inherent challenge in large-scale integrative genomics. To mitigate these effects and enhance generalizability, we have implemented several measures:

1. **Uniform processing**: All raw data incorporated into the Registry of cCREs is processed using standardized, uniform pipelines maintained by the ENCODE Data Coordination Center. This ensures consistency in peak calling, signal normalization, and quality control across experiments and biosamples.
2. **Centralized data generation**: For ENCODE4, assays were largely performed at a single production site per modality to reduce inter-lab variability and batch effects. For example, DNase-seq was generated by the John Stamatoyannopoulos lab (Altius), histone and CTCF ChIP-seq by the Brad Bernstein lab (Broad), and ATAC-seq by the Mike Snyder lab (Stanford), among others.
3. **Cross-sample normalization**: As described in our ENCODE3 publication, we apply a z-score normalization framework to place signals from different biosamples and assays on a comparable scale. This method helps control for technical variability and allows more accurate cross-sample comparisons.
4. **Cross-dataset validation:** To enhance generalizability, we focus our analyses on patterns supported by multiple biosamples and datasets. For instance, the enrichment of H3K27ac signal at REST+ enhancer/silencer cCREs is observed across multiple neuron and neuron-related biosamples, reinforcing the biological relevance of the finding. Similarly, although STARR-seq is more variable across experiments, we observe silencer activity reproducibly across independent datasets from different laboratories. For example key motifs such as GFI1B—a repressive transcription factor—are independently enriched at silencer cCREs across these datasets, further supporting their functional roles.

We acknowledge that some level of variability is unavoidable. To address this, we have added **Supplementary Note 1.9** where we outline and discuss limitations and considerations related to data heterogeneity and generalizability.

- **Supplementary note 2.2 – The differences between two STARR-seq data sets from different labs is discussed, were STARR scores better correlated between two different STARR-seq experiments from the same lab? Is there intrinsic variability between experiments or are differences mainly due to different methodologies between labs?**

Thank you for raising this important question regarding technical variability in STARR-seq data. To systematically evaluate the sources of variation, we performed a principal component analysis (PCA) on 362,591 cCREs with sufficient coverage across all seven STARR-seq DNA input libraries. These included six datasets from the White lab (K562, HepG2, MCF-7, HCT116,

A549, SH-SY5Y) and one from the Reddy lab (K562). Notably, several White lab libraries were generated using the same input DNA pool, allowing us to further dissect batch-specific effects.

Our results reveal both intrinsic and lab-specific sources of variability. The first principal component (PC1), which explained 32% of the variance, captured biological differences between cell types and DNA input batch effects. In contrast, the second principal component (PC2) explained 28% of the variance and primarily reflected inter-lab differences, including the strong GC-content bias we previously reported in the White lab datasets that was absent in the Reddy lab dataset.

These findings suggest that both intrinsic biological variation and methodological differences contribute to discrepancies between STARR-seq experiments. We now discuss these results in **Supplementary Note 2.3** and emphasize that primary comparisons should be made within the same experimental batch whenever possible.

6. Supplementary Note 1.5: "The 894,934 cCREs added to the Registry due to an increase in biosample coverage were also more likely to have STARR-seq activity." This statement is a little mis-leading, since the STARR-seq activity is measured in only one cell type (K562) and the percent of overlap is low. It is also unclear what is driving that activity, are these promoter elements or enhancers? Activators or silencers. Does this explain why a higher percentage of the ENCODE v3 elements have STARR-seq activity given the differences in element content compared to the expanded registry?

Yes, we can see how this statement is a bit confusing. Indeed, this STARR-seq analysis was only performed for one cell type (K562), which limits the extent to which we can interpret activity of cCREs added in the expanded registry. However, the STARR-seq assay is episomal and therefore can reveal elements with intrinsic enhancer potential even if they are not active in the native chromatin environment of K562. In fact, this feature enables us to assess the general functional potential of newly annotated cCREs beyond the specific chromatin context in which they were defined.

What we meant to convey was not that the newly added cCREs are active in K562 per se, but rather that they are more likely to show regulatory potential compared to randomly selected genomic regions. This analysis used lab-defined STARR-seq peaks so it only considered active regions (as opposed to silencers which are discussed later on in the manuscript). This enrichment provides supportive evidence that these new cCREs are not false positives, and instead harbor intrinsic regulatory activity.

Regarding the higher STARR-seq activity observed among cCREs that were present in both ENCODE3 and ENCODE4 ("replicated" category), we agree with your observation. These elements include a larger proportion of promoter cCREs (4% vs < 0.01% in other groups) and highly active enhancers, which tend to be more ubiquitously active across biosamples and more likely to have open chromatin in K562. They also include all cCREs originally identified in K562,

which may further explain their elevated STARR-seq activity. We have added cCRE class counts for each group to **Supplementary Table 5a**.

While we recognize the limitations of using a single-cell-type STARR-seq dataset, we focused on the fold-enrichment over background rather than absolute overlap. All cCRE categories showed ≥4.3-fold enrichment compared to non-cCRE genomic regions, suggesting that the expanded registry retains strong functional signals. We have clarified this point in **Supplementary Note 1.7** to better reflect the limitations and rationale behind this comparison.

Supplementary Note 1.7:

> *To assess the validity of the newly added cCREs, we examined their activity using STARR-seq and their 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.*

7. In results section "Testing cCRE activity using functional assays"

- **Paragraph 2: Given that only 8-12% of STARR seq libraries are used after the overlap restrictions, can the rest of the data be useful in any way? What percentage overlaps 3 cCREs or portions of multiple cCREs that can be used to ask questions about the function of specific portions of each cCRE?**

We agree that although only 8–12% of STARR-seq fragments overlap exactly one or two cCREs and are used in the primary CAPRA pipeline, the remaining fragments—particularly those that partially overlap cCREs or span multiple regulatory elements—can still be highly informative. Motivated by your suggestion, we further developed our CAPRA pipeline to take advantage of more fragments. Roughly 0.1% of fragments fully span three cCREs, and a substantial fraction (53% of fragments) overlap portions of individual cCREs or pairs, allowing investigation of the contribution of specific sequence features such as transcription factor motifs.

We extended CAPRA to include a new "paired sweep" analysis mode (**Supplementary Figure 11**) where we selected fragments anchored in one cCRE and examined changes in activity as fragment size increased to incorporate additional sequence from an adjacent cCRE. While a genome-wide, systematic analysis requires future method development, we showcase one proof-of-concept example in **Supplementary Note 2.2** and **Supplementary Figure 12f,** demonstrating how increases in STARR-seq signal correspond to transcription factor binding motifs.

Supplementary Note 2.2:

> *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 high-coverage cCRE pair EH38E2809119 and EH38E2809120 (**Supplementary Figure 12f**). In the forward sweep (from EH38E2809119 to EH38E2809120), the RNA:DNA ratio declined across the first element, then 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.

- **Last paragraph: The investigation of STARR seq fragments with two cCREs revealing enhancer cooperativity is interesting. The hypothesis that increased diversity of TF binding leads to this is supported by ChIP-seq experiments, but were these enhancers tested in the CRISPR perturbation experiments as well? Is this phenomenon also observed in endogenous chromatin contexts?**

Good question. The current set of CRISPR perturbation experiments in ENCODE4 primarily relied on single guide RNAs (gRNAs) targeting individual regions. As a result, we are unable to directly test cCRE cooperativity in the endogenous chromatin context using the available data. Moreover, the exact resolution of CRISPR interference (CRISPRi) at a given locus is not precisely defined, making it challenging to distinguish the effects of closely spaced elements such as those identified through CAPRA double fragment analysis.

Nevertheless, we agree that this is an important and interesting direction for future study. We are actively exploring the use of additional STARR-seq libraries and the design of multiplexed CRISPR perturbation experiments that can target combinations of cCREs with higher spatial resolution to systematically evaluate enhancer cooperativity in native chromatin contexts.

8. Supplementary Figure 5 and notes discuss that cCREs were more likely to be functionally active than non-cCREs, that around 80% of STARR-seq active regions overlap cCREs and that 50% of the cCREs tested had activity in STARR-seq; but in terms of expectation, wouldn't many of the cCREs be transcriptionally inert (CA-CTCF, CA-TF, etc)? This needs some clarification. Related to this, for Supplementary Figure 6, in the case of "inactive" cCREs, do they overlap CA-CTCF or other transcriptionally silent class of cCREs?

Thank you for raising this important point. As noted, some cCRE classes—particularly CA-CTCF, CA-TF, and CA—are not expected to exhibit enhancer activity in episomal reporter assays such as STARR-seq. In K562, we annotated 173,291 cCREs, of which 172,254 were tested in the STARR-seq assay. Among these, 12% were classified as promoter cCREs and 45% as enhancer cCREs, totaling 57%.

We examined the class composition of the 75,815 cCREs that overlapped STARR-seq peaks in K562 and found that promoter and enhancer cCREs accounted for 72% of active elements—a 1.7-fold enrichment relative to their overall representation. In contrast, other classes such as CA-H3K4me3 (4%), CA-CTCF (8%), CA-TF (15%), and CA (1%) were underrepresented, each showing more than two-fold depletion. These findings support the expectation that cCRE classes lacking enhancer-associated chromatin marks are less likely to be active in STARR-seq.

To facilitate these comparisons, we have added bar plots summarizing the cCRE class composition of active elements across STARR-seq datasets to **Supplementary Figure 10**. Additionally, to ensure consistency with other analyses in the manuscript, we now require that at least 50% of the cCRE overlaps a STARR-seq peak (rather than ≥1 bp), which slightly reduced the estimated proportion of active cCREs in K562 from 50% to 44%.

9. Extended Figure 5, were there any STARR-seq (CAPRA) defined silencers that did not overlap cCREs? To what extent do you expect silencers are missed as a results of 1) not being chromatin accessible or 2) not being covered by available TF chip data? This question relates to the idea that some repressors bind inaccessible DNA.

These are really good questions. Because CAPRA is designed to quantify STARR-seq activity at predefined cCREs, non-cCRE regions are not included in the standard pipeline. To address this directly, we performed an independent CAPRA analysis using size-matched, mappability-controlled genomic regions that do not overlap any annotated cCREs. This analysis identified 161 stringent (0.01%) and 1,743 robust (0.1%) non-cCRE silencers—approximately half the rate observed for silencer cCREs (549 stringent [0.02%] and 5,468 robust [0.2%]). Moreover, we found that non-cCRE silencers exhibited weaker repression on average (mean RNA/DNA log2 ratios of −1.15 and −0.86 for stringent and robust sets, respectively) compared to silencer cCREs (−1.33 and −1.02), suggesting that silencer cCREs are enriched for elements with stronger repressive activity. Motif analysis of non-cCRE silencers revealed enrichment for GFI1B, a repressor also highly enriched in cCRE-defined silencers. While the enrichment was less pronounced (15% vs. 20%), this shared signature suggests that some true silencers may indeed be excluded from the Registry due to a lack of chromatin accessibility or missing TF ChIP-seq data.

These results suggest that restricting CAPRA to cCREs may lead to false negatives, excluding bona fide silencers with weaker or context-specific chromatin signals. However, the silencers identified within the cCRE framework show stronger activity and are supported by orthogonal chromatin or transcription factor binding evidence, increasing confidence in their biological relevance. We have added this analysis to **Supplementary Note 4.7** and believe it provides an interesting direction for future studies—potentially aided by additional prioritization strategies such as repeat element filtering or evolutionary constraint.

10. Supplementary note 4.2:

- **Paragraph starting "Peng et al." – potential typo of percentage in last sentence, reporting the amount of overlap between Jayavelu silencers and silencer cCREs that were also Pang elements, 11/624 would be ~1.7% rather than 93%**

Thank you for catching our mistake. We have updated the text accordingly and apologize for the confusion.

- **Reports of silencer and cCRE overlap in published work could benefit from a visual representation – A Euler plot with overlaps between proposed silencers, cCRE silencers etc. would make this section much easier to digest**

Thank you for this helpful suggestion. We have now added a heatmap summarizing the pairwise overlap of silencer-annotated cCREs across our dataset and previously published studies. This approach enables clearer comparison across all groups and is now included as **Supplementary Figure 18a**.

- **Would processing raw MPRA/STARR-seq data through the same computational pipeline and applying the same silencer defining criteria mitigate the differences observed between these reports?**

Good question. While harmonizing computational pipelines can help minimize certain technical differences, in this case the discordance between silencer annotations primarily reflects differences in the *regions tested*, rather than differences in quantification or analysis strategy. For example, Jayavelu *et al.* tested a distinct library of candidate regions, with only ~3% overlapping our silencer cCREs. However, for the shared regions, we observed strong concordance—93% of our silencer cCREs were also classified as silencers by Jayavelu et al., despite the use of different pipelines (**Supplementary Note 4.2, Supplementary Figure 18b,c**).

In contrast, Pang *et al.* used FAIRE-seq peaks as input, which substantially limits overlap with our silencer sets. Many of our REST+ and STARR-seq silencers exhibit low chromatin accessibility and would not be included in the Pang *et al.* library. Thus, applying uniform pipelines would not resolve the observed differences, which stem largely from minimal overlap in the input libraries across these studies.

To further contextualize these differences, we expanded our comparative analysis to include epigenomic and evolutionary conservation data (**Supplementary Notes 4.5-4.6**). These results suggest that published studies are capturing distinct subclasses of silencers with different sequence and regulatory properties, rather than producing purely conflicting results.

- **None of the STARR-silencer cCREs groups were enriched for repressive chromatin signatures in K562 cells, but this is contradictory to findings in Ngan et al., Cai et al., Pang et al., Hansen et al, etc. This point is somewhat addressed in Supplementary note 4.2 but the idea that the silencer cCREs identified in this study are their own distinct class of novel silencers is an overstatement if only based on the fact that they do not overlap silencers identified in other studies, especially considering the technical and contextual differences in the assays used to defined them. Even if not "enriched", was there any overlap with heterochromatin histone marks or bivalent chromHMM states, for example?**

This is a fair point. We have revised the text to clarify that our silencer cCREs are not necessarily novel in function, but are a distinct subclass that are identified through criteria that differ from previous approaches. As you point out, differences in technical assays, biological

context, and silencer definitions contribute significantly to the limited overlap among silencer datasets.

