nature research

Peer Review Information

Journal: Nature Genetics
Manuscript Title: Genome-wide analysis of cis-regulatory changes underlying metabolic adaptation of cavefish
Corresponding author name(s): Dr Nicolas Rohner

Reviewer Comments & Decisions:

Decision Letter, initial version:

27th January 2021

Dear Nick,

Your Article "Genome-wide analysis of cis-regulatory changes in the metabolic adaptation of cavefish" has been seen by two referees. (Reviewer #1, who accepted the invitation to review the paper, has not returned comments.) You will see from their comments below that, while they find your work of potential interest, they have raised substantial concerns that must be addressed. In light of these comments, we cannot accept the manuscript for publication at this time, but we would be interested in considering a substantially revised version that addresses the referees' concerns.

We hope you will find the referees' comments useful as you decide how to proceed. If you wish to submit a substantially revised manuscript, please bear in mind that we will be reluctant to approach the referees again in the absence of major revisions.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, including with the chief editor, with a view to identifying key priorities that should be addressed in revision, and sometimes overruling referee requests that are deemed beyond the scope of the current study. In this case, we ask that you provide further details about the genomic datasets analyzed in this study as requested by both referees, perform a more thorough exploration of the genetic basis of differential chromatin accessibility between the study populations as requested by Reviewer #2, and clarify and extend the bioinformatics and functional studies where feasible as requested by Reviewer #3. We hope you will find this prioritized set of referee points to be useful when revising your study. Please do not hesitate to get in touch if you would like to discuss these issues further.

If you choose to revise your manuscript taking into account all reviewer and editor comments, please highlight all changes in the manuscript text file. At this stage we will need you to upload a copy of the

manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

If revising your manuscript:

*1) Include a "Response to referees" document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

*2) If you have not done so already please begin to revise your manuscript so that it conforms to our Article format instructions, available here.
Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary:
https://www.nature.com/documents/nr-reporting-summary.pdf
It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.
A revised checklist is essential for re-review of the paper.

Please be aware of our guidelines on digital image standards.

You may use the link below to submit your revised manuscript and related files:

[REDACTED]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

If you wish to submit a suitably revised manuscript, we would hope to receive it within 6 months. If you cannot send it within this time, please let us know. We will be happy to consider your revision so long as nothing similar has been accepted for publication at Nature Genetics or published elsewhere. Should your manuscript be substantially delayed without notifying us in advance and your article is eventually published, the received date would be that of the revised, not the original, version.

Please do not hesitate to contact me if you have any questions or would like to discuss the required revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID

from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

Thank you for the opportunity to review your work.

Sincerely,
Kyle

Kyle Vogan, PhD
Senior Editor
Nature Genetics
https://orcid.org/0000-0001-9565-9665

Referee expertise:

Referee #1: Population genetics, adaptation, epigenomics

Referee #2: Gene regulation, chromatin, metabolism

Referee #3: Epigenomics, development, evolution

Reviewers' Comments:

Reviewer #1:
None

Reviewer #2:
Remarks to the Author:

An interesting manuscript describing differences in gene cis-regulatory modules comparing blind cavefish to their surface counterparts. The authors provide chromatin data (ATAC and Chip-seq) and gene expression (RNA seq) and compare across morphotype. Differential presence of modules is identified and compared to gene expression. In at least some instances, correlation between accessibility and expression is documented.
Exploration of select modules in reporter assays provides anecdotal support for the hypothesis that differential accessibility correlates with evolutionary pressure of the different environments.

I find the manuscript to be interesting and appealing to a broad readership.
I am a bit perplexed as to why the authors appear to have sequenced the cave and surface fish, yet spend virtually no text on exploring the differences - particularly as it relates to the chromatin analysis. I would suggest thorough exploration of the genetic basis of differential chromatin accessibility is critical to this manuscript and urge the authors to perform this important analysis.

Points to help the authors improve their work:

1. Genomics data and its QC is important to many readers. Please provide a table for each data type that includes number of reads per replicate, read quality scores, % uniquely mapped, etc to permit readers to independently evaluate the quality of this data.

2. I am very unclear on the SNP calling in CRMs. This is described in one sentence in the text. The methods are not helpful. Presumaby there is some sequence data of the genomes of the surface and cave fish. Please describe where the data come from, describe how deep the sequence data is to permit readers to evaluate exactly what has been done.

How many total SNPs are there between the different morphotypes of fish? How many fall in known TF binding motifs? How many fall in liver specific CRM's? Please explore this data a bit more - it seems essential to the overall conclusions and is just not described in the current text.

3. I am a bit surprised at the outcome of functional testing of the CRE's. by my reading, less than half of the 25 loci selected for (1) presence of a sequence difference between morphotypes, (2) ATAC bias = RNA bias exhibit differential activity based on morphotype. This, at the very least, merits some discussion.

4. I wonder about the harvest of the fish for liver dissection relative to light/dark cycles and to feeding. Please provide this information in the methods section as it relates to circadian cycle - which was identified as differential across the two fish types.