To address the specific question of chromatin context, we conducted a detailed comparison of chromatin states in K562 cells using the ENCODE4 ChromHMM 15-state model which is now detailed in **Supplementary Note 4.5**, **Supplementary Figure 18e,** and **Supplementary Table 12c**.

Minor comments:

- **Supplementary figure 1 legend: after part d, K563 and Hepg2 typos – K562 and HepG2**

Sorry about this error. We have updated the figure legend accordingly.

- **Supplementary 5i and main Figure 2d are the same data? Same figure?**

Supplementary Figure 5i and Figure 2d were the same data plotted with different coloring schemes to highlight different aspects of the data. Supplementary Figure 5i (now **Supplementary Figure 12b**) is colored to show the density of points and to serve as a comparison with the promoters in **Supplementary Figure 12a**. Figure 2d (now **Figure 2a**) is colored to depict the enhancers that were used for the cell type-specific transcription factor motif analysis. We recognize that this may cause confusion so we have updated the figure legend of **Supplementary 12b** to clarify.

- **Figure 2c is missing a reference in the text. For this figure, it would be helpful to understand the relationship between the STARR-score of a region and the CAPRA quantification of solo and double fragments**.

We thank the reviewer for catching this oversight. Figure 2c has been reworked and moved to **Supplementary Figure 11** and provides a more detailed description of the CAPRA method. The "STARR-score" shown in the original figure is, in fact, the CAPRA-derived quantification of activity based on the relative abundance of RNA to DNA fragments. We apologize for the confusion and have updated the figure legend and manuscript text to use consistent terminology throughout.

- **In this statement, "This cell type-specificity was most dramatic for CRISPR perturbation experiments that disrupt elements in their native chromatin context," is this not expected given that CRISPR targets are chosen a priori?**

You are right. While it is true that many CRISPR-based experiments target regions selected a priori, the dataset we reference—Reilly et al., *Nature Genetics* 2021 ("Direct characterization of cis-regulatory elements and functional dissection of complex genetic associations using HCR–FlowFISH")—includes both targeted and unbiased approaches (**Supplementary Note**

2.1). Specifically, their design incorporated gRNAs selected based on chromatin marks as well as dense, unbiased tiling across broader regions.

To account for potential biases introduced by gRNA selection, we restricted our enrichment analysis to the subset of cCREs that were actually targeted in the CRISPRi assay (i.e., those overlapping at least one gRNA). This internal comparison—analogous to our approach for MPRA and STARR-seq analyses—ensures that enrichment is assessed only among tested elements. The pronounced enrichment observed for K562 cCREs likely reflects the distinctive advantage of CRISPRi in measuring regulatory activity within the native chromatin context, in contrast to episomal assays such as MPRAs and STARR-seq. We have updated the main text and **Supplementary Note 2.1** to clarify this comparison.

Main text:

> *Across all assays, cCREs were much more likely to be functionally active than non-cCRE genomic regions, and 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**).*

Supplementary Note 2.1:

> *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**).*

- **Extended data figure 4f: no embryo shown in this panel, would it be possible to include an example of the negative result?**

Yes! We have updated **Extended Data Figure 4f** to include an image of the embryo for the negative transgenic assay result.

- **In the last results section: Asserting that KLF1 is a causal gene might be overstating the result. The evidence that it is involved is strong but could use additional validation of the connection between KLF1 expression levels and these red blood cell traits. Consider revising the language here.**

Thank you for this suggestion. We have revised the text to more cautiously describe *KLF1* and *PRDX2* as candidate genes rather than asserting causality. The updated text emphasizes that *KLF1* is our top prioritized candidate based on its cell type-specific expression, established role in erythroid differentiation, and prior disease associations. We also discuss *PRDX2* as a

plausible contributor to mature red cell physiology. In response to your suggestion, we have also added additional evidence in support of *KLF1*, including prior reports of coding mutations linked to red blood cell disorders and new data showing allele-specific chromatin accessibility at multiple cCREs connected to *KLF1*. Together, these updates aim to present a more balanced interpretation of the locus and highlight how the Registry of cCREs can guide prioritization of candidates for future functional validation.

- **Extended data figure 6d: is the x-axis cut off? Should it state % of silencer cCREs with high chromatin accessibility?**

Thank you for pointing this out. We have updated the x-axis of Extended Data Figure 6d (now **Extended Data Figure 6a**) to be "% of silencer cCREs with high DNase signal"

- **Supplementary note 5.3: promoter cCREs have been identified for 90% of protein coding genes, how much of an increase is this from the previous cCRE registry?**

Good question. The previous registry covered the promoters of 83% of protein coding genes. We have added new analysis to **Supplementary Note 6.3** that quantifies the increase in annotated protein-coding gene promoters in ENCODE4 relative to ENCODE3 and describes the biological processes enriched among the newly covered genes. This analysis highlights how the addition of new biosamples—particularly neuronal and in vitro-derived cell types—has expanded promoter coverage across previously under-annotated loci.

Supplementary Note 6.3:

> *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.*

Referee #2 (Remarks to the Author):

Moore et al. present an expanded catalog of candidate cis-regulatory elements derived from a combination of existing ENCODE datasets and new datasets released in ENCODE4. The manuscript describes the utility of this new resource, with a number of broad analyses and specific anecdotes. In my opinion, the primary value of the work lies in the resource it creates, which I believe to be well developed. The authors provide extensive supplementary materials, analytical code, and make the resource broadly usable. In this regard, I feel that the manuscript excels. I do, however, have concerns about some of the results presented and suggestions for how the authors could further improve the manuscript. I detail these below in sections for major and minor critique. I apologize if any of my suggestions are difficult to find in the manuscript but line numbers were not provided in the copy I received (this would be helpful in future versions).

Thank you for your encouraging remarks! We have worked hard to address your critiques below. We have also provided line numbers in the revised manuscript.

Major critiques:

1. I'm curious about the regions of low chromatin accessibility that show ChIP-seq signal for TFs. I wonder if the authors could explore this class of regions more in-depth to come up with a rationale for how this happens and whether this represents a distinct biological entity or a technical artifact.

Thank you for this insightful question regarding regions of low chromatin accessibility that exhibit ChIP-seq signal for transcription factors. This is indeed an intriguing class of regulatory elements, and fully understanding their biological relevance will require in-depth computational and experimental follow-up.

As a preliminary investigation, we performed a focused analysis of MAFF/MAFK+ cCREs, a representative subset of low-accessibility, TF cCREs. Multiple lines of evidence suggest that they may be dynamic enhancers that respond to stress or stimuli. These findings are now included in the main text and described in detail in **Supplementary Note 6**.

Main Text:

> *Our analysis revealed that silencer cCREs were enriched within the newly defined CA-TF and TF cCRE classes. However, silencers accounted for only a subset of these groups, prompting further investigation into their broader regulatory potential. One subgroup of particular interest was the MAFF/MAFK+ TF cCREs, which inspired the formal inclusion of TF cCREs into the Registry, as 27% of MAFF and MAFK peaks do not overlap DNase hypersensitive sites (**Supplementary Note 1.1**).*

*We defined MAFF/MAFK+ cCREs in K562 and HepG2 by selecting cCREs that overlapped ChIP-seq peak summits for both transcription factors and analyzed these regions using our multi-modal framework (**Supplementary Note 5, Supplementary Data 7–8, Supplementary Figures 19a–f, Supplementary Table 13**). Although ~30% of MAFF/MAFK+ cCREs were classified as TF cCREs as they lacked detectable DNase hypersensitivity or H3K27ac enrichment in any of the surveyed biosamples (**Supplementary Table 13a**), the remaining 70% of MAFF/MAFK+ cCREs had high epigenetic signals in some biosamples, which provided clues to their possible functions. MAFF and MAFK have been associated with transcriptional repression[69,70]; however, MAFF/MAFK+ cCREs did not exhibit silencer activity based on STARR-seq or expression of nearby genes (**Supplementary Figures 19g,h**).*

*Instead, several lines of evidence suggest that MAFF/MAFK+ cCREs mark a class of latent, stimulus-responsive enhancers, characterized by persistent transcription factor binding and poised for activation under specific cellular conditions. First, MAFF/MAFK+ cCREs were significantly enriched near genes involved in development (e.g., limb, lung, and brain), morphogenesis, signal transduction, and transcriptional regulation (**Supplementary Note 6, Supplementary Table 13b,c**). Second, these elements also exhibited elevated DNase and H3K27ac signals in specific biosample subsets—including myeloid-derived primary cells and fibroblasts—contexts often associated with increased stress or external stimuli (**Supplementary Figure 20c,d, Supplementary Table 13d,e**). Third, conservation patterns of these cCREs were comparable to those of other enhancer classes, suggesting evolutionary pressure to maintain their regulatory potential (**Supplementary Figure 20, Supplementary Table 13f**).*

*To further explore their regulatory context, we compared MAFF/MAFK+ enhancer cCREs to other enhancer cCREs in K562 and HepG2. Transcription factors involved in stimulus-responsive signaling, such as NFE2, BACH1, NFE2L1, JUNB, JUN, and FOS, were significantly enriched in MAFF/MAFK+ enhancer cCREs. (1.7 - 4.2 fold enrichment, $p < 2.2 \times 10^{-16}$), **Supplementary Table 13g**). As a representative example, the MAFF/MAFK+ cCRE EH38E1328964 is classified as an enhancer in K562 but a TF cCRE in HepG2. In K562, this element is bound by 128 transcription factors, including MAFF, MAFK, MAFG, NFE2, NFE2L1, and AP-1 family members such as JUN, JUNB, JUND, and FOSL1 (**Supplementary Table 13h**). In contrast, HepG2 shows binding only by MAFF, MAFK, FOSL2, and NFE2L2, consistent with a latent enhancer state. In HepG2, EH38E1328964 is linked to its nearest protein-coding gene, CLIC4 (~20 kb upstream), via RNAPII ChIA-PET interactions. CLIC4 is a well-characterized stress-responsive gene, transcriptionally induced by oxidative stress, TGF-β, and DNA*

damage, and plays key roles in apoptosis, redox signaling, and transcriptional regulation[71–73].

Together, these findings highlight the regulatory potential of TF cCREs beyond silencing and demonstrate the utility of the Registry framework in uncovering context-specific enhancer classes. MAFF/MAFK+ cCREs exemplify a broader set of dynamic elements where transcription factors may act as sentinels at inaccessible chromatin, enabling rapid transcriptional responses to environmental or physiological cues. We hypothesize that the MAFF/MAFK+ TF cCREs may become active only under specific conditions or cell states, some of which not yet captured in available datasets. Future work will expand these analyses to identify additional subclasses of cCREs and connect predicted regulatory activities to transcription factor binding logic and cell state transitions.

2. The finding that GC content is highly different between protein coding promoters and distal enhancers is interesting. I wonder if the authors could provide concrete evidence that demonstrates that this does not coincide with higher density of CpG regions in the vicinity of promoters. Meaning, is this observation a reflection of the propensity of these regions to be methylated or is it a reflecting of some other GC-related mechanism such as changes to TF binding.

We thank the reviewer for this comment. We have now expanded our analysis to demonstrate that GC content correlates with transcription factor binding preferences across cCRE classes. Our new analysis of 95 sequence-specific transcription factors shows that transcription factors with GC-rich motifs preferentially bind promoter cCREs, while transcription factors with AT-rich motifs are enriched at distal enhancers. These results support the conclusion that GC content reflects meaningful biological differences in transcription factors binding specificity and regulatory element identity, beyond CpG methylation alone. This new analysis has been added to **Supplementary Note 1.8**.

Supplementary Note 1.8:

*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% (79 transcription factors across 122 experiments). This analysis revealed a moderate correlation between motif GC content and promoter preference (Pearson Correlation Coefficient = 0.48, $p = 1.6 \times 10^{-8}$), suggesting that sequence composition contributes to the genomic*

localization of TF binding events. Breaking down by quadrants revealed distinct classes of TFs with consistent sequence and binding preferences. Quadrant I (high GC content, promoter preference) included 79 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[121–124]. In contrast, Quadrant III (low GC content, enhancer preference) included 35 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.

3. Does the CAPRA method for STARR-seq analysis somehow account for the length of the cCRE sequence? If not, wouldn't this confound the analysis (longer sequences have more chance to have more TF binding sites)?

Thank you for asking this question. CAPRA does not normalize STARR scores by the length of the cCRE, as our goal is to quantify the total regulatory activity associated with each element. A longer cCRE does not necessarily have higher STARR-seq activity—activity is determined by the functional content of the sequence, not its length or number of transcription factor sites. Our analysis shows that the mere presence of transcription factor motifs is insufficient to predict regulatory activity. For example, in our combinatorial analysis, we identified cCRE pairs in which a single cCRE was active on its own but became repressed when tested in combination with another element. One such case involved two adjacent liver enhancers where one harbored motifs for repressors such as SMAD3, MAFK, and TCF7L2, which we hypothesize mediated the observed reduction in activity (**Supplementary Figure 14e,f**). Additionally, the new "paired sweep" mode of CAPRA enables us to examine how incremental sequence features—including specific transcription factor motifs—contribute to changes in reporter activity (**Supplementary Figure 12f**). We find that the magnitude and direction of activity shifts depend more on the identity of the binding sites than on the number of sites alone.

4. The enrichment results presented in Figure 2e show that cell type-specific TFs are enriched in the distal enhancer cCREs with differentially high STARR seq scores across K562 and HepG2. What was the background set of regions for this enrichment test? It would seem like the most appropriate set would be the cCRE regions for the individual experiments. Otherwise, the enrichment signal might just represent the enrichment of TF motifs in the cCRE regions used for testing in the assay.

We apologize for the confusion. The results in Figure 2e (now **Figure 2b**) were from analyzing two whole-genome STARR-seq experiments performed on K562 and HepG2 cells, respectively. In total, 1,088,305 distal enhancer cCREs were commonly assayed in these two experiments. We have added additional text to **Figure 2a** to make this distinction more clear. We derived STARR-seq scores via our CAPRA method for these cCREs in each cell line, and Figure 2e (now **Figure 2b**) is a scatterplot between these two scores for each cCRE. The HepG2 STARR+ (green) and K562 STARR+ (pink) enhancers in the figure were defined by comparing the STARR-seq scores in these two cell lines without using any additional background set of regions. Specifically, HepG2 STARR+ enhancers were defined as distal enhancer cCREs (among the 1,088,305 tested in both experiments) that had a positive STARR-seq score in HepG2 and a score at least 1 $\log_2$ units higher than in K562—corresponding to a ≥2-fold difference in normalized RNA:DNA ratio. Conversely, K562 STARR+ enhancers were those with a positive STARR-seq score in K562 that exceeded the HepG2 score by at least 1. The K562 and HepG2 STARR+ cCREs also showed strong enhancer-associated chromatin signatures in the matching cell type (**Supplementary Figure 13**).

5. I feel that the authors have an opportunity to attempt to understand why certain sequences that should not have MPRA/STARR-seq activity do have activity. Could the authors use the tested regions that show "unexpected" activity and try to understand why they have that activity (motifs etc)? These assays have become widely used and such guidance could be very helpful in the design and analysis of future libraries.

This is a great idea. We analyzed cCREs in K562 that showed statistically significant enhancer activity (CAPRA $p < 0.01$, positive STARR score), and divided them into two groups: (1) those with high chromatin accessibility in K562 (DNase z-score > 1.64), and (2) those without chromatin accessibility in K562. We then compared transcription factor motif enrichments between the two groups, and the closed-by-active cCREs are enriched in motifs of transcription factors associated with stress response and pioneer activity. We have updated the main text and added a supplementary note (**Supplementary Note 2.4)** and table (**Supplementary Table 9c**) with the results.

Main text

> *We also analyzed sets of cCREs that lacked chromatin accessibility in their native genomic context, but exhibited strong enhancer activity (i.e., high STARR scores,*

***Supplementary Note 2.4, Supplementary Table 9c**). These "closed-but-active" cCREs were enriched for motifs of transcription factors known or suspected to act in stress response (ATF4[48,49], CEBPG[49,50], SRF[51], and FOXO3[52]) or have pioneer-like functions (CEBPB[53] and CUX1[54]). These findings highlight the utility of episomal assays for uncovering sequence-intrinsic regulatory potential and may inform the design and interpretation of future reporter assays.*

6. I was somewhat confused by some of the writing in the section on REST. Firstly, the authors don't really define the REST-negative subset. Secondly, REST is meant to be active in non-neuronal cells but the authors state that "the REST+ subset of distal enhancer cCREs where more likely to be active in brain tissue…which was further supported by enrichment for H3K27ac signals in neuron-related biosamples". These results seem to make sense to me but only after connecting a lot of dots. If I understand it correctly, the authors are implying that in non-neuronal cells, the REST elements are repressed (no activity in the mouse embryo) but in neuronal cells, the same elements are active because that is what REST is supposed to do (suppress activity of regulatory elements which are normally active in neurons). I feel like the authors could be more explicit in this section to make this easier to follow. I think some of this is compounded by the last sentence of the paragraph where the authors present this as a new finding even though it seems to me that the underlying biology of REST necessitates this observation.

We apologize for the confusion caused by the original version of the REST silencer section. We have extensively revised and clarified this section of the manuscript to more explicitly describe our approach and biological conclusions.

Importantly, we now better emphasize the novel contribution of our analysis: the identification of two distinct classes of REST+ cCREs based on their functional profiles. Specifically, we distinguish between:

1. REST+ enhancer/silencer cCREs, which act as enhancers in mature neurons (which do not express REST) and as silencers in REST-expressing cells, and

2. REST+ silencer cCREs, which lack enhancer activity and function solely as silencers.

The REST sequence motifs in the REST+ enhancer/silencer cCREs are not bound by REST in neurons since REST is not expressed in these cells. Rather, these cCREs harbor the motifs for other activating transcription factors (AP-1 and E2F families) and are bound by these factors, which are expressed in neurons. In contrast, the REST+ silencer cCREs do not have this dual functionality. This classification, supported by transgenic enhancer assays, STARR-seq silencer activity, and gene expression data, reveals that REST+ cCREs do not comprise a homogeneous silencer class but instead include context-dependent regulatory elements with dual functionality.

Additionally, to specially address the reviewer's concern, we now define the REST– subset as cCREs in the same classes (e.g. enhancer) that are not a REST+ cCRE. This control group was used to contextualize the activity of REST+ cCREs in transgenic enhancer assays, chromatin accessibility data, and gene expression analyses.

Main text:

*Silencers are cis-regulatory elements that repress the transcription of target genes[55,56]. Several studies have identified putative human silencers using chromatin and functional data[57–62]; however, there is limited overlap among these datasets, likely due to differences in methodology (**Supplementary Table 10a, b**). To systematically identify silencers within a unified framework, we used the Registry of cCREs and started with a well-characterized subclass: neuron-restrictive silencer elements (NRSEs). NRSEs repress neuronal gene expression in non-neuronal cells[63–65] and are defined by the binding of the repressive transcription factor REST (RE1-Silencing Transcription Factor) to a specific DNA motif (**Figure 3a**). We hypothesized that by characterizing the biochemical and functional activity profiles of candidate NRSEs, we could derive generalizable principles to guide the broader identification of silencer elements.*

*To define candidate NRSEs—hereafter referred to as REST+ cCREs—we selected all non-promoter cCREs that contained a REST motif and overlapped the summits of at least five REST ChIP-seq peaks **(Figure 3a**, **Supplementary Note 3**, **Supplementary Table 10c**). This resulted in 5,850 REST+ cCREs, spanning all cCRE classes **Supplementary Figure 15a, b**, **Supplementary Table 10d**). We then evaluated these elements using multiple orthogonal data types: 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. By stratifying analyses by cCRE class (e.g., distal enhancer, CA-TF), we identified two major categories of REST+ cCREs (**Figure 3b**, **Supplementary Note 3**): (1) REST+ enhancer/silencer cCREs, which function as enhancers in neurons but act as silencers in non-neuronal cells, and (2) REST+ silencer cCREs, which lack enhancer activity and function exclusively as silencers outside the neuronal context.*

*REST+ enhancer/silencer cCREs exhibit context-dependent regulatory activity, functioning as enhancers in neurons and silencers in non-neuronal cells. These elements displayed comparable enhancer activity to other enhancer cCREs (such as those lacking REST binding, REST–) in transgenic mouse assays (59% and 61% validation rates, respectively, **Figure 3b,c**). However, REST+ enhancer/silencer cCREs were significantly more likely to be active in hindbrain and midbrain tissues (2.3 and 2.1 fold enrichments, Fisher's Exact Test, $p = 0.01, 0.04$, **Figure 3d, Supplementary Figure 15c, Supplementary Table 10f)**, consistent with enrichment for H3K27ac in neuron-related*

*biosamples (**Supplementary Note 3, Supplementary Table 10g,h**). In K562 cells, REST+ enhancer/silencer cCREs showed reduced STARR-seq activity (median = –0.10 vs –0.02 for all cCREs, Wilcoxon test, two-sided, $p < 2.2 \times 10^{-16}$, **Figure 3e, Supplementary Figure 15d**), suggesting moderate silencer function in this context. Genes near these cCREs also exhibited reduced expression relative to genes near cCREs lacking open chromatin (median TPM 0.7 vs 2.4, Wilcoxon test, two-sided, $p = 3.0 \times 10^{-6}$, **Supplementary Figure 15e**). Together, these results support a dual-function model in which REST+ cCREs act as neuronal enhancers in the absence of REST binding and as silencers in REST-expressing cell types.*

*REST+ silencer cCREs, on the other hand, lack enhancer activity and function as strong silencers in REST-expressing cell types. In transgenic mouse enhancer assays, only one REST+ silencer cCRE was active (8%), a validation rate significantly lower than that of REST+ enhancer/silencer cCREs (61%, p = 0.04) and class-matched REST– cCREs (40%, Fisher's Exact Test, p = 0.03; **Figure 3c, Supplementary Figure 15c, Supplementary Note 3**). In K562 cells, these elements exhibited strongly negative STARR-seq scores (median = –0.40), significantly lower than those of REST+ enhancer/silencer cCREs (Wilcoxon test, two-sided, $p < 2.2 \times 10^{-16}$; **Figure 3e**, **Supplementary Figure 15d**). Nearby genes also showed minimal expression (median TPM = 0.1, Wilcoxon test, two-sided, $p < 2.2 \times 10^{-16}$; **Supplementary Figure 15e**), suggesting repressive activity in native chromatin context.*

*We illustrate the two types of REST+ cCREs using the enhancer/silencer cCRE EH38E2130108 (**Extended Data Figure 4a**) and the silencer cCRE EH38E4127779 (**Extended Data Figure 4b**). Both cCREs have high REST ChIP-seq signals, moderate DNase signals, and low STARR signals (−0.64 and −0.94, respectively) in K562 cells (**Extended Data Figure 4c,d**). Only EH38E2130108 had enhancer activity, specifically in mouse forebrain, midbrain, and hindbrain (**Extended Data Figure 4e,f**), which was further supported by high H3K27ac and DNase signals in human in vitro-derived neurons. On the other hand, EH38E4127779 lacked enhancer activity in transgenic mice and enhancer signatures in neurons. Neither these example REST+ cCREs nor REST+ cCREs as a whole were enriched for any specific repressive chromatin signatures such as H3K9me3 or H3K27me3 in K562 cells (**Supplementary Table 10i**).*

*Finally we sought to determine distinguishing features between the two categories of REST+ cCREs (**Supplementary Note 3**). As expected, both categories were near genes enriched for neuronal and synaptic processes, but we did not observe any differential enrichment in specific terms between the classes (**Supplementary Table 10j**). REST+ enhancer/silencer cCREs were modestly enriched for enhancer-associated transcription factor motifs, including members of the AP-1, E2F, and MBD families, compared to*

*REST+ silencers (**Supplementary Table 10k**). On average REST+ silencer/enhancer cCREs were more evolutionarily conserved at both the center and in flanking sequences, while REST− silencers displayed a sharp peak of conservation at the center with substantially reduced conservation in flanking regions (**Extended Data Figure 5a**). This pattern, together with a higher proportion of REST+ silencers falling into primate-specific conservation groups (G3: 13% vs. 5% for REST+ enhancers, **Extended Data Figure 5b-e**), suggested potential repeat-associated sequence turnover.*

*To explore this, we examined the overlap of cCREs with repetitive elements. REST+ silencer cCREs had significantly higher overlap with long interspersed nuclear elements (LINEs) compared to REST+ enhancer/silencer cCREs (27% vs. 16%, Fisher's Exact Test, $p < 2.2 \times 10^{-16}$), particularly L1, which are known to be among the most active and evolutionarily dynamic transposable elements in the human genome[66](11% vs 3%, Fisher's Exact Test, $p < 2.2 \times 10^{-16}$, **Supplementary Table 10l**). This enrichment for L1s likely contributes to both the reduced conservation and broader sequence variability observed in REST+ silencer cCREs. These findings suggest that repetitive element insertion, particularly L1-mediated expansion, may underlie the evolutionary origin and sequence turnover of a subset of REST-regulated silencers.*

7. Related to the REST section, I don't really understand the difference in what is presented for the REST+ Distal Enhancers and the REST+ CA-TF regions. Why would REST bind to an element that isnt active in neurons? Is the conclusion (that the CA-TF regions act as strong silencers in REST-bound cell types) different from the conclusion on REST+ Distal Enhancers? In general, I'm finding it very hard to glean the underlying biology and meaning from this section, especially with regard to what is novel.

We appreciate this question and agree that the previous version of the text did not clearly delineate the distinct conclusions for REST+ distal enhancer and CA-TF cCREs. We have now revised the section extensively to clarify both the biological framework and the novelty of our findings (pages 13-15).

To address the reviewer's question directly: REST+ distal enhancer cCREs exhibit dual functionality, acting as enhancers in neurons (when REST is not expressed) and silencers in REST-expressing, non-neuronal cells. These cCREs show enhancer activity in brain tissues in transgenic mouse assays but low STARR-seq activity and reduced expression of nearby genes in non-neuronal cells like K562. We refer to this group as REST+ enhancer/silencer cCREs, highlighting their context-dependent dual functionality.

In contrast, REST+ CA-TF cCREs (and other REST+ non-enhancer cCRE classes, like CA-H3K4me3, CA-CTCF, and TF) show little to no enhancer activity in all cell and tissue types (even in neuronal contexts) and consistently low STARR-seq scores and gene expression in non-neuronal cells. These cCREs appear to act as strong silencers across all contexts and lack

the dual regulatory behavior observed in the REST+ enhancer/silencer group. Thus, while both groups exhibit silencer activity in REST-expressing cells, only the REST+ enhancer/silencer cCREs are active enhancers in neurons.

To clarify this distinction, we now explicitly describe these two classes as:

3. REST+ enhancer/silencer cCREs, with context-dependent activity (enhancer in neurons, silencer in REST+ cells), and

4. REST+ silencer cCREs, which lack enhancer activity altogether and act as silencers in REST+ cells.

This classification—and particularly the identification of REST+ enhancer/silencer cCREs—is a key novel contribution of our study. It highlights that REST+ elements are not functionally homogeneous but instead span a spectrum of regulatory behavior depending on cCRE class and cellular context.

We thank the reviewer again for highlighting the need for greater clarity. These revisions are now reflected in both the main text (pasted above) and **Supplementary Note 3**.

8. Could the authors put their silencer findings in the context of the entire gene regulatory landscape?

Thank you for this suggestion. We have now added a section to the main text further characterizing the silencer cCREs. We have also provided answers to your specific questions below.