Reviewer #3:
Remarks to the Author:

Krishnan et al produced an interesting study that compares Astyanax mexicanus morphotypes on the genomic, epigenomic, and transcriptomic levels. The authors used ATAC-seq to identify regions of differential chromatin accessibility between cave and surface cavefish and went on to dissect those regulatory differences through diverse in vivo, in vitro and biochemical approaches to come up with interesting hypotheses related to evolutionary trade-offs required for metabolic adaptation. The study is generally well-written and easy to understand, however, I do have a number of major concerns related to reasoning and choice of tested sequences. Also, I found the extent of bioinformatics analyses somewhat underwhelming as a lot more could have been done with the data to further support the authors' claims. My points for improvement can be found below.

There is little novelty in Figure 1. In general, I believe that such figures can be made supplementary. Same goes for Figure 2. Instead of showing Venn diagrams that represent presence / absence of peaks it would have been much more informative to see a clustered heatmap of a merged collection of ATAC-seq peaks and their histone PTM state. That would immediately highlight the differences and show in which regions (i.e promoters, enhancers) these differences exist. Also, there is no y axis on the browser snapshots so it is difficult to understand what is actually being shown (signal strength, units i.e cpm, fpkm etc).

There is virtually no description of how ATAC-seq data was processed. How did the authors select for

open chromatin fragments (size filter)? How were the differential peaks identified? A P value is mentioned in Methods, but this needs to explained better.

In Figure 4 it is not clear whether the larvae shown are F0 or stable transgenics. If F0, very little can be concluded given that expression in F0 is generally noisy and not restricted to defined expression domains. A major advantage of using an in vivo system would have been to demonstrate presence / absence of liver expression of selected CREs, however the authors have not done that. It is also not clear which allele is being shown in Figure 4C. Same goes for S4 (zebrafish experiments), which appear noisy and without properly defined expression patterns. Assuming that CRM_15 is the same as CRE_15 it would be interesting to see how transgenic fish for both alleles look like (either in cavefish or zebrafish transgenics), given the apparently highly significant difference in expression observed in the ZFL line for that sequence.

The reasoning behind the choice of CRE_15 is clear (expression difference in Fig 4b), however I am puzzled by the mechanism through which CRE_15 is supposed to act on the Hpdb gene. I have a number of concerns related to this: 1) Again it is not clear due to the lack of y axis labels what is being shown in Fig 5a. 2) That notwithstanding, all three tracks show ATAC-seq signal around the red line (not explained what the red line is), even though that signal is somewhat reduced in "surface", if tracks are comparable to start with. A UCSC screenshot with differentially enriched sites (i.e CRE_15) as a BED track would've been helpful to understand the exact position of this site in relation to the gene. Like this I can only guess that the authors actually refer to the ATAC peak overlayed by the red line. 3) What does this sequence do? Is it an enhancer, as suggested? If so, and if this enhancer were active, one would expect strong enrichment of H3K27ac in Pachon and Tinaja but this is not visible. How do the authors explain that? I also note that the looping distance (~700 bp) would be pretty short. This needs to be assessed through 3C/4C approaches. The authors suggest this difference is due to presence / absence of IRF2 binding site, however no motif analysis for IRF2 is shown and also, as noted above, there is a peak visible in the surface track so independently of the deletion, something is binding to that region in all three tracks. Interestingly, H3K27ac, which correlates with active transcription much better than H3K4me3 appears to be similarly enriched at the TSS in surface and cave morphotypes. This brings into question whether this gene is properly annotated. Finally, the EMSA experiment gel appears overexposed and it is not clear why the exact conditions have not been repeated for Surface (9 lanes) and Pachon oligos (2 lanes.). Overall, more work needs to be done including better bioinformatics analyses and chromatin conformation capture experiments to prove that IRF2 binding and differential chromatin state of CRE_15 are the reason behind Hpdb expression differences.

Finally, more could have been done with the > 2000 differential ATAC-seq peaks. While a correlation with transcription is presented in Figure 3, nothing is known about TFBS motifs at those sites, their chromatin make-up or if more examples (such as the one proposed in Fig. 5) could potentially be found in such a dataset.

Author Rebuttal to Initial comments

Reviewers' Comments:

Reviewer #1:
None

Reviewer #2:
Remarks to the Author:

An interesting manuscript describing differences in gene cis-regulatory modules comparing blind cavefish to their surface counterparts. The authors provide chromatin data (ATAC and Chip-seq) and gene expression (RNA seq) and compare across morphotype. Differential presence of modules is identified and compared to gene expression. In at least some instances, correlation between accessibility and expression is documented.
Exploration of select modules in reporter assays provides anecdotal support for the hypothesis that differential accessibility correlates with evolutionary pressure of the different environments.

I find the manuscript to be interesting and appealing to a broad readership.

We thank the reviewer for the positive assessment of our manuscript and important suggestions and comments that have improved our manuscript significantly. We hope that we have addressed all the reviewer comments satisfactorily.

I am a bit perplexed as to why the authors appear to have sequenced the cave and surface fish yet spend virtually no text on exploring the differences - particularly as it relates to the chromatin analysis. I would suggest thorough exploration of the genetic basis of differential chromatin accessibility is critical to this manuscript and urge the authors to perform this important analysis.