- **How common are silencers compared to enhancers?**

Main text:

> *In total, we annotated 9,972 silencer cCREs—comprising both REST+ cCREs (N=4,787) and STARR silencer cCREs (N=5,468). Of these, 9,182 are predicted to be active in K562 cells (i.e, REST binding or significantly negative STARR score in K562). While silencers represent a smaller fraction of cCREs globally, they constitute a sizable class within a particular cell type. For comparison, in K562 we annotate 20,041 promoters, 41,813 proximal enhancers, and 35,488 distal enhancers. Thus, silencers make up a significant proportion of regulatory activity in this well-characterized cell line. One reason for the smaller global number is that our silencer annotations rely on STARR-seq data from K562; expanding functional assays across additional biosamples and assay sources will likely reveal more silencers and improve coverage across diverse cellular contexts.*

- **Do they show increased or decreased levels of genetic variation compared to enhancers?**

We thank the reviewer for this question. To evaluate genetic variation in silencer cCREs relative to enhancers, we analyzed the density of common human variants using dbSNP. Specifically, we computed the number of common dbSNP variants per base pair for each cCRE group. This metric provides a proxy for evolutionary constraint at the population level, with higher values suggesting increased tolerance for variation.

We found that REST+ enhancer/silencer cCREs exhibited slightly higher variant density than canonical enhancer classes (0.00372 vs. 0.00353–0.00368), but lower than CA-H3K4me3 and TF cCREs. In contrast, REST+ silencer cCREs (0.00398) and both STARR silencer classes (0.00381–0.00389) showed elevated variant densities, suggesting reduced constraint and a greater tolerance for human variation. We have added these results to **Supplementary Note 4.7**.

These results are consistent with our newly added evolutionary conservation and repeat content analyses (**Supplementary Notes 3 and 4.7**). Specifically, REST+ silencer and STARR silencer cCREs show reduced cross-species conservation and are enriched for LINEs, including L1s, which are associated with increased sequence turnover. Taken together, these findings support the conclusion that many silencers are under weaker purifying selection than enhancers and may tolerate greater human genetic variation.

- **Do they tend to be more “universal” / have the same effect across multiple cell types compared to distal enhancers which are highly cell type specific?**

This is an important question and one that will require further investigation in the future. Currently, our ability to systematically assess the universality or cell type-specificity of silencers remains limited by available functional datasets. For enhancers, activity breadth can be estimated using chromatin accessibility (e.g., DNase) or H3K27ac signal across biosamples in the Registry of cCREs. However, these epigenomic signatures are less informative for silencers, which often exhibit low chromatin accessibility and histone acetylation in the same cell types where they function. For example, only 17% of REST+ silencer cCREs show high DNase-seq signal in K562 despite robust REST ChIP-seq binding and evidence of silencing activity, consistent with their repressive function in closed chromatin contexts.

To partially address this question, we assessed the STARR-seq signal for silencer cCREs across six additional STARR-seq datasets from the White lab. We observed that silencer cCREs—including REST+ enhancer/silencer, REST+ silencers and STARR silencers—retained consistent negative STARR scores across multiple cell types. This suggests that, in contrast to enhancers, which typically show strong cell type-specificity, many silencers may exert more universal or broadly repressive activity.

That said, the REST+ silencer category is, by definition, enriched for cCREs with REST binding in at least five biosamples, leading to an overrepresentation of broadly active silencers. For other silencer classes, the observed consistency in STARR-seq signal across diverse cell types supports the hypothesis that silencers may be less restricted to specific cell types compared to enhancers. However, a more definitive assessment will require systematic silencer mapping using perturbation assays (e.g., CRISPRi) across a broader set of cell types.

We have included this new analysis in the main text and as a new supplementary note and figure (**Supplementary Note 4.1, Supplementary Figure 16d**).

Main text:

> *To assess reproducibility and cell type-specificity, we evaluated the activity of STARR silencers across STARR-seq datasets generated from a different lab across six different cell lines including K562 (**Supplementary Note 4.1**). Despite significant differences in dataset characteristics between labs (**Supplementary Note 2.3**), we observed consistent negative STARR scores for STARR silencer cCREs in the K562 dataset, as well as across other cell types (**Supplementary Figure 16e**). Stringent STARR silencers showed the strongest and most consistent repression, supporting the reproducibility of our silencer annotations across experiments. This broad silencing activity across diverse cell lines is consistent with the idea that these silencers repress genes involved in tissue-specific developmental processes such as neuronal and renal differentiation, which are inactive in most transformed cell lines.*

- **Are they more frequently observed in tissues during development compared to adulthood (as implied by the text).**

We appreciate the reviewer's question and the opportunity to clarify. While the manuscript discusses enrichment of chromatin accessibility at silencer cCREs in fetal biosamples, this does not imply that the silencers themselves are specifically active or functioning as silencers during development. Rather, we interpret this enrichment as indicative of a potential window of accessibility during cellular differentiation in which these regions may be accessible and available for transcription factor binding.

For example, although we annotate 9,182 K562 silencers based on REST binding and STARR-seq signal, which is corroborated by repression of nearby gene expression, only ~17% of these elements exhibit open chromatin in K562 as measured by DNase-seq. This is consistent with the role of silencers in establishing repressed or inaccessible chromatin states in mature cell types. The presence of DNase signals at these loci in fetal tissues suggests that they may have been accessible earlier in development, potentially enabling transcriptional repressors like REST to bind and initiate chromatin remodeling or gene repression programs.

Thus, our analysis of fetal samples serves to highlight a possible developmental trajectory—i.e., that silencer cCREs may be accessible transiently during differentiation, rather than implying that they are functionally active as silencers only during fetal life. We have updated the main text and expanded **Supplementary Note 4.3** to clarify this point.

Main text:

> *A major distinction of our silencer annotation lies in our cCRE-based approach. Previous annotations typically centered on open chromatin regions within the specific cell type where silencer activity was assayed or in a limited number of biosamples. Only 14% of our silencer cCREs show high chromatin accessibility in K562 cells (**Supplementary Note 4.4, Supplementary Figure 18d**). Instead, silencer cCREs have high chromatin accessibility in early-stage cell types—embryonic stem cells, induced pluripotent stem cells, and in vitro-derived progenitor cells—and fetal tissues (**Extended Data Figure 6a-c, Supplementary Note 4.4**, **Supplementary Table 12a, b**). Accessibility in fetal or progenitor cell types does not necessarily indicate that these cCREs function as silencers during development. Rather, we interpret this pattern as evidence of transient accessibility during development, potentially enabling transcriptional repressors such as REST to bind, close the chromatin subsequently, and thus establish long-term silencing programs in mature cell types. Therefore, by using chromatin accessibility measurements in a comprehensive collection of biosamples, we are able to identify silencers that would otherwise be missed.*

9. There are some visual complexities to the display of the RTBDN locus in both Ext Data Fig 5 and main Fig 4. Minimally, the directly to the right of RTBDN is labeled "MAST2" in one figure and "MAST1" in the other. MAST2 is on an entirely different chromosome. This raised some major alarm bells for me because I'm unsure how such an error occurs. There are also different naming conventions employed, in one plot antisense genes are labeled with an AS-* name and in the other they appear to use an ENSG name.

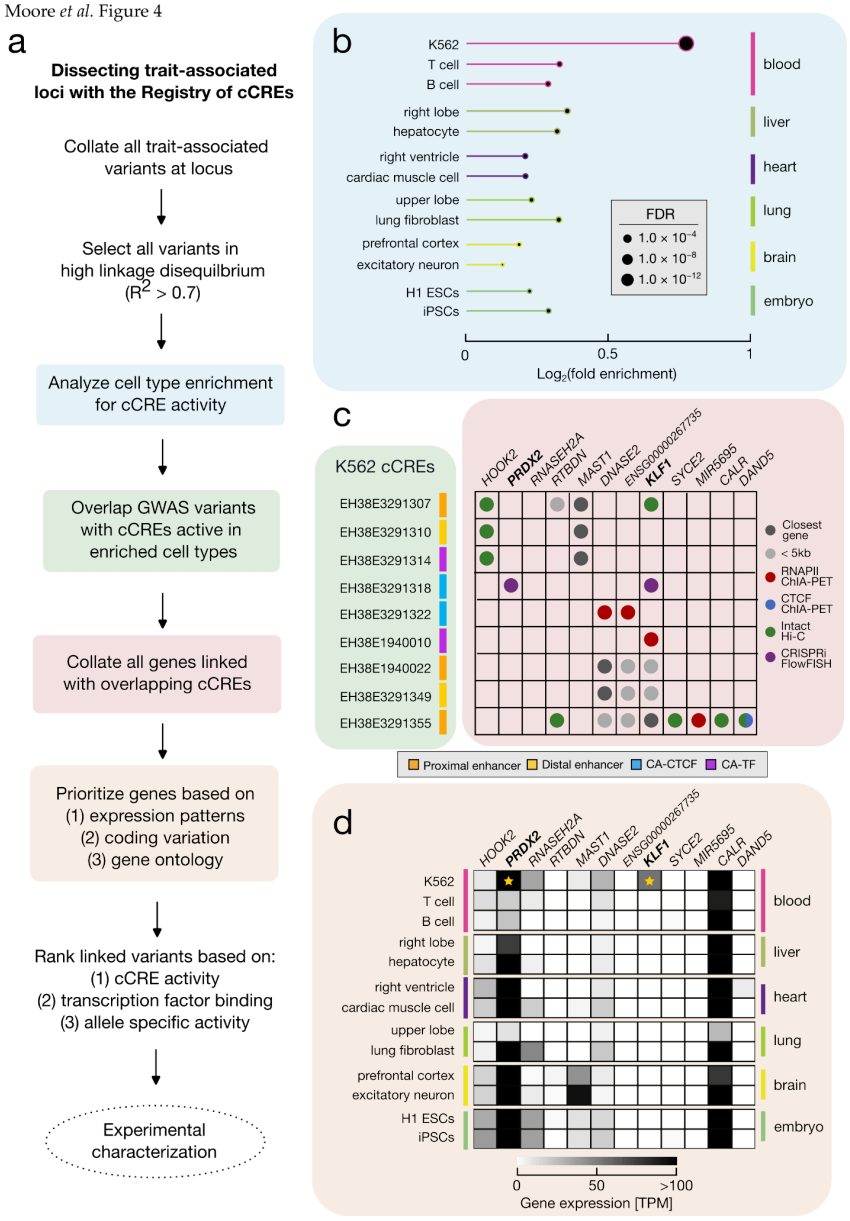
We thank the reviewer for carefully examining the figures and for catching this important oversight. The gene labeled "MAST2" in one of the figures was a typographical error during figure editing and should have read "MAST1." We have corrected this mistake and carefully reviewed all locus plots to ensure consistent and accurate labeling. In addition, we have standardized the gene naming conventions across all figures, using GENCODE V40 gene names throughout, to improve clarity and consistency. Please note that even with the updated and standardized GENCODE V40 naming convention, some lncRNAs will have an ENSG gene name while others have an AS- gene name.

10. The KLF1 GWAS example leaves a little to be desired, especially given that the authors are amongst the world's experts in this type of biological analysis. In my opinion,

this type of variant prioritization analysis in a paper like this should seek to set the standard in the field for plausible effects. The authors put forward rs2290688, stating that it overlaps a CA-CTCF cCRE and implying that it disrupts CTCF binding and thus re-wires 3D interactions in this locus. But does the variant reside in a CTCF motif? If not, what is the proposed mechanism of action? The variant is present in ~30% of individuals (across many ancestries) and thus it seems likely that there are many publicly available datasets of chromatin accessibility (including in ENCODE) where the authors could look in data from blood cells to see if there is any evidence for allele-specific effects (i.e. allelic imbalance). Moreover, this variant and the others proposed as potentially impactful (rs2280742 and rs2072597) aren't shown on the track present in Fig 4a, nor are any 3D chromatin interaction tracks, which we know exist given what is shown in Ext Data Fig 5. Overall, I find this figure to be quite lacking and not to meet the standard in the field for compelling evidence of a GWAS mechanism. While I appreciate the citation of CRISPRi-FlowFISH data, it is unsurprising that CRISPRi of a CTCF site has an effect on gene expression and I feel that much more is needed here. My sentiments are compounded by the discussion of KLF1 as a "novel causal gene" which seems very overstated.

Thank you for this detailed and constructive feedback. We agree that mechanistic interpretation of GWAS loci is complex and that setting high standards for variant prioritization is important. In response, we have revised the text throughout this section to soften the language around *KLF1*, which is now presented as a *prioritized candidate gene* rather than a "causal" one. We also explicitly acknowledge *PRDX2* as an additional candidate supported by genetic and functional evidence, and emphasize that our approach is intended to prioritize genes and regulatory elements for future functional testing rather than to make definitive causal claims.

We have also restructured the variant prioritization section to more clearly reflect the regulatory complexity of the *RTBDN-MAST1* locus. Rather than highlighting individual variants as definitive drivers, we now frame the analysis as a demonstration of how the Registry of cCREs can be used to generate hypotheses and guide variant prioritization in the context of multi-variant, multi-element loci (updated **Figure 4,** see below). This includes explicit acknowledgment that multiple variants may contribute to trait associations through distinct mechanisms, and that combinatorial perturbations may be required to resolve these effects.



Specifically, we scored all variants in high linkage disequilibrium (LD) with RBC GWAS lead variants based on six evidence categories: (1) coding sequence disruption; (2) allele-specific chromatin accessibility in erythroid progenitors; (3) allele-specific chromatin accessibility in other hematopoietic lineages; (4) evidence linking the variant to KLF1 (e.g., CRISPR perturbation, chromatin contacts, linear proximity); (5) transcription factor binding in K562; and (6) allele-specific TF binding in GM12878. Each category was weighted based on biological relevance, and we applied a rank-averaging method to prioritize variants with the strongest overall support (**Supplementary Note 8, Supplementary Table 17e–i**).

This approach identified three top candidates:

- **rs2072597**, a missense variant in *KLF1* that lies within a proximal enhancer cCRE and shows allele-specific chromatin accessibility in multiple hematopoietic cell types;

- **rs2280742**, located within a distal myeloid/erythroid enhancer cCRE near *KLF1*, with supporting evidence from both allele-specific chromatin accessibility and transcription factor binding;

- **rs2290688**, a synonymous variant within *MAST1* that overlaps a constitutive CTCF-anchored cCRE (EH38E3291318) and modulates *KLF1* and *PRDX2* expression in CRISPRi-FlowFISH experiments.