We certainly agree with the reviewer and thank the reviewer for the suggestion. The reason we originally shied away from this analysis, is that even though the two populations are closely related, we still would have expected a large amount of SNPs in the genomic regions between the populations, most of which probably won't necessarily have a function in gene regulation. Nevertheless, we fully agree that this is an important analysis to do. To explore genetic variation in the regulatory regions mapped in the study, we have now used the NGS sequences from the ATAC-seq dataset itself to query for SNPs and indels. The data has now been added to the manuscript. This part is now included in Fig. 3 and Fig. S4. Below is the text that has been added to the manuscript (page 10).

Text in the manuscript:
'To explore the causal role of polymorphisms/mutations in CREs in differential gene regulation, we first did a comprehensive analysis of SNPs and indels within all the putative CREs. Using GATK variant calling tool, we identified a total of 527,644 SNPs and 183,958 indels between morphotypes using the raw reads of the ATAC-seq datasets. We first analyzed our SNP data and found that 27.4% of the peaks had no SNPs while, rest of the peaks had anywhere between 1 and 169 SNPs per peak (Fig. 3a shows distribution for 1-30 SNPs/peak). Similarly, 27.1% peaks had no indels while, rest of the peaks had between 1 and 44 indels (Fig. 3b shows distribution for 1-30 indels/peak). There were 195,833 SNPs between surface and either of the cave populations with 123,611 SNPs between surface and Pachón and 138,278 SNPs between surface

and Tinaja. Only 66,056 SNPs were similarly variant between surface and both the cave populations (Fig. S4b). We observed a similar trend for indels (Fig. S4c). The extent of heterozygosity was greater in surface for both SNPs and indels which is line with earlier observations indicating greater genetic diversity in surface fish and higher inbreeding in cave populations (Bradic et al., 2013) (Fig. 3c, d). To assess if genetic changes in CREs underlie differential chromatin accessibility, we compared the SNP frequency in surface-accessible peaks to that in Pachón- and Tinaja-accessible peaks (Differential peaks with p value < 0.001). To normalize for varying peak width, we focused on a 200bp region spanning the peak center. In addition, we ensured that all SNPs considered for this analysis must be called in both surface and cave sequences to avoid any bias due to lack of SNP call. With all the stringent criteria in place, we observed a small but statistically significant increase in the SNP frequency in both Pachón and Tinaja accessible peaks compared to surface accessible peaks (Fig. 3e, f). This result suggests that sequence differences drive the evolution of newly accessible regions in cavefish.

To delve deeper into the functional consequence of sequence differences in differential peaks, we investigated whether these SNPs could have effects on TF binding. We used an R package 'motifBreakR' and predicted 'altered motifs', which are TF motifs overlapping with SNP(s) that could potentially alter TF binding (Fig 3g) (Coetzee et al., 2015). In our entire dataset, we observed that 48.2% (254,554 SNPs) of the identified SNPs had the potential to alter TF motifs, that resulted in a total of 1,497,297 altered TF binding motifs. Among the differentially accessible CREs, 33% of Pachón accessible CREs consisted of at least one altered motif while 41% of the surface-accessible CREs contained at least one altered motif. Interestingly, HNF1B (hepatocyte nuclear factor 1b) was one of the top 15 TFs with altered motifs in surface-accessible CREs (Fig. 3g) but not in Pachón-accessible CREs (Fig. 3h). As HNF1B is involved in glucose metabolism and is implicated in diabetes (Raile et al., 2008) the presence of altered HNF1B motifs could have significant effects on downstream gene expression patterns between surface and Pachón. This analysis highlights the abundance of altered TF motifs occurring within CREs that could potentially alter CRE activity and thereby the expression of associated genes.'

Points to help the authors improve their work:

1. Genomics data and its QC is important to many readers. Please provide a table for each data type that includes number of reads per replicate, read quality scores, % uniquely mapped, etc. to permit readers to independently evaluate the quality of this data.

We have added QC reports for all of the datasets as Supplementary information SI-6. This includes number of total reads, uniquely mapped reads, unmapped reads and number of peaks obtained in each experiment.

2. I am very unclear on the SNP calling in CRMs. This is described in one sentence in the text. The methods are not helpful. Presumably there is some sequence data of the genomes of the surface and cave fish. Please describe where the data come from, describe how deep the sequence data is to permit readers to evaluate exactly what has been done. How many total SNPs are there between the different morphotypes of fish? How many fall in known TF binding motifs? How many fall in liver specific CRM's? Please explore this data a bit more - it seems essential to the overall conclusions and is just not described in the current text.

We have called variants using the sequence reads obtained from our ATAC-seq datasets. We used the GATK variant calling toolkit from the Broad Institute (Van der Auwera GA & O'Connor BD., 2020) for this purpose. Variant calling was done according to the guidelines for Best Practices for calling variants. To further ensure robust variant calling in open chromatin regions, we restricted our SNP analysis to 200bp where the reads are of best quality. We have added these details to the methods. All the new SNP/indel data are presented in Fig. 3 and Fig. S4. Many of the SNPs overlapped known TF binding motifs, thereby potentially altering their binding. This data is presented in Fig. 3g, h. See also comment above.

3. I am a bit surprised at the outcome of functional testing of the CRE's. by my reading, less than half of the 25 loci selected for (1) presence of a sequence difference between morphotypes, (2) ATAC bias = RNA bias exhibit differential activity based on morphotype. This, at the very least, merits some discussion.

The reviewer is correct in noting that less than half of the loci that we selected for functional testing show a difference in the reporter assay. This can be due to several reasons: 1. These CREs are being tested out of their genomic context. 2. The cell line we used is a zebrafish liver cell line (zebrafish and cavefish diverged roughly 150 million years ago). 3. Not all SNPs/indels will necessarily result in differential TF binding in vivo. Therefore, we feel it is so important to test these elements in a reporter assay. With that being said, despite all these limitations, we found that our assay system is able to capture differential CREs, some of which are even functionally conserved across wide evolutionary distances as seen in the HepG2 cell line experiments.
We have added a similar detailed explanation in the manuscript (page 13).

4. I wonder about the harvest of the fish for liver dissection relative to light/dark cycles and to feeding. Please provide this information in the methods section as it relates to circadian cycle - which was identified as differential across the two fish types.

Both the surface fish and the cavefish are maintained under the exact same lab conditions. They are provided with the same food and both kept under the same 14:10 light:dark cycle. As we are mainly interested in understanding the genetic changes between the morphotypes, maintaining them under the same conditions helps negate out potential environmental effects. For harvesting tissues, the fish were euthanized in MS-222 and immediately dissected. Dissections were performed in the morning (3 hours after light turns on) after fasting the fish overnight. Thus, finding accessible chromatin differentially enriched for circadian rhythm pathway genes indicates causal genetic differences in the pathway between the two fish types. This is supported by earlier studies by us and others which show that while cavefish develop circadian rhythms in the lab, the amplitude and periodicity of the circadian rhythm is altered in cavefish (Beale et al., 2013). We have added these details in the methods sections in the revised version.

Reviewer #3:
Remarks to the Author:

Krishnan et al produced an interesting study that compares Astyanax mexicanus morphotypes on the

genomic, epigenomic, and transcriptomic levels. The authors used ATAC-seq to identify regions of differential chromatin accessibility between cave and surface cavefish and went on to dissect those regulatory differences through diverse in vivo, in vitro and biochemical approaches to come up with interesting hypotheses related to evolutionary trade-offs required for metabolic adaptation. The study is generally well-written and easy to understand, however, I do have a number of major concerns related to reasoning and choice of tested sequences. Also, I found the extent of bioinformatics analyses somewhat underwhelming as a lot more could have been done with the data to further support the authors' claims. My points for improvement can be found below.

We thank the reviewer for the positive assessment of our work. We have now addressed all the reviewer comments, which has improved the manuscript considerably. Regarding additional data analysis, we have now performed an extensive analysis of SNPs/Indels present in the open chromatin regions. We have used the NGS sequences from the ATAC-seq dataset itself to query for SNPs and indels to explore genetic variation in the regulatory regions between the cave and surface populations. The data has now been added to the manuscript. This part in now included in Fig. 3 and Fig. S4. Below is the text that has been added to the manuscript (page 10).

The new text in the manuscript reads:
'To explore the causal role of polymorphisms/mutations in CREs in differential gene regulation, we first did a comprehensive analysis of SNPs and indels within all the putative CREs. Using GATK variant calling tool, we identified a total of 527,644 SNPs and 183,958 indels between morphotypes using the raw reads of the ATAC-seq datasets. We first analyzed our SNP data and found that 27.4% of the peaks had no SNPs while, rest of the peaks had anywhere between 1 and 169 SNPs per peak (Fig. 3a shows distribution for 1-30 SNPs/peak). Similarly, 27.1% peaks had no indels while, rest of the peaks had between 1 and 44 indels (Fig. 3b shows distribution for 1-30 indels/peak). There were 195,833 SNPs between surface and either of the cave populations with 123,611 SNPs between surface and Pachón and 138,278 SNPs between surface and Tinaja. Only 66,056 SNPs were similarly variant between surface and both the cave populations (Fig. S4b). We observed a similar trend for indels (Fig. S4c). The extent of heterozygosity was greater in surface for both SNPs and indels which is line with earlier observations indicating greater genetic diversity in surface fish and higher inbreeding in cave populations (Bradic et al., 2013) (Fig. 3c, d). To assess if genetic changes in CREs underlie differential chromatin accessibility, we compared the SNP frequency in surface-accessible peaks to that in Pachón- and Tinaja-accessible peaks (Differential peaks with p value < 0.001). To normalize for varying peak width, we focused on a 200bp region spanning the peak center. In addition, we ensured that all SNPs considered for this analysis must be called in both surface and cave sequences to avoid any bias due to lack of SNP call. With all the stringent criteria in place, we observed a small but statistically significant increase in the SNP frequency in both Pachón and Tinaja accessible peaks compared to surface accessible peaks (Fig. 3e, f). This result suggests that sequence differences drive the evolution of newly accessible regions in cavefish.'

Further, we have analyzed the newly identified SNPs for their predicted impact on TF binding. The new data is now presented in Fig. 3, S3, S4.