We now explicitly note that rs2290688 lies adjacent to—but does not disrupt—a canonical CTCF motif, and we propose that it may influence 3D chromatin architecture by modulating anchor cCRE activity rather than direct CTCF binding. We have expanded our figures to display all three variants and added 3D chromatin interaction tracks to support interpretation (updated Figure 4). We also now report allele-specific chromatin accessibility at rs2290688 in $CD8^+$ T cells, as well as similar effects at other *KLF1*-linked cCREs in erythroid progenitors, providing additional evidence for regulatory potential.

Overall, we now present this locus as a case study demonstrating how the Registry of cCREs can be used to generate testable hypotheses, rather than as a fully resolved mechanistic analysis. We believe this revised framing better matches the scope of the study and the goals of the manuscript, and thank you for your critiques.

Minor Critiques:

11. Abstract – "STARR-seq, MPRA, CRISPR perturbation, and transgenic mouse assays now cover over 90% of human cCREs". The word "cover" makes it sound like these assays are definitive when in fact most of them have been in immortalized cell lines whose activity may not reflect the primary cell types. Perhaps the authors could say "…assays have profiled over 90% of human cCREs"

Thank you for this feedback. We have updated the abstract accordingly:

Abstract Text

> *Functional characterization data from assays like STARR-seq, MPRA, CRISPR perturbation, and transgenic mouse assays have profiled over 90% of human cCREs, revealing complex regulatory functions.*

13. Suggestion - Define "ENCODE" at first use

Thank you for that suggestion. We have updated the main text accordingly.

Main Text

Over the years, the Encyclopedia of DNA Elements (ENCODE) project has made major contributions towards our understanding of gene regulation …

14. What portion of the cCREs annotated in ENCODE3 are captured in ENCODE4? Given that the methodology for identifying these elements is purportedly novel, I'm curious about the consistency with previous iterations.

The cCRE annotations from ENCODE3 are highly consistent with those in ENCODE4, with 98% of ENCODE3 cCREs retained. Although we have introduced several methodological updates for defining ENCODE4 cCREs—including the incorporation of transcription factor clusters for anchor selection and the addition of new cCRE classes—the core strategy for identifying rDHSs as anchors remained unchanged, which accounts for the high degree of retention between the two versions. We have expanded **Supplementary Note 1.7** and added **Supplementary Table 5a** to provide additional details.

Main Text:

*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.*

15. Ext Data Fig 1 – It is unclear what the numbers in the upper right of the figure refer to and how/why they are different. ("12,800+ experiments across ENCODE & Roadmap Epigenomics" and "5,000+ experiments DNase, ATAC, H3K4me3, H3K27ac, CTCF"). I assume this represents the existing data (ENCODE3) and the data that is new to ENCODE4 but this could be made clearer.

Thank you for this helpful feedback. The blue inset in **Extended Data Figure 1a** refers to the total number of experiments (12,800+) generated by the ENCODE and Roadmap Epigenomics consortia. Of these, a subset of ~5,000 experiments—specifically DNase-seq, ATAC-seq, H3K4me3, H3K27ac, and CTCF ChIP-seq—were used to generate the Registry of cCREs. We have revised the figure and legend to make this distinction clearer. Additionally, the blue inset corresponds to the "body map" figure from the ENCODE4 flagship paper, and our intention was to illustrate how the cCRE registry builds upon the broader ENCODE data production effort.

Extended Data Figure 1 Legend:

***a,** Body map denoting the number of biosamples in the ENCODE4 Registry of cCREs deriving from each major tissue or organ. Tissues/organs in bold denote that biosamples*

in this category are part of the core collection (having DNase, H3K4me3, H3K27ac, and CTCF), and the numbers indicate the number of biosamples belonging to each category. Smaller blue body map represents the entire data production effort of ENCODE and the Roadmap Epigenomics Project (12,800+ experiments) and highlights the subset of experiments used to create the Registry of cCREs (5,000+ DNase, ATAC-seq, H3K4me3, H3K27ac, and CTCF ChIP-seq experiments).

16. First paragraph of results – you use the abbreviation "CRE" instead of "cCREs". Is this intentional?

The use of CRE in the first paragraph of the results section refers to cis-regulatory elements in general. We acknowledge this might be confusing so we have updated the main text accordingly.

Main Text

*During ENCODE4, we leveraged new datasets and improved computational methods to broaden the scope and scale of the Registry of cCREs. The current Registry now encompasses 2,348,854 human cCREs and 926,843 mouse cCREs, covering approximately 21% of the human genome and 9% of the mouse genome, respectively. This represents a threefold increase compared to the ENCODE3 Registry, and establishes the ENCODE4 Registry as one of the most extensive repositories of cis-regulatory elements currently available (**Figure 1, Supplementary Note 1, Supplementary Data 1, 2**).*

17. The extensive Supplementary Notes are commendable. However, at least in my copy, the clickability of the table of contents (specifically the figures/tables/etc) did not work as expected. Given its size, please ensure this works to assist readers.

We sincerely apologize for the inconvenience and have ensured that the links in the supplementary materials work in this version of the resubmission.

18. Sup Fig 5 – Include key for pink/black/grey colors in figure (rather than only explaining this in the legend)

The color key for Supplementary Figure 5a–f (now **Supplementary Figure 10**) was originally included below the panels; however, to improve clarity and visibility, we have repositioned the key above the panels in the revised figure.

19. Figure 2c is not cited in the text

We thank the reviewer for pointing out this oversight. We have double checked to make sure all figures are now referenced in the Main Text. Previous panel Figure 2c has now been moved to **Supplementary Figure 11** which now specifically focuses on the CAPRA method.

20. One number that the authors don't report but would seem to me to be highly citable would be the percentage of the genome covered by the regulatory elements discovered in ENCODE4.

Thank you for this suggestion. We have added this calculation to the main text.

Main text:

> *During ENCODE4, we leveraged new datasets and improved computational methods to broaden the scope and scale of the Registry of cCREs. The current Registry now encompasses 2,348,854 human cCREs and 926,843 mouse cCREs, covering approximately 21% of the human genome and 9% of the mouse genome, respectively. This represents a threefold increase compared to the ENCODE3 Registry, and establishes the ENCODE4 Registry as one of the most extensive repositories of cis-regulatory elements currently available (**Figure 1, Supplementary Note 1, Supplementary Data 1, 2**).*

Referee #3 (Remarks to the Author):

The manuscript by Moore, et al. details a significantly expanded resource of candidate cis-regulatory elements (cCREs) in human and mouse derived from ENCODE datasets, the vast majority of which have supporting data from one or more of four functional assays. Highlighted in this manuscript is their special focus on annotating silencers, which are particularly difficult to validate. Lastly, they demonstrate the utility of this resource researching GWAS variants. This is an extremely valuable and important resource that provides critical annotations to the genomics community. My concerns detailed below are not so much with the identification of these cCREs, but mainly with annotation decisions and some analyses, the lack of some annotations that would significantly enhance the use of the cCREs by the genomics community, and some overly strong claims. In general, the manuscript is well-written, but there are some inconsistencies that show it requires more scrutiny to represent a well-integrated whole, and there is a lack of description in some places that would make this difficult to interpret for scientists without specific expertise. Overall, though, this represents a valuable dataset that will be of great use to a broad spectrum of researchers.

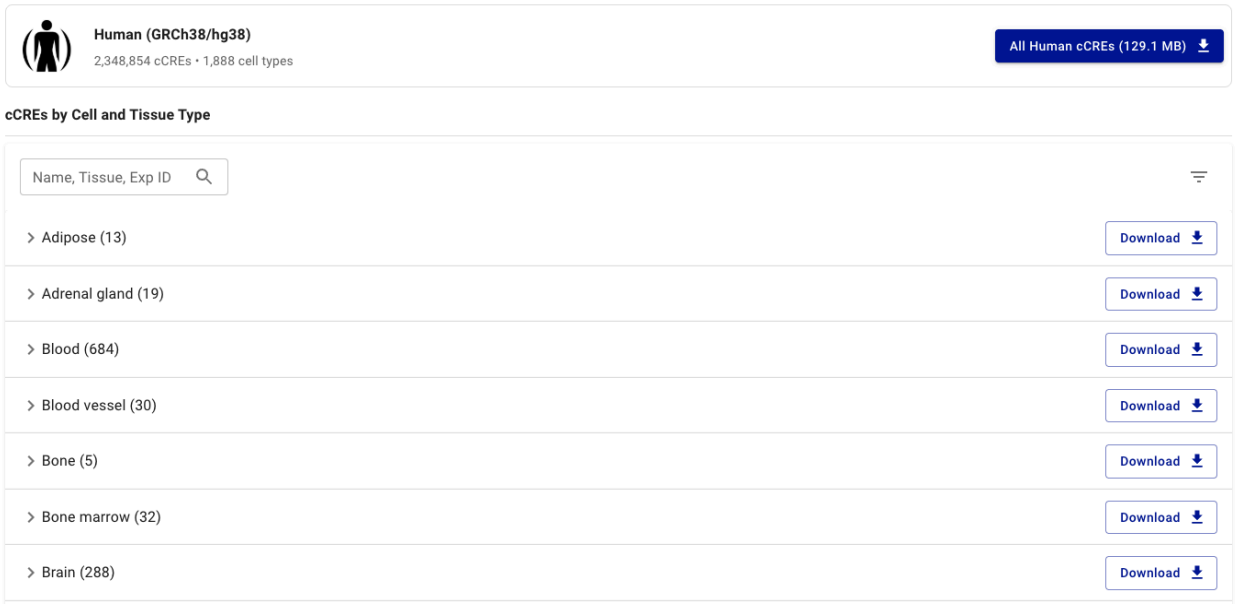
Thank you for your positive assessment of our work. We have made every effort to address your critiques below.

I. Annotation scheme: I appreciate the added categories in this new version that provide more nuance in the function of different cCREs. I understand that this is challenging given the uneven distribution of both biosamples and data available for each biosample, but given that most regulatory elements, especially distal elements, have some form of tissue specificity, it seems important to try and capture that as much as possible, even given the limitations of the data.

1) It is disappointing that annotations are only available either at the aggregated level or at the individual biosample level. I believe most researchers will approach these annotations with a particular tissue of interest, and neither of these annotations will provide accurate (aggregate) or easily summarized (biosample) information about their particular tissue. While it is understandable that some tissues do not have many biosamples and/or comprehensive data across experiments, being able to evaluate tissue-specific information vs the aggregate would be extremely valuable. An excellent example is the dual-function elements that are highlighted later in the paper.

Thank you for this excellent suggestion. We fully agree that tissue-level annotations offer valuable biological resolution that is often lost in cell type-agnostic analyses or hard to extract from individual biosample-level data. In response, we have now generated new collections of cCRE annotations that are summarized at the organ/tissue level (e.g. brain, heart, liver)

Our aggregation strategy is based on the consensus framework we previously developed to define brain cCREs (b-cCREs; Pratt et al., *Science Advances*, 2024) and is described in detail in **Supplementary Note 1.5** and **Supplementary Methods**. These tissue-level cCREs are easily downloadable through a newly designed interface on SCREEN as shown below.



2) Based on the description of how the aggregate annotations were assigned and the fact that the data is highly biased towards originating in a few tissues, especially blood, then the "aggregate" annotations seem like they will primarily reflect these most abundant tissue types. Was there any effort made to try and normalize for this imbalance in tissue? Maybe determine a tissue-specific signal, where possible, and then aggregate over these?

We appreciate the reviewer's question and the opportunity to clarify. The "aggregate" cell type-agnostic annotations referred to in the original manuscript are not generated via a consensus or majority-voting approach across biosamples or tissues. Rather, our cell type-agnostic classifications are based on the maximum Z-score across all individual biosamples. This strategy was intentionally designed to allow even sparsely represented or unique biosamples to contribute to the classification of regulatory elements.

For example, in our manuscript we investigate the silencer activity of EH38E4193243, which regulates RTBN and is classified as a proximal enhancer in our Registry despite exhibiting high H3K27ac signal in only 9 biosamples (<2%)—one of which is the WERI-Rb-1 retinoblastoma cell line. This would not have been possible under a consensus-based or tissue-aggregated classification scheme, where signals from more abundant tissues might overwhelm signals from rarer cell types. By using the maximum Z-score across all available biosamples, we ensure that individual biosamples retain the ability to define elements, even if they are not part of a larger group.

Importantly, all analyses in the manuscript, including signal enrichment and functional associations, are performed at the biosample level to avoid biases introduced by uneven data availability across tissues. We have expanded the supplementary notes to clarify this design choice and its rationale.

We also acknowledge that due to the composition of available biosamples in ENCODE, the depth and coverage of cCRE annotations vary by tissue. For instance, brain and blood have more comprehensive coverage due to ENCODE's targeted profiling efforts, while other tissues remain underrepresented. We now explicitly address these limitations, along with their implications for downstream interpretation, in **Supplementary Note 1.9**.

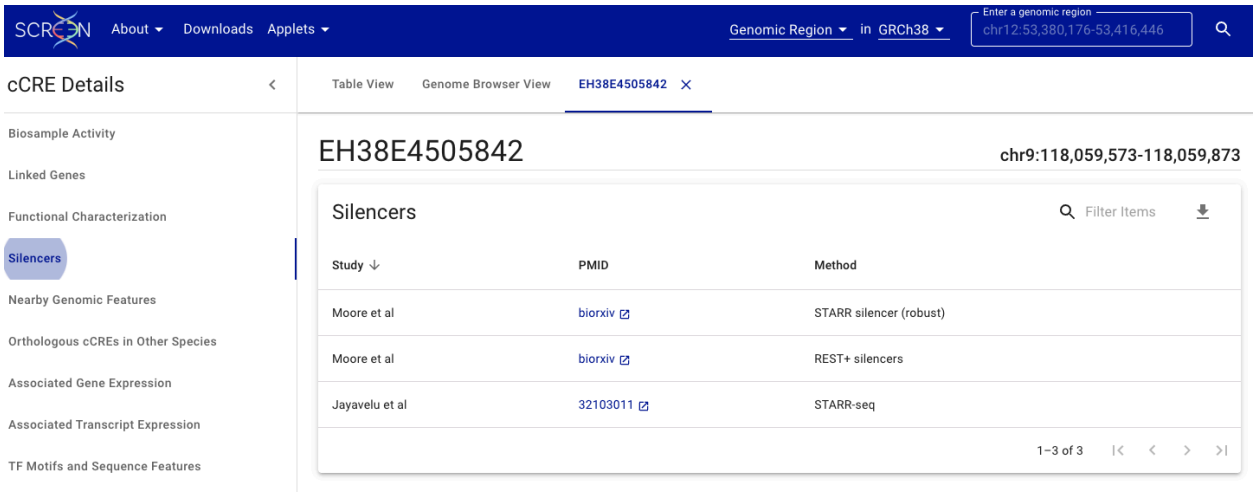
3) Somewhat related to the above, obviously these annotations are limited by the data available, both for the signals used to make these annotations (i.e. H3K4me3, H3K27ac) and uneven number of samples across tissues. There is some discussion about these limitations in reference to the mouse cCREs, but a discussion of how these limitations may affect the human annotations, and their accuracy, seems warranted.

We agree that limitations in data availability—both in terms of the types of assays and their distribution across tissues—impact the completeness of cCRE annotations. While the first submission of our manuscript discussed this for mouse, we have now expanded this discussion to address similar, albeit to a much lesser extent, caveats in the human dataset. Specifically, we evaluated how missing histone modification data (e.g., H3K27ac) can affect cCRE classification. We find that a substantial fraction of CA cCREs may represent putative enhancers cCREs that were not annotated as such due to missing data in the corresponding biosample. In human, we estimate that 67% of CA cCREs lacking H3K27ac data could be classified as enhancers if those data were available. These findings are now included in **Supplementary Note 1.9**, along with a systematic analysis of biosample representation and potential annotation rescue from closely related cell types. This analysis highlights that cCRE annotations are more robust in well-sampled tissues and cell types, and remain limited in sparsely profiled biological contexts.

4) Given the emphasis on discovery of silencers in the paper, I find it curious that none of the annotations identify silencer elements. And there doesn't seem to be a way in the Screen tool to known which elements have evidence of acting as a silencer. I understand that these span the annotation categories and that traditional repressive histone marks are not available for most samples, but it would seem providing that information would be valuable.

Thank you for this feedback. We completely agree that silencer annotations should have been more prominently featured in SCREEN. To address this, we added silencer-related information in two locations. First, on each cCRE's detail page, we now include a "Silencer" tab indicating whether the cCRE is annotated as a REST+ silencer or STARR+ silencer in this study, or if it has been previously identified as a silencer in other published datasets. Second, we have

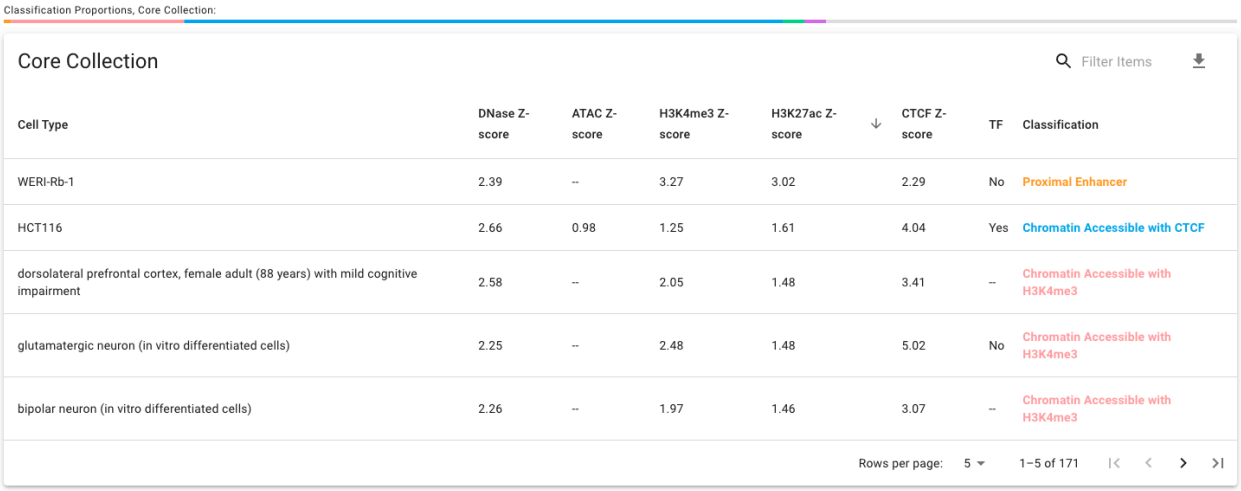
added a new section under the “Downloads” tab in SCREEN, where users can access lists of cCREs annotated as silencers in this study (REST+ and STARR-seq) as well as from external studies. Screenshot of this update is provided below for reference.



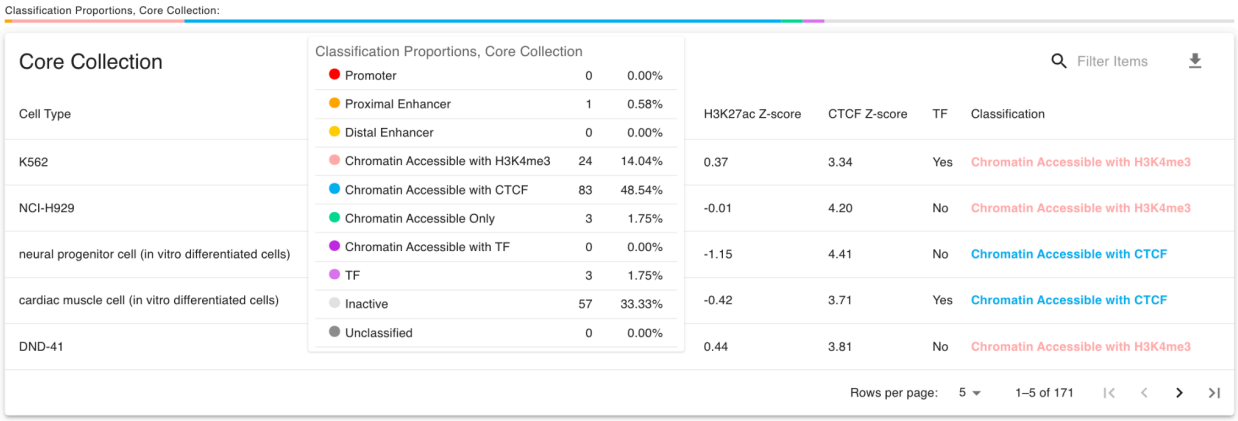
5) At the very least, I think it would be informative to know how ubiquitous a cCRE is across the tissues, even if only considering the open chromatin data. As the cCREs expand in number and span larger portions of the genome, it becomes more important to be able to see which are relevant to a particular tissue of interest. More on this below.

This is a fantastic suggestion. We have added two new "summary bars" for each cCRE in SCREEN to better represent biosample-specific activity. The first summarizes each cCRE’s classification across the 170 Core Collection biosamples. The second shows the fraction of the 1,154 Partial Data Collection biosamples in which the cCRE exhibits high DNase-seq signal. These additions are intended to give users a quick and informative view of cCRE ubiquity and activity across biosamples. We have included two sample screenshots of these bars below: the first with the static bar and the second displaying the hover-over feature which gives a summary of each category.

New bar chart w/ hover over selected:



New bar chart w/ hover over selected:



6) The GC content analysis is not surprising. I wonder whether this mainly reflects that promoters are more ubiquitously accessible, and high GC content reflects the DNA sequence environment that helps maintain their accessibility. If you were to, say, focus only on promoters and label these based on how ubiquitous they were across tissues, would you also see a similar relationship with GC content? More generally, is GC content more associated with how ubiquitous an element is rather than its annotated function?

We agree that GC content could reflect both regulatory function and ubiquity of activity, and we sought to disentangle these effects by stratifying promoter and enhancer cCREs and assessing their chromatin accessibility across biosamples. Our results indicate that while GC content is modestly correlated with ubiquity, particularly for enhancer cCREs, it does not fully explain observed sequence differences across functional categories. We have added these findings to **Supplementary Note 1.8** to clarify the relationship between GC content, transcriptional activity, and accessibility.

Supplementary Note 1.8:

*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.*

7) The idea of a "novel TSS" is suddenly introduced in Ext Data Fig 2d without previously being mentioned or defined. "eRNAs" are likewise mentioned without definition of these data or annotations.

Thank you for this feedback. We have updated the main text to clarify these annotations and added an additional reference to **Supplementary Note 1.8** where this analysis is detailed.

Main text:

*In each cell type, promoter and distal enhancer cCREs segregated along the first dimension (**Extended Data Figure 2b, Supplementary Figure 7**), which strongly correlated with the percentage of guanine and cytosine (GC) nucleotides of the sequences (R = 0.92, **Extended Data Figure 2c**). This GC-associated axis further distinguished promoters of protein-coding genes from those of lncRNAs, as well as subsets of distal enhancers—specifically, those overlapping the 5' ends of long-read RNA-seq transcripts (putative novel TSSs) and those overlapping PRO-cap peaks (candidate eRNAs) (**Extended Data Figure 2d, Supplementary Note 1.8**). These results suggest that GC content is a primary sequence feature underlying distinctions among transcribed cCREs, and that our classification scheme captures these differences. These results suggest that GC content is one of the primary sequence features differentiating transcribed cCREs and that our current classification scheme is able to detect such differences.*

Supplementary Note 1.6:

*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 Wilcoxon rank sum test with FDR correction, FDR < 2.2 × 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), and (3) all other distal enhancers (N=29,677). Enhancer cCREs that overlapped novel TSSs had the lowest dimension one values (5.5, FDR < 1 x 10^{-6}) followed by enhancer cCREs with eRNAs (6.4, FDR < 2.2 × 10^{-16}) and all other distal enhancer cCREs (8.3).

8) It would be great to know the distribution of sizes of the cCREs, perhaps broken down by annotation category, and the total genome coverage of cCREs.

Thank you for these suggestions. All cCREs range in size from 150 bp to 350 bp in width and we added **Supplementary Figure 3** to show the distributions of widths for each cCRE category.

We have also added text to **Supplementary Note 1.4** describing this data.

Supplementary Note 1.4

*In addition to their biochemical signatures, cCREs also differ in their genomic properties, including element size, which may reflect underlying biological features such as TF 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.*

We have also added the total genome coverage in the main text.

Main text:

*During ENCODE4, we leveraged new datasets and improved computational methods to broaden the scope and scale of the Registry of cCREs. The current Registry now encompasses 2,348,854 human cCREs and 926,843 mouse cCREs, covering approximately 21% of the human genome and 9% of the mouse genome, respectively. This represents a threefold increase compared to the ENCODE3 Registry, and establishes the ENCODE4 Registry as one of the most extensive repositories of cis-regulatory elements currently available (**Figure 1, Supplementary Note 1, Supplementary Data 1, 2**).*

9) In the ENCODE4 cCRE Pipeline, there are several heuristic parameters that understandably cannot be exhaustively evaluated but for which no real justification is provided nor a sense for how much or little they affect the annotations *(i.e. DHSs are first filtered by width (>= 150 bp), signal (> than 10%tile), and FDR(<= 0.001); DHSs are first filtered by width (>= 150 bp), signal (> than 10%tile), and FDR(<= 0.001); We then select all rPeaks that overlap the summits of at least 5 peaks and do not overlap an rDHS)*. It would be great to show at least some bit of sensitivity analysis to indicate these parameter decisions are fairly robust.

We appreciate your thoughtful comment regarding heuristic parameter choices in the cCRE pipeline. Many of these thresholds (e.g., DHS width ≥150 bp, signal >10th percentile, FDR ≤ 0.001) were originally selected during development of the ENCODE3 cCRE Registry (Moore et al., 2020), where we performed extensive internal benchmarking to balance sensitivity and specificity using chromatin marks, transcription start sites, and conservation as external validation criteria. To ensure consistency across versions of the Registry, we retained these parameters for ENCODE4.

To provide additional justification for an ENCODE4-specific heuristic—requiring ≥5 transcription factor peaks to define non-rDHS TF clusters—we analyzed conservation across TF factor clusters supported by different numbers of ChIP-seq peaks (1 to 20). We observed a clear elbow at five supporting transcription factors based on cluster count, a reversal in conservation trends favoring rDHS-overlapping clusters at this threshold, and a drop in great ape-specific content for non-rDHS peaks supported by fewer transcription factors. These findings provide empirical support for selecting ≥5 TFs as a conservative and robust cutoff, balancing specificity and evolutionary constraint. We have added this analysis to **Supplementary Note 1.1** and **Supplementary Figure 1e-g**.

Supplementary Note 1.1

> *We selected a threshold of five ChIP-seq peaks to define non-rDHS TF clusters by performing several empirical analyses aimed at balancing sensitivity and specificity. First, we generated an elbow plot of the number of rPeak clusters as a function of supporting transcription factors, observing a clear inflection point at five TFs, consistent with diminishing returns in signal-to-noise beyond this cutoff (**Supplementary Figure 1e**). Second we compared evolutionary conservation (PhastCons scores) for TF clusters that did or did not overlap rDHSs, stratified by the number of supporting transcription factor 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, but this pattern reversed at five or more transcription factors, where clusters 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 rPeaks 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 ≥5 TFs as a robust criterion for defining non-rDHS TF clusters, balancing evolutionary conservation, cross-species constraint, and regulatory signal complexity.

Additionally, to facilitate user-specific exploration of parameter thresholds, we retain all z-scores and intermediate signal values for each cCRE across biosamples, which are available via our downloads page and API. While the default z-score threshold of 1.64 was originally calibrated in ENCODE3 based on concordance with DNase peaks and chromatin state annotations, users can readily apply more stringent or lenient thresholds depending on their biological question or tolerance for false positives. This flexibility allows the Registry to support a wide range of downstream applications and analytical goals.

II. Testing with functional assays: This was a monumental effort that is much appreciated. Interestingly, this section also emphasizes the cell-type specific nature of cCREs, especially distal elements. The paired cCRE analysis was interesting.

Thank you for your encouraging remarks.

1) It states that 93% of cCREs were tested, but what percentage showed activity in at least one assay?