The new text on page 10 in the manuscript reads:
'To delve deeper into the functional consequence of sequence differences in differential peaks, we investigated whether these SNPs could have effects on TF binding. We used an R package 'motifBreakR'

and predicted 'altered motifs', which are TF motifs overlapping with SNP(s) that could potentially alter TF binding (Fig 3g) (Coetzee et al., 2015). In our entire dataset, we observed that 48.2% (254,554 SNPs) of the identified SNPs had the potential to alter TF motifs, that resulted in a total of 1,497,297 altered TF binding motifs. Among the differentially accessible CREs, 33% of Pachón accessible CREs consisted of at least one altered motif while 41% of the surface-accessible CREs contained at least one altered motif. Interestingly, HNF1B (hepatocyte nuclear factor 1b) was one of the top 15 TFs with altered motifs in surface-accessible CREs (Fig. 3g) but not in Pachón-accessible CREs (Fig. 3h). As HNF1B is involved in glucose metabolism and is implicated in diabetes (Raile et al., 2008) the presence of altered HNF1B motifs could have significant effects on downstream gene expression patterns between surface and Pachón. This analysis highlights the abundance of altered TF motifs occurring within CREs that could potentially alter CRE activity and thereby the expression of associated genes.'

There is little novelty in Figure 1. In general, I believe that such figures can be made supplementary.

We agree with the reviewer and have moved Fig. 1 to the supplement.

Same goes for Figure 2. Instead of showing Venn diagrams that represent presence / absence of peaks it would have been much more informative to see a clustered heatmap of a merged collection of ATAC-seq peaks and their histone PTM state. That would immediately highlight the differences and show in which regions (i.e promoters, enhancers) these differences exist.

We apologize for not bringing the message out clearly enough. In the original Figure 2b we intended to highlight the phenomena of parallel or independent evolution that was observed at the level of chromatin features. We observe that CREs that were differentially accessible between surface and one of the cave morphs were differentially accessible between surface and the other cave morph as well indicating high similarity or parallel evolution in the gain and loss of chromatin features between the two cave morphs. We believe that this data is critical to the main message of the manuscript, however, we thank the reviewer for the suggestion to use heat maps instead of Venn diagrams. We have now included heatmaps for representing this observation. We have also added detailed explanation of the phenomena in the text to better aid the readers' understanding.

The revised text on page5 reads:
'We further analyzed highly differentially accessible regions mapping close (<10 Kb) to genes and observed that 74.4% of the regions that were accessible in surface fish (surface-accessible CREs) lost accessibility in both the cave populations (Fig. 1d, f). Similarly, 77.4% of the regions that gained accessibility in Pachón as compared to surface also gained accessibility in Tinaja (cave-accessible CREs) (Fig. 1e, g). In other words, as shown in the heatmaps (Fig. 1d, e lower panels), CREs that were differentially accessible between surface and one of the cave morphs were differentially accessible between surface and the other cave morph as well. These comparisons indicate that both the cave populations have gained or lost accessible chromatin states with regulatory potential in a very similar set of regions of the genome during evolution.'

Also, there is no y axis on the browser snapshots so it is difficult to understand what is actually being shown (signal strength, units i.e cpm, fpkm etc).

Thank you for pointing out the missing y axes. We have added y axes, denoting rpm, to all browser shots in the manuscript. As before, each chromatin feature represented on the browser shot has the same y axis to enable comparison.

There is virtually no description of how ATAC-seq data was processed. How did the authors select for open chromatin fragments (size filter)? How were the differential peaks identified? A P value is mentioned in Methods, but this needs to explained better.

We have added additional details to the methods section "ChIP-seq and ATAC-seq analyses**".**

In Figure 4 it is not clear whether the larvae shown are F0 or stable transgenics. If F0, very little can be concluded given that expression in F0 is generally noisy and not restricted to defined expression domains. A major advantage of using an in vivo system would have been to demonstrate presence / absence of liver expression of selected CREs, however the authors have not done that. It is also not clear which allele is being shown in Figure 4C. Same goes for S4 (zebrafish experiments), which appear noisy and without properly defined expression patterns. Assuming that CRM_15 is the same as CRE_15 it would be interesting to see how transgenic fish for both alleles look like (either in cavefish or zebrafish transgenics), given the apparently highly significant difference in expression observed in the ZFL line for that sequence.

Original Figure 4 depicted transient transgenics in the F0 generation. Although, transient transgenic approaches are routinely used to assess CRE activity (Fisher et al., 2006; Parker et al., 2014), we completely agree with the reviewer that transient transgenics are noisy and interpretations from these have to be taken with a grain of salt. We want to point out that we only used the F0 transgenics to assess general CRE activity. Despite all limitations of transient transgenics our analysis provides evidence that the CREs are functional and conserved across species and across developmental stages. To assess differential CRE activity, we have solely relied on our in vitro reporter assays, which we found to be highly reproducible and quantitative. To better reflect this, we have made this point clearer in the manuscript and moved the zebrafish and *Astyanax* transgenics to the supplement (Fig S5). In addition, as requested by the reviewer, we have generated stable F1 transgenics for CRE_15 and observed GFP expression in the liver/gut region consistent with our expectations (Fig. S5d). However, due to the inherent technical limitations of random integrations in transgenics, we refrained from quantitative comparisons.

The reasoning behind the choice of CRE_15 is clear (expression difference in Fig 4b), however I am puzzled by the mechanism through which CRE_15 is supposed to act on the Hpdb gene. I have a number of concerns related to this: 1) Again it is not clear due to the lack of y axis labels what is being shown in Fig 5a.

We have changed the browser shot to include the y axis and genomic coordinates.