In deeply profiled cell types such as K562, we found that 91% of promoter cCREs and 65% of enhancer cCREs exhibited significant activity in at least one functional assay. However, because comprehensive functional characterization is only available for a limited number of cell types, the global validation rate across all cCREs is lower—28% of tested cCREs show evidence of activity in at least one assay.

We have summarized this in a newly added table, **Table 1**, and have added a detailed tally:

Table 1 | Summary of cCREs tested by ENCODE4 functional characterization assays.

Assay	# surveyed cCREs	# of cCREs with activity*	# experiments	# cell types/tissues
STARR-seq	2,277,726	567,059	6	6
MPRA	193,658	79,321	62	6
CRISPR perturbations	227,169	25,362	83	15
Transgenic mouse enhancer assays	6,063	3,590	1,947	22
Total	**2,289,918**	**621,213**	**2,100**	**39**

We have updated the main test accordingly and have included both STARR-seq peaks and CAPRA quantifications as part of our testing calculation resulting in 97% of cCRE being tested.

Main text:

> *Nearly all of the human cCREs (97%) were tested by at least one assay in at least one cellular context, with 28% having significant activity in at least one assay (**Table 1, Supplementary Note 2.1, Supplementary Table 7**). This overall rate likely underestimates the true fraction of functional cCREs, as most assays were performed in a limited number of cell types and are biased toward detecting enhancer activity; for example, in K562—where the most extensive data are available—91% of promoter cCREs and 65% of enhancer cCREs show significant activity.*

2) A clearer explanation of CAPRA is needed. Unless you are familiar with MPRAs, simply knowing how it calculates a p-value doesn't help you understand the significance of the p-value. The Methods description does not shed light on this either.

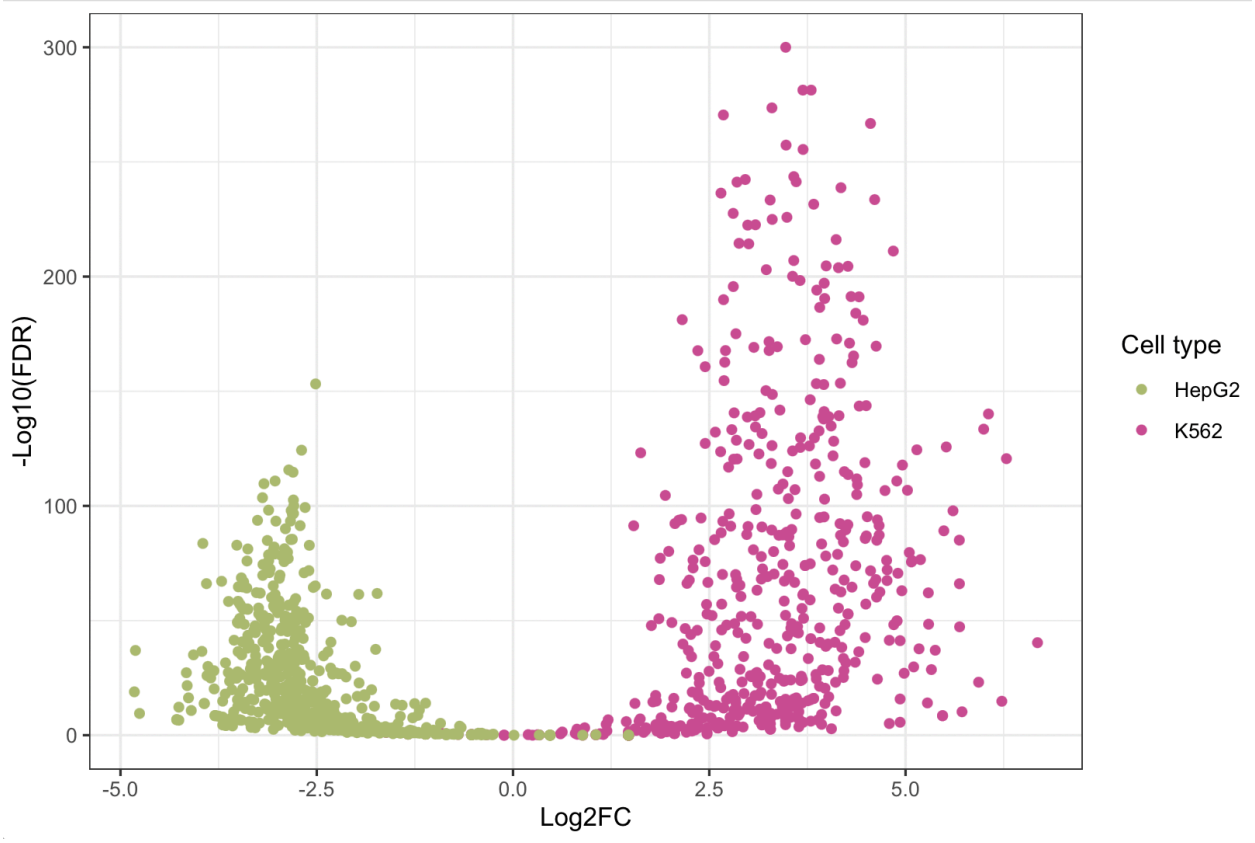
Thank you for pointing this out. We have added **Supplementary Note 2.2** to provide a more detailed explanation of the CAPRA pipeline, including its rationale, input data, and interpretation of the resulting p-values. To further clarify how CAPRA works, we have also added **Supplementary Figure 11** that visually illustrates the approach. We hope this additional context makes the method more accessible to readers unfamiliar with STARR-seq assays.

3) Given the amount of TF ChIP-seq data generated for the K562 and HepG2 cells, why was this not used to investigate cell-specific STARR+ cCREs instead of just motifs?

Thank you for this suggestion. Our initial decision to focus on sequence motifs was based on the fact that STARR-seq is an episomal assay, which tests regulatory activity outside of the native chromatin context. In contrast, ChIP-seq reflects transcription factor binding in endogenous chromatin, where access is highly influenced by chromatin state, including accessibility, histone modifications, and nucleosome positioning. As such, transcription factor binding in vivo may not reflect the intrinsic sequence-driven activity measured by STARR-seq. Motif analysis is better suited to capture sequence-encoded features that are unmasked in the episomal setting.

To further investigate this point, we compared the overlap of transcription factor ChIP-seq peak summits and STARR+ cCREs in both K562 and HepG2 cells. We found that STARR+ cCREs in K562 were broadly enriched for nearly all TFs profiled in K562, and likewise, STARR+ cCREs in HepG2 were broadly enriched for TFs profiled in HepG2. Of the 538 ChIP-seq datasets in K562, 490 (91%) were significantly enriched in K562 STARR+ cCREs (FDR < 0.05) and of the 626 ChIP-seq datasets in HepG2, 556 (89%) were enriched in HepG2 STARR+ cCREs. Only one experiment—STAG1 in K562—was enriched for STARR+cCREs in the opposite cell type (FDR = 0.02). These results suggest that TF ChIP-seq enrichment is highly dependent on chromatin accessibility in the native context, and therefore does not isolate the specific sequence features

that drive differential STARR-seq activity between cell types. We have added a brief summary of this analysis and its implications to the main text of the revised manuscript and **Supplementary Table 9b**.

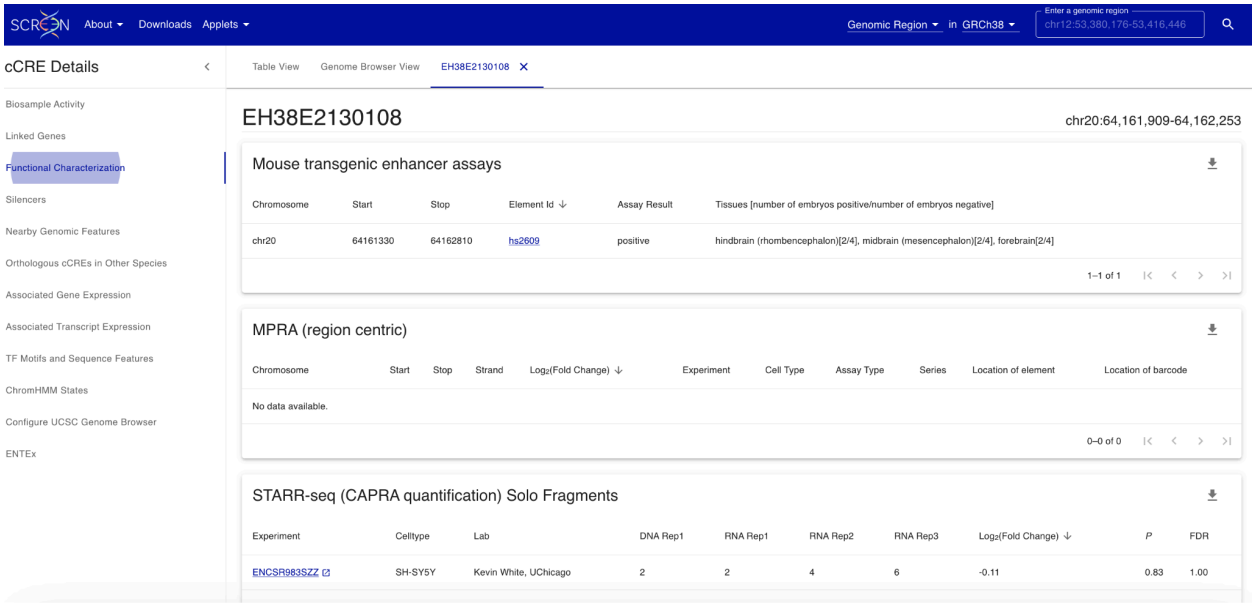


Main text:

*To determine what sequence features were responsible for the cell type-specific activity of enhancer cCREs, we investigated which transcription factor motifs were enriched in the distal enhancer cCREs with differentially high STARR scores in K562 cells (K562 STARR+) versus HepG2 cells (HepG2 STARR+) (**Figure 2a**). Both the K562 and HepG2 STARR+ cCREs were enriched for cell type-relevant transcription factor motifs and were more likely to be annotated as enhancers in their respective cell types (**Figure 2b, Supplementary Figure 13, Supplementary Table 9a**). We also evaluated transcription factor ChIP-seq enrichment in STARR+ cCREs and found widespread enrichment for nearly all profiled TFs in the corresponding cell type, indicating that in vivo TF binding is largely influenced by chromatin accessibility and less informative for distinguishing the sequence-intrinsic drivers of cell type-specific STARR activity (**Supplementary Table 9b**).*

4) This functional data is powerful validation, but it doesn't seem to be integrated into the SCREEN application? It would be great for users to know which elements have been validated and using what assay.

Thank you for this suggestion. The functional validation data is currently available on SCREEN via the *Functional Data* tab on each *cCRE Details* page. However, we agree that this information could be made more prominent and accessible. In response, we plan to enhance the visibility of validated elements by providing a bulk download option for functionally tested cCREs, allowing users to explore and integrate this data more easily into their own analyses. Screenshots of these updates are provided below for reference.



III. Silencer analysis: These are difficult to functionally validate, and so a focus on this is greatly appreciated.

1) An intuition for how the REST+ cCREs were tested using "enhancer" transgenic assays is needed. Presumably, the hypothesis is that since REST is only active in neural cells, then you would not expect to see transgenic activity in non-neural cells, but you would in neurons.

We agree that clarifying the rationale for using transgenic enhancer assays is important. REST is a transcriptional repressor active in non-neuronal cells, where it binds to neuron-restrictive silencer elements (NRSEs) to repress neuronal gene expression. In neurons, REST is not expressed and therefore cannot bind these elements, permitting the expression of neuronal genes.

We identified candidate NRSEs—referred to as REST+ cCREs—based on the presence of REST motifs and ChIP-seq binding. These elements showed silencer activity in REST-expressing cell types, as indicated by low STARR scores and repression of nearby gene expression. Notably, a subset of REST+ cCREs were annotated as enhancers in the Registry of cCREs, prompting us to test whether they ever exhibited bona fide enhancer activity.

We used transgenic mouse enhancer assays, which provide functional readouts across multiple tissues, including brain regions—particularly relevant for interrogating REST biology. In these assays, a subset of REST+ cCREs, specifically those classified as enhancers, exhibited robust activity in brain tissues. This suggests that some REST+ cCREs are context-dependent regulatory elements that act as silencers when bound by REST and as enhancers in its absence. These results are further supported by enrichment of active chromatin marks (e.g., H3K27ac) and binding motifs for transcriptional activators associated with neuronal enhancers.
Together, these findings support a model in which REST+ cCREs fall into two distinct functional classes:

- **REST+ enhancer/silencer cCREs**, which act as silencers in REST-expressing cells but function as enhancers in neurons; and

- **REST+ silencer cCREs**, which lack enhancer activity and function solely as silencers in REST-expressing contexts.

This classification—particularly the identification of REST+ enhancer/silencer cCREs—is a key contribution of our study. It highlights that REST-bound elements are not functionally homogeneous, but instead exhibit diverse regulatory behaviors depending on cell type and chromatin context.

2) Similarly for STARR-seq, since this is typically used to detect enhancers, how do you test for silencer activity? How do you calculate a p-value for these? This does not appear to be described in the main paper nor the supplementary notes.