2) That notwithstanding, all three tracks show ATAC-seq signal around the red line (not explained what the red line is), even though that signal is somewhat reduced in "surface", if tracks are comparable to start with. A UCSC screenshot with differentially enriched sites (i.e CRE_15) as a BED track would've been helpful to understand the exact position of this site in relation to the gene. Like this I can only guess that the authors actually refer to the ATAC peak overlayed by the red line.

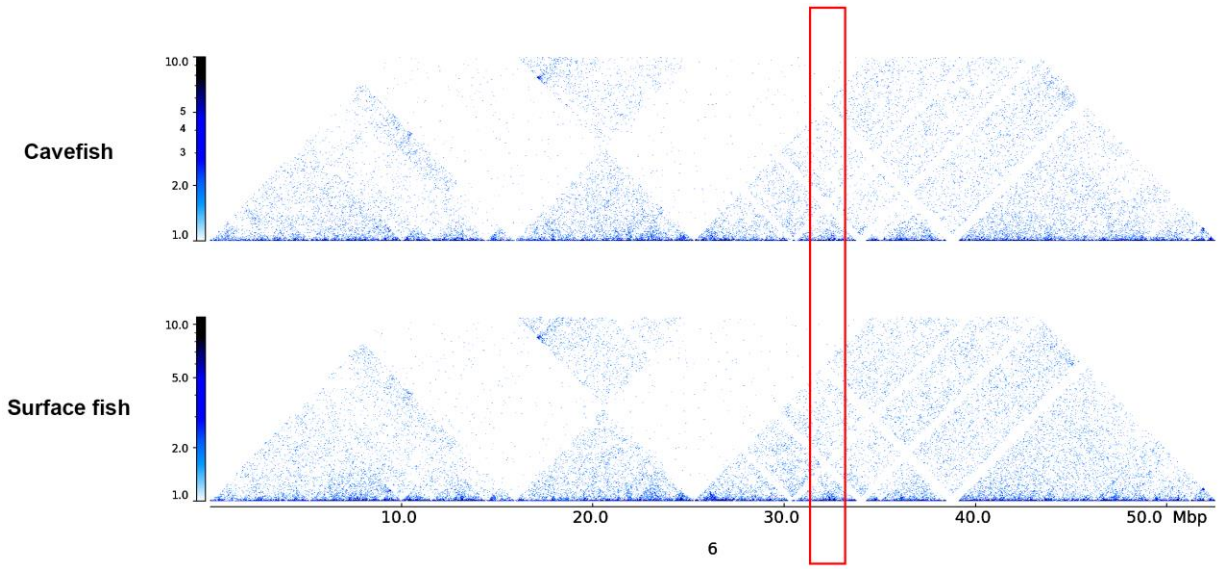
The figure has been rectified to include the genomic coordinates for more clarity. In addition, we mention the exact distance of the CRE from the transcription start site of *Hpdb* gene in the legend.

3) What does this sequence do? Is it an enhancer, as suggested? If so, and if this enhancer were active, one would expect strong enrichment of H3K27ac in Pachon and Tinaja but this is not visible. How do the authors explain that?

CRE_15, as the reviewer has also noticed, is just 744bp away from the gene start site. Hence by convention, this CRE would classify as a promoter element. In the text we originally call this element an enhancer, because the characterization was done via an 'enhancer assay'. We realize that this annotation is confusing, hence we stick to calling this element as a CRE in the revised version. Regarding the H3K27ac mark around CRE_15, we have replaced the browser shot with one having clearer visualization. H3K27ac is known to mark broad regions around open chromatin and as can be seen in the browser view, the mark is much more pronounced in cave populations than in surface not only at the TSS but in the flanking regions as well.

I also note that the looping distance (~700 bp) would be pretty short. This needs to be assessed through 3C/4C approaches.

This is a very important question raised by the reviewer that we are also interested in addressing. However, as later pointed out by the reviewer, the looping distance between CRE_15 and *Hpdb*-TSS is very short. Most chromosome conformation capture techniques have a resolution of ~1-2kb at the best (Dekker et al., 2013) and hence it is technically very challenging to assess this particular CRE-TSS interaction. Nevertheless, we agree with the reviewer that it would be interesting to address this question in the long term. We are currently generating a genome-wide Hi-C map for surface and cavefish livers. The study, however, is highly data-intensive with ~400 million reads per sample. To give a picture of the ongoing analysis - currently, we have very encouraging results from the analysis of 20 million reads (out of 400 million reads) from each sample to standardize the analysis pipeline and perform quality controls. Below are pyramid heatmaps generated to visualize the topologically associated domains of chromosome 6 for surface and Pachón genome. As the reviewer will notice, there are multiple possible cis interactions in the genomic region surrounding CRE-15 (Red box). A complete analysis of this dataset to discover differential cis interactions requires several more months of intensive computation and will enable us to investigate 3D interactions at a genome wide scale. This will be beyond the scope of this study and thus will be published as a separate study in the near future.



The authors suggest this difference is due to presence / absence of IRF2 binding site, however no motif analysis for IRF2 is shown and also, as noted above, there is a peak visible in the surface track so independently of the deletion, something is binding to that region in all three tracks.