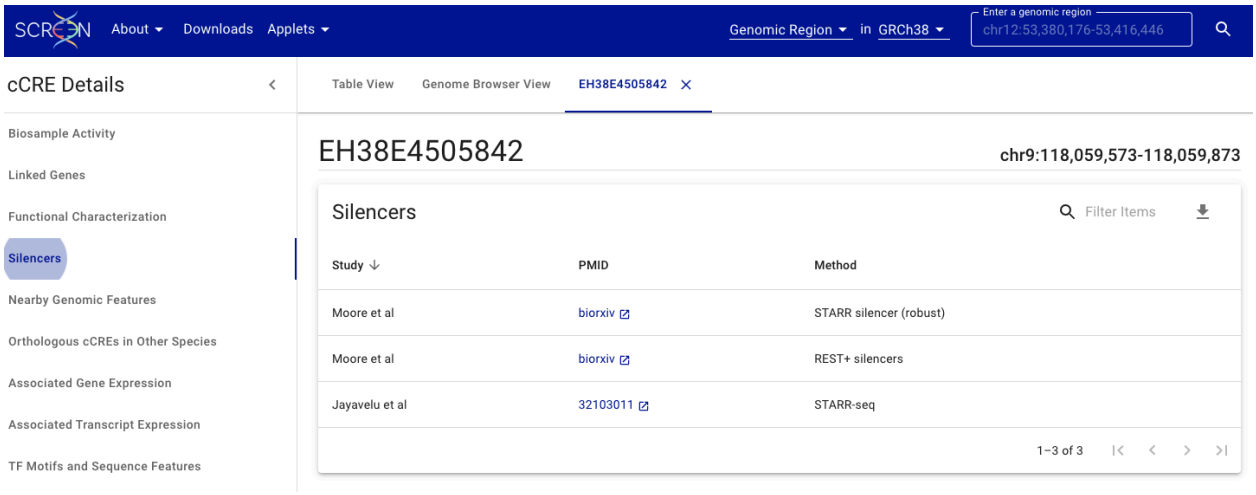
Thank you for raising this important point. While STARR-seq is often used to detect enhancer activity, it can also detect silencer activity by quantifying elements that decrease transcriptional output relative to their input abundance. To explicitly address this, we have added **Supplementary Note 2.2** detailing our method, CAPRA, which quantifies both enhancer and silencer activity based on RNA-to-DNA fragment ratios anchored on cCREs.

In brief, CAPRA compares the number of RNA and DNA fragments overlapping each cCRE using the DESeq2 framework. This yields both a log2 fold-change and a p-value for each element. A significant positive $\log_2$ fold-change reflects enhancer-like activity (higher-than-expected RNA count), while a significant negative $\log_2$ fold-change reflects silencer-like activity (lower-than-expected RNA count). The p-value tests whether the observed RNA abundance deviates from what would be expected based on DNA input, thereby enabling the detection of both activating and repressive elements.

We have also clarified this approach in the main text and expanded **Supplementary Note 2.2** and added **Supplementary Figure 11** to more explicitly describe how silencer activity is assessed from STARR-seq data.

3) Again it would be great to have results from these silencer assays incorporated into SCREEN.

Thank you for this suggestion. On SCREEN, silencer annotations now have their own tab on a cCRE's detail page and are available for bulk download under the "Downloads" tab (screenshot below for reference).



IV: Disease associated genes

1) Since ENCODE datasets are so highly enriched in assays performed in blood cell types, it is unsurprising that the example hear is related to a blood cell trait. Given as well that only aggregate annotations are given without tissue information, it brings into question whether the current cCRE annotation is nearly as useful for non-blood cell traits as is demonstrated for a blood cell trait. While the example provided shows the power of what can be done when all of these types of data are available for a relevant tissue, unless it can be similarly demonstrated in a non-blood cell trait, or even attempted, then this limitation must be acknowledged.

We appreciate your concern and the opportunity to clarify. While ENCODE has generated a large number of datasets in blood-derived cell types, the Registry of cCREs is not inherently biased toward blood or immune traits, as all analyses are performed at the level of individual biosamples (as opposed to at the aggregate level). This per-biosample approach ensures that every profiled tissue is used for the annotation and interpretation of regulatory elements.

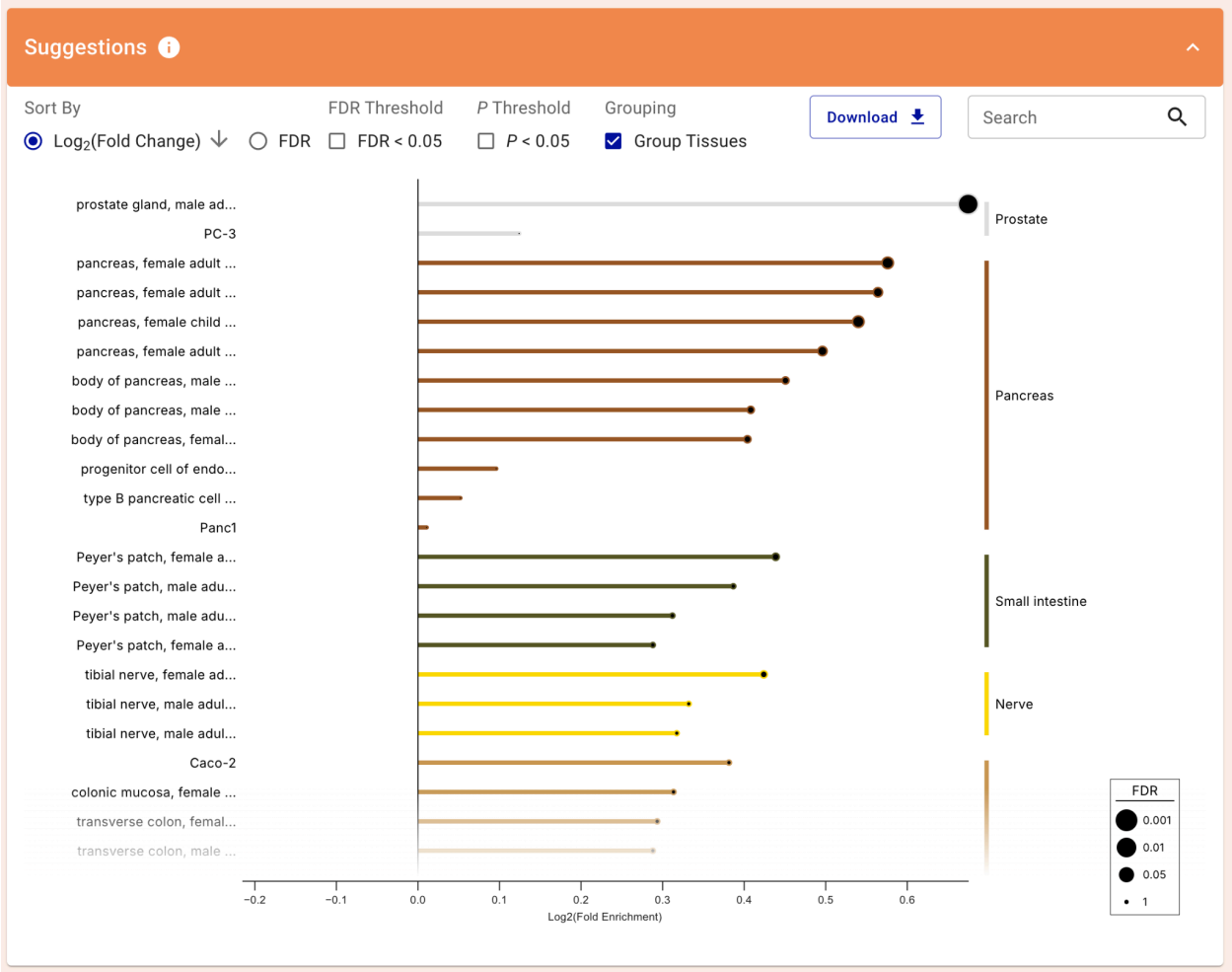
Indeed, we observe enrichment of GWAS variants for diverse traits in biologically relevant cell types and tissues—for example, brain cCREs for neuropsychiatric disorders, liver cCREs for cholesterol traits, breast cCREs (e.g., MCF-7 and primary breast tissue) for breast cancer loci and prostate cCREs for prostate cancer (images from SCREEN are shown below). This highlights the utility of the Registry beyond blood-related phenotypes.

Autism spectrum disorder or schizophrenia - Anney et al 28540026

LDL cholesterol - Surakka et al 25961943

Breast Cancer - Michailidou et al 25751625

Prostate cancer - Schumacher et al 29892016



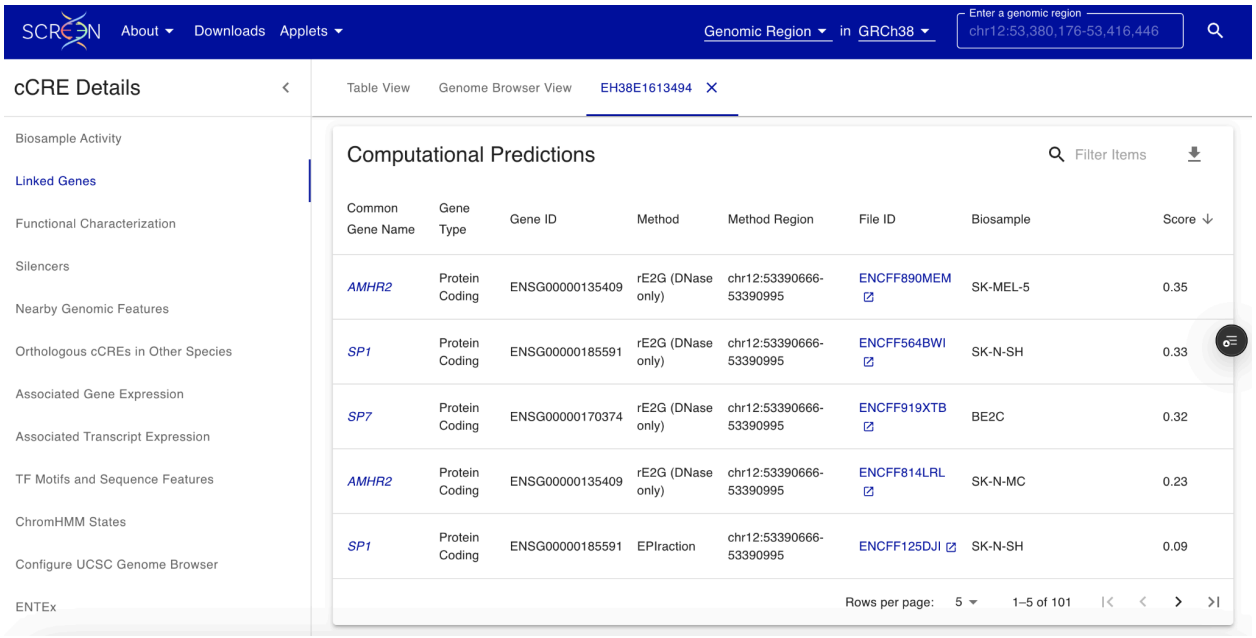
In the specific example we presented (RBC traits at RTBDN-MAST1 locus), nearly all analysis was conducted entirely within K562—a single deeply-characterized cell line—and thus biosample composition has no bearing on the results. We selected this example because ENCODE has generated extensive functional genomic data in K562, including CRISPRi-FlowFISH and 3D chromatin interactions, which allowed us to demonstrate the power of integrative analyses when deep datasets exist for a relevant biosample.

Finally, we acknowledge that some tissues remain underrepresented in ENCODE, and that this may limit interpretability in certain contexts. We now explicitly discuss this limitation in **Supplementary Note 1.9**.

V: Other comments

1) An obviously important characteristic of a regulatory element is its target gene(s). While I understand within ENCODE this is the focus of a different effort within the Interactions group, I note that some of the leaders of this effort are co-authors on this manuscript. It is disappointing that these target annotations are not better integrated within the SCREEN resource. Right now, SCREEN does have a "Linked Genes" drop down for each cCRE and provides evidence for links based on individual assays, but predictions from the ENCODE Enhancer-Gene predictions group (rE2G, ABC, GraphReg, EPIraction) are not integrated. As these efforts are part of the same consortium, and the Abstract of this paper even touts that "Integrating the Registry with other ENCODE annotations facilitates genetic variation interpretation and trait-associated gene identification", there should be better integration across groups to more easily provide all relevant information in a single resource.

Thank you for pointing this out. In response we have incorporated the latest sets of computational predictions for rE2G, ABC, GraphReg and EPIraction into SCREEN in cCRE details and also available for download. Screenshots detailing these updates are shown below for reference:



2) The analysis comparing the overlap in these new ENCODE cCREs with independently defined brain-specific regulatory elements does support the these cCREs are more comprehensive than the previous version, but this again is performed in a well-represented tissue in the datasets used. The statement "These results demonstrate that the ENCODE4 Registry is comprehensive while still maintaining biological specificity." has some truth, but the comprehensiveness is necessarily uneven across tissues. This should either be acknowledged, or the comprehensiveness should be demonstrated in a larger range of tissues.

This is a fair point. While the brain was a well-represented tissue in ENCODE4, we agree that demonstrating Registry improvements across a broader range of biological contexts is important for evaluating comprehensiveness. To address this, we expanded our comparison in **Supplementary Note 1.7** to include two additional large-scale datasets highlighting expansion of cCREs in the myeloid lineage and various fetal tissues.

Supplementary Note 1.7:

> *Liu et al. defined 1,032,273 accessible candidate cis-regulatory elements (acCREs) across 17 fetal organ systems using SHARE-seq[34]. 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 and fetal 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 cis-regulatory elements (tCREs) defined by Zhang et al. using single-cell 5' RNA-seq across 23 human tissues[35]. Of 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**).*

We now explicitly address the unevenness of biosample sampling, along with their implications for downstream interpretation, in **Supplementary Note 1.9** and in the main text:

Main text:

> *While the ENCODE Registry of cCREs provides extensive coverage across diverse tissues, its annotations are inherently influenced by the availability and quality of underlying functional genomics data (**Supplementary Note 1.9, Supplementary Figure 9**). Biosample-rich tissues such as brain and blood are better represented than other tissue types, and varying signal quality and sequencing depth also affect rDHS detection and cCRE inclusion. To mitigate these biases, we recommend biosample-level analyses over aggregate annotations, particularly when assessing tissue-specific activity or enrichment patterns.*

Referee #3 (Remarks on code availability):

We have reviewed the code, though not thoroughly. The authors should be commended for comprehensively providing all of their scripts, though overall these are poorly annotated, except for previously published software like CAPRA. The readme for the overall repository (https://github.com/weng-lab/ENCODE-cCREs/blob/master/README.md) is not informative.

Thank you for taking the time to examine our code and for acknowledging the comprehensive availability of our scripts. We appreciate the feedback regarding documentation and agree that clearer annotation is essential for usability and reproducibility. In response, we have significantly improved the README documentation in the main repository to provide a clearer overview of the project structure, key scripts, and usage instructions. Additionally, we have reviewed and updated inline comments in several key scripts to clarify their purpose and function. We are committed to maintaining this repository as a useful resource for the community and welcome further suggestions for improvement.