We have now included the motif logo for IRF2 within Fig. 5a. As supplementary data (SI-5), we also provide the motif search result for the entire CRE_15 from surface and Pachón sequence. We agree that the surface allele shows a lesser intense open chromatin peak at the CRE_15 locus. As shown in Fig. 5e, there are other mutations that govern the differential CRE output. While IRF2 binding is completely abolished due to the deletion, the other mutations will likely only decrease/increase TF binding and not abolish it altogether. Hence, the small ATAC peak noticed in surface genome, could be due to the other proteins binding. It can be noticed, however, that the active ChIP-seq marks are absent from the surface samples which points to a functional impairment of the CRE activity.

Interestingly, H3K27ac, which correlates with active transcription much better than H3K4me3 appears to be similarly enriched at the TSS in surface and cave morphotypes. This brings into question whether this gene is properly annotated.

We apologize for the confusion which was caused by poor quality of the browser view. We have rectified this and hope that the differential chromatin marks are more clearly visible now. Regarding incorrect annotation of the gene, our RNA-seq data conforms to the existing annotation in Ensembl (as shown in Fig. 5a).

Finally, the EMSA experiment gel appears overexposed and it is not clear why the exact conditions have not been repeated for Surface (9 lanes) and Pachon oligos (2 lanes.).

We have repeated the experiments and have replaced the gel picture with a less exposed version. In addition, we have performed all the control experiments for the Pachón oligo as well. All this new data can now be found in Fig 5d.

Overall, more work needs to be done including better bioinformatics analyses and chromatin conformation capture experiments to prove that IRF2 binding and differential chromatin state of CRE_15 are the reason behind Hpdb expression differences.

As mentioned in above comment, the proximity of CRE_15 to the TSS of *Hpdb* poses a major limitation for analyzing physical interaction by any 'C' method. We now provide complete motif analysis results of surface and Pachón CRE_15 for more information about the locus.

Finally, more could have been done with the > 2000 differential ATAC-seq peaks. While a correlation with transcription is presented in Figure 3, nothing is known about TFBS motifs at those sites, their chromatin make-up or if more examples (such as the one proposed in Fig. 5) could potentially be found in such a dataset.

To address this comment, we have performed motif analysis for the differential CREs. This data can now be found in Fig. 2g and h. In addition, we provide examples of genes regulated by the morphotype-specific CREs that belong to the select pathways/Reactomes depicted in Fig. 2c and d. This new data is now in Fig. 2e and f. We have also performed a 'broken motif' analysis that predicts if SNPs in TF binding sites can potentially alter its binding to DNA. This data highlights the abundance of broken TF motifs between surface and cavefish and that such mutations could be causal to differences in CRE function. In addition, we cite a recent study that reports a mutation in HNF4a, a TF whose motifs are enriched in surface-accessible CREs and show how the regulatory networks of genes downstream to HNF4a are differentially expressed between the morphotypes. This data is presented in Fig S3b, c.
Lastly, as mentioned earlier, we have performed in silico analysis to understand how SNPs can affect TF binding (Fig. 3g, h)

The revised text in the manuscript reads:
We observed a significant correlation ($p<0.05$; see methods for details) between morphotype-biased CREs and expression of nearby genes, with s-CREs associated to a larger proportion of surface-upregulated genes and c-CREs associated to more cave-upregulated genes (Fig. 2a, b). We observed the s-CREs associated genes were enriched in circadian clock categories, lipid metabolism, and TGF-b signaling pathways (Fig. 2c), while c-CREs displayed enrichment of multiple pathways involved in lipid metabolism and immune function (Fig.2d). Interestingly, lipid metabolism pathway genes enriched near s-CREs comprised of catabolic genes (lipases and fatty acid binding proteins) (Fig. 2e) that are upregulated in surface, while lipid metabolism pathway genes near c-CREs highlighted lipid signaling and

anabolic genes (fatty acid synthase and acyl CoA synthetases) that are upregulated in the cavefish (Fig. 2f). These findings are in line with previous studies showing increased fat accumulation in cavefish (Riddle et al., 2018; Xiong et al., 2018). Network analysis of the genes associated with c-CREs showed upregulation of lipid synthesis pathways; specifically, genes such as fasn, which is a fatty acid synthase gene, glucose 6 phosphate dehydrogenase (g6pd) and slc30a8 which is a zinc transporter involved in insulin function were upregulated in cavefish (Fig. S3a) (Eissing et al., 2013; Flannick et al., 2014; Ham et al., 2016).

Beale, A., Guibal, C., Tamai, T.K., Klotz, L., Cowen, S., Peyric, E., Reynoso, V.H., Yamamoto, Y., and Whitmore, D. (2013). Circadian rhythms in Mexican blind cavefish Astyanax mexicanus in the lab and in the field. Nat Commun *4*, 2769.

Bradic, M., Teotonio, H., and Borowsky, R.L. (2013). The population genomics of repeated evolution in the blind cavefish Astyanax mexicanus. Mol Biol Evol *30*, 2383-2400.
Coetzee, S.G., Coetzee, G.A., and Hazelett, D.J. (2015). motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. Bioinformatics *31*, 3847-3849.
Dekker, J., Marti-Renom, M.A., and Mirny, L.A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet *14*, 390-403.
Fisher, S., Grice, E.A., Vinton, R.M., Bessling, S.L., Urasaki, A., Kawakami, K., and McCallion, A.S. (2006). Evaluating the biological relevance of putative enhancers using Tol2 transposon-mediated transgenesis in zebrafish. Nat Protoc *1*, 1297-1305.
Parker, H.J., Bronner, M.E., and Krumlauf, R. (2014). A Hox regulatory network of hindbrain segmentation is conserved to the base of vertebrates. Nature *514*, 490-493.
Raile, K., Klopocki, E., Wessel, T., Deiss, D., Horn, D., Muller, D., Ullmann, R., and Gruters, A. (2008). HNF1B abnormality (mature-onset diabetes of the young 5) in children and adolescents: high prevalence in autoantibody-negative type 1 diabetes with kidney defects. Diabetes Care *31*, e83.

Decision Letter, first revision:

Our ref: NG-A55628R

19th August 2021

Dear Nick,

Your revised manuscript "Genome-wide analysis of cis-regulatory changes in the metabolic adaptation of cavefish" (NG-A55628R) has been seen by the original referees. As you will see from their comments below, they find that the paper has improved in revision and they have no remaining requests, and therefore we will be happy in principle to publish your study in Nature Genetics as an Article pending final revisions to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and we will send you a checklist detailing our editorial and formatting requirements in about a week. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Thank you again for your interest in Nature Genetics. Please do not hesitate to contact me if you have any questions.

Sincerely,
Kyle

Kyle Vogan, PhD

Senior Editor
Nature Genetics
https://orcid.org/0000-0001-9565-9665

Reviewer #2 (Remarks to the Author):

The authors have been transparent in addressing my major concerns.
I like the new section on the genome sequence and believe it adds to the interest of the manuscript.
I also applaud the authors for providing cogent explanations of why the reporter assays do not meet with predictions - allowing readers to evaluate the data in appropriate light.
I have no further request for additional data or explanation.
Paul Wade

Reviewer #3 (Remarks to the Author):

The authors have successfully addressed all of my concerns as well as concerns of other reviewers. In my opinion the manuscript is ready for publication.

Final Decision Letter:

In reply please quote: NG-A55628R1 Rohner

9th March 2022

Dear Nick,

I am delighted to say that your manuscript "Genome-wide analysis of cis-regulatory changes underlying metabolic adaptation of cavefish" has been accepted for publication in an upcoming issue of Nature Genetics.

Over the next few weeks, your paper will be copyedited to ensure that it conforms to Nature Genetics style. Once your paper is typeset, you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

After the grant of rights is completed, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately.

You will not receive your proofs until the publishing agreement has been received through our system.

Due to the importance of these deadlines, we ask that you please let us know now whether you will be difficult to contact over the next month. If this is the case, we ask you provide us with the contact information (email, phone and fax) of someone who will be able to check the proofs on your behalf,

and who will be available to address any last-minute problems.

Your paper will be published online after we receive your corrections and will appear in print in the next available issue. You can find out your date of online publication by contacting the Nature Press Office (press@nature.com) after sending your e-proof corrections. Now is the time to inform your Public Relations or Press Office about your paper, as they might be interested in promoting its publication. This will allow them time to prepare an accurate and satisfactory press release. Include your manuscript tracking number (NG-A55628R1) and the name of the journal, which they will need when they contact our Press Office.

Before your paper is published online, we will be distributing a press release to news organizations worldwide, which may very well include details of your work. We are happy for your institution or funding agency to prepare its own press release, but it must mention the embargo date and Nature Genetics. Our Press Office may contact you closer to the time of publication, but if you or your Press Office have any enquiries in the meantime, please contact press@nature.com.

Acceptance is conditional on the data in the manuscript not being published elsewhere, or announced in the print or electronic media, until the embargo/publication date. These restrictions are not intended to deter you from presenting your data at academic meetings and conferences, but any enquiries from the media about papers not yet scheduled for publication should be referred to us.

Please note that Nature Genetics is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles), then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including <a href="https://www.nature.com/nature-portfolio/editorial-policies/self-archiving-and-license-to-publish. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

Please note that Nature Research offers an immediate open access option only for papers that were first submitted after 1 January 2021.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles, and download a record of your refereeing activity for the Nature journals.

An online order form for reprints of your paper is available at https://www.nature.com/reprints/author-reprints.html. Please let your coauthors and your institutions' public affairs office know that they are also welcome to order reprints by this method.

If you have not already done so, we invite you to upload the step-by-step protocols used in this manuscript to the Protocols Exchange, part of our on-line web resource, natureprotocols.com. If you complete the upload by the time you receive your manuscript proofs, we can insert links in your article that lead directly to the protocol details. Your protocol will be made freely available upon publication of your paper. By participating in natureprotocols.com, you are enabling researchers to more readily reproduce or adapt the methodology you use. Natureprotocols.com is fully searchable, providing your protocols and paper with increased utility and visibility. Please submit your protocol to https://protocolexchange.researchsquare.com/. After entering your nature.com username and password you will need to enter your manuscript number (NG-A55628R1). Further information can be found at https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards#protocols

Sincerely,
Kyle

Kyle Vogan, PhD
Senior Editor
Nature Genetics
https://orcid.org/0000-0001-9565-9665